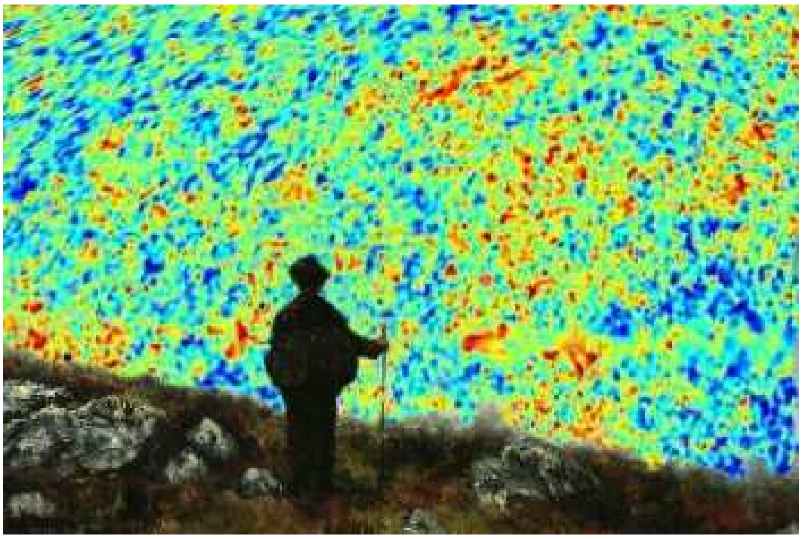
Lecture Notes for FY3464 & 3466 Quantum Field Theory I and II

An Introduction to Quantum Field Theory with Applications in Cosmology and Astroparticle Physics

M. Kachelrieß



M. Kachelrieß
Institutt for fysikk
NTNU, Trondheim
Norway
email: Michael.Kachelriess@ntnu.no

For a final version, see *Quantum Fields – from the Hubble to the Planck Scale*,
Graduate Text book series, Oxford University Press, ISBN: 9780198802877.

Contents

1. Classical mechanics

As a starter we review in this chapter those concepts of classical mechanics which are essential for progressing towards quantum theory. We recall first briefly the Lagrangian and Hamiltonian formulation of classical mechanics and their derivation from an action principle. We illustrate also the Green function method using as example the driven harmonic oscillator and recall the action of a relativistic point particle.

1.1. Action principle

Variational principles Fundamental laws of Nature as Newton's axioms or Maxwell's equations were discovered in the form of differential equations. Starting from Leibniz and Euler, it was realised that one can re-express differential equations in the form of variational principles: In this approach, the evolution of a physical system is described by the extremum of an appropriately chosen functional. Various versions of such variational principles exist, but they have in common that the functionals used have the dimension of "energy × time", i.e. the functionals have the same dimension as Planck's constant $\hbar$. A quantity with this dimension is called action S. An advantage of using the action as main tool to describe dynamical systems is that this allows us to implement easily both space-time and internal symmetries. For instance, choosing as ingredients of the action local functions that transform as scalars under Lorentz transformations leads automatically to relativistically invariant field equations. Moreover, the action S summarises economically the information contained typically in a set of various coupled differential equations.

If the variational principle is formulated as an integral principle, then the functional S will depend on the whole path $q(t)$ described by the system between the considered initial and final time. In the formulation of quantum theory we will pursue, we will look for a direct connection from the classical action $S[q]$ of the path $[q(t)\colon q'(t')]$ to the transition amplitude $\langle q', t'|q, t\rangle$. Thus the use of the action principle will not only simplify the discussion of symmetries of a physical system, but lies also at the heart of the approach to quantum theory we will follow.

1.1.1. Hamilton's principle and Lagrange's equations

A functional $F[f(x)]$ is a map from a certain space of functions $f(x)$ into the real or complex numbers. We will consider mainly functionals from the space of (at least) twice differentiable functions between fixed points a and b. More specifically, Hamilton's principle uses as functional the action S defined by

$$S[q^i] = \int_a^b \mathrm{d}t\, L(q^i, \dot{q}^i, t)\,, \tag{1.1}$$

where L is a function of the $2n$ independent functions q^i and $\dot{q}^i = \mathrm{d}q^i/\mathrm{d}t$ as well as of the parameter t. In classical mechanics, we call L the Lagrange function of the system, q^i are its

n generalised coordinates, $\dot{q}^i$ the corresponding velocities and t is the time. The extrema of this action give those paths $q(t)$ from a to b which are solutions of the equations of motion for the system described by L.

How do we find those paths that extremize the action S? First of all, we have to prescribe which variables are kept constant, which are varied and which constraints the variations have to obey. Depending on the variation principle we choose, these conditions and the functional form of the action will differ. Hamilton's principle corresponds to a smooth variation of the path,

$$q^i(t,\varepsilon) = q^i(t,0) + \varepsilon\eta^i(t)\,,$$

that keeps the endpoints fixed, $\eta^i(a) = \eta^i(b) = 0$, but is otherwise arbitrary. The scale factor ε determines the magnitude of the variation for the one-parameter family of paths $\varepsilon\eta^i(t)$. The notation $S[q^i]$ stresses that we consider the action as a functional only of the coordinates q^i: The velocities $\dot{q}^i$ are not varied independently because ε is time-independent. Since the time t is not varied in Hamilton's principle, varying the path $q^i(t,\varepsilon)$ requires only to calculate the resulting change of the Lagrangian L. Following this prescription, the action has an extremum if

$$0 = \left.\frac{\partial S[q^i(t,\varepsilon)]}{\partial\varepsilon}\right|_{\varepsilon=0} = \int_a^b \mathrm{d}t\, \left(\frac{\partial L}{\partial q^i}\frac{\partial q^i}{\partial\varepsilon} + \frac{\partial L}{\partial\dot{q}^i}\frac{\partial\dot{q}^i}{\partial\varepsilon}\right) = \int_a^b \mathrm{d}t\, \left(\frac{\partial L}{\partial q^i}\eta^i + \frac{\partial L}{\partial\dot{q}^i}\dot{\eta}^i\right)\,. \tag{1.2}$$

Here we applied—as always in the following—Einstein's convention to sum over a repeated index pair. Thus e.g. the first term in the bracket equals

$$\frac{\partial L}{\partial q^i}\eta^i \equiv \sum_{i=1}^n \frac{\partial L}{\partial q^i}\eta^i$$

for a system described by n generalised coordinates. We can eliminate $\dot{\eta}^i$ in favour of η^i, integrating the second term by parts, arriving at

$$\left.\frac{\partial S[q^i(t,\varepsilon)]}{\partial\varepsilon}\right|_{\varepsilon=0} = \int_a^b \mathrm{d}t\, \left[\frac{\partial L}{\partial q^i} - \frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial L}{\partial\dot{q}^i}\right)\right]\eta^i + \left[\frac{\partial L}{\partial\dot{q}^i}\eta^i\right]_a^b\,. \tag{1.3}$$

The boundary term $[\ldots]_a^b$ vanishes, because we required that the functions η^i are zero at the endpoints a and b. Since these functions are otherwise arbitrary, each individual term in the first bracket has to vanish for an extremal curve. The n equations resulting from the condition $\partial S[q^i(t,\varepsilon)]/\partial\varepsilon = 0$ are called the (Euler-) Lagrange equations of the action S,

$$\frac{\partial L}{\partial q^i} - \frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial L}{\partial\dot{q}^i} = 0\,, \tag{1.4}$$

and give the equations of motion for the system specified by L. In the future, we will use a more concise notation, calling

$$\delta q^i \equiv \lim_{\varepsilon\to 0}\frac{q^i(t,\varepsilon) - q^i(t,0)}{\varepsilon} = \left.\frac{\partial q^i(t,\varepsilon)}{\partial\varepsilon}\right|_{\varepsilon=0} \tag{1.5}$$

the variation of q^i, and similarly for functions and functionals of q^i. Thus we can re-write e.g. Eq. (1.2) in a more evident form as

$$0 = \delta S[q^i] = \int_a^b \mathrm{d}t\, \delta L(q^i,\dot{q}^i,t) = \int_a^b \mathrm{d}t\, \left(\frac{\partial L}{\partial q^i}\delta q^i + \frac{\partial L}{\partial\dot{q}^i}\delta\dot{q}^i\right)\,. \tag{1.6}$$

We close this paragraph with three remarks. First, we note that Hamilton's principle is often called the principle of least action. This name is somewhat misleading, since the extremum of the action can be also a maximum or a saddle point. Second, observe that the Lagrangian L is not uniquely fixed: Adding a total time-derivative, $L \to L' = L + \mathrm{d}f(q,t)/\mathrm{d}t$, does not change the resulting Lagrange equations,

$$S' = S + \int_a^b \mathrm{d}t\, \frac{\mathrm{d}f}{\mathrm{d}t} = S + f(q(b), t_b) - f(q(a), t_a)\,, \tag{1.7}$$

since the last two terms vanish varying the action with the restriction of fixed endpoints a and b. Finally, note that we used a Lagrangian that depends only on the coordinates and their *first* derivatives. Such a Lagrangian leads to second-order equations of motion and thus to a mechanical system specified by the $2n$ pieces of information $\{q_i, \dot{q}_i\}$. Ostrogradsky showed 1850 that a stable ground-state is impossible, if the Lagrangian contains higher derivatives $\ddot{q}, q^{(3)}, \ldots$, cf. problem 1.**??**. Therefore such theories contradict our experience that the vacuum is stable. Constructing Lagrangians for the fundamental theories describing Nature, we should restrict ourselves thus to Lagrangians that lead to second-order equations of motion.

Lagrange function We illustrate now how one can use symmetries to constrain the possible form of a Lagrangian L. As example, we consider the case of a free non-relativistic particle with mass m subject to the Galilean principle of relativity. More precisely, we use that the homogeneity of space and time forbids that L depends on $\boldsymbol{x}$ and t, while the isotropy of space implies that L depends only on the norm of the velocity vector $\boldsymbol{v}$, but not on its direction. Thus the Lagrange function of a free particle can be only a function of v^2, $L = L(v^2)$.

Let us consider two inertial frames moving with the infinitesimal velocity $\boldsymbol{\varepsilon}$ relative to each other. (Recall that an inertial frame is defined as a coordinate system where a force-free particle moves along a straight line.) Then a Galilean transformation connects the velocities measured in the two frames as $\boldsymbol{v}' = \boldsymbol{v} + \boldsymbol{\varepsilon}$. The Galilean principle of relativity requires that the laws of motion have the same form in both frames, and thus the Lagrangians can differ only by a total time-derivative. Expanding the difference δL in ε gives with $\delta v^2 = 2\boldsymbol{v} \cdot \boldsymbol{\varepsilon} + \mathcal{O}(\varepsilon^2)$

$$\delta L = \frac{\partial L}{\partial v^2}\, \delta v^2 = 2\boldsymbol{v} \cdot \boldsymbol{\varepsilon} \frac{\partial L}{\partial v^2}\,. \tag{1.8}$$

Since $v^i = \mathrm{d}x^i/\mathrm{d}t$, the term $\partial L/\partial v^2$ has to be independent of v such that the difference δL is a total time-derivative. Hence, the Lagrangian of a free particle has the form $L = av^2 + b$. The constant b drops out of the equations of motion, and we can set it therefore to zero. To be consistent with usual notation, we call the proportionality constant $m/2$, and the total expression kinetic energy T,

$$L = T = \frac{1}{2} m v^2\,. \tag{1.9}$$

For a system of non-interacting particles, the Lagrange function L is additive, $L = \sum_a \frac{1}{2} m_a v_a^2$. If there are interactions (assumed for simplicity to depend only on the coordinates), then we subtract a function $V(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots)$ called potential energy. One confirms readily that this choice for L reproduces Newton's law of motion.

Energy The Lagrangian of a closed system does not depend on time because of the homogeneity of time. Its total time derivative is

$$\frac{\mathrm{d}L}{\mathrm{d}t} = \frac{\partial L}{\partial q^i}\,\dot{q}^i + \frac{\partial L}{\partial \dot{q}^i}\,\ddot{q}^i\,. \tag{1.10}$$

Using the equations of motion and replacing $\partial L/\partial q^i$ by $(\mathrm{d}/\mathrm{d}t)\partial L/\partial \dot{q}^i$, it follows

$$\frac{\mathrm{d}L}{\mathrm{d}t} = \dot{q}^i\,\frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial L}{\partial \dot{q}^i} + \frac{\partial L}{\partial \dot{q}^i}\,\ddot{q}^i = \frac{\mathrm{d}}{\mathrm{d}t}\left(\dot{q}^i\,\frac{\partial L}{\partial \dot{q}^i}\right)\,. \tag{1.11}$$

Hence the quantity

$$E \equiv \dot{q}^i\,\frac{\partial L}{\partial \dot{q}^i} - L \tag{1.12}$$

remains constant during the evolution of a closed system. This holds also more generally, e.g. in the presence of static external fields, as long as the Lagrangian is not time-dependent.

We have still to show that E coincides indeed with the usual definition of energy. Using as Lagrange function $L = T(q,\dot{q}) - V(q)$, where the kinetic energy T is quadratic in the velocities, we have

$$\dot{q}^i\,\frac{\partial L}{\partial \dot{q}^i} = \dot{q}^i\,\frac{\partial T}{\partial \dot{q}^i} = 2T \tag{1.13}$$

and thus $E = 2T - L = T + V$.

Conservation laws In a general way, we can derive the connection between a symmetry of the Lagrangian and a corresponding conservation law as follows: Let us assume that under a change of coordinates $q^i \to q^i + \delta q^i$, the Lagrangian changes at most by a total time derivative,

$$L \to L + \delta L = L + \frac{\mathrm{d}\delta F}{\mathrm{d}t}\,. \tag{1.14}$$

In this case, the equation of motions are unchanged and the coordinate change $q^i \to q^i + \delta q^i$ is a symmetry of the Lagrangian. But the change $\mathrm{d}\delta F/\mathrm{d}t$ has to equal δL induced by the variation δq^i,

$$\frac{\partial L}{\partial q^i}\delta q^i + \frac{\partial L}{\partial \dot{q}^i}\delta \dot{q}^i - \frac{\mathrm{d}\delta F}{\mathrm{d}t} = 0\,. \tag{1.15}$$

Replacing again $\partial L/\partial q^i$ by $(\mathrm{d}/\mathrm{d}t)\partial L/\partial \dot{q}^i$ and applying the product rule gives as conserved quantity

$$Q = \frac{\partial L}{\partial \dot{q}^i}\delta q^i - \delta F\,. \tag{1.16}$$

Thus any continuous symmetry of a Lagrangian system results in a conserved quantity. In particular, energy conservation follows for a system invariant under time-translations with $\delta q^i = \dot{q}^i\delta t$. Other conservation laws are discussed in problem 1.??.

1.1.2. Palatini's principle and Hamilton's equations

Legendre transformation and the Hamilton function In the Lagrange formalism, we describe a system specifying its generalised coordinates and velocities using the Lagrangian,

$L = L(q^i, \dot{q}^i, t)$. An alternative is to use generalised coordinates and their canonically conjugated momenta p_i defined as

$$p_i = \frac{\partial L}{\partial \dot{q}^i}\,. \tag{1.17}$$

The passage from $\{q^i, \dot{q}^i\}$ to $\{q^i, p_i\}$ is a special case of a Legendre transformation[1]: Starting from the Lagrangian L we define a new function $H(q^i, p_i, t)$ called Hamiltonian or Hamilton function via

$$H(q^i, p_i, t) = \frac{\partial L}{\partial \dot{q}^i}\,\dot{q}^i - L(q^i, \dot{q}^i, t) = p_i \dot{q}^i - L(q^i, \dot{q}^i, t)\,. \tag{1.18}$$

Here we assume that we can invert the definition (1.17) and are thus able to substitute velocities $\dot{q}^i$ by momenta p_i in the Lagrangian L.

The physical meaning of the Hamiltonian H follows immediately comparing its defining equation with the one for the energy E. Thus the numerical value of the Hamiltonian equals the energy of a dynamical system; we insist however that H is expressed as function of coordinates and their conjugated momenta. A coordinate q_i that does not appear explicitly in L is called cyclic. The Lagrange equations imply then $\partial L/\partial \dot{q}_i = const.$, so that the corresponding canonically conjugated momentum $p_i = \partial L/\partial \dot{q}^i$ is conserved.

Palatini's formalism and Hamilton's equations Previously, we considered the action S as a functional only of q^i. Then the variation of the velocities $\dot{q}^i$ is not independent and we arrive at n second order differential equations for the coordinates q^i. An alternative approach is to allow independent variations of the coordinates q^i and of the velocities $\dot{q}^i$. We trade the latter against the momenta $p_i = \partial L/\partial \dot{q}^i$ and rewrite the action as

$$S[q^i, p_i] = \int_a^b \mathrm{d}t\ \left[p_i \dot{q}^i - H(q^i, p_i, t)\right]\,. \tag{1.19}$$

The independent variation of coordinates q^i and momenta p_i gives

$$\delta S[q^i, p_i] = \int_a^b \mathrm{d}t\ \left[p_i \delta \dot{q}^i + \dot{q}^i \delta p_i - \frac{\partial H}{\partial q^i}\delta q^i - \frac{\partial H}{\partial p_i}\delta p_i\right]\,. \tag{1.20}$$

The first term can be integrated by parts, and the resulting boundary terms vanishes by assumption. Collecting then the δq^i and δp_i terms and requiring that the variation is zero, we obtain

$$0 = \delta S[q^i, p_i] = \int_a^b \mathrm{d}t\ \left[-\left(\dot{p}_i + \frac{\partial H}{\partial q^i}\right)\delta q^i + \left(\dot{q}^i - \frac{\partial H}{\partial p_i}\right)\delta p_i\right]\,. \tag{1.21}$$

As the variations δq^i and δp_i are independent, their coefficients in the round brackets have to vanish separately. Thus we obtain in this formalism directly Hamilton's equations,

$$\dot{q}^i = \frac{\partial H}{\partial p_i}\,, \qquad and \qquad \dot{p}_i = -\frac{\partial H}{\partial q^i}\,. \tag{1.22}$$

Consider now an observable $O = O(q^i, p_i, t)$. Its time-dependence is given by

$$\frac{\mathrm{d}O}{\mathrm{d}t} = \frac{\partial O}{\partial q^i}\dot{q}^i + \frac{\partial O}{\partial p_i}\dot{p}_i + \frac{\partial O}{\partial t} = \frac{\partial O}{\partial q^i}\frac{\partial H}{\partial p_i} - \frac{\partial O}{\partial p_i}\frac{\partial H}{\partial q^i} + \frac{\partial O}{\partial t}\,, \tag{1.23}$$

[1] The concept of a Legendre transformation may be familiar from thermodynamics, where it is used to change between extensive variables (e.g. the entropy S) and their conjugate intensive variables (e.g. the temperature T).

where we used Hamilton's equations. If we define the Poisson brackets $\{A,B\}$ between two observables A and B as

$$\{A,B\} = \frac{\partial A}{\partial q^i}\frac{\partial B}{\partial p_i} - \frac{\partial A}{\partial p_i}\frac{\partial B}{\partial q^i}\,, \tag{1.24}$$

then we can rewrite Eq. (1.23) as

$$\frac{\mathrm{d}O}{\mathrm{d}t} = \{O,H\} + \frac{\partial O}{\partial t}\,. \tag{1.25}$$

This equations gives us a formal correspondence between classical and quantum mechanics: The time-evolution of an operator O in the Heisenberg picture is given by the same equation as in classical mechanics, if the Poisson bracket is changed to a commutator. Since the Poisson bracket is antisymmetric, we find

$$\frac{\mathrm{d}H}{\mathrm{d}t} = \frac{\partial H}{\partial t}\,. \tag{1.26}$$

Hence the Hamiltonian H is a conserved quantity, if and only if H is time-independent.

1.2. Green functions and the response method

We can test the internal properties of a physical system, if we impose an external force $J(t)$ on it and compare its measured to its calculated response. If the system is described by linear differential equations, then the superposition principle is valid: We can reconstruct the solution $x(t)$ for an arbitrary applied external force $J(t)$, if we know the response to a normalised delta function-like kick $J(t) = \delta(t-t')$. Mathematically, this corresponds to the knowledge of the Green function $G(t-t')$ for the differential equation $D(t)x(t) = J(t)$ describing the system. Even if the system is described by a non-linear differential equation, we can often use a linear approximation in case of a sufficiently small external force $J(t)$. Therefore the Green function method is extremely useful and we will apply it extensively discussing quantum field theories.

We illustrate this method with the example of the harmonic oscillator which is the prototype for a quadratic, and thus exactly solvable, action. In classical physics, causality implies that the knowledge of the external force $J(t')$ at times $t' < t$ is sufficient to determine the solution $x(t)$ at time t. We define therefore two Green functions $\widetilde{G}$ and G_R by

$$x(t) = \int_{-\infty}^{t} \mathrm{d}t'\, \widetilde{G}(t-t')J(t') = \int_{-\infty}^{\infty} \mathrm{d}t' G_R(t-t')J(t')\,, \tag{1.27}$$

where the retarded Green function G_R satisfies $G_R(t-t') = \widetilde{G}(t-t')\vartheta(t-t')$. The definition (1.27) is motivated by the trivial relation $J(t) = \int \mathrm{d}t'\, \delta(t-t')J(t')$: An arbitrary force $J(t)$ can be seen as a superposition of delta functions $\delta(t-t')$ with weight $J(t')$. If the Green function $G_R(t-t')$ determines the response of the system to a delta function-like force, then we should obtain the solution $x(t)$ integrating $G_R(t-t')$ with the weight $J(t')$.

We convert the equation of motion $m\ddot{x} + m\omega^2 x = J$ of a forced harmonic oscillator into the form $D(t)x(t) = J(t)$ by writing

$$D(t)x(t) \equiv m\left(\frac{\mathrm{d}^2}{\mathrm{d}t^2} + \omega^2\right)x(t) = J(t)\,. \tag{1.28}$$

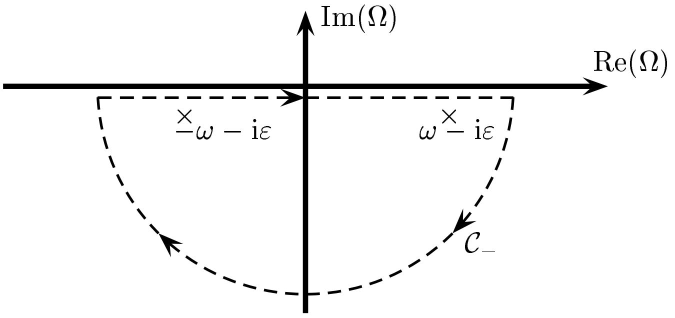

$\mathrm{Im}(\Omega)$
$\mathrm{Re}(\Omega)$
$-\omega - \mathrm{i}\varepsilon$
$\omega - \mathrm{i}\varepsilon$
$\mathcal{C}_-$

Figure 1.1.: Poles and contour in the complex Ω plane used for the integration of the retarded Green function.

Inserting (1.27) into (1.28) gives

$$\int_{-\infty}^{\infty} \mathrm{d}t' \, D(t) G_R(t-t') J(t') = J(t) \,. \tag{1.29}$$

For an arbitrary external force $J(t)$, this relation can be only valid if

$$D(t) G_R(t-t') = \delta(t-t') \,. \tag{1.30}$$

Thus a Green function $G(t-t')$ is the inverse of its defining differential operator $D(t)$. As we will see, Eq. (1.30) does not specify uniquely the Green function, and thus we will omit the index "R" for the moment. Performing a Fourier transformation,

$$G(t-t') = \int \frac{\mathrm{d}\Omega}{2\pi} G(\Omega) \mathrm{e}^{-\mathrm{i}\Omega(t-t')} \quad and \quad \delta(t-t') = \int \frac{\mathrm{d}\Omega}{2\pi} \mathrm{e}^{-\mathrm{i}\Omega(t-t')} \,, \tag{1.31}$$

we obtain

$$\int \frac{\mathrm{d}\Omega}{2\pi} G(\Omega) D(t) \mathrm{e}^{-\mathrm{i}\Omega(t-t')} = \int \frac{\mathrm{d}\Omega}{2\pi} \mathrm{e}^{-\mathrm{i}\Omega(t-t')} \,. \tag{1.32}$$

The action of $D(t)$ on the plane-waves $\mathrm{e}^{-\mathrm{i}\Omega(t-t')}$ can be evaluated easily, since the differentiation has become equivalent with multiplication, $\mathrm{d}/\mathrm{d}t \to -\mathrm{i}\Omega$. Comparing then the coefficients of the plane-waves on both sides of this equation, we have to invert only an algebraic equation, arriving at

$$G(\Omega) = \frac{1}{m} \frac{1}{\omega^2 - \Omega^2} \,. \tag{1.33}$$

For the back-transformation with $\tau = t - t'$,

$$G(\tau) = \int \frac{\mathrm{d}\Omega}{2\pi m} \frac{\mathrm{e}^{-\mathrm{i}\Omega\tau}}{\omega^2 - \Omega^2} \,, \tag{1.34}$$

we have to specify how the poles at $\Omega^2 = \omega^2$ are avoided. It is this choice by which we select the appropriate Green function. In classical physics, we implement causality ("cause always precedes its effect") selecting the retarded Green function.

We will use Cauchy's residue theorem, $\oint \mathrm{d}z \, f(z) = 2\pi\mathrm{i} \sum res_{z_0} f(z)$, to calculate the integral. Its application requires to close the integration contour adding a path which gives a vanishing contribution to the integral. This is achieved, when the integrand $G(\Omega)\mathrm{e}^{-\mathrm{i}\Omega\tau}$ vanishes fast enough along the added path. Thus we have to choose for positive τ the contour $\mathcal{C}_-$ in the lower plane, $\mathrm{e}^{-\mathrm{i}\Omega\tau} = \mathrm{e}^{-|\Im(\Omega)|\tau} \to 0$ for $\Im(\Omega) \to -\infty$, while we have to close the contour in the upper plane for negative τ. If we want to obtain the retarded Green function $G_R(\tau)$ which vanishes for $\tau < 0$, we have to shift therefore the poles $\Omega_{1/2} = \pm\omega$ into the lower plane as shown in Fig. 1.1 by adding a small negative imaginary part, $\Omega_{1/2} \to \Omega_{1/2} = \pm\omega - \mathrm{i}\varepsilon$, or

$$G_R(\tau) = -\frac{1}{2\pi m}\int \mathrm{d}\Omega \, \frac{\mathrm{e}^{-\mathrm{i}\Omega\tau}}{(\Omega - \omega + \mathrm{i}\varepsilon)(\Omega + \omega + \mathrm{i}\varepsilon)} \,. \tag{1.35}$$

The residue $\mathrm{res}_{z_0} f(z)$ of a function f with a single pole at z_0 is given by

$$res_{z_0} f(z) = \lim_{z\to z_0} (z - z_0) f(z) \,. \tag{1.36}$$

Thus we pick up at $\Omega_1 = -\omega - \mathrm{i}\varepsilon$ the contribution $2\pi\mathrm{i}\, \mathrm{e}^{+\mathrm{i}\omega\tau}/(-2\omega)$, while we obtain $2\pi\mathrm{i}\, \mathrm{e}^{-\mathrm{i}\omega\tau}/(2\omega)$ from $\Omega_2 = \omega - \mathrm{i}\varepsilon$. Combining both contributions and adding a minus sign because the contour is clockwise, we arrive at

$$G_R(\tau) = \frac{\mathrm{i}}{2m\omega}\left[\mathrm{e}^{-\mathrm{i}\omega\tau} - \mathrm{e}^{\mathrm{i}\omega\tau}\right]\vartheta(\tau) = \frac{1}{m}\,\frac{\sin(\omega\tau)}{\omega}\,\vartheta(\tau) \tag{1.37}$$

as result for the retarded Green function of the forced harmonic oscillator.

We can now obtain a particular solution solving (1.27). For instance, choosing $J(t') = \delta(t - t')$, results in

$$x(t) = \frac{1}{m}\,\frac{\sin(\omega t)}{\omega}\,\vartheta(t) \,. \tag{1.38}$$

Thus the oscillator was at rest for $t < 0$, got a kick at $t = 0$, and oscillates according $x(t)$ afterwards. Note the following two points: First, the fact that the kick proceeds the movement is the result of our choice of the retarded (or causal) Green function. Second, the particular solution (1.38) for an oscillator initially at rest can be generalised by adding the solution to the homogeneous equation $\ddot{x} + \omega^2 x = 0$.

1.3. Relativistic particle

In special relativity, we replace the Galilean transformations as symmetry group of space and time by Lorentz transformations. The latter are all those coordinate transformations $x^\mu \to \tilde{x}^\mu = \Lambda^\mu_{\ \nu} x^\nu$ that keep the squared distance

$$s_{12}^2 \equiv (t_1 - t_2)^2 - (x_1 - x_2)^2 - (y_1 - y_2)^2 - (z_1 - z_2)^2 \tag{1.39}$$

between two space-time events x_1^μ and x_2^μ invariant. The distance of two infinitesimally close space-time events is called the line-element $\mathrm{d}s$ of the space-time. In Minkowski space, it is given by

$$\mathrm{d}s^2 = \mathrm{d}t^2 - \mathrm{d}x^2 - \mathrm{d}y^2 - \mathrm{d}z^2 \tag{1.40}$$

using a Cartesian inertial frame. We can interpret the line-element $\mathrm{d}s^2$ as a scalar product, if we introduce the metric tensor $\eta_{\mu\nu}$ with elements

$$\eta_{\mu\nu} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{pmatrix} \tag{1.41}$$

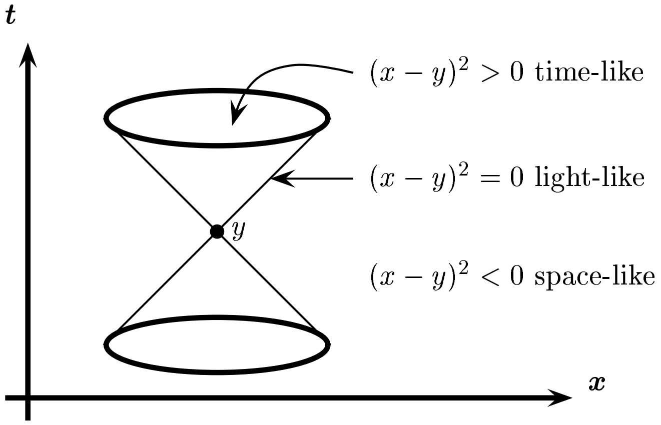


Figure 1.2.: Light-cone at the point $P(y^\mu)$ generated by light-like vectors. Contained in the light-cone are the time-like vectors, outside the space-like ones.

and a scalar product of two vectors as

$$a \cdot b \equiv \eta_{\mu\nu} a^\mu b^\nu = a_\mu b^\mu = a^\mu b_\mu \,. \tag{1.42}$$

In Minkowski space, we call a four-vector any four-tuple V^μ that transforms as $\tilde{V}^\mu = \Lambda^\mu{}_\nu V^\nu$. By convention, we associate three-vectors with the spatial part of four-vectors with upper indices, e.g. we set $x^\mu = \{t, x, y, z\}$ or $A^\mu = \{\phi, \boldsymbol{A}\}$. Lowering then the index by contraction with the metric tensor result in a minus sign of the spatial components of a four-vector, $x_\mu = \eta_{\mu\nu} x^\mu = \{t, -x, -y, -z\}$ or $A_\mu = \{\phi, -\boldsymbol{A}\}$. Summing over an index pair, typically one index occurs in an upper and one in a lower position. Note that in the denominator, an upper index counts as a lower index and vice versa, cf. e.g. with Eqs. (1.18) and (1.17). Additionally to four-vectors, we will meet tensors $T^{\mu_1 \cdots \mu_n}$ of rank n which transform as $\tilde{T}^{\mu_1 \cdots \mu_n} = \Lambda^{\mu_1}{}_{\nu_1} \cdots \Lambda^{\mu_n}{}_{\nu_n} T^{\nu_1 \cdots \nu_n}$.

Since the metric $\eta_{\mu\nu}$ is indefinite, the norm of a vector a^μ can be

$$a_\mu a^\mu > 0, \quad \text{time-like,} \tag{1.43a}$$

$$a_\mu a^\mu = 0, \quad \text{light-like or null-vector,} \tag{1.43b}$$

$$a_\mu a^\mu < 0, \quad \text{space-like.} \tag{1.43c}$$

The cone of all light-like vectors starting from a point P is called light-cone, cf. Fig. 1.2. The time-like region inside the light-cone consists of two parts, past and future. Only events inside the past light-cone can influence the physics at point P, while P can influence only the interior of its future light-cone. The proper-time τ is the time displayed by a clock moving with the observer. With our conventions—negative signature of the metric and $c = 1$—the proper-time elapsed between two space-time events equals the integrated line-element between them,

$$\tau_{12} = \int_1^2 \mathrm{d}s = \int_1^2 [\eta_{\mu\nu} \mathrm{d}x^\mu \mathrm{d}x^\nu]^{1/2} = \int_1^2 \mathrm{d}t [1 - v^2]^{1/2} < t_2 - t_1 \,. \tag{1.44}$$

The last part of this equation, where we introduced the three-velocity $v^i = \mathrm{d}x^i/\mathrm{d}t$ of the clock, shows explicitly the relativistic effect of time dilation, as well as the connection between coordinate time t and the proper-time τ of a moving clock, $\mathrm{d}\tau = (1-v^2)^{1/2}\mathrm{d}t \equiv \mathrm{d}t/\gamma$. The line describing the position of an observer is called world-line. Parametrising the worldline by the parameter σ, $x = x(\sigma)$, the proper-time is given by

$$\tau = \int \mathrm{d}\sigma \left[\eta_{\mu\nu}\frac{\mathrm{d}x^\mu}{\mathrm{d}\sigma}\frac{\mathrm{d}x^\nu}{\mathrm{d}\sigma}\right]^{1/2} . \tag{1.45}$$

Note that τ is invariant under a reparameterisation $\tilde{\sigma} = f(\sigma)$.

The only invariant differential we have at our disposal to form an action for a free point-like particle is the line-element, or equivalently the proper-time,

$$S_0 = \alpha \int_a^b \mathrm{d}s = \alpha \int_a^b \mathrm{d}\sigma \,\frac{\mathrm{d}s}{\mathrm{d}\sigma} \tag{1.46}$$

with $L = \alpha \mathrm{d}s/\mathrm{d}\sigma = \alpha \mathrm{d}\tau/\mathrm{d}\sigma$. We check now if this choice which implies the Lagrangian

$$L = \alpha\frac{\mathrm{d}\tau}{\mathrm{d}\sigma} = \alpha\left[\eta_{\mu\nu}\,\frac{\mathrm{d}x^\mu}{\mathrm{d}\sigma}\,\frac{\mathrm{d}x^\nu}{\mathrm{d}\sigma}\right]^{1/2} \tag{1.47}$$

for a free particle is sensible: The action has the correct non-relativistic limit,

$$S_0 = \alpha\int_a^b \mathrm{d}s = \alpha\int_a^b \mathrm{d}t\,\sqrt{1-v^2} = \int_a^b \mathrm{d}t\,\left(-m + \frac{1}{2}mv^2 + \mathcal{O}(v^4)\right) , \tag{1.48}$$

if we set $\alpha = -m$. The mass m corresponds to a potential energy in the non-relativistic limit and has therefore a negative sign in the Lagrangian. Moreover, a constant drops out of the equations of motion, and thus the term $-m$ can be omitted in the non-relativistic limit. The time t enters the relativistic Lagrangian in a Lorentz invariant way as one of the dynamical variables, $x^\mu = (t, \boldsymbol{x})$, while σ assumes now t's purpose to parametrise the trajectory, $x^\mu(\sigma)$. Since a moving clock goes slower than a clock at rest, solutions of this Lagrangian maximise the action.

Example 1.1: Relativistic dispersion relation: We extend the non-relativistic definition of the momentum, $p_i = \partial L/\partial\dot{x}^i$, to four dimensions setting $p_\alpha = -\partial L/\partial\dot{x}^\alpha$. Note the minus sign that reflects the minus in the spatial components of a covariant vector, $p_\alpha = (E, -\boldsymbol{p})$. Then

$$p_\alpha = -\frac{\partial L}{\partial\dot{x}^\alpha} = m\,\frac{\mathrm{d}x_\alpha/\mathrm{d}\sigma}{\mathrm{d}\tau/\mathrm{d}\sigma} = m\,\frac{\mathrm{d}x_\alpha}{\mathrm{d}\tau} \equiv mu_\alpha\,. \tag{1.49}$$

In the last step, we defined the four-velocity $u^\alpha = \mathrm{d}x^\alpha/\mathrm{d}\tau$. Using $\mathrm{d}t = \gamma\mathrm{d}\tau$, it follows $u_\alpha u^\alpha = 1$ and $p_\alpha p^\alpha = m^2$. The last relation expresses the relativistic dispersion relation $E^2 = m^2 + \boldsymbol{p}^2$.

The Lagrange equations are

$$\frac{\mathrm{d}}{\mathrm{d}\sigma}\frac{\partial L}{\partial(\mathrm{d}x^\alpha/\mathrm{d}\sigma)} = \frac{\partial L}{\partial x^\alpha}\,. \tag{1.50}$$

Consider e.g. the x^1 component, then

$$\frac{\mathrm{d}}{\mathrm{d}\sigma}\frac{\partial L}{\partial(\mathrm{d}x^1/\mathrm{d}\sigma)} = \frac{\mathrm{d}}{\mathrm{d}\sigma}\left(\frac{1}{L}\frac{\mathrm{d}x^1}{\mathrm{d}\sigma}\right) = 0\,. \tag{1.51}$$

Since $L = -m\mathrm{d}\tau/\mathrm{d}\sigma$, Newton's law follows for the x^1 coordinate after multiplication with $\mathrm{d}\sigma/\mathrm{d}\tau$,

$$\frac{d^2x^1}{\mathrm{d}\tau^2} = 0\,, \tag{1.52}$$

and similar for the other coordinates.

An equivalent, but often more convenient form for the Lagrangian of a free particle is

$$L = -m\eta_{\mu\nu}\dot{x}^\mu\dot{x}^\nu\,, \tag{1.53}$$

where we set $\dot{x}^\mu = \mathrm{d}x^\mu/\mathrm{d}\tau$. If there are no interactions (except gravity), we can neglect the mass m of the particle and one often sets $m \to -1$.

Next we want to add an interaction term S_{em} between a particle with charge q and an electromagnetic field. The simplest possible action is to integrate the potential A_μ along the world-line $x^\mu(\sigma)$ of the particle,

$$S_{\mathrm{em}} = -q\int \mathrm{d}x^\mu\, A_\mu(x) = -q\int \mathrm{d}\sigma\,\frac{\mathrm{d}x^\mu}{\mathrm{d}\sigma}\,A_\mu(x)\,. \tag{1.54}$$

Using the choice $\sigma = \tau$, we can view $q\dot{x}^\mu$ as the current j^μ induced by the particle and thus the interaction has the form $L_{\mathrm{em}} = -j^\mu A_\mu$. Any candidate for S_{em} should be invariant under a gauge transformation of the potential,

$$A_\mu(x) \to A_\mu(x) - \partial_\mu\Lambda(x)\,. \tag{1.55}$$

This is the case, since the induced change in the action,

$$\delta_\Lambda S_{\mathrm{em}} = q\int_1^2 \mathrm{d}\sigma\,\frac{\mathrm{d}x^\mu}{\mathrm{d}\sigma}\,\frac{\partial\Lambda(x)}{\partial x^\mu} = q\int_1^2 \mathrm{d}\Lambda = q[\Lambda(2)-\Lambda(1)]\,, \tag{1.56}$$

depends only on the endpoints. Thus $\delta_\Lambda S_{\mathrm{em}}$ vanishes keeping the endpoints fixed. Assuming that the Lagrangian is additive,

$$L = L_0 + L_{\mathrm{em}} = -m\left[\eta_{\mu\nu}\frac{\mathrm{d}x^\mu}{\mathrm{d}\sigma}\frac{\mathrm{d}x^\nu}{\mathrm{d}\sigma}\right]^{1/2} - q\frac{\mathrm{d}x^\mu}{\mathrm{d}\sigma}\,A_\mu(x) \tag{1.57}$$

the Lagrange equations give now

$$\frac{\mathrm{d}}{\mathrm{d}\sigma}\left[\frac{m\mathrm{d}x_\alpha/\mathrm{d}\sigma}{[\eta_{\mu\nu}\mathrm{d}x^\mu/\mathrm{d}\sigma\;\mathrm{d}x^\nu/\mathrm{d}\sigma]^{1/2}} + qA_\alpha\right] = q\,\frac{\mathrm{d}x^\lambda}{\mathrm{d}\sigma}\,\frac{\partial A_\lambda(x)}{\partial x^\alpha}\,. \tag{1.58}$$

Performing then the differentiation of $A(x(\sigma))$ with respect to σ and moving it to the RHS, we find

$$m\frac{\mathrm{d}}{\mathrm{d}\sigma}\left[\frac{\mathrm{d}x_\alpha/\mathrm{d}\sigma}{\mathrm{d}\tau/\mathrm{d}\sigma}\right] = q\left(\frac{\mathrm{d}x^\lambda}{\mathrm{d}\sigma}\,\frac{\partial A_\lambda}{\partial x^\alpha} - \frac{\mathrm{d}x^\lambda}{\mathrm{d}\sigma}\,\frac{\partial A_\alpha}{\partial x^\lambda}\right) = q\,\frac{\mathrm{d}x^\lambda}{\mathrm{d}\sigma}\,F_{\alpha\lambda}\,, \tag{1.59}$$

where we introduced the electromagnetic field-strength tensor $F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu$. Choosing $\sigma = \tau$ we obtain the covariant version of the Lorentz equation,

$$m\frac{\mathrm{d}^2x^\alpha}{\mathrm{d}\tau^2} = q\,F^\alpha{}_\lambda\,u^\lambda\,. \tag{1.60}$$

You should work through problem 1.??, if this equation and the covariant formulation of the Maxwell equations are not familiar to you.

Summary

The Lagrange and Hamilton function are connected by a Legendre transformation, $L(q^i, \dot{q}^i, t) = p_i \dot{q}^i - H(q^i, p_i, t)$. Lagrange's and Hamilton's equations follow extremizing the action $S[q^i] = \int_a^b \mathrm{d}t\, L(q^i, \dot{q}^i, t)$ and $S[q^i, p_i] = \int_a^b \mathrm{d}t\, \left[p_i \dot{q}^i - H(q^i, p_i, t)\right]$, respectively, keeping the endpoints a and b in coordinate space fixed. Knowing the Green function $G(t-t')$ of a linear system, we can find the solution $x(t)$ for an arbitrary external force $J(t)$ by integrating $G(t-t')$ with the weight $J(t)$.

Further reading

The series of Landau and Lifshitz on theoretical physics is a timeless resource for everybody studying and working in this field; its volume 1 [LL76] presents a succinct treatment of classical mechanics.

2. Quantum mechanics

The main purpose of this chapter is to introduce Feynman's path integral as an alternative to the standard operator approach to quantum mechanics. Most of our discussion of quantum fields will be based on this approach, and thus becoming familiar with this technique using the simpler case of quantum mechanics is of central importance. Instead of employing the path integral directly, we will use as basic tool the vacuum persistence amplitude $\langle 0,\infty|0,-\infty\rangle_J$: This quantity is the probability amplitude that a system under the influence of an external force J stays in its ground-state. Since we can apply an arbitrary force J, the amplitude $\langle 0,\infty|0,-\infty\rangle_J$ contains all information about the system. Moreover, it serves as a convenient tool to calculate Green functions which will become our main target studying quantum field theories.

2.1. Reminder of the operator approach

A classical system described by a Hamiltonian $H(q^i,p_i,t)$ can be quantised promoting q^i and p_i to operators[1] $\hat{q}^i$ and $\hat{p}_i$ which satisfy the canonical commutation relations $\left[\hat{q}^i,\hat{p}_j\right]=\mathrm{i}\delta^i_j$. The latter are the formal expression of Heisenberg's uncertainty relation. Apart from ordering ambiguities, the Hamilton operator $H(\hat{q}^i,\hat{p}_i,t)$ can be directly read from the Hamiltonian $H(q^i,p_i,t)$. The basic features of any quantum theory can be synthesised into a few principles.

General principles A physical system in a pure state is fully described by a probability amplitude

$$\psi(a,t)=\langle a|\psi(t)\rangle\in\mathbb{C}\,, \tag{2.1}$$

where $\{a\}$ is a set of quantum numbers specifying the system and the states $|\psi(t)\rangle$ form a complex Hilbert space. The probability p to find the specific values a_* in a measurement is given by $p(a_*)=|\psi(a_*,t)|^2$. The possible values a_* are the eigenvalues of Hermitian operators $\hat{A}$ whose eigenvectors $|a\rangle$ form an orthogonal, complete basis. In Dirac's bra-ket notation, we can express these statements by

$$\hat{A}|a\rangle=a|a\rangle\,,\qquad\langle a|a'\rangle=\delta\left(a-a'\right),\qquad\int\mathrm{d}a\,|a\rangle\langle a|=1\,, \tag{2.2}$$

In general, operators do not commute. Their commutation relations can be obtained by the replacement $\{A,B\}\to\mathrm{i}[\hat{A},\hat{B}]$ in the definition (1.24) of the Poisson brackets.

The state of a particle moving in one dimension in a potential $V(q)$ can be described either by the eigenstates of the position operator $\hat{q}$ or of the momentum operator $\hat{p}$. Both form a complete, orthonormal basis, and they are connected by a Fourier transformation which we

[1] When there is the danger of an ambiguity, operators will be written with a "hat"; otherwise we drop it.

choose to be asymmetric,

$$\psi(q) = \langle q|\,\psi\rangle = \int \frac{\mathrm{d}p}{2\pi}\,\mathrm{e}^{\mathrm{i}px}\,\psi(p) = \int \frac{\mathrm{d}p}{2\pi}\,\langle q|\,p\rangle\,\langle p|\,\psi\rangle \tag{2.3a}$$

$$\psi(p) = \langle p|\,\psi\rangle = \int \mathrm{d}q\,\mathrm{e}^{-\mathrm{i}kx}\,\psi(x) = \int \mathrm{d}q\,\langle p|\,q\rangle\,\langle q|\,\psi\rangle\,. \tag{2.3b}$$

Choosing this normalisation has the advantage that the factor $1/(2\pi)$ in the Fourier transformation is the same as in the density of free states, $\mathrm{d}pL/(2\pi)$, which will enter quantities like decat rates or cross sections. From Eq. (2.3), it follows that the asymmetry in the Fourier transformation is reflected in the completeness relation of the states,

$$\int \mathrm{d}q\,|q\rangle\,\langle q| = 1 \quad and \quad \int \frac{\mathrm{d}p}{2\pi}\,|p\rangle\,\langle p| = 1\,. \tag{2.4}$$

Time-evolution Since the states $|\psi(t)\rangle$ form a complex Hilbert space, the superposition principle is valid: If ψ_1 and ψ_2 are possible states of the system, then also

$$\psi(t) = c_1\psi_1(t) + c_2\psi_2(t)\,, \qquad c_i \in \mathbb{C}\,. \tag{2.5}$$

In quantum mechanics, a stronger version of this principle holds which states that if $\psi_1(t)$ and $\psi_2(t)$ describe the possible time-evolution of the system, then so does also the superposed state $\psi(t)$. This implies that the time-evolution is described by a linear, homogeneous differential equation. Choosing it as first-order in time, we can write the evolution equation as

$$\mathrm{i}\partial_t|\psi(t)\rangle = D|\psi(t)\rangle\,, \tag{2.6}$$

where the differential operator D on the RHS has to be still determined.

We call the operator describing the evolution of a state from $\psi(t)$ to $\psi(t')$ the time-evolution operator $U(t',t)$. This operator is unitary, $U^{-1} = U^\dagger$, in order to conserve probability and forms a group, $U(t_3,t_1) = U(t_3,t_2)U(t_2,t_1)$ with $U(t,t) = 1$. For an infinitesimal time step δt,

$$|\psi(t+\delta t)\rangle = U(t+\delta t,t)\,|\psi(t)\rangle\ , \tag{2.7}$$

we can set with $U(t,t) = 1$

$$U(t+\delta t,t) = 1 - \mathrm{i}H\delta t\,. \tag{2.8}$$

Here we introduced the generator of infinitesimal time-translations H. The analogy to classical mechanics suggests that H is the operator version of the classical Hamilton function $H(q,p)$. Inserting Eq. (2.8) into (2.7) results in

$$\frac{|\psi(t+\delta t)\rangle - |\psi(t)\rangle}{\delta t} = -\mathrm{i}H\,|\psi(t)\rangle\ . \tag{2.9}$$

Comparing then Eqs. (2.6) and (2.9) reveals that the operator D on the RHS of Eq. (2.6) coincides with the Hamiltonian H. We call a time-evolution equation of this type for arbitrary H Schrödinger equation.

Next we want to determine the connection between H and U. Plugging $\psi(t) = U(t,0)\psi(0)$ in the Schrödinger equation gives

$$\left[\mathrm{i}\frac{\partial U(t,0)}{\partial t} - HU(t,0)\right]\psi(0) = 0\,. \tag{2.10}$$

Since this equation is valid for an arbitrary state $\psi(0)$, we can rewrite it as an operator equation,

$$\mathrm{i}\partial_{t'}U(t',t) = HU(t',t)\,. \tag{2.11}$$

Integrating it, we find as formal solution

$$U(t',t) = 1 - \mathrm{i}\int_t^{t'} \mathrm{d}t'' H(t'')U(t'',t) \tag{2.12}$$

or, if H is time-independent,

$$U(t,t') = \exp(-\mathrm{i}H(t-t'))\,. \tag{2.13}$$

Up-to now, we have considered the Schrödinger picture where operators are constant and the time-evolution is given by the change in the state vectors $|\psi(t)\rangle$. In the Heisenberg picture, the time evolution is driven completely by the one of the operators. States and operators in the two pictures are connected by

$$O_{\mathrm{S}}(t) = U(t,t_0)O_{\mathrm{H}}(t)U^\dagger(t,t_0)\,, \tag{2.14a}$$

$$|\psi_{\mathrm{S}}(t)\rangle = U(t,t_0)\,|\psi_{\mathrm{H}}(t)\rangle\,, \tag{2.14b}$$

if they agree at the time t_0.

Propagator We insert the solution of U for a time-independent H into $|\psi(t')\rangle = U(t',t)|\psi(t)\rangle$ and multiply from the left with $\langle q'|$,

$$\psi(q',t') = \langle q'|\psi(t')\rangle = \langle q'|\exp[-\mathrm{i}H(t'-t)]|\psi(t)\rangle\,. \tag{2.15}$$

Then we insert $1 = \int \mathrm{d}^3q|q\rangle\langle q|$,

$$\psi(q',t') = \int \mathrm{d}^3q\,\langle q'|\exp[-\mathrm{i}H(t'-t)]|q\rangle\langle q|\psi(t)\rangle = \int \mathrm{d}^3q\,K(q',t';q,t)\psi(q,t)\,. \tag{2.16}$$

In the last step we introduced the propagator or Green function K in its coordinate representation,

$$K(q',t';q,t) = \langle q'|\exp[-\mathrm{i}H(t'-t)]|q\rangle\,. \tag{2.17}$$

The Green function K equals the probability amplitude for the propagation between two space-time points; $K(q',t';q,t)$ is therefore also called more specifically two-point Green function. We can express the propagator K by the solutions of the Schrödinger equation, $\psi_n(q,t) = \langle q|n(t)\rangle = \langle q|n\rangle\exp(-\mathrm{i}E_nt)$ as

$$\begin{aligned} K(q',t';q,t) &= \sum_{n,n'}\langle q'|n\rangle\underbrace{\langle n|\exp(-\mathrm{i}H(t'-t))|n'\rangle}_{\delta_{n,n'}\exp(-\mathrm{i}E_n(t'-t))}\langle n'|q\rangle \\ &= \sum_n \psi_n(q')\psi_n^*(q)\exp(-\mathrm{i}E_n(t'-t))\,, \end{aligned} \tag{2.18}$$

where n represents the complete set of quantum numbers specifying the energy eigenvalues of the system. Note that this result is very general and holds for any time-independent Hamiltonian.

Let us compute the propagator of a free particle in one dimension, described by the Hamiltonian $H = p^2/2m$. We write with $\tau = t' - t$

$$\begin{aligned} K(q',t';q,t) &= \langle q'|\,\mathrm{e}^{-\mathrm{i}H\tau}\,|q\rangle = \langle q'|\,\mathrm{e}^{-\mathrm{i}\tau\hat{p}^2/2m}\int\frac{\mathrm{d}p}{2\pi}\,|p\rangle\,\langle p|\,q\rangle \\ &= \int\frac{\mathrm{d}p}{2\pi}\,\mathrm{e}^{-\mathrm{i}\tau p^2/2m}\,\langle q'|\,p\rangle\,\langle p|\,q\rangle = \int\frac{\mathrm{d}p}{2\pi}\,\mathrm{e}^{-\mathrm{i}\tau p^2/2m+\mathrm{i}(q'-q)p}\,, \end{aligned} \tag{2.19}$$

where we used $\langle q'|\,p\rangle = \exp(\mathrm{i}q'p)$ in the last step. The integral is Gaussian, if we add an infinitesimal factor $\exp(-\varepsilon p^2)$ to the integrand in order to ensure the convergence of the integral. Thus the physical value of the energy $E = p^2/(2m)$ seen as a complex variable is approached from the negative imaginary plane, $E \to E - \mathrm{i}\varepsilon$. Taking afterwards the limit $\varepsilon \to 0$, we obtain

$$K(q',t';q,t) = \left(\frac{m}{2\pi\mathrm{i}\tau}\right)^{1/2}\mathrm{e}^{\mathrm{i}m(q'-q)^2/2\tau}\,. \tag{2.20}$$

Knowing the propagator, we can calculate the solution $\psi(t')$ at any time t' for a given initial state $\psi(t)$ via Eq. (2.16).

Example 2.1: Calculate the integrals $A = \int\mathrm{d}x\exp(-x^2/2)$, $B = \int\mathrm{d}x\exp(-ax^2/2 + bx)$, and $C = \int\mathrm{d}x\cdots\mathrm{d}x_n\exp(-x^TAx/2 + J^Tx)$ for a symmetric $n \times n$ matrix A.
a.) We square the integral and calculate then A^2 introducing polar coordinates, $r^2 = x^2 + y^2$,

$$A^2 = \int_{-\infty}^{\infty}\mathrm{d}x\int_{-\infty}^{\infty}\mathrm{d}y\exp(-(x^2+y^2)/2) = 2\pi\int_0^{\infty}\mathrm{d}r\,r\mathrm{e}^{-r^2/2} = 2\pi\int_0^{\infty}\mathrm{d}t\,\mathrm{e}^{-t} = 2\pi\,,$$

where we substituted $t = r^2/2$. Thus the result for the basic Gaussian integral is $A = \sqrt{2\pi}$. All other solvable variants of Gaussian integrals can be reduced to this result.
b.) We complete the square in the exponent,

$$-\frac{a}{2}\left(x^2 - \frac{2b}{a}x\right) = -\frac{a}{2}\left(x - \frac{b}{a}\right)^2 + \frac{b^2}{2a}\,,$$

and shift then the integration variable to $x' = x - b/a$. The result is

$$B = \int_{-\infty}^{\infty}\mathrm{d}x\exp(-ax^2/2 + bx) = \mathrm{e}^{b^2/2a}\int_{-\infty}^{\infty}\mathrm{d}x'\exp(-ax'^2/2) = \sqrt{\frac{2\pi}{a}}\,\mathrm{e}^{b^2/2a}\,. \tag{2.21}$$

c.) We should complete again the square and try $X' = X - A^{-1}J$. With

$$\begin{aligned} (X - A^{-1}J)^TA(X - A^{-1}J) &= X^TAX - X^TAA^{-1}J - J^TA^{-1}AX + J^TA^{-1}AA^{-1}J \\ &= X^TAX - 2J^TX + J^TA^{-1}J\,, \end{aligned}$$

we obtain after shifting the integration vector,

$$C = \int\mathrm{d}x_1\cdots\mathrm{d}x_n\exp(-X^TAX/2 + J^TX) = \exp(J^TA^{-1}J/2)\int\mathrm{d}x_1'\cdots\mathrm{d}x_n'\exp(-X'^TAX'/2)\,.$$

Since the matrix A is symmetric, we can diagonalise A via an orthogonal transformation $D = OAO^T$. This corresponds to a rotation of the integration variables, $Y = OX'$. The Jacobian of this transformation is one, and thus the result is

$$C = \exp(J^TA^{-1}J/2)\prod_{i=1}^{n}\int\mathrm{d}y_i\exp(-a_iy_i^2/2) = \sqrt{\frac{(2\pi)^n}{\det A}}\,\exp\left(\frac{1}{2}J^TA^{-1}J\right)\,. \tag{2.22}$$

In the last step we expressed the product of eigenvalues a_i as the determinant of A.

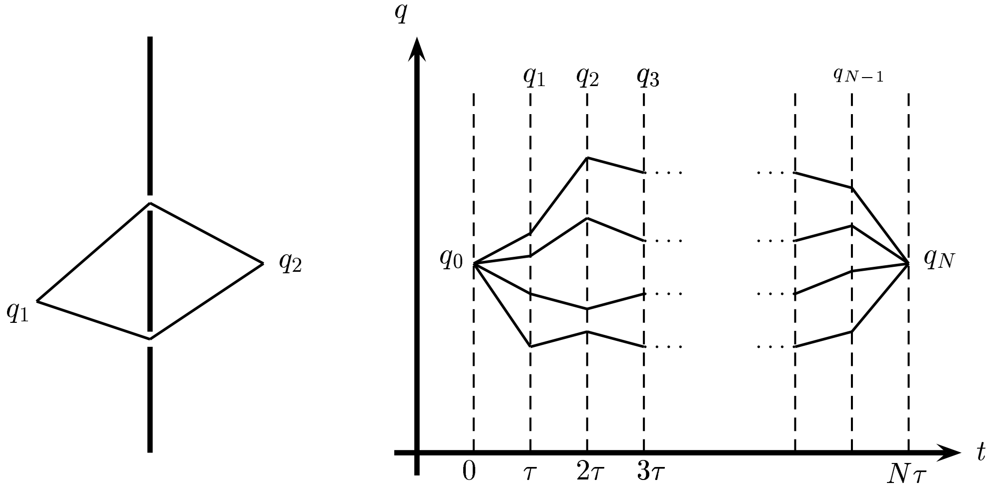


Figure 2.1.: Left: The double slit experiment. Right: The propagator $K(q_N,\tau;q_0,0)$ expressed as a sum over all N-legged continuous paths.

2.2. Path integrals in quantum mechanics

In problem 2.?? you are asked to calculate the classical action of a free particle and of a harmonic oscillator and to compare them to the corresponding propagators found in quantum mechanics. Surprisingly, you will find that in both cases the propagator can be written as $K(q',t';q,t) = N\,\exp(\mathrm{i}S)$ where S is the classical action along the path $[q(t):q'(t')]$ and N a normalisation constant. This suggests that we can reformulate quantum mechanics, replacing the standard operator formalism used to evaluate the propagator (2.17) "somehow" by the classical action.

To get an idea how to proceed, we look at the famous double-slit experiment sketched in Fig. 2.1: According to the superposition principle, the amplitude A for a particle to move from the source at q_1 to the detector at q_2 is the sum of the amplitudes A_i for the two possible paths,

$$A = K(q_2,t_2;q_1,t_1) = \sum_{\text{paths}} A_i\,. \tag{2.23}$$

Clearly, we could add in a Gedankenexperiment more and more screens between q_1 and q_2, increasing at the same time the number of holes. Although we replace in this way continuous space-time by a discrete lattice, the differences between these two descriptions should vanish for sufficiently small spacing τ. Moreover, for $\tau\to 0$, we can expand $U(\tau)=\exp(-\mathrm{i}H\tau)\simeq 1-\mathrm{i}H\tau$. Applying then $H=\hat p^2/(2m)+V(\hat q)$ to eigenfunctions $|q\rangle$ of $V(\hat q)$ and $|p\rangle$ of $\hat p^2$, we can replace the operator H by its eigenvalues. In this way, we hope to express the propagator as a sum over paths, where the individual amplitudes A_i contain only classical quantities.

We apply now this idea to a particle moving in one dimension in a potential $V(q)$. The transition amplitude A for the evolution from the state $|q,0\rangle$ to the state $|q',t'\rangle$ is

$$A \equiv K(q',t';q,0) = \left\langle q'\right|\mathrm{e}^{-\mathrm{i}Ht'}\left|q\right\rangle\,. \tag{2.24}$$

This amplitude equals the matrix-element of the propagator K for the evolution from the initial point $q(0)$ to the final point $q'(t')$. Let us split the time evolution into two smaller steps, writing $\mathrm{e}^{-\mathrm{i}Ht'} = \mathrm{e}^{-\mathrm{i}H(t'-t_1)}\mathrm{e}^{-\mathrm{i}Ht_1}$. Inserting also $\int \mathrm{d}q_1 \left|q_1\right\rangle \left\langle q_1\right| = 1$, the amplitude becomes

$$A = \int \mathrm{d}q_1 \left\langle q'\right| e^{-\mathrm{i}H(t'-t_1)} \left|q_1\right\rangle \left\langle q_1\right| \mathrm{e}^{-\mathrm{i}Ht_1} \left|q\right\rangle = \int \mathrm{d}q_1 \, K(q', t'; q_1, t_1) K(q_1, t_1; q, 0). \tag{2.25}$$

This formula expresses simply the group property, $U(t', 0) = U(t', t_1)U(t_1, 0)$, of the time-evolution operator U evaluated in the basis of the continuous variable q. More physically, we can view this equation as an expression of the quantum mechanical rule for combining amplitudes: If the same initial and final states can be connected by various ways, the amplitudes for each of these processes should be added. A particle propagating from q to q' must be somewhere at the intermediate time t_1. Labelling this intermediate position as q_1, we compute the amplitude for propagation via the point q_1 as the product of the two propagators in Eq. (2.25) and integrate over all possible intermediate positions q_1.

We continue to divide the time interval t' into a large number N of time intervals of duration $\tau = t'/N$. Then the propagator becomes

$$A = \left\langle q'\right| \underbrace{\mathrm{e}^{-\mathrm{i}H\tau}\mathrm{e}^{-\mathrm{i}H\tau} \cdots \mathrm{e}^{-\mathrm{i}H\tau}}_{N \text{ times}} \left|q\right\rangle . \tag{2.26}$$

We insert again a complete set of states $\left|q_i\right\rangle$ between each exponential, obtaining

$$\begin{aligned} A &= \int \mathrm{d}q_1 \cdots \mathrm{d}q_{N-1} \left\langle q'\right| \mathrm{e}^{-\mathrm{i}H\tau} \left|q_{N-1}\right\rangle \left\langle q_{N-1}\right| \mathrm{e}^{-\mathrm{i}H\tau} \left|q_{N-2}\right\rangle \cdots \left\langle q_1\right| \mathrm{e}^{-\mathrm{i}H\tau} \left|q\right\rangle \\ &\equiv \int \mathrm{d}q_1 \cdots \mathrm{d}q_{N-1} K_{q_N,q_{N-1}} K_{q_{N-1},q_{N-2}} \cdots K_{q_2,q_1} K_{q_1,q_0} , \end{aligned} \tag{2.27}$$

where we have defined $q_0 = q$ and $q_N = q'$. Note that these initial and final positions are fixed and therefore are not integrated over. Figure 2.1 illustrates that we can view the amplitude A as the integral over the partial amplitudes A_{path} of the individual N-legged continuous paths.

We ignore the problem of defining properly the limit $N \to \infty$, keeping N large but finite. We rewrite the amplitude as sum over the amplitudes for all possible paths, $A = \sum_{\text{paths}} A_{\text{path}}$, with

$$\sum_{\text{paths}} = \int dq_1 \cdots dq_{N-1}, \qquad A_{\text{path}} = K_{q_N,q_{N-1}} K_{q_{N-1},q_{N-2}} \cdots K_{q_2,q_1} K_{q_1,q_0} .$$

Let us look at the last expression in detail. We can expand the exponential in each propagator $K_{q_{j+1},q_j} = \left\langle q_{j+1}\right| \mathrm{e}^{-\mathrm{i}H\tau} \left|q_j\right\rangle$ for a single sub-interval, because τ is small,

$$\begin{aligned} K_{q_{j+1},q_j} &= \left\langle q_{j+1}\right| \left(1 - \mathrm{i}H\tau - \frac{1}{2}H^2\tau^2 + \cdots\right) \left|q_j\right\rangle \\ &= \left\langle q_{j+1} \middle| q_j\right\rangle - \mathrm{i}\tau \left\langle q_{j+1}\right| H \left|q_j\right\rangle + \mathcal{O}(\tau^2) . \end{aligned} \tag{2.28}$$

In the second term of (2.28), we insert a complete set of momentum eigenstates between H and $\left|q_j\right\rangle$. This gives

$$\begin{aligned} &- \mathrm{i}\tau \left\langle q_{j+1}\right| \left(\frac{\hat{p}^2}{2m} + V(\hat{q})\right) \int \frac{\mathrm{d}p_j}{2\pi} \left|p_j\right\rangle \left\langle p_j \middle| q_j\right\rangle \\ &= -\mathrm{i}\tau \int \frac{\mathrm{d}p_j}{2\pi} \left(\frac{p_j^2}{2m} + V(q_{j+1})\right) \left\langle q_{j+1} \middle| p_j\right\rangle \left\langle p_j \middle| q_j\right\rangle \\ &= -\mathrm{i}\tau \int \frac{\mathrm{d}p_j}{2\pi} \left(\frac{p_j^2}{2m} + V(q_{j+1})\right) \mathrm{e}^{\mathrm{i}p_j(q_{j+1}-q_j)} . \end{aligned} \tag{2.29}$$

The expression (2.29) is not symmetric in q_j and q_{j+1}. The reason for this asymmetry is that we could have inserted the factor 1 either to the right or to the left of the Hamiltonian H. In the latter case, we would have obtained p_{j+1} and $V(q_j)$ in (2.29). Since the difference $[V(q_{j+1}) - V(q_j)]\tau \simeq V'(q_j)(q_{j+1}-q_j)\tau \simeq V'(q_j)\dot{q}_j\tau^2$ is of order τ^2, the ordering problem should not matter in the continuum limit which we will take eventually; we set therefore $V(q_{j+1}) \simeq V(q_j)$.

The first term of (2.28) gives a delta function, which we can express as

$$\langle q_{j+1}|\, q_j\rangle = \delta(q_{j+1}-q_j) = \int \frac{\mathrm{d}p_j}{2\pi}\, \mathrm{e}^{\mathrm{i}p_j(q_{j+1}-q_j)}\,. \tag{2.30}$$

Now we can combine the two terms, obtaining as propagator for the step $q_j \to q_{j+1}$

$$K_{q_{j+1},q_j} = \int \frac{\mathrm{d}p_j}{2\pi}\, \mathrm{e}^{\mathrm{i}p_j(q_{j+1}-q_j)} \left[1 - \mathrm{i}\tau\left(\frac{{p_j}^2}{2m} + V(q_j)\right) + \mathcal{O}(\tau^2)\right]. \tag{2.31}$$

Since we work at $\mathcal{O}(\tau)$, we can exponentiate the factor in the square bracket,

$$1 - \mathrm{i}\tau\; H(p_j,q_j) + \mathcal{O}(\tau^2) = \mathrm{e}^{-\mathrm{i}\tau H(p_j,q_j)}\,. \tag{2.32}$$

Next we rewrite the exponent in the first factor of Eq. (2.31) using $\dot{q}_j = (q_{j+1}-q_j)/\tau$, such that we can factor out the time-interval τ. The amplitude A_{path} consists of N such factors. Combining them, we obtain

$$A_{\text{path}} = \prod_{j=0}^{N-1} \int \frac{\mathrm{d}p_j}{2\pi} \exp \mathrm{i}\tau \sum_{j=0}^{N-1} [p_j\dot{q}_j - H(p_j,q_j)]\,. \tag{2.33}$$

We recognise the argument of the exponential as the discrete approximation of the action $S[q,p]$ in the Palatini form of a path passing through the points $q_0 = q, q_1, \cdots, q_{N-1}, q_N = q'$. The propagator $K = \int \mathrm{d}q_1 \cdots \mathrm{d}q_{N-1} A_{\text{path}}$ becomes then

$$K = \prod_{j=1}^{N-1} \int \mathrm{d}q_j \prod_{j=0}^{N-1} \int \frac{\mathrm{d}p_j}{2\pi} \exp \mathrm{i}\tau \sum_{j=0}^{N-1} [p_j\dot{q}_j - H(p_j,q_j)]\,. \tag{2.34}$$

For $N \to \infty$, this expression approximates an integral over all functions $p(t)$, $q(t)$ consistent with the boundary conditions $q(0) = q$, $q(t') = q'$. We adopt the notation $\mathcal{D}p\mathcal{D}q$ for the *functional* or *path integral* over all functions $p(t)$ and $q(t)$,

$$K \equiv \int \mathcal{D}p(t)\mathcal{D}q(t)\mathrm{e}^{\mathrm{i}S[q,p]} = \int \mathcal{D}p(t)\mathcal{D}q(t) \exp\left(\mathrm{i}\int_0^{t'} \mathrm{d}t\,(p\dot{q} - H(p,q))\right). \tag{2.35}$$

This result expresses the propagator as a path integral in phase-space. It allows us to obtain for any classical system which can be described by a Hamiltonian the corresponding quantum dynamics.

If the Hamiltonian is of the form $H = p^2/2m + V$, as we have assumed[2] in our derivation, we can carry out the quadratic momentum integrals in (2.34). We can rewrite this expression as

$$K = \prod_{j=1}^{N-1} \int \mathrm{d}q_j \exp -\mathrm{i}\tau \sum_{j=0}^{N-1} V(q_j) \prod_{j=0}^{N-1} \int \frac{\mathrm{d}p_j}{2\pi} \exp \mathrm{i}\tau \sum_{j=0}^{N-1} \left(p_j\dot{q}_j - p_j^2/2m\right).$$

[2] Since we evaluated $\exp(-\mathrm{i}H\tau)$ for infinitesimal τ, the result (2.35) holds also for a time-dependent potential $V(q,t)$.

The p integrals are all uncoupled Gaussians. One such integral gives

$$\int \frac{\mathrm{d}p}{2\pi}\,\mathrm{e}^{\mathrm{i}\tau(p\dot{q}-p^2/2m)} = \sqrt{\frac{m}{2\pi\mathrm{i}\tau}}\,\mathrm{e}^{\mathrm{i}\tau m\dot{q}^2/2}\,, \tag{2.36}$$

where we added again an infinitesimal factor $\exp(-\varepsilon p^2)$ to the integrand. Thence the propagator becomes

$$\begin{aligned} K &= \prod_{j=1}^{N-1}\int \mathrm{d}q_j \exp -\mathrm{i}\tau \sum_{j=0}^{N-1} V(q_j) \prod_{j=0}^{N-1}\left(\sqrt{\frac{m}{2\pi\mathrm{i}\tau}}\exp \mathrm{i}\tau\frac{m\dot{q}_j^2}{2}\right) \\ &= \left(\frac{m}{2\pi\mathrm{i}\tau}\right)^{N/2}\prod_{j=1}^{N-1}\int \mathrm{d}q_j \exp \mathrm{i}\tau\sum_{j=0}^{N-1}\left(\frac{m\dot{q}_j^2}{2} - V(q_j)\right)\,. \end{aligned} \tag{2.37}$$

The argument of the exponential is again a discrete approximation of the action $S[q]$ of a path passing through the points $q_0 = q, q_1, \cdots, q_{N-1}, q_N = q'$, but now seen as functional of only the coordinate q. As above, we can write this in a more compact form as

$$K = \langle q_f, t_f | q_i, t_i\rangle = \int \mathcal{D}q(t)\mathrm{e}^{\mathrm{i}S[q]} = \int \mathcal{D}q(t)\exp\left(\mathrm{i}\int_{t_i}^{t_f}\mathrm{d}t\, L(q,\dot{q})\right)\,, \tag{2.38}$$

where the integration includes all paths satisfying the boundary condition $q(t_i) = q_i$ and $q(t_f) = q_f$. This is the main result of this section, and is known as the *path integral in configuration space.* It will serve us as starting point discussing quantum field theories of bosonic fields.

Knowing the path integral and thus the propagator is sufficient to solve scattering problems in quantum mechanics. In a relativistic theory, the particle number during the course of a scattering process is however not fixed, since energy can be converted into matter. In order to prepare us for such more complex problems, we will generalise in the next section the path integral to a *generating functional* for *n-point Green functions*: In this formalism, the usual propagator giving the probability amplitude that a single particle moves from $q_i(t_i)$ to $q_f(t_f)$ becomes the special case of a 2-point Green function, while Green functions with $n > 2$ describe processes involve more points. For instance, the 4-point Green function will be the essential ingredient to calculate 2 → 2 scattering processes in a quantum field theory (QFT). The corresponding generating functional is the quantity which n.th derivative returns the n-point Green functions.

2.3. Generating functional for Green functions

Having re-expressed the transition amplitude $\langle q_f, t_f | q_i, t_i\rangle$ of a quantum mechanical system as a path integral, we want to generalise first this result to the matrix elements of an arbitrary potential $V(q)$ between the states $|q_i, t_i\rangle$ and $|q_f, t_f\rangle$. For all practical purposes, we can assume that we can expand $V(q)$ as a power-series in q; thus it is sufficient to consider the matrix elements $\langle q_f, t_f | q^m | q_i, t_i\rangle$. In a QFT, the initial and final states are generally free particles which are described mathematically as harmonic oscillators. In this case, we are able to reconstruct all excited states $|n\rangle$ from the ground-state,

$$|n\rangle = \frac{1}{\sqrt{n!}}\,(a^\dagger)^n\,|0\rangle\,.$$

Therefore it will be sufficient to study matrix elements between the ground-state $|0\rangle$. With this choice, we are able to extend the integration limit in the path integral (2.38) to $t = \pm\infty$. This will not only simplify its evaluation, but also avoid the need to choose a specific inertial frame. As a result, the generating functional will have an obviously Lorentz invariant form in a relativistic theory.

Time-ordered products of operators and the path integral In a first step, we try to include the operator q^m into the transition amplitude $\langle q_f, t_f | q_i, t_i\rangle$. We can reinterpret our result for the path integral as follows,

$$\langle q_f, t_f | \mathbf{1} | q_i, t_i\rangle = \int \mathcal{D}q(t)\, 1\, \mathrm{e}^{\mathrm{i}S[q]} \,. \tag{2.39}$$

Thus we can see the LHS as matrix element of the unit operator $\mathbf{1}$, while the RHS corresponds to the path integral average of the classical function $f(q, \dot{q}) = 1$. Now we want to generalise this rather trivial statement to two operators $\hat{A}(t_a)$ and $\hat{B}(t_b)$ given in the Heisenberg picture. In evaluating the unknown function f on the RHS of

$$\int \mathcal{D}q(t)\; A(t_a)B(t_b)\, \mathrm{e}^{\mathrm{i}S[q(t)]} = \langle q_f, t_f |\; f\{A(t_a)B(t_b)\}\; | q_i, t_i\rangle \,, \tag{2.40}$$

we go back to Eq. (2.27) and insert $\hat{A}(t_a)$ and $\hat{B}(t_b)$ at the correct intermediate times,

$$= \begin{cases} \int \mathrm{d}q_1 \cdots \mathrm{d}q_{N-1} \ldots \langle q_{a+1}, t_{a+1} |\, \hat{A}\, | q_a, t_a\rangle \cdots \langle q_{b+1}, t_{b+1} |\, \hat{B}\, | q_b, t_b\rangle \ldots & \text{f}or \quad t_a > t_b \,, \\ \int \mathrm{d}q_1 \cdots \mathrm{d}q_{N-1} \ldots \langle q_{b+1}, t_{b+1} |\, \hat{B}\, | q_b, t_b\rangle \cdots \langle q_{a+1}, t_{a+1} |\, \hat{A}\, | q_a, t_a\rangle \ldots & \text{f}or \quad t_a < t_b \,. \end{cases} \tag{2.41}$$

Since the time along a classical path increases, the matrix elements of the operators $\hat{A}(t_a)$ and $\hat{B}(t_b)$ are also ordered with time increasing from the left to the right. If we define the *time-ordered product* of two operators as

$$T\{\hat{A}(t_a)\hat{B}(t_b)\} = \hat{A}(t_a)\hat{B}(t_b)\vartheta(t_a - t_b) + \hat{B}(t_b)\hat{A}(t_a)\vartheta(t_b - t_a) \,, \tag{2.42}$$

then the path integral average of the classical quantities $A(t_a)$ and $B(t_b)$ corresponds to the matrix-element of the time-ordered product of these two operators,

$$\langle q_f, t_f | T\{\hat{A}(t_a)\hat{B}(t_b)\} | q_i, t_i\rangle = \int \mathcal{D}q(t)\; A(t_a)B(t_b)\, \mathrm{e}^{\mathrm{i}S[q(t)]} \,, \tag{2.43}$$

and similar for more than two operators.

External sources We want to include next in our formalism the possibility that we can change the state of our system by applying an external driving force or source term $J(t)$. In quantum mechanics, we could imagine e.g. a harmonic oscillator in the ground-state $|0\rangle$, making a transition under the influence of an external force J to the state $|n\rangle$ at the time t and back to the ground-state $|0\rangle$ at the time $t' > t$. Including such transitions, we can mimic the relativistic process of particle creation and annihilation as follows: We identify the vacuum (i.e. the state containing zero real particles) with the ground-state of the quantum mechanical system, and the creation and annihilation of n particle with the (de-) excitation of the n.th energy level by an external source J. Schwinger realised that adding a linear coupling to an external source,

$$L \to L + J(t)q(t) \,, \tag{2.44}$$

leads also to an efficient way to calculate the matrix elements of an arbitrary polynomial of operators $q(t_n)\cdots q(t_1)$: If the source $J(t)$ would be a simple number instead of a time-dependent function in the augmented path-integral

$$\langle q_f,t_f|q_i,t_i\rangle_J \equiv \int \mathcal{D}q(t)\mathrm{e}^{\mathrm{i}\int_{t_i}^{t_f}\mathrm{d}t(L+Jq)}\,, \tag{2.45}$$

then we could obtain $\langle q_f,t_f|q^m|q_i,t_i\rangle_J$ simply by differentiating $\langle q_f,t_f|q_i,t_i\rangle_J$ m-times with respect to J. However, the LHS is a functional of $J(t)$ and thus we need to perform instead functional derivatives with respect to $J(t)$. By analogy with the rules for the differentiation of functions, e.g. $\partial x^l/\partial x^k=\delta^l_k$, we define[3] a functional derivative as

$$\frac{\delta}{\delta J(x)}1=0 \quad \mathrm{a}nd \qquad \frac{\delta J(x)}{\delta J(x')}=\delta(x-x')\,. \tag{2.46}$$

Thus we replace for a continuous index the Kronecker delta by a delta function. Moreover, we assume that the Leibniz and the chain rule holds for sufficiently nice functions $J(x)$.

Now we are able to differentiate $\langle q_f,t_f|q_i,t_i\rangle_J$ with respect to the source J. Starting from

$$\frac{\delta}{\delta J(t_1)}\int \mathcal{D}q(t)\,\mathrm{e}^{\mathrm{i}\int_{-\infty}^{\infty}\mathrm{d}tJ(t)q(t)}=\mathrm{i}\int \mathcal{D}q(t)\,q(t_1)\mathrm{e}^{\mathrm{i}\int_{-\infty}^{\infty}\mathrm{d}tJ(t)q(t)}\,, \tag{2.47}$$

we obtain

$$\langle q_f,t_f|T\{\hat q(t_1)\cdots\hat q(t_n)\}|q_i,t_i\rangle=(-\mathrm{i})^n\frac{\delta^n}{\delta J(t_1)\cdots\delta J(t_n)}\langle q_f,t_f|q_i,t_i\rangle_J\bigg|_{J=0}\,. \tag{2.48}$$

Thus the source $J(t)$ is a convenient tool to obtain the functions $q(t_1)\cdots q(t_n)$ in front of $\exp(\mathrm{i}S)$. Having performed the functional derivatives, we set the source $J(t)$ to zero, coming back to the usual path integral. Physically, the expression (2.48) corresponds to the probability amplitude that a particles moves from $q_i(t_i)$ to $q_f(t_f)$, having the intermediate positions $q(t_1),\ldots,q(t_n)$.

Vacuum persistence amplitude As last step, we want to eliminate the initial and final states $|q_i,t_i\rangle_J$ and $|q_f,t_f\rangle$ in favour of the ground state or vacuum, $|0\rangle$. In this way, we convert the transition amplitude $\langle q_f,t_f|q_i,t_i\rangle_J$ into the probability amplitude that a system which was in the ground-state $|0\rangle$ at $t_i\to-\infty$ remains in this state at $t_f\to\infty$ despite the action of the source $J(t)$. Inserting a complete set of energy eigenstates, $1=\sum_n|n\rangle\langle n|$, into the propagator, we obtain

$$\langle q',t'|q,t\rangle=\sum_n\psi_n(q')\psi_n^*(q)\exp(-\mathrm{i}E_n(t'-t))\,. \tag{2.49}$$

We can isolate the ground-state $n=0$ by adding either to the energies E_n or to the time difference $\tau=(t'-t)$ a small negative imaginary part. In this case, all terms are exponentially damped in the limit $\tau\to\infty$, and the ground-state as state with the smallest energy dominates more and more the sum. Alternatively, we can add a term $+\mathrm{i}\varepsilon q^2$ to the Lagrangian.

[3] As the notation suggest, the variation of a functional defined in Eq. (1.5) is the special case of a *directional* functional derivative, cf. problem 1.??.

Remark 2.1: Wick rotation and Euclidean action:
Instead of adding the infinitesimal small term $\mathrm{i}\varepsilon q^2$ to the Lagrangian, we can do a more drastic change, rotating in the action the time axis clockwise by 90 degrees in the complex plane. Inserting $t_E = \mathrm{i}t$ into $x_\mu x^\mu$, we see that this procedure called Wick rotation corresponds to the transition from Minkowski to Euclidean space,

$$x^2 = t^2 - \boldsymbol{x}^2 = (-\mathrm{i}t_E)^2 - \boldsymbol{x}^2 = -[t_E^2 + \boldsymbol{x}^2] = -x_E^2 \,.$$

Performing the changes $t = -\mathrm{i}t_E$ and $\mathrm{d}t = -\mathrm{i}\mathrm{d}t_E$ in the action of a particle moving in an one-dimensional potential gives

$$S = -\mathrm{i}\int \mathrm{d}t_E \left(-\frac{1}{2}m\dot{q}_E^2 - V(q)\right) \equiv \mathrm{i}S_E \,. \tag{2.50}$$

Note that the Euclidean action $S_E = T + V$ is bounded from below. The phase factor in the path-integral transforms as $\mathrm{e}^{\mathrm{i}S} = \mathrm{e}^{-S_E}$, and thus contributions with large S_E are exponentially damped in the Euclidean path-integral.

Finally, we have only to connect the results we obtained so far. Adding a coupling to an external source $J(t)$ and a damping factor $+\mathrm{i}\varepsilon q^2$ to the Lagrangian gives us the ground-state or vacuum persistence amplitude

$$Z[J] \equiv \langle 0, \infty | 0, -\infty\rangle_J = \int \mathcal{D}q(t)\, \mathrm{e}^{\mathrm{i}\int_{-\infty}^{\infty} \mathrm{d}t (L + Jq + \mathrm{i}\varepsilon q^2)} \tag{2.51}$$

in the presence of a classical source J. This amplitude is a functional of J which we denote by $Z[J]$. Taking derivatives w.r.t. the external sources J, and setting them afterwards to zero, we obtain

$$\left.\frac{\delta^n Z[J]}{\delta J(t_1)\cdots\delta J(t_n)}\right|_{J=0} = \mathrm{i}^n \int \mathcal{D}q(t)\; q(t_1)\cdots q(t_n) \mathrm{e}^{\mathrm{i}\int_{-\infty}^{\infty}\mathrm{d}t (L+\mathrm{i}\varepsilon q^2)} \,. \tag{2.52}$$

The RHS corresponds to the path integral in Eqs. (2.43), augmented by the factor $\mathrm{i}\varepsilon q^2$. But this factor damps in the limit of large t everything except the ground state. Thus we found that $Z[J]$ is the generating functional for the vacuum expectation value of the time-ordered product of operators $\hat{q}(t_i)$,

$$(-\mathrm{i})^n \left.\frac{\delta^n Z[J]}{\delta J(t_1)\cdots\delta J(t_n)}\right|_{J=0} = \langle 0, \infty | T\{\hat{q}(t_1)\cdots\hat{q}(t_n)\} | 0, -\infty\rangle = G_F(t_1, \ldots, t_n) \,. \tag{2.53}$$

In the last step, we defined also the n-point Green function $G_F(t_1, \ldots, t_n)$. The subscript F indicates that the $\mathrm{i}\varepsilon q^2$ prescription selects from the set of possible Green functions (retarded, advanced, ...) the ones suggested by Feynman. These functions will be the main building block we will use to perform calculations in quantum field theory, and the formula corresponding to Eq. (2.53) will be our master formula in field theory.

2.4. Oscillator as a one-dimensional field theory

Canonical quantisation A one-dimensional harmonic oscillator can be viewed as a free quantum field theory in one time and zero space dimensions. In order to exhibit this equivalence clearer, we rescale the usual Lagrangian

$$L(x, \dot{x}) = \frac{1}{2}m\dot{x}^2 - \frac{1}{2}m\omega^2 x^2 \,, \tag{2.54}$$

where m is the mass of the oscillator and ω its frequency as

$$\phi(t) \equiv \sqrt{m}x(t)\,. \tag{2.55}$$

We call the variable $\phi(t)$ a "scalar field," and the Lagrangian now reads

$$L(\phi,\dot{\phi}) = \frac{1}{2}\dot{\phi}^2 - \frac{1}{2}\omega^2\phi^2\,. \tag{2.56}$$

After the rescaling, the kinetic term $\dot{\phi}^2$ has the dimensionless coefficient 1/2. This choice is standard in field theory and therefore such a field is called "canonically normalised."

We derive the corresponding Hamiltonian, determining first the conjugate momentum π as $\pi(t) = \partial L/\partial\dot{\phi} = \dot{\phi}(t)$. Thus the classical Hamiltonian follows as

$$H(\phi,\pi) = \frac{1}{2}\pi^2 + \frac{1}{2}\omega^2\phi^2\,. \tag{2.57}$$

The transition to quantum mechanics is performed by promoting ϕ and π to operators which satisfy the canonical commutation relations $[\phi,\pi] = \mathrm{i}$. The harmonic oscillator is solved most efficiently introducing creation and annihilation operators, $a^\dagger$ and a. They are defined by

$$\phi = \frac{1}{\sqrt{2\omega}}\left(a^\dagger + a\right) \quad and \quad \pi = \mathrm{i}\sqrt{\frac{\omega}{2}}\left(a^\dagger - a\right)\,, \tag{2.58}$$

and satisfy $\left[a,a^\dagger\right] = 1$. The Hamiltonian follows as

$$H = \frac{\omega}{2}\left(aa^\dagger + a^\dagger a\right) = \left(a^\dagger a + \frac{1}{2}\right)\omega\,. \tag{2.59}$$

We interpret $N \equiv a^\dagger a$ as the number operator, counting the number n of quanta with energy ω in the state $|n\rangle$.

We now work in the Heisenberg picture where operators are time dependent. The time evolution of the operator $a(t)$ can be found from the Heisenberg equation,

$$\mathrm{i}\frac{\mathrm{d}a}{\mathrm{d}t} = [a,H] = \omega a\,, \tag{2.60}$$

from which we deduce that

$$a(t) = a(0)\mathrm{e}^{-\mathrm{i}\omega t} = a_0\mathrm{e}^{-\mathrm{i}\omega t}\,. \tag{2.61}$$

As a consequence, the field operator $\phi(t)$ can be expressed in terms of the creation and annihilation operators as

$$\phi(t) = \frac{1}{\sqrt{2\omega}}\left(a_0\mathrm{e}^{-\mathrm{i}\omega t} + a_0^\dagger\mathrm{e}^{\mathrm{i}\omega t}\right)\,. \tag{2.62}$$

If we look at $\phi(t)$ as a classical variable, then a_0 and $a_0^\dagger$ have to satisfy $a_0 = a_0^\dagger \equiv a_0^*$ in order to make ϕ real: Thus they are simply the Fourier coefficients of the single eigenmode $\sin(\omega t)$. This suggests that we can short-cut the quantisation procedure as follows: We write down the field as sum over its eigenmodes $i = 1,\ldots,k$. Then we re-interpret the Fourier coefficients as creation and annihilation operators, requiring $[a_i,a_j^\dagger] = \delta_{ij}$.

Path integral approach We solve now the same problem, the rescaled Lagrangian (2.56), in the path integral approach. Using this method, we have argued that it is convenient to include a coupling to an external force J. Let us define therefore the effective action S_{eff} as the sum of the classical action S, the coupling to the external force J and a small imaginary part $\mathrm{i}\varepsilon\phi^2$ to make the path integral well-defined,

$$S_{\mathrm{e}ff} = S + \int_{-\infty}^{\infty} \mathrm{d}t \left(J\phi + \mathrm{i}\varepsilon\phi^2\right) = \int_{-\infty}^{\infty} \mathrm{d}t \left[\frac{1}{2}\dot{\phi}^2 - \frac{1}{2}\omega^2\phi^2 + J\phi + \mathrm{i}\varepsilon\phi^2\right] . \tag{2.63}$$

The function $\mathrm{e}^{\mathrm{i}S_{eff}}$ is the integrand of the path integral. We start our work by massaging S_{eff} into a form such that the path integral can be easily performed. The first two terms in S_{eff} can be viewed as the action of a differential operator $D(t)$ on $\phi(t)$, writing

$$\frac{1}{2}\left(\dot{\phi}^2 - \omega^2\phi^2\right) = -\frac{1}{2}\phi(t)\left(\frac{\mathrm{d}^2}{\mathrm{d}t^2} + \omega^2\right)\phi(t) = -\frac{1}{2}\phi(t)D(t)\phi(t) . \tag{2.64}$$

Here we performed a partial integration and dropped the boundary term: This is admissible, because the boundary term vanishes varying the action.

We can evaluate this operator going to Fourier space,

$$\phi(t) = \int \frac{\mathrm{d}E}{2\pi}\, \mathrm{e}^{-\mathrm{i}Et}\, \phi(E) \quad \mathrm{a}nd \quad J(t) = \int \frac{\mathrm{d}E}{2\pi}\, \mathrm{e}^{-\mathrm{i}Et}\, J(E) . \tag{2.65}$$

To keep the action real, we have to write all bilinear quantities as $\phi(E)\phi(-E')$, etc. Since only the phases depend on time, the time integration gives a factor $2\pi\delta(E - E')$, expressing energy conservation,

$$S_{\mathrm{e}ff} = \frac{1}{2}\int \frac{\mathrm{d}E}{2\pi} \left[\phi(E)(E^2 - \omega^2 + \mathrm{i}\varepsilon)\phi(-E) + J(E)\phi(-E) + J(-E)\phi(E)\right] . \tag{2.66}$$

In the path integral, this expression corresponds to a Gaussian integral of the type (2.21), where we should "complete the square." Shifting the integration variable to

$$\tilde{\phi}(E) = \phi(E) + \frac{J(E)}{E^2 - \omega^2 + \mathrm{i}\varepsilon} ,$$

we obtain

$$S_{\mathrm{e}ff} = \frac{1}{2}\int \frac{\mathrm{d}E}{2\pi} \left[\tilde{\phi}(E)(E^2 - \omega^2 + \mathrm{i}\varepsilon)\tilde{\phi}(-E) - J(E)\,\frac{1}{E^2 - \omega^2 + \mathrm{i}\varepsilon}\,J(-E)\right] . \tag{2.67}$$

Here we see that the "damping rule" for the path integral makes also the integral over the energy denominator well-defined. The physical interpretation of this way of shifting the poles—which differs from our treatment of the retarded Green function in the classical case—will be postponed to the next chapter, where we will discuss this issue in detail.

We are now in the position to evaluate the generating functional $Z[J]$. The path integral measure is invariant under a simple shift of the integration variable, $\mathcal{D}\tilde{\phi} = \mathcal{D}\phi$, and we omit the tilde from now on. Furthermore, the second term in S_{eff} does not depend on ϕ and can be factored out,

$$\begin{aligned} Z[J] = \exp&\left(-\frac{\mathrm{i}}{2}\int \frac{\mathrm{d}E}{2\pi} J(E)\,\frac{1}{E^2 - \omega^2 + \mathrm{i}\varepsilon}\,J(-E)\right) \\ &\times \int \mathcal{D}\phi\, \exp\frac{\mathrm{i}}{2}\int \frac{\mathrm{d}E}{2\pi}\, \left[\phi(E)(E^2 - \omega^2 + \mathrm{i}\varepsilon)\phi(-E)\right] . \end{aligned} \tag{2.68}$$

Setting the external force to zero, $J = 0$, the first factor becomes one and the generating functional $Z[0]$ becomes equal to the path integral in the second line. But for $J = 0$, the oscillator remains in the ground-state and thus $Z[0] = \langle 0, \infty | 0, -\infty \rangle = 1$. Therefore

$$Z[J] = \exp\left(-\frac{\mathrm{i}}{2}\int \frac{\mathrm{d}E}{2\pi} J(E)\,\frac{1}{E^2 - \omega^2 + \mathrm{i}\varepsilon}\,J(-E)\right) . \tag{2.69}$$

Inserting the Fourier transformed quantities, we arrive at

$$Z[J] = \exp\left(-\frac{\mathrm{i}}{2}\int \mathrm{d}t'\,\mathrm{d}t\; J(t')G_F(t'-t)J(t)\right) , \tag{2.70}$$

where the Feynman propagator

$$G_F(t-t') = \int \frac{\mathrm{d}E}{2\pi}\,\mathrm{e}^{-\mathrm{i}E(t-t')}\,\frac{1}{E^2 - \omega^2 + \mathrm{i}\varepsilon} \tag{2.71}$$

differs from the retarded propagator G_R defined in Eq. (1.35) by the position of its poles.

The generating functional $Z[J]$ given by (2.70) is in the form most suitable for deriving arbitrary n-point Green functions using our master formula (2.53). Note that finding $Z[J]$ required only to determine the inverse of the differential operator $D(t)$, accounting for the right boundary conditions induced by the $\mathrm{i}\varepsilon\phi^2$ term. This inverse is the Feynman propagator or two-point function $G_F(t'-t)$ which we can determine directly solving

$$-\,D(t)G_F(t'-t) = \delta(t'-t) . \tag{2.72}$$

Going to Fourier space, we find immediately

$$G_F(E) = \frac{1}{E^2 - \omega^2 + \mathrm{i}\varepsilon} . \tag{2.73}$$

This suggests that we can short-cut the calculation of $Z[J]$ by determining the Feynman propagator and using then directly (2.69) or (2.70).

These results allow us also to calculate arbitrary matrix elements between oscillator states. For instance, we obtain the expectation value $\langle 0|\,\phi^2\,|0\rangle$ from

$$\langle 0|\,|T\{\phi(t')\phi(t)\}\,|0\rangle = (-\mathrm{i})^2 \frac{\delta^2 Z[J]}{\delta J(t')\delta J(t)}\bigg|_{J=0} = \mathrm{i}G_F(t'-t) = \frac{1}{2\omega}\,\mathrm{e}^{\mathrm{i}\omega|t-t'|} . \tag{2.74}$$

Here, we used in the last step the explicit expression for G_F which you should check in problem 2.??. Taking the limit $t' \searrow t$ and replacing $\phi^2 \to mx^2$, we reproduce the standard result $\langle 0|\,x^2\,|0\rangle = 1/(2m\omega)$. Matrix elements between excited states $|n\rangle = (n!)^{-1/2}(a^\dagger)^n\,|0\rangle$ are obtained by expressing the creation operator $a^\dagger$ using $\pi(t) = \dot\phi(t)$ as

$$a^\dagger = \sqrt{\frac{\omega}{2}}\left(1 - \frac{\mathrm{i}}{\omega}\frac{\mathrm{d}}{\mathrm{d}t}\right)\phi(t) . \tag{2.75}$$

2.5. The need for quantum fields

We have already argued that any relativistic quantum theory has to be a many-particle theory. Such a theory has to include infinitely many degrees of freedom—as field theories like electrodynamics do. Before we move on to introduce the simplest quantum field theory in the next chapter, we present an argument that relativity and the single particle picture are incompatible.

In classical mechanics, the principle of relativity implies that all trajectories of massive particles are time-like, while massless particles move along light-like trajectories. This implements causality, i.e. the requirement that no signal can be transmitted faster than light. How should we translate this principle into a quantum theory? Causality would be clearly satisfied, if the relativistic propagator $K(\boldsymbol{x}',t';\boldsymbol{x},t)$ vanishes for space-like distances. Another, less restrictive translation of the principle of relativity would be to ask that measurements performed at space-like separated points do not influence each other. This is achieved, if all observables $O(x)$ commute for space-like distances,

$$[\hat{O}(\boldsymbol{x},t),\hat{O}(\boldsymbol{x}',t')]=0 \qquad \text{for} \quad (t-t')^2<(\boldsymbol{x}-\boldsymbol{x}')^2\,. \tag{2.76}$$

In quantum mechanics, the Heisenberg operators $\hat{\boldsymbol{x}}(t)$ and $\hat{\boldsymbol{p}}(t)$ depend, however, only on time. Therefore we can not implement the condition (2.76) in such a framework.

The only rescue for causality in relativistic quantum mechanics is therefore the vanishing of the propagator $K(t',\boldsymbol{x}';t,\boldsymbol{x})$ outside the light-cone. We evaluate the propagator as in the non-relativistic case,

$$K(\boldsymbol{x}',t';\boldsymbol{x},t)=\langle\boldsymbol{x}'|\,\mathrm{e}^{-\mathrm{i}H(t'-t)}\,|\boldsymbol{x}\rangle=\int\frac{\mathrm{d}^3p}{(2\pi)^3}\,\langle\boldsymbol{x}'|\,\mathrm{e}^{-\mathrm{i}E_{\boldsymbol{p}}(t'-t)}\,|\boldsymbol{p}\rangle\,\langle\boldsymbol{p}|\,\boldsymbol{x}\rangle \tag{2.77}$$

inserting however the relativistic dispersion relation, $E_{\boldsymbol{p}}=\sqrt{m^2+\boldsymbol{p}^2}$. Next we use that the momentum operator $\hat{\boldsymbol{p}}$ generates space translations, $\exp(-\mathrm{i}\hat{\boldsymbol{p}}\boldsymbol{x})|\boldsymbol{0}\rangle=|\boldsymbol{x}\rangle$, to obtain

$$K(\boldsymbol{x}',t';\boldsymbol{x},t)=K(x'-x)=\int\frac{\mathrm{d}^3p}{(2\pi)^3}\,|\langle\boldsymbol{0}|\boldsymbol{p}\rangle|^2\,\mathrm{e}^{-\mathrm{i}p(x'-x)}\,. \tag{2.78}$$

Here we introduced also the four-vector $p^\mu=(E_p,\boldsymbol{p})$, rewriting the plan-wave thereby in a Lorentz-invariant way. In order that the complete propagator is invariant, we have to choose as integration measure $\propto \mathrm{d}^3p/E_{\boldsymbol{p}}$, cf. problem 2.??, and we set therefore $|\langle\boldsymbol{0}|\boldsymbol{p}\rangle|^2=1/(2E_{\boldsymbol{p}})$. Knowing its explicit expression, it is a straight-forward exercise to show that the propagator does not vanish outside the light-cone, but goes only exponentially to zero, $K(\boldsymbol{x}',0;\boldsymbol{x},0)\propto\exp(-m|\boldsymbol{x}'-\boldsymbol{x}|)$. Thus we failed to implement both versions of causality into relativistic quantum mechanics. Instead, we will develop quantum field theory with the aim to implement causality via the condition (2.76).

Before starting this endeavour, we can draw still some important conclusion from Eq. (2.78). For space-like distances, $(x-x')^2<0$, a Lorentz boost can change the time order of two space-time events, cf. problem 1.??. Consistency requires thus to include both time-orderings: If a particle is created at t and absorbed at $t'>t$, then it can be created necessarily also at t' and absorbed at $t>t'$. We extend therefore the propagator as

$$K(x'-x)=\int\frac{\mathrm{d}^3p}{(2\pi)^3 2E_{\boldsymbol{p}}}\left[\vartheta(t'-t)\mathrm{e}^{-\mathrm{i}p(x'-x)}+\vartheta(t-t')\mathrm{e}^{\mathrm{i}p(x'-x)}\right]\,, \tag{2.79}$$

where we chose the opposite sign for the plan-wave in the second factor: In this way, the phase of the plane-waves observed in both frames agree, $-E_{\boldsymbol{p}}\tau\vartheta(\tau)<0$ and $+E_{\boldsymbol{p}}\tau\vartheta(-\tau)<0$, and similarly for the momenta. If we imagine that the propagating particle carries a conserved charge, then we can associate the positive frequencies to the propagation of a particle (with charge q) and the negative frequencies to the propagation of an antiparticle (with charge $-q$). Then the resulting current is frame-independent, if the antiparticle has the same mass but the opposite additive charges. This prediction of relativistic quantum field theory is experimentally confirmed with extreme accuracy: For instance, the limits on the mass and charge difference of electrons and positrons are smaller than 8×10^{-9} and 4×10^{-8}, respectively [O$^+$14].

Finally, we should mention an alternative way to implement causality: Instead of defining quantum fields $\hat{\phi}(x^\mu)$ on classical space-time, we could promote time t to an operator, parametrising the world-line $\hat{x}^\mu(\tau)$ of a particle e.g. by its proper-time τ. Considering then the surface $\hat{x}^\mu(\tau, \sigma)$ generated by a set of world-lines is the starting point of string theory.

Summary

Using Feynman's path integral approach, we can express a transition amplitude as a sum over all paths weighted by a phase which is determined by the classical action, $\langle q_f, t_f | q_i, t_i \rangle = \int \mathcal{D}q(t) \exp(\mathrm{i}S[q])$. Adding a linear coupling to an external source J and a damping term to the Lagrangian, we obtain the ground-state persistence amplitude $\langle 0, \infty | 0, -\infty \rangle_J$. This quantity serves as the generating functional $Z[J]$ for n-point Green functions $G(t_1, \dots, t_n)$ which are the time-ordered vacuum expectation values of the operators $\hat{q}(t_1), \dots, \hat{q}(t_n)$.

Further reading

For additional examples for the use of the path integral and Green functions in quantum mechanics see e.g. [Mac00] or [Das06]. [Sch05] sketches the historical development that lead to Schwinger's Green functions, including his quantum action principle.

3. Free scalar field

We extend in this chapter the path integral approach from quantum mechanics to the simplest field theory, containing a single real scalar field $\phi(x)$. Such a field may either represent an elementary particle like the Higgs scalar, a bound-state like a scalar meson, or a scalar parameter describing a specific property of a more complex theory. Proceeding similar to our approach in quantum mechanics, we will introduce the generating functional $Z[J] = \langle 0+|0-\rangle_J$ of n-point Green functions as our main tool to calculate the time-ordered vacuum expectation value of a product of fields $\phi(x_1)\cdots\phi(x_n)$. Calculating the vacuum energy of the scalar field, we will encounter for the first time that many calculations in quantum field theories return a formally infinite result. In order to extract sensible predictions, we have to introduce therefore the concepts of regularisation and renormalisation.

3.1. Lagrange formalism and path integrals for fields

A field is a map which associates to each space-time point x a k-tupel of values $\phi_a(x)$, $a = 1,\ldots,k$. The space of field values $\phi_a(x)$ can be characterised by its transformation properties under Poincaré transformations, i.e. the group of translations and Lorentz transformations, and internal symmetry groups. The latter are in practically all physical applications Lie groups like U(1), SU(n) or SO(n). Except for a real scalar field ϕ, these fields have several components. Thus we have to generalise Hamilton's principle to a collection of fields $\phi_a(x)$, where the index a includes all internal as well as space-time indices. Moreover, the Lagrangian for a field $\phi_a(x)$ will contain not only time but also space derivatives.

To ensure Lorentz invariance, we consider a scalar Lagrange density $\mathscr{L}(x)$ that depend as a local function on the fields and their derivatives. By analogy to $L(q,\dot{q})$, we restrict ourselves to fields $\phi_a(x)$ and their first derivatives $\partial_\mu\phi_a(x)$. We include no explicit time-dependence, since "everything" should be explained by the fields and their interactions. The Lagrangian $L(\phi_a, \partial_\mu\phi_a)$ is obtained by integrating the density $\mathscr{L}$ over a given space volume V. The action S is thus the four-dimensional integral

$$S[\phi_a] = \int_{t_a}^{t_b} \mathrm{d}t\, L(\phi_a, \partial_\mu\phi_a) = \int_\Omega \mathrm{d}^4x\, \mathscr{L}(\phi_a, \partial_\mu\phi_a) \tag{3.1}$$

with $\Omega = V \times [t_a : t_b]$. A variation $\delta\phi_a(x)$ of the fields leads to a variation of the action,

$$\delta S = \int_\Omega \mathrm{d}^4x \left[\frac{\partial\mathscr{L}}{\partial\phi_a}\,\delta\phi_a + \frac{\partial\mathscr{L}}{\partial(\partial_\mu\phi_a)}\,\delta(\partial_\mu\phi_a)\right], \tag{3.2}$$

where we have to sum over field components ($a = 1,\ldots,k$) and the Lorentz index $\mu = 0,\ldots,3$. The correspondence $q(t) \to \phi(x^\mu)$ implies that the scale factor ε parametrising the variations $\phi_a(x^\mu, \varepsilon)$ depends not on x^μ. We can therefore eliminate again the variation of the field gradients $\partial_\mu\phi_a$ by a partial integration using Gauss' theorem,

$$\delta S = \int_\Omega \mathrm{d}^4x \left[\frac{\partial\mathscr{L}}{\partial\phi_a} - \partial_\mu\left(\frac{\partial\mathscr{L}}{\partial(\partial_\mu\phi_a)}\right)\right]\delta\phi_a = 0\,. \tag{3.3}$$

The surface term vanishes, since we require that the variation is zero on the boundary $\partial\Omega$. Thus the Lagrange equations for the fields ϕ_a are

$$\frac{\partial\mathscr{L}}{\partial\phi_a} - \partial_\mu\left(\frac{\partial\mathscr{L}}{\partial(\partial_\mu\phi_a)}\right) = 0\,. \tag{3.4}$$

If the Lagrange density $\mathscr{L}$ is changed by a four-dimensional divergence, $\delta\mathscr{L} = \partial_\mu K^\mu$, and surface terms can be dropped, the same equations of motion result. Note also that it is often more efficient to perform directly the variation $\delta\phi_a$ in the action $S[\phi_a]$ than to use the Lagrange equations.

The path integral becomes now a functional integral over the k fields ϕ_a,

$$K = \int \mathcal{D}\phi_1\cdots\mathcal{D}\phi_k\,\mathrm{e}^{\mathrm{i}S[\phi_a]} = \int \mathcal{D}\phi_1\cdots\mathcal{D}\phi_k\,\mathrm{e}^{\mathrm{i}\int_\Omega \mathrm{d}^4x\,\mathscr{L}(\phi_a,\partial_\mu\phi_a)}\,. \tag{3.5}$$

A major problem we have to address later is that the k fields ϕ_a are often not independent: For instance, in electrodynamics all potentials A^μ connected by a gauge transformation describe the same physics. This redundancy makes the path integral (3.5) ill-defined. We start therefore with the simplest case of a single, real scalar field ϕ where such problems are absent. Moreover, we restrict ourselves in this chapter to a free field without interactions.

3.2. Generating functional for a scalar field

Lagrangian The (free) Schrödinger equation $\mathrm{i}\partial_t\psi = H_0\psi$ can be obtained substituting $\omega \to \mathrm{i}\partial_t$ and $\boldsymbol{k} \to -\mathrm{i}\boldsymbol{\nabla}_x$ into the non-relativistic energy-momentum relation $\omega = \boldsymbol{k}^2/(2m)$. With the same replacements, the relativistic $\omega^2 = m^2 + \boldsymbol{k}^2$ becomes the Klein-Gordon equation

$$(\Box + m^2)\phi = 0 \qquad with \quad \Box = \eta_{\mu\nu}\partial^\mu\partial^\nu = \partial_\mu\partial^\mu\,. \tag{3.6}$$

The relativistic energy-momentum relation implies that the solutions to the free Klein-Gordon equation consist of plane-waves with positive and negative energies $\pm\sqrt{\boldsymbol{k}^2+m^2}$. For the stability of a quantum system it is however essential that its energy eigenvalues are bounded from below. Otherwise, we could generate e.g. in a scattering process $\phi+\phi \to n\phi$ an arbitrarily high number of ϕ particles with sufficiently low energy, and no stable form of matter could exist. Interpreting the Klein-Gordon equation as a relativistic wave equation for a single particle can be therefore not fully satisfactory, since the energy of its solutions is not bounded from below.

How do we guess the correct Lagrange density $\mathscr{L}$? Plane waves can be seen as a collection of coupled harmonic oscillators at each space-time point. The correspondence $\dot{q} \to \partial_\mu\phi$ means that the kinetic field energy is quadratic in the field derivatives. Relativistic invariance implies that the Lagrange density is a scalar, leaving as the only two possible terms containing derivatives

$$\eta_{\mu\nu}(\partial^\mu\phi)(\partial^\nu\phi) \quad and \quad \phi\Box\phi\,.$$

Using the action principle to derive the equation of motions, we can however drop boundary terms performing partial integrations. Thus these two terms are equivalent, up to a minus sign. The Klein-Gordon equation $\Box\phi = -m^2\phi$ suggests that the mass term is also quadratic in the field ϕ. Therefore we try as Lagrange density

$$\mathscr{L} = \frac{1}{2}\eta_{\mu\nu}\,(\partial^\mu\phi)\,(\partial^\nu\phi) - \frac{1}{2}m^2\phi^2 \equiv \frac{1}{2}\eta_{\mu\nu}\partial^\mu\phi\partial^\nu\phi - \frac{1}{2}m^2\phi^2\,. \tag{3.7}$$

From now, we will drop the parenthesis around $\partial^\mu\phi$ and it should be understood from the context that the derivative ∂^μ acts only on the first field ϕ. Even shorter alternative notations are $(\partial_\mu\phi)^2$ and the concise $(\partial\phi)^2$. Swapping the indices in the the Lagrangian (3.7), we obtain for the second part of the Lagrange equation

$$\frac{\partial}{\partial(\partial_\alpha\phi)}\left(\eta^{\mu\nu}\partial_\mu\phi\partial_\nu\phi\right) = \eta^{\mu\nu}\left(\delta^\alpha_\mu\partial_\nu\phi + \delta^\alpha_\nu\partial_\mu\phi\right) = \eta^{\alpha\nu}\partial_\nu\phi + \eta^{\mu\alpha}\partial_\mu\phi = 2\partial^\alpha\phi\,. \tag{3.8}$$

Hence the Lagrange equation becomes

$$\frac{\partial\mathscr{L}}{\partial\phi} - \partial_\alpha\left(\frac{\partial\mathscr{L}}{\partial(\partial_\alpha\phi)}\right) = -m^2\phi - \partial_\alpha\partial^\alpha\phi = 0\,, \tag{3.9}$$

and the Lagrange density (3.7) leads indeed to the Klein-Gordon equation. We can understand the relative sign in the Lagrangian splitting the relativistic kinetic energy into the "proper" kinetic energy $(\partial_t\phi)^2/2$ and the gradient energy density $(\boldsymbol{\nabla}\phi)^2/2$,

$$\mathscr{L} = \frac{1}{2}\dot{\phi}^2 - \frac{1}{2}(\boldsymbol{\nabla}\phi)^2 - \frac{1}{2}m^2\phi^2\,. \tag{3.10}$$

The last two terms correspond to a potential energy and carry therefore the opposite sign of the first one.

Instead of guessing, we can derive the correct Lagrangian $\mathscr{L}$ as follows: We multiply the free field equation for ϕ by a variation $\delta\phi$ that vanishes on $\partial\Omega$. Then we integrate over Ω, perform a partial integration of the kinetic term, use $\partial_\mu\delta = \delta\partial_\mu$, the Leibniz rule and ask that the variation vanishes,

$$A\int_\Omega \mathrm{d}^4x\,\delta\phi\,(\Box + m^2)\phi = A\int_\Omega \mathrm{d}^4x\,\left[-\delta(\partial_\mu\phi)\partial^\mu\phi + \delta\phi\phi m^2\right] = \tag{3.11a}$$

$$= A\int_\Omega \mathrm{d}^4x\,\delta\left[-\frac{1}{2}(\partial_\mu\phi)^2 + \frac{1}{2}\phi^2 m^2\right] = 0\,. \tag{3.11b}$$

The term in the square brackets agrees with our guess (3.7), taking into account that the source-free field equation fixes the Lagrangian only up to the overall factor A. In analogy with a quantum mechanical oscillator, we want that the coefficients of the two terms are $\pm 1/2$ and thus we set $|A| = 1$.

We can determine the correct overall sign of $\mathscr{L}$ by calculating the energy density ρ of the scalar field and requiring that it is bounded from below and stable against small perturbations. We identify the energy density ρ of the scalar field with its Hamiltonian density $\mathscr{H}$, and use the connection between the Lagrangian and the Hamiltonian known from classical mechanics. The transition from a system with a finite number of degrees of freedom to one with an infinite number of degrees of freedom proceeds as follows,

$$p_i = \frac{\partial L}{\partial\dot{q}^i} \qquad\Rightarrow\qquad \pi_a = \frac{\partial\mathscr{L}}{\partial\dot{\phi}_a}\,, \tag{3.12a}$$

$$H = p_i\dot{q}^i - L \qquad\Rightarrow\qquad \mathscr{H} = \sum_a \pi_a\dot{\phi}_a - \mathscr{L}\,. \tag{3.12b}$$

The canonically conjugated momentum π of a real scalar field is

$$\pi = \frac{\partial\mathscr{L}}{\partial\dot{\phi}} = \dot{\phi}\,. \tag{3.13}$$

Thus the Hamilton density is

$$\mathscr{H} = \pi\dot{\phi} - \mathscr{L} = \pi^2 - \mathscr{L} = \frac{1}{2}\dot{\phi}^2 + \frac{1}{2}(\nabla\phi)^2 + \frac{1}{2}m^2\phi^2 \geq 0 \tag{3.14}$$

and thus obviously positive definite. Moreover, generating fluctuations $\delta\phi$ costs energy and thus the system is stable against small perturbations. Hence the transition from a single particle interpretation of the Klein-Gordon equation to a field theory has been sufficient to cure the problem of the negative energy solutions.

Note that we could subtract a constant ρ_0 from the Lagrangian which would drop out of the equation of motion. From Eq. (3.14) we see that such a constant corresponds to a uniform energy density of empty space. Such a term would act as an additional source of the gravitational field, but would be otherwise unobservable. Next we generalise the Lagrangian by subtracting a polynomial in the fields, $V(\phi)$, subject to the stability constraint discussed above. Hence the potential V should be bounded from below, and we can expand it around its minimum at $\phi \equiv v$,

$$\left.\frac{\mathrm{d}V}{\mathrm{d}\phi}\right|_{\phi=v} = 0\,, \qquad \left.\frac{\mathrm{d}^2V}{\mathrm{d}\phi^2}\right|_{\phi=v} \equiv m^2 > 0\,. \tag{3.15}$$

The term $V''(v)$ acts as mass term the field ϕ. We will see soon that terms ϕ^n with $n \geq 3$ generate interactions between n particles, as expected from the analogy of a quantum field to coupled quantum mechanical oscillators. The field ϕ has the non-zero value $\phi = v$ everywhere, if the minimum v of $V(\phi)$ is not at zero, $v \neq 0$. If the value of $V(\phi)$ at the minimum v is not zero, $V(v) \neq 0$, then the non-zero potential implies a non-zero uniform energy density $\rho = V(v)$.

Generating functional Now we move on to the quantum theory of a scalar field, which we define by the path-integral over $\exp \mathrm{i}S[\phi]$. The Green functions which encode all information about this theory can be obtained from the generating functional

$$Z[J] = \langle 0+|0-\rangle_J = \mathcal{N}\int \mathcal{D}\phi \exp \mathrm{i}\int \mathrm{d}^4x \,\left(\frac{1}{2}\partial_\mu\phi\partial^\mu\phi - \frac{1}{2}m^2\phi^2 + J\phi\right)\,, \tag{3.16}$$

where we appended to the action a linear coupling between the field and an external source. To ensure the convergence of the integral, we add an infinitesimal small imaginary part to the squared mass of the particle, $m^2 \to m^2 - \mathrm{i}\varepsilon$. Next we perform an integration by part of the first term, exploiting the fact that the boundary term vanishes,

$$Z[J] = \mathcal{N}\int \mathcal{D}\phi \exp \mathrm{i}\int \mathrm{d}^4x \,\left(-\frac{1}{2}\phi(\Box + m^2)\phi + J\phi\right)\,. \tag{3.17}$$

The first two terms, $A = -(\Box + m^2)$, are quadratic and symmetric in the field ϕ,

$$-\frac{1}{2}\int \mathrm{d}^4x\, \phi(x)(\Box_x + m^2)\phi(x) = \frac{1}{2}\int \mathrm{d}^4x\mathrm{d}^4x'\, \phi(x)A(x,x')\delta(x-x')\phi(x')\,. \tag{3.18}$$

Note that the operator A is local, $A(x) \propto A(x,x')\delta(x-x')$: Since special relativity forbids action at a distance, non-local terms like $\phi(x')A(x,x')\phi(x)$ should not appear in a relativistic Lagrangian.

If we discretise continuous space-time x^μ into a lattice, we can use Eq. (2.22) to perform the path integral,

$$Z[J] = \mathcal{N} \left(\frac{(2\pi \mathrm{i})^N}{\det[A]}\right)^{1/2} \exp\left(-\frac{1}{2}\,\mathrm{i}\,JA^{-1}J\right) \equiv \mathcal{N} Z[0] \exp(\mathrm{i}W[J])\,. \tag{3.19}$$

The prefactor of the exponential function does not depend on J and is thus given by $\mathcal{N}Z[0] = \langle 0+|0-\rangle$. The vacuum should be stable and normalised to one in the absence of sources, $\langle 0+|0-\rangle = 1$. Therefore the proper normalisation of $Z[J]$ implies that $\mathcal{N}^{-1} = Z[0]$. Thus we can omit the normalisation factor, if we normalise the path integral measure $\mathcal{D}\phi$ such that the Gaussian integral over a free field is one. In the last step of Eq. (3.19), we defined a new functional $W[J]$ that depends only quadratically on the source J; therefore it should be easier to handle than $Z[J]$. Going for $N \to \infty$ back to continuous space-time, the matrix multiplications become integrations,

$$Z[J] = \exp(\mathrm{i}W[J]) = \exp\left(-\frac{\mathrm{i}}{2}\int \mathrm{d}^4x\mathrm{d}^4x'\, J(x)A^{-1}(x,x')J(x')\right) \tag{3.20}$$

and

$$W[J] = -\frac{1}{2}\int \mathrm{d}^4x\mathrm{d}^4x'\, J(x)A^{-1}(x,x')J(x')\,. \tag{3.21}$$

Propagator In order to evaluate the functional $W[J]$ we have to find the inverse $\Delta(x,x') \equiv A^{-1}(x,x')$ of the differential operator A, defined by

$$-(\Box + m^2)\Delta(x,x') = \delta(x-x')\,. \tag{3.22}$$

Because of translation invariance, the Green function $\Delta(x,x')$ can depend only on the difference $x - x'$. Therefore it is advantageous to perform a Fourier transformation and to go to momentum space,

$$-\int \frac{\mathrm{d}^4k}{(2\pi)^4}(\Box + m^2)\Delta(k)\mathrm{e}^{-\mathrm{i}k(x-x')} = \int \frac{\mathrm{d}^4k}{(2\pi)^4}\mathrm{e}^{-\mathrm{i}k(x-x')}\,, \tag{3.23}$$

or

$$\Delta_F(k) = \frac{1}{k^2 - m^2 + \mathrm{i}\varepsilon}\,, \tag{3.24}$$

where the pole at $k^2 = m^2$ is avoided by the iε. Thus the $m^2 \to m^2 - \mathrm{i}\varepsilon$ prescription introduced to ensure the convergence of the path integral tells us also how to handle the poles of the Green function. The index F specifies that the propagator Δ_F is the Green function obtained with the $m^2 - \mathrm{i}\varepsilon$ prescription proposed by Feynman. (Some authors use instead D_F for the propagator of massive bosons and Δ_F for the propagator of massless bosons.)

Note that the four momentum components k^μ are independent. Therefore $\Delta_F(k)$ describes the propagation of a *virtual* particle that has—in contrast to a real or external particle—not to be on "mass-shell:" in general

$$k_0 \neq \pm\omega_k \equiv \pm\sqrt{\boldsymbol{k}^2 + m^2}\,.$$

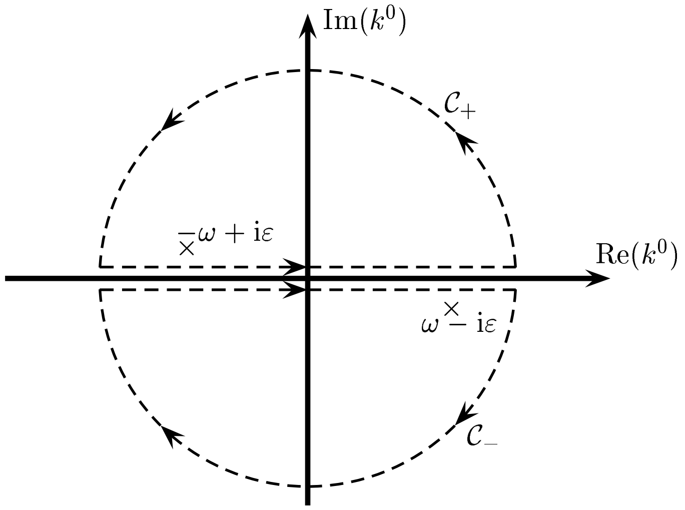


Figure 3.1.: Poles and contours in the complex k^0 plane used for the integration of the Feynman propagator.

We can evaluate the k_0 integral in the coordinate representation of $\Delta_F(x - x')$ explicitly,

$$\Delta_F(x-x') = \int \frac{\mathrm{d}^4k}{(2\pi)^4} \frac{e^{-\mathrm{i}k(x-x')}}{k_0^2 - \boldsymbol{k}^2 - m^2 + \mathrm{i}\varepsilon} \tag{3.25a}$$

$$= \int \frac{\mathrm{d}^3k}{(2\pi)^3} \int \frac{\mathrm{d}k_0}{2\pi} \frac{\mathrm{e}^{-\mathrm{i}k_0(t-t')}\mathrm{e}^{\mathrm{i}\boldsymbol{k}(\boldsymbol{x}-\boldsymbol{x}')}}{(k_0 - \omega_k + \mathrm{i}\varepsilon)(k_0 + \omega_k - \mathrm{i}\varepsilon)}\,, \tag{3.25b}$$

using Cauchy's theorem[1]. The integrand has two simple poles at $+\omega_k - \mathrm{i}\varepsilon$ and $-\omega_k + \mathrm{i}\varepsilon$, cf. Fig. 3.1. For negative $\tau = t - t'$, we can close the integration contour $\mathcal{C}_+$ on the upper half-plane, including the pole at $-\omega_k$,

$$\int \mathrm{d}k_0 \frac{\mathrm{e}^{-\mathrm{i}k_0\tau}}{(k_0 - \omega_k + \mathrm{i}\varepsilon)(k_0 + \omega_k - \mathrm{i}\varepsilon)} = 2\pi\mathrm{i}\,\mathrm{r}es_{-\omega_k} = 2\pi\mathrm{i}\,\frac{\mathrm{e}^{\mathrm{i}\omega_k\tau}}{-2\omega_k} \qquad \mathrm{f}or \quad \tau < 0\,. \tag{3.26}$$

For positive τ, we have to choose the contour $\mathcal{C}_-$ in the lower plane, picking up $2\pi\mathrm{i}\ \mathrm{e}^{-\mathrm{i}\omega_k\tau}/(2\omega_k)$ and an additional minus sign since the contour is clockwise. Combining both results, we obtain

$$\mathrm{i}\Delta_F(x) = \int \frac{\mathrm{d}^3k}{(2\pi)^3 2\omega_k} \left[\mathrm{e}^{-\mathrm{i}\omega_k t}\vartheta(x^0) + \mathrm{e}^{\mathrm{i}\omega_k t}\vartheta(-x^0)\right] \mathrm{e}^{\mathrm{i}\boldsymbol{kx}}\,, \tag{3.27}$$

or after shifting the integration variable $\boldsymbol{k} \to -\boldsymbol{k}$ in the second term

$$\mathrm{i}\Delta_F(x) = \int \frac{\mathrm{d}^3k}{(2\pi)^3 2\omega_k} \left[\mathrm{e}^{-\mathrm{i}(\omega_k t - \boldsymbol{kx})}\vartheta(x^0) + \mathrm{e}^{\mathrm{i}(\omega_k t - \boldsymbol{kx})}\vartheta(-x^0)\right]\,. \tag{3.28}$$

Comparing this expression to our guess (2.79) at the end of the last chapter, we see that our intuitive arguments about the structure of a Lorentz invariant propagator in a quantum theory

[1] Since ε is infinitesimal and $\omega_k > 0$, we can set $2\mathrm{i}\omega_k\varepsilon + \varepsilon^2 \to \mathrm{i}\varepsilon$.

were correct. We stress once again the salient features of the Feynman propagator: First, the propagator contains positive and negative frequencies, as expected from the existence of solutions to the Klein-Gordon equation with positive and negative energies. Second, positive frequencies propagate forward in time, while negative frequencies propagate backward. This implies the existence of antiparticles. Third, the relativistic normalisation of (on-shell) plane waves includes a factor $1/\sqrt{2\omega_k}$, or

$$\langle k|k'\rangle = 2\omega_k(2\pi)^3\delta(\boldsymbol{k}-\boldsymbol{k}')\,, \tag{3.29}$$

while the non-relativistic normalisation uses $\langle \boldsymbol{k}|\boldsymbol{k}'\rangle = \delta(\boldsymbol{k}-\boldsymbol{k}')$.

Remark 3.1: Other Green functions are obtained, if we choose different prescriptions for the handling of the poles. For positive $\tau = t - t'$, we have to close the circle on the lower half-plane. Shifting both poles to the lower half-plane, $\pm\omega_k - \mathrm{i}\varepsilon$, gives thus the retarded propagator $\Delta_{ret}(x)$ vanishing for all $\tau < 0$. In the opposite case, we shift both poles to the upper half-plane, $\pm\omega_k + \mathrm{i}\varepsilon$, and obtain the advanced propagator $\Delta_{adv}(x)$. Both propagators are real-valued, propagating a real solution of the wave equation into another real one at a different time, as required in classical physics. Moreover, both Green functions have support inside the light-cone, the retarded in the forward and the advanced in the backward part of the light cone. This behaviour should be contrasted with the Feynman propagator Δ_F which is complex-valued and non-zero in $\mathbb{R}(1,3)$.
Another way to handle the singularities is to use Cauchy's Principal value prescription, obtaining $\Delta(x) = \frac{1}{2}[\Delta_{adv}(x) + \Delta_{ret}(x)]$. This choice corresponds to an action-at-distance which seems to have no relevance in physics. Finally, we can shift one pole up and the other one down. The choice $\pm(\omega_k - \mathrm{i}\varepsilon)$ used in the Feynman propagator allows us to rotate the integration contour anti-clockwise to $-\mathrm{i}\infty : +\mathrm{i}\infty$ avoiding both poles in the complex k_0 plane. Since $k_0 = \mathrm{i}\partial_t$, this transformation is consistent with the clockwise rotation in coordinate space required to obtain an Euclidean action bounded from below. Thus the Feynman prescription is the only one in which the physics in Minkowski and Euclidean space are analytically connected, cf. with the remark 2.1.

We are now in the position to evaluate the generating functional

$$W[J] = -\frac{1}{2}\int \mathrm{d}^4x\mathrm{d}^4x' J(x)\Delta_F(x-x')J(x')\,. \tag{3.30}$$

in Fourier space. Inserting the Fourier transformations for the propagator as well as for the sources J gives

$$W[J] = -\frac{1}{2}\int \mathrm{d}^4x\mathrm{d}^4x' \int \frac{\mathrm{d}^4k}{(2\pi)^4}\frac{\mathrm{d}^4k'}{(2\pi)^4}\frac{\mathrm{d}^4\tilde{k}}{(2\pi)^4} J(k)^*\mathrm{e}^{\mathrm{i}kx}\,\frac{\mathrm{e}^{-\mathrm{i}\tilde{k}(x-x')}}{\tilde{k}^2-m^2+\mathrm{i}\varepsilon}\,J(k')\mathrm{e}^{-\mathrm{i}k'x'}\,. \tag{3.31}$$

Exchanging the integration order and performing the space-time integrations leads to the conservation of the four-momenta entering and leaving the two interaction points, $(2\pi)^8\delta(k-\tilde{k})\delta(\tilde{k}-k')$: The source $J(k)^*$ produce a scalar particle with momentum k, and thus only the Fourier component k of the scalar propagator contributes. This is a very general behaviour, based solely on the translation invariance of the free particle states we are using. In the final step, we cancel two of the three momentum integrations with the two momentum delta functions and are left with

$$W[J] = -\frac{1}{2}\int \frac{\mathrm{d}^4k}{(2\pi)^4} J(k)^*\,\frac{1}{k^2-m^2+\mathrm{i}\varepsilon}\,J(k)\,. \tag{3.32}$$

The functional $W[J]$ has the same structure as the one for the harmonic oscillator found in the last chapter. We will see that it contains, as in the one-dimensional case, all information about a free scalar field, not only about its ground state. Note that the contribution of Fourier modes to $W[J]$ increases, the closer they are on-shell, $k^2 \to m^2$. For $k^2 = m^2$, the propagator diverges finally. In order to interpret this unphysical result, we can compare it to a classical harmonic oscillator: If an external source is applied in resonance with the eigen-frequency ω of the oscillator, the oscillation amplitude will increase until friction cannot be longer neglected. In our case at hand, the non-zero life-time of unstable particles plays the rôle of friction: Including in our formalism that the exchanged particle is unstable, the infinitesimal ε would be replaced by half its decay-width, $\varepsilon \to \mathrm{i}\Gamma/2$.

Attractive Yukawa potential by scalar exchange From our macroscopic experience, we know the two cases of electromagnetism, where equal electric charges repel each other, and of gravity where two masses attract each other. The first physics question we want to answer with our newly developed formalism is if the scalar field falls into the category of a fundamentally attractive or repulsive interaction.

In order to address this question, we consider two static point charges as external sources, $J = J_1(x_1) + J_2(x_2)$ with $J_i = \delta(\boldsymbol{x} - \boldsymbol{x}_i)$, in $W[J]$. Multiplying out the terms in $J(x)\Delta_F(x - x')J(x')$ gives four contributions, $W_{ij} \propto J_i J_j$: The terms $W_{11}[J]$ and $W_{22}[J]$ correspond to the emission and re-absorption of the particle by the same source J_i. They are examples for self-interactions that we neglect for the moment. The interaction between two different charges is given by

$$W_{12}[J] = W_{21}[J] = -\frac{1}{2}\int \mathrm{d}^4x\mathrm{d}^4x' \int \frac{\mathrm{d}^4k}{(2\pi)^4} J_1(x)\, \frac{\mathrm{e}^{-\mathrm{i}k(x-x')}}{k^2 - m^2 + \mathrm{i}\varepsilon}\, J_2(x') \tag{3.33}$$

$$= -\frac{1}{2}\int \mathrm{d}t\mathrm{d}t' \int \frac{\mathrm{d}^4k}{(2\pi)^4} \frac{\mathrm{e}^{-\mathrm{i}k^0(t-t')}\,\mathrm{e}^{\mathrm{i}\boldsymbol{k}(\boldsymbol{x}_1-\boldsymbol{x}_2)}}{k^2 - m^2 + \mathrm{i}\varepsilon}\,. \tag{3.34}$$

Performing one of the two time integrals, e.g. the one over t', gives $2\pi\delta(k_0)$. Hence our assumption of static sources implies that the virtual particle carries zero energy and is space-like, $k^2 = -\boldsymbol{k}^2 < 0$. Eliminating then the k_0 integral with the help of the delta function, we obtain next

$$W_{12}[J] = \frac{1}{2}\int \mathrm{d}t \int \frac{\mathrm{d}^3k}{(2\pi)^3} \frac{\mathrm{e}^{\mathrm{i}\boldsymbol{k}\boldsymbol{r}}}{\boldsymbol{k}^2 + m^2} \tag{3.35}$$

with $\boldsymbol{r} = \boldsymbol{x}_1 - \boldsymbol{x}_2$. The denominator is always positive, and we can therefore omit the iε. Before we can go on, we have to make sense out of the infinite time integral: Looking at

$$Z[J] = \langle 0|\exp(-\mathrm{i}H[J]\tau)|0\rangle = \exp(\mathrm{i}W[J])\,, \tag{3.36}$$

we see that $W[J] = -E\tau$ with $\tau = t - t'$ as the considered time interval. Hence the potential energy V of two static point charges separated by the distance $\boldsymbol{r}$ is

$$V = -(W_{12} + W_{21})/\tau = -\int \frac{\mathrm{d}^3k}{(2\pi)^3} \frac{\mathrm{e}^{\mathrm{i}\boldsymbol{k}\boldsymbol{r}}}{\boldsymbol{k}^2 + m^2} = -\frac{\mathrm{e}^{-mr}}{4\pi r} < 0\,. \tag{3.37}$$

Thus the potential energy of two equal charges is reduced by the exchange of a scalar particle, which means that the scalar force between them is attractive. If the exchanged particle is massive, the range of the force is of order $1/m$. These two observations were the basic motivation for Yukawa to suggest in 1935 the exchange of scalar particles as model for the nuclear force. Note also that we obtain in the limit $m \to 0$ a $1/r$ potential as in Newton's and Coulomb's law. Thus we learnt the important fact that the only two known forces of infinite range, the electromagnetic and the gravitational force, are transmitted by massless particles, the photon and the graviton, respectively. The result $V \propto 1/r$ for $m = 0$ and $n = 4$ space-time dimensions, or more generally $V \propto 1/r^{n-3}$ for $n \geq 4$, follows from simple dimensional analysis: For $m = 0$, the only remaining dimensionfull parameter after the integration over $\boldsymbol{k}$ is the distance r. From Eq. (3.37), we read off that the potential energy V has the dimension $[V] = k^{n-3}$. Thus the potential energy due to the exchange of a massless particle scales as r^{-n+3}. Finally, note that the amplitude $W_{12} + W_{21}$ or $J(x)\Delta_F(x - x')J(x') = J(x')\Delta_F(x' - x)J(x)$ is symmetric against the exchange $1 \leftrightarrow 2$ or $x_1 \leftrightarrow x_2$, reflecting that the scalar propagator is an even function. Thus scalar particles are bosons and follow Bose-Einstein statistics.

3.3. Green functions for a free scalar field

In the last section, we obtained the scalar Feynman propagator or two-point Green function as the inverse of the Klein-Gordon operator. As next step, we want to derive n-point Green functions from the their generating functional. Moreover, we will introduce two types of Green functions, namely disconnected n-point functions $\mathcal{G}(x_1, \dots, x_n)$ and connected n-point functions $G(x_1, \dots, x_n)$. Consider the expansion of the exponential in Eq. (3.19),

$$Z[J] = \mathrm{e}^{\mathrm{i}W[J]} = \sum_{n=0}^{\infty} \frac{\mathrm{i}^n}{n!} W^n = \sum_{n=0}^{\infty} \frac{\mathrm{i}^n}{n!} \int \mathrm{d}^4x_1 \cdots \mathrm{d}^4x_n \mathcal{G}_n(x_1, \cdots, x_n) J(x_1) \cdots J(x_n)\,, \tag{3.38}$$

where we assume that $Z[J]$ is normalised so that $Z[0] = 1$. The RHS serves as definition of the disconnected n-point Greens function $\mathcal{G}(x_1, \cdots, x_n)$. They can be calculated as the functional derivatives of $Z[J]$,

$$\mathcal{G}(x_1, \dots, x_n) = \frac{1}{\mathrm{i}^n} \frac{\delta^n}{\delta J(x_1) \cdots \delta J(x_n)} Z[J] \bigg|_{J=0}\,. \tag{3.39}$$

For $n = 2$ we should re-derive the Feynman propagator. Starting from

$$\begin{aligned} \frac{1}{\mathrm{i}} \frac{\delta Z[J]}{\delta J(x)} &= \frac{1}{\mathrm{i}} \frac{\delta}{\delta J(x)} \exp\left(-\frac{\mathrm{i}}{2} \int \mathrm{d}^4x_1 \mathrm{d}^4x_2 J(x_1) \Delta_F(x_1 - x_2) J(x_2) \right) \\ &= -\int \mathrm{d}^4x_1 \Delta_F(x - x_1) J(x_1) \exp(\mathrm{i}W[J])\,, \end{aligned} \tag{3.40}$$

we obtain

$$\begin{aligned} &\frac{1}{\mathrm{i}} \frac{\delta}{\delta J(y)} \frac{1}{\mathrm{i}} \frac{\delta}{\delta J(x)} Z[J] = \mathrm{i}\Delta_F(x - y) \exp(\mathrm{i}W[J]) \\ &+ \left(\int \mathrm{d}^4x_1 \Delta_F(x - x_1) J(x_1) \right) \left(\int \mathrm{d}^4x_1 \Delta_F(y - x_1) J(x_1) \right) \exp(\mathrm{i}W[J])\,. \end{aligned} \tag{3.41}$$

Setting $J = 0$ gives the desired result for the 2-point function,

$$\mathcal{G}(x, y) = \mathrm{i}\Delta_F(x - y)\,. \tag{3.42}$$

It is straightforward to continue: Another functional derivative gives the 3-point function,

$$\begin{aligned}\frac{\delta}{\mathrm{i}\delta J(x_1)}\frac{\delta}{\mathrm{i}\delta J(x_2)}\frac{\delta}{\mathrm{i}\delta J(x_3)}Z[J] = &-\left(\int \mathrm{d}^4x\Delta_F(x_1-x)J(x)\right)\\ &\times\left(\int \mathrm{d}^4x\Delta_F(x_2-x)J(x)\right)\left(\int \mathrm{d}^4x\Delta_F(x_3-x)J(x)\right)\exp(\mathrm{i}W[J])\\ &-\mathrm{i}\Delta_F(x_2-x_3)\int \mathrm{d}^4x\Delta_F(x_1-x)J(x)\exp(\mathrm{i}W[J])\\ &-\mathrm{i}\Delta_F(x_2-x_1)\int \mathrm{d}^4x\Delta_F(x_3-x)J(x)\exp(\mathrm{i}W[J])\\ &-\mathrm{i}\Delta_F(x_3-x_1)\int \mathrm{d}^4x\Delta_F(x_2-x)J(x)\exp(\mathrm{i}W[J])\,. \end{aligned} \tag{3.43}$$

For n odd, we obtain always a source J in the prefactor because $W[J]$ is an even polynomial in J. Hence all odd n-point functions are zero. We continue with the 4-point function: Taking another derivative and setting $J=0$, only terms linear in J in Eq. (3.43) contribute and thus we obtain

$$\begin{aligned}\mathcal{G}(x_1,x_2,x_3,x_4) = &-[\Delta_F(x_1-x_2)\Delta_F(x_3-x_4)\\ &+\Delta_F(x_1-x_3)\Delta_F(x_2-x_4)\\ &+\Delta_F(x_1-x_4)\Delta_F(x_2-x_3)]\,.\end{aligned} \tag{3.44}$$

We see that the 4-point function is the sum of all permutations of products of two 2-point functions. For instance, the first term $\Delta_F(x_1-x_2)\Delta_F(x_3-x_4)$ in the 4-point function describes the independent propagation of a scalar particle from x_1 to x_2 and of another one from x_3 to x_4. Thus our approach leads indeed to a many-particle theory. Since we did not include interactions, particles are propagating independently and the n-point function factorises into products of two-point functions. Thus the functional $Z[J]$ generates disconnected Green functions. The statement that the n-point function is the sum of all permutations of the product of the $n/2$ two-point functions holds for all n and is called "Wick's theorem".

We introduce next the connected n-point functions $G(x_1,\dots,x_n)$. Their generating functional is $W[J]$,

$$G(x_1,\dots,x_n) = \frac{1}{\mathrm{i}^n}\frac{\delta^n}{\delta J(x_1)\cdots\delta J(x_n)}\mathrm{i}W[J]\bigg|_{J=0}\,. \tag{3.45}$$

For a free theory, W is quadratic in the sources J. Hence, all connected n-point functions $G(x_1,\dots,x_n)$ with $n>2$ vanish and the only non-zero one is the two-point function with

$$G(x,y) = \mathrm{i}\Delta_F(x-y) = \mathcal{G}(x,y)\,. \tag{3.46}$$

To summarise: There exists only one non-zero connected n-point function in a free theory which is determined by the Feynman propagator, $G(x,y)=\mathrm{i}\Delta_F(x-y)$. All non-zero disconnected n-point functions can be obtained by permuting the product of $n/2$ two-point functions ("Wick's theorem"). Hence any higher-order Green function can be constructed out of a single building block, the Feynman propagator. In perturbation theory, we will recast the interacting theory—loosely speaking—in "interaction vertices times free propagators". This enables us to derive simple Feynman rules that tell us how one constructs an arbitrary Green function out of vertices and propagators.

Causality and the Feynman propagator We already discussed in section 2.5 that any valid relativistic theory should implement the requirement of causality: No signal using ϕ particles as carrier should travel with a speed larger than the one of light. We also saw that the Feynman prescription leads to a relativistic consistent interpretation of the propagator, although the propagator does not vanish outside the light-cone but goes only exponentially to zero, cf. problem 1.??. One may therefore wonder, if this means that the uncertainty principle makes the light-cone "fuzzy" and thus the axiom of special relativity that no signal can be transmitted with $v > c$ is violated on scales smaller $\lesssim 1/m$.

We can address this question considering the field $\phi(x)$ as operator $\hat{\phi}(x)$ and asking then when a measurement of $\hat{\phi}(x)$ influences $\hat{\phi}(x')$. Recall first that the Feynman propagator equals the 2-point Green function which in turn corresponds to the vacuum expectation value of the time-ordered product of field operators,

$$\mathcal{G}(x_1,x_2) = \langle 0|T\{\hat{\phi}(x_1)\hat{\phi}(x_2)\}|0\rangle = \mathrm{i}\Delta_F(x_1-x_2)\,. \tag{3.47}$$

The property $\Delta_F(x_1-x_2) = \Delta_F(x_2-x_1)$ implies that the field operators $\hat{\phi}(x_1)$ and $\hat{\phi}(x_2)$ commute,

$$\begin{aligned}\langle 0|T\{\hat{\phi}(x_1)\hat{\phi}(x_2)\}|0\rangle &= \langle 0|\hat{\phi}(x_1)\hat{\phi}(x_2)|0\rangle\vartheta(t_1-t_2)\\ &+ \langle 0|\hat{\phi}(x_2)\hat{\phi}(x_1)|0\rangle\vartheta(t_2-t_1)\,.\end{aligned} \tag{3.48}$$

Using the analogy of a free quantum field to an infinite set of oscillators, we try to express the field operator $\hat{\phi}(x)$ through annihilation and creation operators. Comparing to the expansion (2.62) of an oscillator in $d=1$, e.g. to $\phi(t) = (2\omega)^{-1/2}(a\mathrm{e}^{-\mathrm{i}\omega t} + a^\dagger \mathrm{e}^{\mathrm{i}\omega t})$, suggests the ansatz

$$\hat{\phi}(x) = \int \frac{\mathrm{d}^3k}{\sqrt{(2\pi)^3 2\omega_k}}\left[a(\boldsymbol{k})\mathrm{e}^{-\mathrm{i}kx} + a^\dagger(\boldsymbol{k})\mathrm{e}^{+\mathrm{i}kx}\right]\,, \tag{3.49}$$

with $k^0 = \omega_k$, $a(\boldsymbol{k})$ and $a^\dagger(\boldsymbol{k})$ as annihilation and creation operators that satisfy

$$a(\boldsymbol{k})\,|0\rangle = 0\,, \qquad a^\dagger(\boldsymbol{k})\,|0\rangle = |\boldsymbol{k}\rangle \qquad \mathrm{a}nd \qquad [a(\boldsymbol{k}), a^\dagger(\boldsymbol{k}')] = \delta(\boldsymbol{k}-\boldsymbol{k}')\,. \tag{3.50}$$

Hence the vacuum state $|0\rangle$ is defined by $a(\boldsymbol{k})\,|0\rangle = 0$ for all $\boldsymbol{k}$.

If this ansatz is correct, then we should be able to reproduce the known form of the Feynman propagator: Inserting our ansatz for the field into $\langle 0|\hat{\phi}(\boldsymbol{x},t)\hat{\phi}(0)|0\rangle$ for $t>0$, we obtain four terms containing the products aa, $aa^\dagger$, $a^\dagger a$, $a^\dagger a^\dagger$. Only $aa^\dagger$ survives, resulting into

$$\langle 0|\int \frac{\mathrm{d}^3k\,\mathrm{d}^3k'}{(2\pi)^3\sqrt{2\omega_k 2\omega_{k'}}} a(\boldsymbol{k})\mathrm{e}^{-\mathrm{i}kx}a^\dagger(\boldsymbol{k}')\vartheta(t)|0\rangle = \int \frac{\mathrm{d}^3k}{(2\pi)^3 2\omega_k}\mathrm{e}^{-\mathrm{i}kx}\vartheta(t)\,. \tag{3.51}$$

In the second step, we used the commutation rule $[a(\boldsymbol{k}), a^\dagger(\boldsymbol{k}')] = \delta(\boldsymbol{k}-\boldsymbol{k}')$. Performing the same exercise for $t<0$, we see that we reproduce also in this case the corresponding term of the Feynman propagator. Thus we conclude that our ansatz for the field and commutation rules for the annihilation and creation operators are consistent. Note that we could create in Eq. (3.51) alternatively one-particle states, $\langle 0|a(\boldsymbol{k})a^\dagger(\boldsymbol{k}')|0\rangle = \langle \boldsymbol{k}|\boldsymbol{k}'\rangle$, and consistency requires thus that the states $|\boldsymbol{k}\rangle$ are non-relativistically normalised, $\langle \boldsymbol{k}|\boldsymbol{k}'\rangle = \delta(\boldsymbol{k}-\boldsymbol{k}')$. This should come as no surprise, since we started from the analogy to the non-relativistic oscillator. If one prefers states satisfying the relativistic normalisation, one can rescale the creation and annihilation operators such that $[\tilde{a}(k), \tilde{a}^\dagger(k')] = (2\pi)^3 2\omega_k \delta(\boldsymbol{k} - \boldsymbol{k}')$ and the field-operator becomes

$$\hat{\phi}(x) = \int \frac{\mathrm{d}^3k}{(2\pi)^3 2\omega_k} \left[\tilde{a}(k)\mathrm{e}^{-\mathrm{i}kx} + \tilde{a}^\dagger(k)\mathrm{e}^{+\mathrm{i}kx}\right] . \tag{3.52}$$

Both normalisations lead to canonical commutation relations between the field $\hat{\phi}$ and its canonically conjugated momentum density $\hat{\pi} = \dot{\hat{\phi}}$ at equal times,

$$[\hat{\phi}(\boldsymbol{x}, t), \hat{\pi}(\boldsymbol{x}', t)] = \mathrm{i}\delta(\boldsymbol{x} - \boldsymbol{x}') . \tag{3.53a}$$

$$[\hat{\phi}(\boldsymbol{x}, t), \hat{\phi}(\boldsymbol{x}', t)] = [\hat{\pi}(\boldsymbol{x}, t), \hat{\pi}(\boldsymbol{x}', t)] = 0 . \tag{3.53b}$$

We come back to the question if the commutator of two fields vanishes for space-like separation. We evaluate first

$$\begin{aligned}[\hat{\phi}(x), \hat{\phi}(x')] &= \int \frac{\mathrm{d}^3k\, \mathrm{d}^3k'}{(2\pi)^3 \sqrt{2\omega_k 2\omega_{k'}}} \left[a(\boldsymbol{k})\mathrm{e}^{-\mathrm{i}kx} + a^\dagger(\boldsymbol{k})\mathrm{e}^{+\mathrm{i}kx}, a(\boldsymbol{k}')\mathrm{e}^{-\mathrm{i}k'x'} + a^\dagger(\boldsymbol{k}')\mathrm{e}^{+\mathrm{i}k'x'}\right] \\ &= \int \frac{\mathrm{d}^3k}{(2\pi)^3 2\omega_k} \left(\mathrm{e}^{-\mathrm{i}k(x-x')} - \mathrm{e}^{+\mathrm{i}k(x-x')}\right) \equiv D(x-x') . \end{aligned} \tag{3.54}$$

For equal times, $t = t'$, exchanging the dummy variable $\boldsymbol{k} \to -\boldsymbol{k}$ in the second term shows that the contribution from positive and negative energies cancel. Thus the equal-time commutator of two fields is zero, as claimed in (3.53b). For space-like distances, $(x - x')^2 < 0$, we can find a Lorentz boost which changes the ordering of the space-time events, $x - x' \to -(x - x')$. Since the function $D(x - x')$ is the sum of two Lorentz invariant expressions, its value has to be the same in all inertial frames. But for space-like distances, we can transform $D(x)$ into $-D(x)$, and therefore $D(x)$ has to vanish, if x is outside the light-cone of x' and vice versa. Thus we have shown that also the commutator of two space-like separated fields vanishes,

$$[\hat{\phi}(x), \hat{\phi}(x')] = 0 \qquad \text{for} \quad (x - x')^2 < 0 , \tag{3.55}$$

which is the condition for causality: The transmission of a signal corresponds not only to the propagation of a virtual particle, but includes its measurement. Thus the fact that the Feynman propagator does not vanish outside the light-cone does not contradict causality by itself.

There are two main differences between the Feynman propagator and the commutator of two fields: First, $[\hat{\phi}(x), \hat{\phi}(x')]$ is an operator, while $\mathrm{i}\Delta(x_1 - x_2)$ is a vacuum expectation value. The quantum vacuum fluctuates, and these fluctuations are correlated also on space-like distances, similar to the ERP correlations in quantum mechanics. The Feynman propagator $\mathrm{i}\Delta_F(x_1 - x_2)$ is designed to describe not only the propagation of time-like particles, but includes also the space-like propagation of virtual particles: The most "extreme" case is the *instantaneous* exchange of particles transmitting the Coulomb or Yukawa force between static sources, cf. Eq. (3.35). Second, in $[\hat{\phi}(x), \hat{\phi}(x')]$ we subtract the contribution of positive and negative frequencies, while we add them in the Feynman propagator. As a result, the contributions from a particle travelling the distance $\boldsymbol{x}$ and from an antiparticle travelling the distance $-\boldsymbol{x}$ cancel in the commutator, while they add up in the Feynman propagator. Since causality relies on the cancellation between positive and negative energy modes in $[\hat{\phi}(x), \hat{\phi}(x')]$, we conclude that a relativistic quantum theory has to incorporate antiparticles.

3.4. Vacuum energy and the Casimir effect

Vacuum energy We now aim at calculating the energy of the vacuum state of a free scalar quantum field. The energy density ρ of the field ϕ is given by the vacuum expectation value of its Hamiltonian density $\mathscr{H}$,

$$\rho = \langle 0|\mathscr{H}|0\rangle = \rho_0 + \frac{1}{2}\langle 0|\pi^2 + (\boldsymbol{\nabla}\phi)^2 + m^2\phi^2|0\rangle = \rho_0 + \rho_1 \,. \tag{3.56}$$

Here we added the constant energy density ρ_0 to (3.14) and used that the vacuum is normalised, $\langle 0|0\rangle = 1$. For the calculation of ρ_1, we can recycle our result for the propagator of a scalar field by considering $\phi^2(x)$ as the limit of two fields at nearby points,

$$\langle 0|\phi(x')\phi(x)|0\rangle_{x'\searrow x} = \int \frac{\mathrm{d}^3k}{(2\pi)^3 2\omega_k} \mathrm{e}^{-\mathrm{i}k(x'-x)}\bigg|_{x'\searrow x} = \int \frac{\mathrm{d}^3k}{(2\pi)^3 2\omega_k} \,. \tag{3.57}$$

We perform first the differentiation in $\langle \pi^2\rangle = \langle \dot{\phi}^2\rangle$ and $\langle (\boldsymbol{\nabla}\phi)^2\rangle$ and send then $x' \searrow x$. Thus π^2 and $(\boldsymbol{\nabla}\phi)^2$ add a ω_k^2 and $\boldsymbol{k}^2$ term, respectively,

$$\rho = \langle 0|\mathscr{H}|0\rangle = \rho_0 + \int \frac{\mathrm{d}^3k}{(2\pi)^3 2\omega_k}\left[\frac{1}{2}(\omega_k^2 + \boldsymbol{k}^2 + m^2)\right] = \rho_0 + \int \frac{\mathrm{d}^3k}{(2\pi)^3}\frac{1}{2}\omega_k \,. \tag{3.58}$$

If we insert $\hbar$ and c into this expression, we see that ρ_0 as a classical contribution to the energy density of the vacuum is $\propto \hbar^0$, while the second term $\rho_1 \propto \hbar\omega_k/V$ as a quantum correction is linear in $\hbar$. The total energy density ρ of the vacuum state of a free scalar field has a very intuitive interpretation: Additionally to the classical energy density ρ_0, it sums up the zero-point energies of all individual modes $\boldsymbol{k}$ of a free field. Despite its simplicity, we cannot make sense out of this result: Since both the density of modes and their energy increases with $|\boldsymbol{k}|$, the integral diverges. This is the first example that momentum integrals in quantum field theories are often ill-defined and require some care and cure. One calls momentum integrals which are divergent for $k \to 0$ infrared (IR) divergent, while one calls integrals which diverge for $k \to \infty$ ultraviolet (UV) divergent.

Let us now consider the case that the Hamiltonian (3.14) describes physics correctly only up to the energy scale Λ, while the modes with $|\boldsymbol{k}| \gtrsim \Lambda$ do not contribute to ρ_1. Such a possibility exists e.g. in supersymmetric theories where the contributions of different particle types cancel each other above the scale Λ_{SUSY} where supersymmetry is broken. Integrating the contribution to the vacuum energy density by field modes up to the cutoff scale Λ, we find

$$\rho_1 = \int_0^\Lambda \frac{\mathrm{d}k\, k^2}{2\pi^2}\frac{1}{2}\omega_k \sim \Lambda^4 \tag{3.59}$$

in the limit $\Lambda \gg m$. Since only the total energy density ρ is observable, the unknown ρ_0 can be always chosen such that $\rho_0 + \rho_1$ agrees with observations, even if $|\rho_0|, |\rho_1| \gg |\rho|$. Nevertheless, the strong sensitivity of ρ_1 on the value of the cutoff scale Λ is puzzling for two reasons: First, cosmological observations determine the total vacuum energy density ρ_Λ to which all types of fields contribute as $\rho_\Lambda \sim (\mathrm{m}eV)^4$. On the other hand, accelerator experiments give no indications that a cancellation mechanism as supersymmetry works at energy below few TeV. Thus we expect naively at least $\rho_\Lambda \sim (\Lambda_{SUSY})^4 \gtrsim (\mathrm{f}ewTeV)^4$, which is 60 orders of magnitude larger than observed, if no strong cancellation of the various contributions to ρ_Λ takes place. Second, the behaviour $\rho_1 \sim \Lambda^4$ implies that all scalar particles with mass $m \lesssim \Lambda$ contribute equally to ρ. This poses the question, if we have to know the physics at energy scales much larger than those we probe experimentally in order to make predictions using QFT. Such a behaviour would be in contradiction of developing successfully chemistry, atomic or nuclear physics using only the experimental data and models of the corresponding relevant energy scale E. Something similar should happen in QFT too and we will study later the conditions that heavy particles with mas m "decouple" at energies $E \ll m$: In this case, their effects are either suppressed by factors E/m, or are hidden in unobservable quantities like ρ_1.

Casimir effect Although we cannot calculate unambiguously the vacuum energy, we can determine the energy difference of different vacua. As a concrete example, we consider the suggestion by Casimir that the vacuum between two conducting plates is disturbed. As a result, the vacuum energy density between the plates becomes a function of their distance d. The difference of the vacuum energy density inside and outside the plates is finite and leads to a measurable force between them.

Let us consider two parallel, uncharged, perfectly conducting plates at distance d. Standing waves between them have the form $\sin(n\pi x/d)$ with discrete energies $\omega_n = n\pi/d$. The vacuum fluctuations of a photon have the same form as the one of massless scalar field, except that there is an additional factor two due to its two spin degrees of freedom. Thus the vacuum energy inside the box of volume dL_yL_z per single polarisation mode is given by

$$E = L_yL_z \sum_{n=1}^{\infty} \int \frac{\mathrm{d}k_y\mathrm{d}k_z}{(2\pi)^2}\, \frac{1}{2}\sqrt{\left(\frac{n\pi}{d}\right)^2 + k_y^2 + k_z^2}\,. \tag{3.60}$$

To simplify the calculations, we consider a $1+1$ dimensional system of two plates separated by the distance d. Then the energy density $\rho = E/d$ of a massless field per polarisation mode inside the plates is

$$\rho(d) = \frac{\pi}{2d^2}\sum_{n=1}^{\infty} n\,. \tag{3.61}$$

Next we introduce a cutoff function $f(a) = \exp(-an\pi/d)$ which suppresses the high-energy modes,

$$\rho(d) \to \rho(a,d) = \frac{\pi}{2d^2}\sum_{n=1}^{\infty} n\mathrm{e}^{-an\pi/d}\,. \tag{3.62}$$

This procedure is called *regularisation*: For $a > 0$, we obtain a well-defined mathematical sum which we can manipulate following the usual rules of analysis, while we recover for $a \to 0$ the original divergent sum. We have chosen as argument of the exponential $an\pi/d$, because the physically relevant quantities are the energy levels $\omega_n = n\pi/d$ of the system. Now we can evaluate the regularised sum, rewriting it as a geometrical sum,

$$\rho(a,d) = \frac{\pi}{2d^2}\sum_{n=1}^{\infty} n\mathrm{e}^{-an\pi/d} = -\frac{1}{2d}\frac{\partial}{\partial a}\sum_{n=0}^{\infty}\mathrm{e}^{-an\pi/d} \tag{3.63a}$$

$$= -\frac{1}{2d}\frac{\partial}{\partial a}\frac{1}{1-\mathrm{e}^{-a\pi/d}} = \frac{\pi}{2d^2}\frac{\mathrm{e}^{-a\pi/d}}{(1-\mathrm{e}^{-a\pi/d})^2}\,. \tag{3.63b}$$

Then we use $\mathrm{e}^x(1-\mathrm{e}^{-x})^2 = 4\sinh^2(x/2)$ and expand $\rho(a,d)$ for small a in a Laurent series,

$$\rho(a,d) = \frac{\pi}{8d^2}\,\frac{1}{\sinh^2(a\pi/2d)} = \frac{1}{2\pi a^2} - \frac{\pi}{24d^2} + \mathcal{O}(a^2d^{-4})\,. \tag{3.64}$$

Note that we isolated thereby the divergence into a term which does not depend on the distance d of the plates. Thus the divergence cancels in the difference of the vacuum energy with and without plates,

$$\rho_{\mathrm{Cas}}(d) \equiv \lim_{a\to 0}\left[\rho(a,d) - \rho(a,d\to\infty)\right] = -\frac{\pi}{24d^2}\,. \tag{3.65}$$

This final step in order to obtain a finite result is called *renormalisation*. One can verify that the result is not only independent of the cutoff parameter a, but also of the shape of a reasonable[2] cutoff function $f(a)$. In contrast, the a dependent terms in Eq. (3.64) may depend on the form of $f(a)$.

The quantity measured in actual experiments is the force F with which the plates attract (or repel) each other. This force is given by

$$-F = \frac{\partial E}{\partial d} = \frac{\partial(d\rho_{\mathrm{Cas}})}{\partial d} = \frac{\pi}{24d^2}\,. \tag{3.66}$$

Thus two parallel plates attract each other. The experimentally relevant case of electromagnetic waves between two parallel plates in 3+1 dimensions can be calculated analogously. The experimental confirmation of the Casimir effect has been achieved only in the 1990s, with a precision on the 1% level.

How can we understand that the Casimir force is independent on the details of the regularisation procedure? Let us compare the impact of the two plates on modes with different wave-number $k = 2\pi/\lambda$: In a typical experimental set-up, the plates are separated by a distance of the order $d \sim 1\,\mathrm{mm}$ and thus $k_0 \equiv 2\pi/d \sim \mathrm{meV}$. The plates eliminate all low-energy modes with $k < k_0$ between them, while the modes with $k > k_0$ attain a discrete spectrum. However, for $k_n \gg k_0$, the spacing between the modes becomes negligible and experimentally one cannot distinguish the discrete spectrum from a continuous one. In particular, we can approximate the sum over the discrete energies by an integral and the contributions of modes with $k \gg k_0$ with and without plates cancel calculating the energy difference. Since the main contribution to the Casimir energy comes from cutting off modes with $k \lesssim 2\pi/d \sim \mathrm{meV}$, we conclude that the Casimir energy is an IR effect. Therefore the details of the UV regularisation should not influence the result and any reasonable cutoff function that makes the mathematical manipulations (3.63a-3.64) well-defined should lead to the same result.

Summary

The exchange of time-like quanta with zero energy between two static sources leads to the Yukawa potential. The corresponding force mediated by a scalar field is attractive. The Feynman propagator obtained by the $m^2 - \mathrm{i}\varepsilon$ prescription is the unique Green function which can be analytically continued to an Euclidean Green function. It propagates particles (with positive frequencies) forward in time, while anti-particles (with negative frequencies) propagate backward in time. While these two contributions add up in the scalar Feynman propagator, they cancel in the commutator of field-operators at space-like distances, as required by causality. Disconnected n-point Green functions are generated by the functional $Z[J]$, while

[2] Reasonable means that $f(a)$ is normalised, $f(0) = 1$, and that all its derivatives vanish for large a, $\lim_{a\to\infty} f^{(n)}(a) = 0$.

$\mathrm{i}W[J] = \ln Z[J]$ generates connected Green functions. Wick's theorem says that a n-point function can be obtained as the permutation of products of 2-point functions. The Casimir effect shows that the zero-point energies of quantum fields have real, measurable consequences.

Further reading

The quantisation of free fields using both canonical quantisation and the path integral approach is discussed extensively by [GR08].

4. Scalar field with $\lambda\phi^4$ interaction

We know from quantum mechanics that adding an anharmonic term to an oscillator forces us to use either perturbative or numerical methods. The same happens in field theory: No analytic solution for a realistic interacting theory is at present known in $n = 4$ space-time dimensions. Therefore we develop in this chapter a perturbative method to evaluate the generating functionals $Z[J]$ and $W[J]$. We continue to work with the simplest case of a single real scalar field and choose as interaction a $\lambda\phi^4$ term. Then the coupling constant λ is dimensionless in the for us interesting case $n = 4$. If λ is small enough, we may hope that a perturbative series expansion in λ provides a useful approximation scheme. As motivation, we note that a scalar field with $\lambda\phi^4$ interaction can not only model a wide range of phenomena in statistical physics but describes also the Higgs field of the SM and its self-interactions.

4.1. Perturbation theory for interacting fields

General formalism The Lagrange density $\mathscr{L}$ in the functional $Z[J]$ for the scalar field considered up to now was at most quadratic in the fields and its derivatives. On one hand, this allowed us to evaluate the path integral, while on the other hand this means that the field has no interactions: Two wave packets described by the free propagator just pass each other without interaction, as the superposition principle prescribes. As next step we add therefore an interaction term $\mathscr{L}_I$ to the free Lagrangian $\mathscr{L}_0$, i.e. we set $\mathscr{L} = \mathscr{L}_0 + \mathscr{L}_I$. Then the generating functional $Z[J]$ for an interacting real scalar field ϕ becomes

$$Z[J] = \int \mathcal{D}\phi \exp \mathrm{i} \int \mathrm{d}^4x \; (\mathscr{L}_0 + \mathscr{L}_I + J\phi) \,, \tag{4.1}$$

while we denote the free functionals we considered so far from now on as $Z_0[J]$ and $W_0[J]$.

Starting from the full generating functional $Z[J]$ we can define *exact* Green functions which we denote by boldface letters: For instance, the exact 2-point function or propagator is given analogous to Eq. (3.45) by

$$\boldsymbol{G}(x_1, x_2) = \frac{1}{\mathrm{i}^2} \frac{\delta^2 Z[J]}{\delta J(x_1)\delta J(x_2)}\bigg|_{J=0} = \int \mathcal{D}\phi \; \phi(x_1)\phi(x_2) \mathrm{e}^{\mathrm{i}\int \mathrm{d}^4x\,[\mathscr{L}_0 + \mathscr{L}_I]} \,. \tag{4.2}$$

In general, we are not able to calculate these exact Green functions, and we will apply therefore perturbation theory. We assume that the interaction term $\mathscr{L}_I$ is a polynomial $\mathcal{P}(\phi)$ of degree ≥ 3 in the field ϕ and contains an expansion parameter λ which is small in the considered kinematic regime, $\mathscr{L}_I = \lambda\mathcal{P}(\phi)$ with $\lambda \ll 1$. This suggests to expand the interaction term,

$$\exp \mathrm{i} \int \mathrm{d}^4x \mathscr{L}_I(\phi) = 1 + \mathrm{i}\lambda \int \mathrm{d}^4x \mathcal{P}(\phi(x)) + \frac{(\mathrm{i}\lambda)^2}{2!} \int \mathrm{d}^4x_1 \mathrm{d}^4x_2 \mathcal{P}(\phi(x_1))\mathcal{P}(\phi(x_2)) + \ldots \tag{4.3}$$

Since

$$\mathrm{i}\phi(x) \mathrm{e}^{\mathrm{i}\int \mathrm{d}^4x' (\mathscr{L}_0 + J\phi)} = \frac{\delta}{\delta J(x)} \, \mathrm{e}^{\mathrm{i}\int \mathrm{d}^4x' (\mathscr{L}_0 + J\phi)} \,, \tag{4.4}$$

we can perform the replacement

$$\mathscr{L}_I(\phi(x)) \to \mathscr{L}_I\left(\frac{1}{\mathrm{i}}\frac{\delta}{\delta J(x)}\right). \tag{4.5}$$

Then the interaction $\mathscr{L}_I$ does not depend longer on ϕ and can be pulled out of the functional integral,

$$Z[J] = \exp \mathrm{i}\int \mathrm{d}^4x \mathscr{L}_I\left(\frac{1}{\mathrm{i}}\frac{\delta}{\delta J(x)}\right)\int \mathcal{D}\phi \exp \mathrm{i}\int \mathrm{d}^4x\ (\mathscr{L}_0 + J\phi) \tag{4.6a}$$

$$= \exp \mathrm{i}\int \mathrm{d}^4x \mathscr{L}_I\left(\frac{1}{\mathrm{i}}\frac{\delta}{\delta J(x)}\right) Z_0[J] = \exp \mathrm{i}\int \mathrm{d}^4x \mathscr{L}_I\left(\frac{1}{\mathrm{i}}\frac{\delta}{\delta J(x)}\right) \mathrm{e}^{\mathrm{i}W_0[J]}. \tag{4.6b}$$

The solution of the free functionals Z_0 and W_0 was given in Eq. (3.19) as

$$Z_0[J] = Z_0[0]\exp\left(-\frac{\mathrm{i}}{2}\int \mathrm{d}^4x\mathrm{d}^4x' J(x)\Delta_F(x-x')J(x')\right) = Z_0[0]\sum_{n=0}^{\infty}\frac{\mathrm{i}^n}{n!}W_0^n. \tag{4.7}$$

Perturbation theory consists in a double expansion of the two exponentials in $Z[J]$: One in the coupling constant λ and one in the number of external sources J. The latter is fixed by the number of external particles in a scattering process, while the former is chosen according to the desired precision of the calculation.

Choosing the interaction term Let us recall from our discussion of the free Lagrangian the physical requirements we should impose on the Lagrangian: Each term should be a Lorentz scalar which is local in the fields. The corresponding Hamiltonian has to be bounded from below, stable against small perturbations and real. These conditions assure that the vacuum in the absence of external sources is stable.

Additional restrictions follow from a surprisingly simple argument employing dimensional analysis: Using natural units, $\hbar = c = 1$, the dimension of all physical quantities can be expressed as powers of one basic unit which we choose as mass m. Then we use that the action has dimension zero, $[S] = m^0$, and thus the Lagrangian $[\mathscr{L}] = m^4$ in four space-time dimensions. We consider next the free Lagrangian: From the kinetic term, we conclude that the dimension of a scalar field is $[\phi] = m^1$. Thus simple dimensional analysis shows that the term m^2 in front of ϕ^2 has the interpretation of a mass squared. Furthermore, we can order possible self-couplings of a scalar field according to their dimension as

$$\mathscr{L}_I = g_3 M\phi^3 + g_4\phi^4 + \frac{g_5}{M}\phi^5 + \dots, \tag{4.8}$$

where the coupling constants g_i are dimensionless and we introduced the mass scale M to ensure $[\mathscr{L}] = m^4$. We call ϕ^d an operator of dimension d. Similar as in the case of the Fermi constant, $G_F = \sqrt{2}g^2/(8m_W^2)$, the scale M could be connected to the exchange of a heavy particle.

Let us now estimate by dimensional analysis which energy scaling of the interaction probability we expect for the different coupling terms in $\mathscr{L}_I$. At lowest order perturbation theory, the interaction probability is $\mathrm{d}W \propto |\mathscr{L}_I|^2$. Hence the interaction probability scales as $\propto (g_d/M^{d-4})^2$. Now we consider the ultra-relativistic limit, so that we can neglect the mass m of the scalar particle compared to the center-of-mass (cms) energy $\sqrt{s}$. A probability has to be dimensionless, and for $s \gg m^2$ the only remaining dimensionfull variable that can enter the total interaction probability W is s. Thus W has to scale as $g_d^2(s/M^2)^{d-4}$ in the limit $s \gg m^2$. Let us now distinguish the two ranges $m^2 \ll s \ll M^2$ and $s \gg M^2$. In the latter case, the interaction terms with $d > 4$ contain the large factors $(\sqrt{s}/M))^{d-4} \gg 1$ and perturbation theory becomes thus unreliable. In contrast, these terms are smaller than one below the scale M and thus suppressed relative to the operators with dimension $d \leq 4$. Therefore we neglect in a first approach all operators with dimension $d \geq 5$. Simplifying further $\mathscr{L}_I$, we want to include only one interaction term. In this case, a ϕ^3 term would lead to an unstable vacuum. Therefore our choice for the scalar self-interaction is $\mathscr{L}_I = -\lambda\phi^4/4!$, where the factor $1/4!$ was added for later convenience. If this choice of interaction is realised in Nature for a specific particle has to be decided by experiment.

4.2. Green functions for the $\lambda\phi^4$ theory

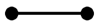
We start now with the perturbative evaluation of Eq. (4.6a) for a $\lambda\phi^4/4!$ interaction. From

$$\begin{aligned} Z[J] &= \left(1 - \frac{\mathrm{i}\lambda}{4!}\int \mathrm{d}^4x \, \frac{\delta^4}{\mathrm{i}^4\delta J(x)^4} + \ldots\right) Z_0[J] = Z_0[J] - \frac{\mathrm{i}\lambda}{4!}\int \mathrm{d}^4x \, \frac{\delta^4 Z_0[J]}{\delta J(x)^4} + \ldots \\ &= Z_0[J]\left(1 + \lambda z_1[J] + \lambda^2 z_2[J] + \ldots\right) \end{aligned} \tag{4.9}$$

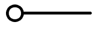
we see that we will generate a series of the type free Green functions plus higher order corrections in λ. The calculation of the first-order correction is very similar to the calculation of the free four-point function, with the difference that now the four sources sit at the same point. You should find in problem **??** as result

$$\begin{aligned} \left(\frac{\delta}{\mathrm{i}\delta J(x)}\right)^4 \exp(\mathrm{i}W_0[J]) = &\left[3(\mathrm{i}\Delta_F(0))^2 + 6\mathrm{i}\Delta_F(0)\left(\int \mathrm{d}^4y\Delta_F(x-y)J(y)\right)^2\right. \\ &\left. + \left(\int \mathrm{d}^4y\Delta_F(x-y)J(y)\right)^4\right]\exp(\mathrm{i}W_0[J])\,. \end{aligned} \tag{4.10}$$

Next we introduce a graphical representation for the various terms in Eq. (4.10). Each Feynman propagator $\Delta_F(x-y)$ is represented by

$\mathrm{i}\Delta_F(x-y) =x$  y (4.11)

a source term $J(x)$ by

$\mathrm{i}\int \mathrm{d}^4x \; J(x) =$  (4.12)

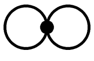
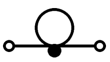
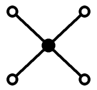
and an interaction vertex by

$$-\,\mathrm{i}\lambda\int \mathrm{d}^4x = \bullet \tag{4.13}$$

Each source and vertex has its own coordinates and an integration over all coordinates is implied. In the case of the ϕ^4 interaction, a vertex connects four lines. Using this notation[1],

[1] This graphical notation first introduced by Stückelberg was made popular by Feynman. The graphs are therefore often called Feynman diagrams or Feynman graphs.

we can express Z_1 as

$Z_1[J] = \frac{1}{4!}\Big(3$  $+\,6$  $+$  $\Big) \exp\left(\frac{1}{2}\ \circ\!\!-\!\!-\!\!\circ\right)\,. \qquad (4.14)$

A graph consists of lines and dots, where the latter may be vertices or sources. We distinguish internal and external lines: A line which ends on both sides at a dot with at least two lines attached is called internal; otherwise it is an external line. The three graphs contained in $Z_1[J]$ differ by the number of loops, i.e. by the number L of closed lines. A graph with loop number $L = 0$ (as the third one in $Z_1[J]$) is called a tree graph, otherwise it is a loop graph. An inspection of the three graphs shows their loop number L is connected to the number n of lines and d of dots as $L = n - d + 1$. Expressing L via the number of vertices and sources, $d = V + S$, and the number of internal and external lines, $n = I + E$, we have $L = I + E - V - S + 1$. Since each external line comes with one source, we can express therefore the loop number also as $L = I - V + 1$, a formula which is valid for all types of interactions. Note also that the first and the second graph contained in $Z_1[J]$ can be obtained from the third one by joining two and one lines, respectively. There are six ways to join one line, and three ways to join two lines. Thus the prefactors of the various graphs can be derived by simple symmetry arguments.

Knowing $Z_1[J]$, we can derive disconnected Green functions valid at $\mathcal{O}(\lambda)$ by performing functional derivatives,

$$\mathcal{G}^{(n)}(x_1,\ldots,x_n) = \frac{1}{\mathrm{i}^n}\frac{\delta^n}{\delta J(x_1)\cdots\delta J(x_n)} Z_0[J]\,(1+\lambda z_1[J])\bigg|_{J=0}\,. \qquad (4.15)$$

In the graphical notation, differentiating with respect to $J(x)$ amounts to replace the open dot denoting the source $\mathrm{i}\int \mathrm{d}^4y J(y)$ by its position x,

$$\frac{1}{\mathrm{i}}\frac{\delta}{\delta J(x)}\ \circ\!\!-\!\!-\!\!-\!\!- = \ x\ -\!\!-\!\!-\!\!- \qquad (4.16)$$

Vacuum diagrams We call terms in the perturbative evaluation of $Z[J]$ which contain no source vacuum diagrams. Since setting $J = 0$ eliminates all graphs containing at least one source, the vacuum diagrams correspond to loops without external lines. The corresponding Green functions are the "zero-point" Green functions $\mathcal{G}^{(0)}$.

Let us assume that the path integral measure $\mathcal{D}\phi$ is chosen such that the free vacuum is normalised, $Z_0[0] = 1$. Switching on interactions will change the free vacuum into the true vacuum of the interacting theory. Therefore the true vacuum and thus $Z[J]$ are not normalised. As example, we obtain setting $J = 0$ in our result (4.14) for $Z[J]$ at lowest order perturbation theory,

$$\mathcal{G}^{(0)} \equiv Z[0] = 1 - \frac{\mathrm{i}\lambda}{8}\int \mathrm{d}^4x(\mathrm{i}\Delta_F(0))^2 \neq 1\,. \qquad (4.17)$$

Because of $\mathcal{N} = \exp\ln(\mathcal{N})$, a normalisation different from one is equivalent to adding a constant term to the Lagrangian,

$$\mathscr{L} \to \mathscr{L} + \ln(\mathcal{N})/(VT) = \mathscr{L} - \rho\,, \qquad (4.18)$$

where VT is the four-dimensional integration volume in the action.

Since vacuum diagrams only change the vacuum energy density ρ but do not contribute to scattering processes, one often prefers to eliminate these diagrams multiplying $Z[J]$ with the normalisation constant

$$\mathcal{N}^{-1} = Z[0] = \int \mathcal{D}\phi\, \mathrm{e}^{\mathrm{i}S}\,. \tag{4.19}$$

Thus one uses the normalised generating functional $\widetilde{Z}[J] \equiv \mathcal{N}Z[J] = Z[J]/Z[0]$ which corresponds to a vacuum with zero energy density ρ. We now show that this normalisation eliminates all vacuum graphs. Expanding the numerator and denominator of $\widetilde{Z}[J]$ up to $\mathcal{O}(\lambda)$, we have at lowest order perturbation theory

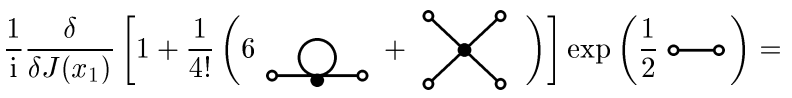
$$\widetilde{Z}[J] = \frac{Z[J]}{Z[0]} = \frac{1+\lambda z_1[J]+\mathcal{O}(\lambda^2)}{1+\lambda z_1[0]+\mathcal{O}(\lambda^2)}\, Z_0[J] = \{1+\lambda(z_1[J]-z_1[0])\}\; Z_0[J] + \mathcal{O}(\lambda^2)\,. \tag{4.20}$$

Thus dividing $Z[J]$ by the source-free functional subtracts indeed the vacuum graph $\mathcal{O}(\lambda)$. It becomes obvious that this procedure works at any order perturbation theory, if we look at the generating functional for connected graphs, $W[J]$. As dividing $Z[J]$ by the source-free functional $Z[0]$ corresponds to

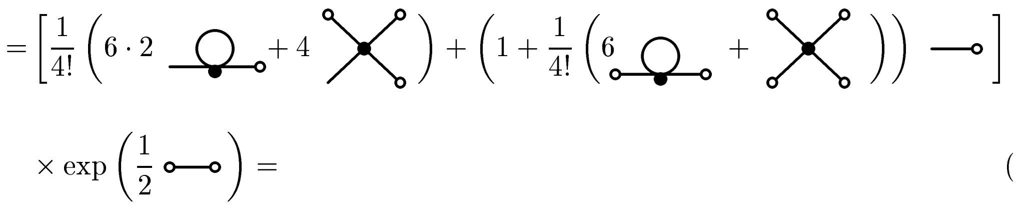
$$\mathrm{i}\widetilde{W}[J] = \ln \widetilde{Z}[J] = \ln Z[J] - \ln Z[0]\,, \tag{4.21}$$

it is clear that this procedure eliminates indeed all vacuum graphs.

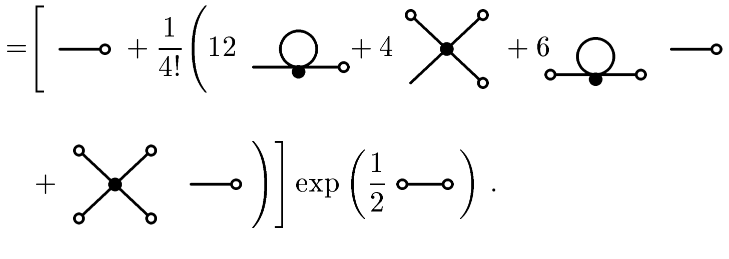
2-point functions We start by taking one derivative of the normalised generating functional,



(4.22a)



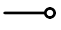
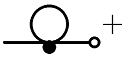
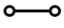
(4.22b)

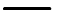
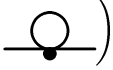
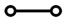


(4.22c)

Every term in this expression contains at least one source J, and the one-point function $\mathcal{G}^{(1)}(x)$ vanishes therefore. If we proceed to the two-point function $\mathcal{G}^{(2)}(x_1,x_2)$, we have to differentiate only those terms with one source,

$$\frac{1}{\mathrm{i}^2}\frac{\delta^2}{\delta J(x_1)\delta J(x_2)}\widetilde{Z}[J] =$$

$$= \frac{1}{\mathrm{i}}\frac{\delta}{\delta J(x_2)}\Bigg[$$  $$+\frac{1}{4!}\Bigg(12$$  $$+ \begin{matrix}\text{vanishing}\\ \text{terms}\\ J=0\end{matrix} \quad \text{for}\Bigg)\Bigg]\exp\Bigg(\frac{1}{2}$$  $$\Bigg)$$

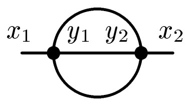
$$= \Bigg($$  $$+\frac{1}{2}$$  $$\Bigg)\exp\Bigg(\frac{1}{2}$$  $$\Bigg)\,. \tag{4.23}$$

Setting then the sources J to zero, the exponential factor becomes one. Converting the graphical formula back into standard notation, we find the 2-point function $\mathcal{G}^{(2)}(x_1,x_2)$ at order $\mathcal{O}(\lambda)$ as the sum of the free 2-point function $\mathcal{G}_0^{(2)}(x_1,x_2)$ and a correction term,

$$\mathcal{G}^{(2)}(x_1,x_2) = \mathcal{G}_0^{(2)}(x_1,x_2) - \frac{\mathrm{i}\lambda}{2}\int \mathrm{d}^4x\, \mathrm{i}\Delta_F(x_1-x)\mathrm{i}\Delta_F(x-x)\mathrm{i}\Delta_F(x-x_2)\,. \tag{4.24}$$

This correction is called the self-energy $\Sigma(x_1,x_2)$ of the scalar particle. Note that the prefactors combine to $6\times 2/4! = 1/2$, so that there appears an extra factor 1/2. Such factors are called symmetry factors. They appear because we included a factor 1/4! in $\mathscr{L}_I$ to compensate for the 4! permutations of four sources. This cancellation works in most diagrams however only partially, and a non-zero prefactor is left over.

Example 4.1: Let us illustrate how one can determine the symmetry factor of more complicated diagrams. As first step, we express the Green function that corresponds to a given Feynman diagram as the time-ordered product of fields. Consider e.g. the so-called "sunrise diagram", which is a second order diagram, corresponding to the term



$$= \frac{1}{2!}\left(\frac{-\mathrm{i}\lambda}{4!}\right)^2\int \mathrm{d}^4y_1\mathrm{d}^4y_2\langle 0|T[\phi(x_1)\phi(x_2)\phi^4(y_1)\phi^4(y_2)]|0\rangle + (y_1\leftrightarrow y_2)$$

in the perturbative expansion. The exchange graph $y_1\leftrightarrow y_2$ is identical to the original one, cancelling the factor $1/2!$ from the Taylor expansion. This cancellation takes place in general: The $1/n!$ factor from the Taylor expansion of a n.th order contains n interaction points which leads to $n!$ permutations. Next consider how the fields ϕ are combined in $T[\cdots]$: The internal points y_1 and y_2 denote interaction points, which have four fields attached. In contrast, the external points x_1 and x_2 carry each only one field. We have to count the number of possible ways to combine the fields in the time-ordered product into the five propagators of the graph. As shorthand notation, we mark a possible combination as $\underbracket{\phi(x_1)\phi(y_1)}$. We have four possibilities to combine $\phi(x_1)$ with $\phi(y_1)$, $\underbracket{\phi(x_1)\phi(y_1)}$. Similarly, there are four possibilities for $\underbracket{\phi(x_2)\phi(y_2)}$. The remaining six fields can be combined in 3! ways into pairs, as e.g. in $\phi(y_1)\phi(y_1)\phi(y_1)\phi(y_2)\phi(y_2)\phi(y_2)$. Thus the symmetry factor of this diagram is given by

$$S = \left(\frac{1}{2!}\times 2!\right)\left(\frac{1}{4!}\right)^2(4\times 4\times 3!) = \frac{1}{3!}\,.$$

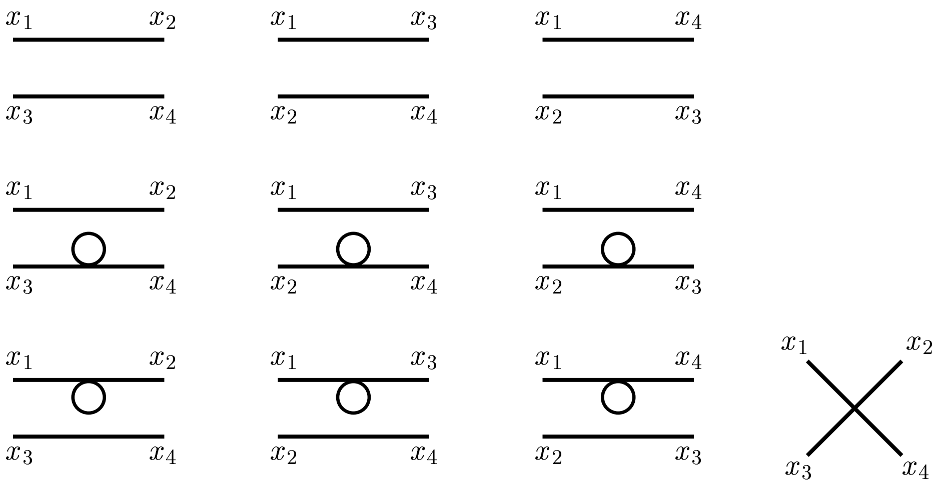
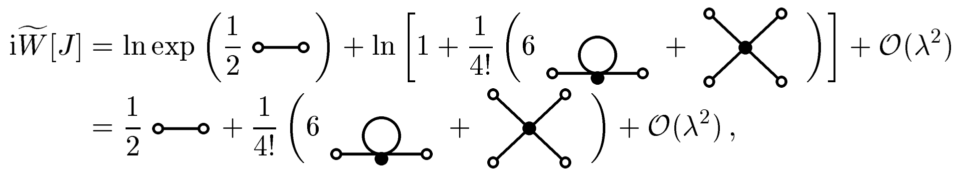


Figure 4.1.: Graphs contributing to the disconnected four-point function $\mathcal{G}(x_1, x_2, x_3, x_4)$.

4-point functions The disconnected 4-point function $\mathcal{G}(x_1, x_2, x_3, x_4)$ is shown graphically in Fig. 4.1. The first three graphs correspond to the free 4-point function from (3.44), the next six graphs are the corresponding $\mathcal{O}(\lambda)$ corrections. Finally, the last diagram corresponds to the connected 4-point function $G(x_1, x_2, x_3, x_4)$. Next we want to derive $G(x_1, x_2, x_3, x_4)$ from its generating functional $\widetilde{W}[J]$. We insert $\widetilde{Z}[J] = \exp(\mathrm{i}\widetilde{W}[J])$ into


$\mathrm{i}\widetilde{W}[J] = \ln\exp\left(\frac{1}{2}\right) + \ln\left[1 + \frac{1}{4!}\left(6 \quad + \right)\right] + \mathcal{O}(\lambda^2)$
$= \frac{1}{2} + \frac{1}{4!}\left(6 \quad + \right) + \mathcal{O}(\lambda^2)\,,$
(4.25)

where we expanded the logarithm, $\ln(1 + x) \simeq x$. Taking four derivatives with respect to J, only the last term survives and we obtain as connected Green function

$$G(x_1, x_2, x_3, x_4) = -\mathrm{i}\lambda \int \mathrm{d}^4x\, \mathrm{i}\Delta_F(x_1 - x)\mathrm{i}\Delta_F(x_2 - x)\mathrm{i}\Delta_F(x_3 - x)\mathrm{i}\Delta_F(x_4 - x)\,. \tag{4.26}$$

Feynman rules for the $\lambda\phi^4$ theory We can summarise our results in a few simple rules which allow us to write down Green functions directly, without the need to derive them from their generating functional. The Feynman rules for connected Green functions $G(x_1, \dots, x_n)$ in coordinate space are as follows:

1. Draw all topologically different diagrams for the chosen order $\mathcal{O}(\lambda^n)$ and number of external coordinates or particles.
2. To each line connecting the points x and x' we associate a propagator $\mathrm{i}\Delta_F(x - x')$.
3. Each vertex has a factor $-\mathrm{i}\lambda$ and connects n lines for a $\lambda\phi^n$ interaction.

4. Integrate over all intermediate points.

5. Determine and add the symmetry factor.

Translation invariance of Minkowski space implies that the propagators depend like $\exp(\pm ikx)$ on the position of the interaction point. Therefore the space-time integrations result in four-momentum conservation at each vertex, $\int d^4x\, e^{-i\sum_j k_j x} = (2\pi)^4\delta(\sum_j k_j)$. In the case of tree-level diagrams, all the four-momentum integrals of the $I = V - 1$ propagators will be eliminated by delta functions, leaving over one delta function expressing overall momentum conservation. In contrast, L integrations over loop momenta $\int d^4k/(2\pi)^4$ remain in the case of loop diagrams. Accounting for the result of the space-time integrations, we can give the Feynman rules directly in momentum space. Defining the Fourier transformed n-point function as

$$G(k_1, \dots, k_n) = \int \prod_{i=1}^{n} d^4x_i \ \exp\Big(i\sum_i k_i x_i\Big) G(x_1, \dots, x_n)\,, \tag{4.27}$$

the Feynman rules for these Green functions have the following changes:

2. To each line associate a propagator $i\Delta_F(k) = i/(k^2 - m^2 + i\varepsilon)$.

3. Fix the external momenta and impose 4-momentum conservation at each vertex.

4. Integrate over all unconstrained momenta k with $\int d^4k/(2\pi)^4$. The number of independent momenta we have to integrate over equals the loop number L of the graph.

We will see in the next section for $\mathcal{G}^{(2)}(p)$ as example how these rules work out.

4.3. Loop diagrams

In the Fourier transformed Green functions $G(k_1, \dots, k_n)$ of tree-level graphs all integrals about the propagator momenta have been cancelled by delta functions, and $G(k_1, \dots, k_n)$ is a mathematically well-defined distribution. In contrast, the integration over momenta in loop graphs are often divergent, requiring to regularise and to renormalise these expressions. Aim of this section is to illustrate this procedure. We concentrate first on the technicalities involved in the evaluation of these loop diagrams, before we interpret the results. We will have time to digest these examples, before we will come back to the problem of renormalisation in chapter 11. The basic steps in the evaluation of simple Feynman integrals are summarised in the appendix 4.A.

4.3.1. Self-energy

We consider first the only one-loop diagram contained in $Z_1[J]$, the 2-point function of a scalar particle at $\mathcal{O}(\lambda)$,

$$G^{(2)}(x_1, x_2) = i\Delta_F(x_1 - x_2) - \frac{\lambda}{2}\Delta_F(0)\int d^4x \Delta_F(x_1 - x)\Delta_F(x - x_2)\,. \tag{4.28}$$

The calculation of $\mathcal{G}^{(2)}(x_1, x_2)$ consists of three steps: First, we have to combine its two pieces into a single, modified propagator: As it stands, the expression seems to describes the propagation of two modes, a free one plus one consisting of the $\mathcal{O}(\lambda)$ correction, while $\mathcal{G}^{(2)}(x_1, x_2)$ should describe the propagation of a single particle with properties modified by the self-interactions. Second, we have to calculate the loop diagram $\mathrm{i}\Delta_F(0)$ which will turn out to be infinite. Thus the final task is the question how we should interpret this result.

We start concentrating on the correction term, which we call the self-energy $\Sigma(x_1, x_2)$ of the scalar particle, and insert the Fourier representation of the two propagators into the integral,

$$\Sigma(x_1, x_2) = -\frac{\lambda}{2}\Delta_F(0) \int \mathrm{d}^4x \frac{\mathrm{d}^4p}{(2\pi)^4} \frac{\mathrm{d}^4p'}{(2\pi)^4} \frac{\mathrm{e}^{-\mathrm{i}p(x_1-x)}}{p^2 - m^2 + \mathrm{i}\varepsilon} \frac{\mathrm{e}^{-\mathrm{i}p'(x-x_2)}}{p'^2 - m^2 + \mathrm{i}\varepsilon} \,. \tag{4.29}$$

The d^4x integration results in $(2\pi)^4\delta(p - p')$, then one of the momentum integrations can be performed. Together this gives

$$\Sigma(x_1 - x_2) = -\frac{\lambda}{2}\Delta_F(0) \int \frac{\mathrm{d}^4p}{(2\pi)^4} \frac{\mathrm{e}^{-\mathrm{i}p(x_1-x_2)}}{(p^2 - m^2 + \mathrm{i}\varepsilon)^2} \,. \tag{4.30}$$

Inserting also for the free Green function its Fourier representation, we arrive at

$$G^{(2)}(x_1 - x_2) = \int \frac{\mathrm{d}^4p}{(2\pi)^4} \, \mathrm{e}^{-\mathrm{i}p(x_1-x_2)} \left[\frac{\mathrm{i}}{p^2 - m^2 + \mathrm{i}\varepsilon} - \frac{\frac{\lambda}{2}\Delta_F(0)}{(p^2 - m^2 + \mathrm{i}\varepsilon)^2}\right] . \tag{4.31}$$

The Green function $G^{(2)}(p)$ in momentum space is thus given by the expression in the square bracket, which we could have written down immediately using the Feynman rules in momentum space. Next we factor out one propagator,

$$G^{(2)}(p) = \frac{\mathrm{i}}{p^2 - m^2 + \mathrm{i}\varepsilon} \left[1 + \frac{\frac{\mathrm{i}\lambda}{2}\Delta_F(0)}{p^2 - m^2 + \mathrm{i}\varepsilon}\right] . \tag{4.32}$$

Assuming that perturbation theory is justified, the second term in the parenthesis should be small. Thus $[1 + \lambda a] = [1 - \lambda a]^{-1} + \mathcal{O}(\lambda^2)$, and we obtain

$$G^{(2)}(p) = \frac{\mathrm{i}}{p^2 - m^2 - \frac{\mathrm{i}\lambda}{2}\Delta_F(0) + \mathrm{i}\varepsilon} \,. \tag{4.33}$$

The residue of the free propagator $\mathrm{i}/(p^2 - m^2 + \mathrm{i}\varepsilon)$ defines the "bare" particle mass m at zero order in λ. Switching on interactions, we continue to define the physical (or renormalised) mass $m_{\mathrm{p}hys}$ of the scalar particle by the residue of $G^{(2)}(p)$. Thus at order λ,

$$m_{\mathrm{p}hys}^2 = m^2 + \delta m^2 = m^2 + \frac{\mathrm{i}\lambda}{2}\Delta_F(0) \,. \tag{4.34}$$

Hence interactions shift or "renormalise" the "bare" mass m used initially in the classical Lagrangian $\mathscr{L}$. It is important to realise that such a renormalisation does not appertain to QFTs but happens in any interacting theory. A familiar example in a classical context is the Debye screening of the electric charge in a plasma. As next step, we have to calculate (and to interpret properly)

$$\mathrm{i}\Delta_F(0) = \int \frac{\mathrm{d}^4k}{(2\pi)^4} \frac{\mathrm{i}}{k^2 - m^2 + \mathrm{i}\varepsilon} \,. \tag{4.35}$$

Since the mass correction is $\delta m^2 = \mathrm{i}\lambda\Delta_F(0)/2$, the Feynman propagator at coincident points $\Delta_F(0)$ has to be purely imaginary. Otherwise the $\lambda\phi^4$ theory would contain no stable particles.

Wick rotation The integrals appearing in loop graphs can be easier integrated, if one performs a Wick rotation from Minkowski to Euclidean space: Rotating the integration contour anti-clockwise to $-\mathrm{i}\infty : +\mathrm{i}\infty$ avoids both poles in the complex k_0 plane and is thus admissible. Introducing as new integration variable $\mathrm{i}k_4 = k_0$, it follows

$$\int_{-\infty}^{\infty} \mathrm{d}k_0 \, \frac{1}{k^2 - m^2 + \mathrm{i}\varepsilon} = \int_{-\mathrm{i}\infty}^{\mathrm{i}\infty} \mathrm{d}k_0 \, \frac{1}{k^2 - m^2 + \mathrm{i}\varepsilon} = \mathrm{i} \int_{-\infty}^{\infty} \mathrm{d}k_4 \, \frac{1}{k^2 - m^2 + \mathrm{i}\varepsilon} \, . \tag{4.36}$$

We next combine $\boldsymbol{k}$ and k_4 into a new four-vector $k_E = (\boldsymbol{k}, k_4)$. Since

$$k^2 = -(|\boldsymbol{k}|^2 + k_4^2) = -k_E^2 \tag{4.37}$$

we work now (apart from the overall sign) in an Euclidean space. In particular, the denominator never vanishes and we can omit the iε. Moreover, the integrand is now spherically symmetric. Thus we have

$$\mathrm{i}\Delta_F(0) = \int \frac{\mathrm{d}^4 k_E}{(2\pi)^4} \, \frac{1}{k_E^2 + m^2} \, . \tag{4.38}$$

As required by our interpretation $\delta m^2 = \mathrm{i}\Delta_F(0)$, the propagator $\Delta_F(0)$ is imaginary. Because the relative sign of the momenta and the mass term indicates if we work in the Euclidean or Minkowski space, we will omit the index E in the following. Introducing furthermore spherical coordinates, we see that $\Delta_F(0)$ diverges quadratically for large k,

$$\lambda \mathrm{i}\Delta_F(0) \propto \int_0^{\Lambda} \mathrm{d}k \, k^3 \frac{1}{k^2 + m^2} \propto \Lambda^2 \, . \tag{4.39}$$

Dimensional regularisation Using the integral representation

$$\frac{1}{k^2 + m^2} = \int_0^{\infty} \mathrm{d}s \, \mathrm{e}^{-s(k^2 + m^2)} \tag{4.40}$$

and interchanging the integrals, we can reduce the momentum integral to a Gaussian integral. Manipulations like interchanging the order of integrations or a change of integration variables in divergent expressions as Eq. (4.39) are however ambiguous. Before we can proceed, we have to "regularise" therefore the integral, similar as we did introducing a cutoff function into the expression of the zero-point energy.

We will use dimensional regularisation (DR), i.e. we will calculate integrals in $d = 4 - 2\varepsilon$ dimensions where they are finite. Then we find

$$\mathrm{i}\Delta_F(0) = \int_0^{\infty} \mathrm{d}s \int \frac{\mathrm{d}^d k}{(2\pi)^d} \, \mathrm{e}^{-s(k^2 + m^2)} = \frac{1}{(4\pi)^{d/2}} \int_0^{\infty} \mathrm{d}s \, s^{-d/2} \mathrm{e}^{-sm^2} \, . \tag{4.41}$$

The substitution $x = sm^2$ transforms the integral into one of the standard representations of the Gamma function (see the appendix 4.A for some useful formula),

$$\mathrm{i}\Delta_F(0) = \frac{(m^2)^{\frac{d}{2}-1}}{(4\pi)^{d/2}} \, \Gamma\left(1 - \frac{d}{2}\right) \, . \tag{4.42}$$

This expression diverges for $d = 2, 4, 6, \ldots$, but is as announced finite for $d = 4 - 2\varepsilon$ and small ε. In the next step, we would like to expand the expression in a Laurent series, separating pole terms in ε and a finite remainder.

Appearance of a dimensionfull scale As the expression stands, we can not expand the prefactor of the Gamma function, because it is dimensionfull. In order to make the factor m^{d-2} dimensionless, we should supply a new mass scale. More physically, we can understand the need for an additional dimensionfull scale by the requirement that the action $S = \int \mathrm{d}^d x \mathscr{L}$ remains dimensionless if we deviate from $d = 4$ dimensions. From the kinetic term, we deduce that the scalar field has the mass dimension $[\phi] = d/2 - 1$. The interaction term implies then that λ acquires the dimension $[\lambda] = 4-d$. In order to retain a dimensionless coupling constant, we introduce therefore a mass μ called the renormalisation scale as follows,

$$S_I = \int \mathrm{d}^4 x \mathscr{L}_I = -\int \mathrm{d}^4 x \frac{\lambda}{4!}\phi^4 \to -\mu^{4-d} \int \mathrm{d}^d x \frac{\lambda}{4!}\phi^4 \,. \tag{4.43}$$

Adding the factor μ^{4-d} to our previous result, we obtain

$$\lambda \mu^{4-d} \mathrm{i}\Delta_F(0) = \lambda \, \frac{m^2}{(4\pi)^2} \left(\frac{4\pi\mu^2}{m^2}\right)^{2-d/2} \Gamma(1-d/2)\,. \tag{4.44}$$

Now we expand the dimensionless last two factors in this expression around $d = 4$ using Eq. (A.42) for the Gamma function,

$$\Gamma(1-d/2) = \Gamma(-1+\varepsilon) = -\frac{1}{\varepsilon} - 1 + \gamma + \mathcal{O}(\varepsilon) \tag{4.45}$$

and

$$a^\varepsilon = \mathrm{e}^{\varepsilon \ln a} = 1 + \varepsilon \ln a + \mathcal{O}(\varepsilon^2)\,. \tag{4.46}$$

Note that we require the expansion of the prefactor of the Gamma function up to $\mathcal{O}(\varepsilon^2)$ because of the pole term in (4.45). Thus the mass correction is given by

$$\lambda \mu^{4-d} \, \mathrm{i}\Delta_F(0) \propto m^2 \left[-\frac{1}{\varepsilon} - 1 + \gamma + \mathcal{O}(\varepsilon)\right] \left[1 + \varepsilon \ln\left(\frac{4\pi\mu^2}{m^2}\right) + \mathcal{O}(\varepsilon^2)\right]\,. \tag{4.47}$$

or

$$\delta m^2 = \frac{\mathrm{i}\lambda}{2} \mu^{4-d} \, \Delta_F(0) = \frac{\lambda}{2} \frac{m^2}{(4\pi)^2} \left[-\frac{1}{\varepsilon} - 1 + \gamma + \ln\left(\frac{m^2}{4\pi\mu^2}\right) + \mathcal{O}(\varepsilon)\right]\,. \tag{4.48}$$

This expansion has allowed us to separate the correction into a divergent term $\propto 1/\varepsilon$ and a finite remainder. The latter contains an analytic part, $-1+\gamma$, and an non-analytic piece that depends on the renormalisation scale, $\ln(m^2/4\pi\mu^2)$. This result is typical for DR: First, all divergences appear in the limit $d \to 4$ as poles of the Gamma function. Second, the renormalisation scale μ enters always via Eq. (4.46) in a logarithm. Thus the only dimensionfull parameter which can set the scale of the mass correction δm^2 is the mass of the particle in the loop. In the case of a theory with a single particle, the correction must have therefore the form $\delta m^2 \propto m^2$ using DR. You should contrast this behaviour with the one using as regularisation scheme an Euclidean cutoff Λ: Integrating up to momenta $\Lambda \gg m$, the particle mass m can be neglected, and the correction diverges as a power-law $\delta m^2 \propto \Lambda^2$.

Let us now discuss in turn the three different kind of terms present in Eq. (4.48). First, the form of the divergent terms depends on the regularisation scheme applied. However, in any scheme we can eliminate them using Eq. (4.34), requiring that the unobservable bare mass m^2 contains the same divergent terms with the opposite sign. Second, the finite analytic terms depend also on the scheme, since we can always shift a finite part from m^2 to δm^2. In order to specify precisely δm^2 we have to fix therefore a *renormalisation scheme* , i.e. a set of rules resolving these ambiguities. Finally, the finite, non-analytic terms are important predictions, which are independent of the regularisation scheme (apart from a rescaling of the arbitrary parameter μ). As we will see in chapter 9 such non-analytic terms are necessary that the S-matrix is unitary, or in other words that the theory preserves probability.

4.3.2. Vacuum energy density

We can generate out of the self-energy diagram new one-loop graphs by adding or subtracting two external lines. Subtracting two lines generates an one-loop graph without external lines[2], the "zero-point" Green function $G^{(0)}$ at order λ^0. One way to calculate this quantity is to evaluate directly $Z_0[0]$ using

$$\mathrm{D}et A = \exp\ln\mathrm{D}et A = \exp\mathrm{T}r\ln A \tag{4.49}$$

what gives

$$Z_0[0] = \exp\left[-\frac{1}{2}\mathrm{T}r\ln(\Box - m^2)\right]. \tag{4.50}$$

We postpone the question how such an expression can be evaluated and use instead another approach, recycling our result for the self-energy. Vacuum diagrams are generated by the functional $Z[J]$ setting $J = 0$,

$$\langle 0+|0-\rangle = Z_0[0] = \int\mathcal{D}\phi\exp\mathrm{i}\int\mathrm{d}^4x\left(\frac{1}{2}\partial_\mu\phi\partial^\mu\phi - \frac{1}{2}m^2\phi^2\right). \tag{4.51}$$

We saw in Eq. (3.57) that the zero-point energy is related to the propagator at coincident points. Since we suspect a connection between vacuum diagrams and the zero-point energy, we try to relate $Z[0]$ and $\Delta_F(0)$. Taking a derivative with respect to m^2 gives

$$\frac{\partial}{\partial m^2}\langle 0+|0-\rangle = -\frac{\mathrm{i}}{2}\int\mathrm{d}^4x\,\langle 0+|\phi(x)^2|0-\rangle = -\frac{\mathrm{i}}{2}\int\mathrm{d}^4x\,\mathrm{i}\Delta_F(0)\langle 0+|0-\rangle. \tag{4.52}$$

The additional factor $\langle 0+|0-\rangle = \mathcal{N}^{-1}$ on the RHS takes into account that we defined the Feynman propagator with respect to a normalised vacuum. Translation invariance implies that $\langle 0+|0-\rangle$ does not depend on x. Thus we obtain

$$\frac{\partial}{\partial m^2}\ln\langle 0+|0-\rangle = -\frac{\mathrm{i}}{2}\int\mathrm{d}^4x\,\mathrm{i}\Delta_F(0) = -\frac{\mathrm{i}}{2}VT\,\mathrm{i}\Delta_F(0) \tag{4.53}$$

with VT as the four-dimensional integration volume. Integrating and exponentiating the resulting formal solution, we obtain

$$\langle 0+|0-\rangle = \exp\left\{-\frac{\mathrm{i}}{2}VT\int\mathrm{d}m^2\,\mathrm{i}\Delta_F(0)\right\}. \tag{4.54}$$

Comparing this result to

$$\langle 0+|0-\rangle = \langle 0+|\exp(-\mathrm{i}HT)|0-\rangle = \exp(-\mathrm{i}\rho VT), \tag{4.55}$$

[2] Although $G^{(0)}$ is often represented as a closed loop, it has also no internal line; this is in agreement with our general formula $l = n - V + 1$.

we see that we should associate

$$\rho = \frac{1}{2}\int \mathrm{d}m^2\, \mathrm{i}\Delta_F(0) \tag{4.56}$$

with the energy density of the vacuum. On the other hand, we can connect ρ to the source-free generating functionals as

$$\rho = \frac{\mathrm{i}\ln Z[0]}{VT} = \frac{-W[0]}{VT}\,, \tag{4.57}$$

Thus the contribution of quantum fluctuations to the energy density of the vacuum is given by the sum of connected vacuum graphs, in accordance with (4.21).

Next we evaluate (4.56) which gives the contribution of a free scalar field to the vacuum energy density. Using our result (4.42) for the propagator, $\mathrm{i}\Delta_F(0) = C(m^2)^{d/2-1}$, we can perform the integration over m^2,

$$\rho = C\,\frac{m^d}{d} - \rho_0 = \frac{m^d}{(4\pi)^{d/2}d}\,\Gamma\left(1-\frac{d}{2}\right) - \rho_0\,, \tag{4.58}$$

where we introduced the integration constant ρ_0.

The energy density given by Eq. (4.58) diverges for $d = 2, 4, 6\ldots$ as $1/\varepsilon$. We can make ρ finite and equal to the observed value ρ_Λ, if we choose ρ_0 as

$$\rho_0 = \mu^{d-4}\left[\frac{1}{4}\,\frac{m^4}{(4\pi)^2}\,\frac{1}{\varepsilon} - \rho_\Lambda\right]. \tag{4.59}$$

The prefactor μ^{d-4} ensures again that the action remains dimensionless also for $d \neq 4$.

Note that this implies that we should start off with $\mathscr{L} - \rho_0$ instead of $\mathscr{L}$. Even if we dismiss a non-zero vacuum energy in the classical Lagrangian, it will appear automatically by quantum corrections. More generally, every possible term that is not forbidden by a symmetry in $\mathscr{L}$ will show up calculating loop corrections. We have seen that we can absorb the vacuum energy density ρ into the normalisation of the path integral,

$$\int \mathcal{D}\phi\, \mathrm{e}^{-\int \mathrm{d}^4x\rho} = \mathcal{N}\int \mathcal{D}\phi\,.$$

Therefore, one may wonder if ρ has a real physical meaning or could be eliminated by a simple redefinition of the integration measure. The answer is no: First, we used our freedom to define the path integral measure setting $Z_0[0] = 1$. Second, ρ depends on the parameters (masses, coupling constants) of the considered theory, but the path integral measure should be independent of the details of the Lagrangian we integrate.

Remark 4.1: Equivalence to the zero-point energy:
Performing the k^0 integral in Eq. (4.35) or using (3.28)

$$\mathrm{i}\Delta_F(0) = \int \frac{\mathrm{d}^3k}{(2\pi)^3 2\omega_k} = \int \frac{\mathrm{d}^3k}{(2\pi)^3 2\sqrt{m^2+\boldsymbol{k}^2}} \tag{4.60}$$

and integrating then with respect to m^2,

$$\rho = \frac{1}{2}\int \mathrm{d}m^2\, \mathrm{i}\Delta_F(0) = \frac{1}{2}\int \frac{\mathrm{d}^3k}{(2\pi)^3}\,\sqrt{m^2+\boldsymbol{k}^2}\,, \tag{4.61}$$

shows that the present expression for the vacuum energy agrees with the sum over zero-point energy evaluated in Eq. (3.58). However, the results for the unrenormalised ρ differ: While Eq. (3.59) shows that $\rho \propto \Lambda^4$ using a cutoff, we have obtained $\rho \propto m^4$ in the case of dimensional regularisation. Thus in this scheme a massless particle as the photon would give a zero contribution to the cosmological constant. We will come back to this difference in chapter 25.

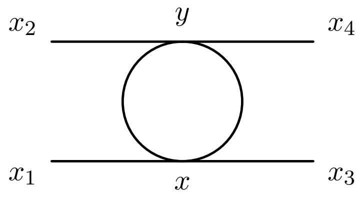
4.3.3. Vertex correction

Feynman amplitude For our last example we add two external lines to the self-energy diagram. The corresponding Green function can be used to describe 2 → 2 scattering at $\mathcal{O}(\lambda^2)$. A scattering process corresponds however to a transition between an initial state at $t = -\infty$ and a final state at $t = \infty$ which contain both real, on-shell particles. In order to obtain scattering amplitudes, we should therefore replace the propagators of external lines—which describe virtual particles—by on-shell wave functions. This rule will be derived later in chapter 9.2. For the moment, we simply anticipate that we obtain the Feynman amplitude $\mathrm{i}\mathcal{A}$ describing a scattering process using the Feynman rules for momentum space, but writing for scalar external particles simply the prefactor of a plane-wave without normalisation factor, i.e. simply "1". Moreover, we omit the delta function expressing the conservation of the external momenta. Thus we add to the Feynman rules in momentum space:

6. The Feynman amplitude $\mathrm{i}\mathcal{A}$ describing scattering processes is obtained omitting the factor $(2\pi)^4\delta(\sum_k k_i - \sum_f k_f)$ expressing the conservation of the external momenta, and the propagators on external lines.

Determining the Feynman amplitude Instead of calculating the order λ^2 term in the perturbative expansion of the generating functional $Z[J]$ we use directly the Feynman rules to obtain the Feynman amplitude for this process. According to these rules, the first steps in the calculation of the Feynman amplitude are to draw all Feynman diagrams, to find the symmetry factor and to associate then the corresponding mathematical expressions to the graphical symbols.

We determine first the symmetry factor, following the same procedure as in example 4.1. In coordinate space, we have to connect four external points (say $x_1, \ldots, x_4$) with the help of two vertices (say at x and y) which combine each four lines. An example is shown here



Two additional diagrams are obtained connecting x_1 with x_2 or x_4. In order to determine the symmetry factor, we consider the expression for the four-point function corresponding to the graph shown above,

$$\frac{1}{2!}\left(-\mathrm{i}\frac{\lambda}{4!}\right)^2 \int \mathrm{d}^4x\mathrm{d}^4y \langle 0|T\{\phi(x_1)\phi(x_2)\phi(x_3)\phi(x_4)\phi^4(x)\phi^4(y)\}|0\rangle + (x \leftrightarrow y)\,, \tag{4.62}$$

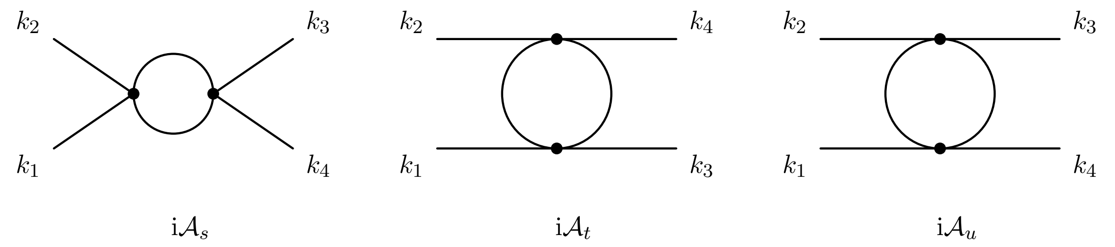

k_2
k_3
k_1
k_4
$\mathrm{i}\mathcal{A}_s$
k_2
k_4
k_1
k_3
$\mathrm{i}\mathcal{A}_t$
k_2
k_3
k_1
k_4
$\mathrm{i}\mathcal{A}_u$

Figure 4.2.: Three graphs contributing to $\phi(k_1)\phi(k_2) \to \phi(k_3)\phi(k_4)$ at order λ^2.

and count the number of possible contractions: We can connect $\phi(x_1)$ with each one of the four $\phi(x)$, and then $\phi(x_3)$ with one of the three remaining $\phi(x)$. This gives 4×3 possibilities. Another 4×3 possibilities come by the same reasoning from the upper part of the graph. The remaining pairs $\phi^2(x)$ and $\phi^2(y)$ can be combined in two possibilities. Finally, the factor 1/2! from the Taylor expansion is cancelled by the exchange graph. Thus the symmetry factor is

$$S = \frac{1}{2!} 2! \left(\frac{4 \times 3}{4!} \right)^2 2 = \frac{1}{2} \,. \tag{4.63}$$

Next we associate mathematical expressions to the symbols of the graphs in momentum space: We replace internal propagators by $\mathrm{i}\Delta(k)$, external lines by 1 and vertices by $-\mathrm{i}\lambda$. Imposing four-momentum conservation at the two vertices leaves one free loop momentum, which we call p. The momentum of the other propagator is then fixed to $p - q$, where $q^2 = s = (p_1 + p_2)^2$, $q^2 = t = (p_1 - p_3)^2$, and $q^2 = u = (p_1 - p_4)^2$ for the three graphs shown in Fig. 4.2. Thus the Feynman amplitudes in $n = 4$ are at order $\mathcal{O}(\lambda^2)$

$$\mathrm{i}\mathcal{A}_q^{(2)} = \frac{1}{2}\lambda^2 \int \frac{\mathrm{d}^4 p}{(2\pi)^4} \frac{1}{[p^2 - m^2 + \mathrm{i}\varepsilon]} \frac{1}{[(p-q)^2 - m^2 + \mathrm{i}\varepsilon]} \,. \tag{4.64}$$

The squared cms energy s and the two variables describing the momentum transfer t and u are called Mandelstam variables. For 2 → 2 scattering, they are connected by $s + t + u = m_1^2 + m_2^2 + m_3^2 + m_4^2$, see problem 4.**??**. According to the value of q^2 one calls the diagrams the s, t and u channel.

Performing again a simple counting of the powers of loop momenta, we find that the amplitude is logarithmically divergent,

$$\mathcal{A}_q^{(2)} \propto \int \frac{\mathrm{d}^4 p}{p^4} \propto \ln(\Lambda) \,. \tag{4.65}$$

If we consider the infinite number of one-loop graphs characterised by $n = V \geq 0$, then we see that adding two external lines increases the number of propagators in the loop by one. As a result, the convergence of the loop integral improves from a quartic divergence (vacuum energy), over a quadratic divergence (self-energy energy) to a logarithmic divergence for the vertex correction. Adding two or more external lines to the vertex correction would therefore produce a finite diagram. At one-loop, the $\lambda\phi^4$ theory contains thus only three divergent Feynman graphs.

Calculating the loop integral The path to be followed in the evaluation of simple loop integrals as (4.64) can be sketched schematically as follows: Regularise the integral (and add a mass scale if you use DR). Combine then the denominators, and shift the integration variable to eliminate linear terms in the denominator by completing the square. Performing the same shift of variables in the numerator, linear terms can be dropped as they vanish after integration. Finally, Wick rotate the integrand, and reduce the integral to a known one by a suitable variable substitution. We do the last steps once in the appendix 4.A where we derive a list of useful Feynman integrals which we simply look up in the future.

We start by rewriting the integral for $d = 4 - 2\varepsilon$ dimensions as

$$\mathrm{i}\mathcal{A}_q^{(2)} = \frac{1}{2}\lambda^2(\mu^2)^{4-d}\int \frac{\mathrm{d}^d p}{(2\pi)^d}\,\frac{1}{D}\,, \tag{4.66}$$

where we introduced also the short-cut D for the denominator in the integrand. Next we use

$$\frac{1}{ab} = \int_0^1 \frac{\mathrm{d}z}{[az + b(1-z)]^2} \tag{4.67}$$

to combine the two denominators, setting $a = p^2 - m^2$ and $b = (p-q)^2 - m^2$,

$$\mathcal{D} \equiv az + b(1-z) = p^2 - m^2 - 2pq(1-z) + q^2(1-z)\,. \tag{4.68}$$

Then we eliminate the term linear in p substituting $p'^2 = [p - q(1-z)]^2$,

$$\mathcal{D} = p'^2 - m^2 + q^2 z(1-z)\,. \tag{4.69}$$

Since $\mathrm{d}^d p = \mathrm{d}^d p'$, we can drop the primes and find

$$\mathrm{i}\mathcal{A}_q^{(2)} = \frac{1}{2}\lambda^2(\mu^2)^{4-d}\int_0^1 \mathrm{d}z \int \frac{\mathrm{d}^d p}{(2\pi)^d}\,\frac{1}{[p^2 - m^2 + q^2 z(1-z)]^2}\,. \tag{4.70}$$

Performing a Wick rotation requires that $q^2 z(1-z) < m^2$ for all $z \in [0:1]$, or $q^2 < 4m^2$. The integral is of the type $I(\omega, 2)$ calculated in the appendix and equals

$$I(\omega,2) = \mathrm{i}\,\frac{1}{(4\pi)^\omega}\,\frac{\Gamma(2-\omega)}{\Gamma(2)}\,\frac{1}{[m^2 - q^2 z(1-z)]^{2-\omega}}\,. \tag{4.71}$$

Inserting the result into the Feynman amplitude gives

$$\mathcal{A}_q^{(2)} = \frac{1}{2}\lambda^2(\mu^2)^{4-d}\frac{\Gamma(2-d/2)}{(4\pi)^{d/2}}\int_0^1 \mathrm{d}z\,[m^2 - q^2 z(1-z)]^{d/2-2} \tag{4.72a}$$

$$= \frac{\lambda^2}{32\pi^2}(\mu^2)^{2-d/2}\Gamma(2-d/2)\int_0^1 \mathrm{d}z\,\underbrace{\left[\frac{m^2 - q^2 z(1-z)}{4\pi\mu^2}\right]}_{f}^{\frac{d}{2}-2}. \tag{4.72b}$$

In the last step, we made the function f dimensionless. Now we take the limit $\varepsilon = 2 - d/2 \to 0$, expanding both the Gamma function, $\Gamma(2-d/2) = \Gamma(\varepsilon) = 1/\varepsilon - \gamma + \mathcal{O}(\varepsilon)$, and $f^{-\varepsilon}$. From

$$\frac{\lambda^2}{32\pi^2}\mu^{2\varepsilon}\left(\frac{1}{\varepsilon} - \gamma + \mathcal{O}(\varepsilon)\right)\left[1 - \varepsilon\int_0^1 \mathrm{d}z\,\ln f + \mathcal{O}(\varepsilon^2)\right] \tag{4.73}$$

we see that all diagrams give the same divergent part, while we have to replace q^2 by the value $\{s, t, u\}$ appropriate for the three diagrams,

$$\mathcal{A} = \mathcal{A}^{(1)} + \mathcal{A}_s^{(2)} + \mathcal{A}_t^{(2)} + \mathcal{A}_u^{(2)} + \mathcal{O}(\lambda^3) \tag{4.74a}$$

$$= -\lambda\mu^\varepsilon + \frac{3\lambda^2\mu^{2\varepsilon}}{32\pi^2\varepsilon} - \frac{\lambda^2\mu^{2\varepsilon}}{32\pi^2}\left[3\gamma + F(s,m,\mu) + F(t,m,\mu) + F(u,m,\mu)\right], \tag{4.74b}$$

with

$$F(q^2, m, \mu) = \int_0^1 \mathrm{d}z \, \ln\left[\frac{m^2 - q^2 z(1-z)}{4\pi\mu^2}\right]. \tag{4.75}$$

Note that t and u are in the physical region negative and thus the condition $q^2 < 4m^2$ is always satisfied for these two diagrams, cf. with problem 4.**??**. By contrast, for the s channel diagram the relation $q^2 = s > 4m^2$ holds: In this case, we have to continue analytically the result (4.75) into the physical region. We will postpone this task to chapter 9 and note for the moment only that thereby the argument of the logarithm in (4.75) changes sign. Additionally, an imaginary part of the scattering amplitude is generated.

4.3.4. Basic idea of renormalisation

The regularisation of loop integrals has introduced as a new parameter the renormalisation scale μ. As we perform perturbation theory at order λ^n, we have to connect the parameters $\{m_n, \lambda_n, \rho_n\}$ of the truncated theory with the physical ones of the full theory. This process is called renormalisation and will replace the undetermined parameter μ by a physical momentum scale relevant for the considered process.

Renormalisation of the coupling Let us try to connect the amplitude i$\mathcal{A}$ to a physical measurement. We assume that experimentalists measured $\phi\phi \to \phi\phi$ scattering. It is sufficient that they provide us with a single value, e.g. with the value of the differential cross section $\mathrm{d}\sigma/\mathrm{d}\Omega$ at zero-momentum transfer close to threshold $s = 4m^2$. Then the function $F(s, m, \mu)$ is given by

$$F(4m^2, m, \mu) = C\lambda^2 \ln\left[\frac{4m^2}{\mu^2}\right] + const\,, \tag{4.76}$$

while it becomes for $s \gg m^2$

$$F(s, m, \mu) = C\lambda^2 \ln\left[\frac{s}{\mu^2}\right] + const\,. \tag{4.77}$$

Subtracting the infinite parts (and the constant term) from $\mathcal{A}$, we obtain for $s, t, u \gg m^2$,

$$\mathcal{A} = -\lambda - C\lambda^2\left[\ln(s/\mu^2) + \ln(t/\mu^2) + \ln(u/\mu^2)\right] + \mathcal{O}(\lambda^3) \equiv -\lambda - C\lambda^2 L(s/\mu^2)\,, \tag{4.78}$$

where we introduced also the sloppy notation L for the three log terms in the square bracket. This expression for $\mathcal{A}$ is finite but still arbitrary since it contains μ.

We use now the experimental measurement at the scale $s = 4m^2$ to connect via

$$\mathcal{A} = -\lambda - C\lambda^2 L(4m^2/\mu^2) \tag{4.79}$$

the measured value $\lambda_{\mathrm{p}hys}$ of the coupling to our calculation,

$$-\lambda_{\mathrm{p}hys} = -\lambda - C\lambda^2 L(4m^2/\mu^2) + \mathcal{O}(\lambda^3)\,. \tag{4.80}$$

Here we indicated also that our perturbative calculation is only valid up to $\mathcal{O}(\lambda^3)$ terms. Now we solve for λ,

$$\begin{aligned}-\lambda &= -\lambda_{phys} + C\lambda^2 L(4m^2/\mu^2) + \mathcal{O}(\lambda^3) && \text{(4.81a)}\\ &= -\lambda_{phys} + C\lambda^2_{phys} L(4m^2/\mu^2) + \mathcal{O}(\lambda^3_{phys})\,. && \text{(4.81b)}\end{aligned}$$

In the second line, we could replace λ^2 by λ^2_{phys}, because their difference is of $\mathcal{O}(\lambda^3)$. Next we insert λ back into the matrix element $\mathcal{A}$ for general s and replace then again λ^2 by λ^2_{phys},

$$\begin{aligned}\mathcal{A} &= -\lambda - C\lambda^2 L(s/\mu^2) + \mathcal{O}(\lambda^3) && \text{(4.82a)}\\ &= -\lambda_{phys} + C\lambda^2_{phys} L(4m^2/\mu^2) - C\lambda^2_{phys} L(s/\mu^2) + \mathcal{O}(\lambda^3_{phys}) && \text{(4.82b)}\\ &= -\lambda_{phys} - C\lambda^2_{phys} L(s/4m^2) + \mathcal{O}(\lambda^3_{phys})\,. && \text{(4.82c)}\end{aligned}$$

Combining the log's, the scale μ has cancelled and we find

$$\mathcal{A} = -\lambda_{phys} - \frac{\lambda^2_{phys}}{32\pi^2}\left[\ln(s/4m^2) + \ln(t/4m^2) + \ln(u/4m^2)\right] + \mathcal{O}(\lambda^3_{phys})\,. \tag{4.83}$$

Thus the amplitude is finite and depends only on the measured value λ_{phys} of the coupling constant and the kinematical variables s, t and u.

Running coupling We look now from a somewhat different point of view at the problem of the apparent μ dependence of physical observables. Assume again that we have subtracted the infinite parts (and the constant term) of the amplitude $\mathcal{A}$, obtaining Eq. (4.78). We now demand that the scattering amplitude $\mathcal{A}$ as a physical observable is independent of the arbitrary scale μ, $\mathrm{d}\mathcal{A}/\mathrm{d}\mu = 0$. If we did a perturbative calculation up to $\mathcal{O}(\lambda^n)$, the condition $\mathrm{d}\mathcal{A}/\mathrm{d}\mu = 0$ can hold only up to terms $\mathcal{O}(\lambda^{n+1})$. The explicit μ dependence of the amplitude, $\partial\mathcal{A}/\partial\mu \neq 0$, can be only cancelled by a corresponding change of the parameters m and λ contained in the classical Lagrangian, converting them into "running" parameters $m(\mu)$ and $\lambda(\mu)$. Then the condition[3] $\mathrm{d}\mathcal{A}/\mathrm{d}\mu = 0$ becomes

$$\left(\frac{\partial}{\partial\mu} + \frac{\partial m^2}{\partial\mu}\frac{\partial}{\partial m^2} + \frac{\partial\lambda}{\partial\mu}\frac{\partial}{\partial\lambda}\right)\mathcal{A}(s,t,m(\mu),\lambda(\mu),\mu) = 0\,. \tag{4.84}$$

The only explicit μ dependence of $\mathcal{A}$ is contained in the $F(q^2,m,\mu)$ functions, giving in the limit $\varepsilon \to 0$

$$\frac{\partial}{\partial\mu}\mathcal{A}(s,t,m(\mu),\lambda(\mu),\mu) = -3\,\frac{\lambda^2}{32\pi^2}\,\frac{\partial}{\partial\mu}F(q^2,m,\mu) = \frac{3\lambda^2}{16\pi^2\mu}\,. \tag{4.85}$$

Since the change of $m(\mu)$ and $\lambda(\mu)$ is given by loop diagrams, it includes at least an additional factor λ. Therefore the action of the derivatives ∂_{m^2} and ∂_λ on the 1-loop contribution $\mathcal{A}^{(2)}$ leads to a term of $\mathcal{O}(\lambda^3)$ which can be neglected. Thus the only remaining term will be given by ∂_μ acting on the tree-level term $\mathcal{A}^{(1)}$. Note that this holds also at higher orders and ensures that $\partial_\mu\lambda$ at $\mathcal{O}(\lambda^{n+1})$ is determined by the parameters calculated at $\mathcal{O}(\lambda^n)$. Combining the two contributions we find

$$\mu\frac{\partial\lambda}{\partial\mu} = \frac{3\lambda^2}{16\pi^2} + \mathcal{O}(\lambda^3)\,. \tag{4.86}$$

[3] Equations of the type (4.84) that describe the change of observables as function of the scale μ are called renormalisation group equations (RGE).

Thus the scattering amplitude $\mathcal{A}$ is independent of the scale μ, if we transform the coupling constant λ into a scale dependent "running" coupling $\lambda(\mu)$ whose evolution is given by Eq. (4.86). Since we truncate the perturbation series at a finite order, the cancellation of the scale dependence is incomplete and a residual dependence of physical quantities on μ remains. Separating variables in Eq. (4.86), we find

$$\lambda(\mu) = \frac{\lambda_0}{1 - 3\lambda_0/16\pi^2 \ln(\mu/\mu_0)} \tag{4.87}$$

with $\lambda_0 \equiv \lambda(\mu_0)$ as initial condition. Thus the running coupling $\lambda(\mu)$ increases logarithmically for increasing μ in the $\lambda\phi^4$ theory.

Comparing (4.87) to our result for the scattering amplitude (4.83),

$$\mathcal{A} = -\lambda_0 \left(1 + \frac{\lambda_0}{32\pi^2}\left[\ln(s/4m^2) + \ln(t/4m^2) + \ln(u/4m^2)\right]\right) + \mathcal{O}(\lambda_0^3)\,, \tag{4.88}$$

we see that we can rewrite the amplitude using a symmetric point $q^2 = s = t = u$ and $\mu_0^2 = 4m^2$ as

$$\mathcal{A} = -\lambda_0 \left(1 + \frac{3\lambda_0}{32\pi^2} \ln(q^2/\mu_0^2)\right) = -\lambda(q^2)\,. \tag{4.89}$$

This shows that the q^2 dependence of the amplitude $\mathcal{A}$ in the limit $q^2 \gg m^2$ is determined completely by the scale dependence of the running coupling $\lambda(\mu)$. Therefore, we should set the renormalisation scale μ in general equal to the physical momentum scale q that characterises the considered process. We will come back to this topic in chapter 12, giving a formal definition of the running coupling.

4.A. Appendix: Evaluation of Feynman integrals

Combination of propagators The standard strategy in the evolution of loop integrals is the combination of the n propagator denominators into a single propagator-like denominator of higher power. One uses either Schwinger's proper-time representation

$$\frac{\mathrm{i}}{p^2 - m^2 + \mathrm{i}\varepsilon} = \int_0^\infty \mathrm{d}s\, \mathrm{e}^{\mathrm{i}s(p^2 - m^2 + \mathrm{i}\varepsilon)} \tag{4.90}$$

or the Feynman parameter integral

$$\begin{aligned}\frac{1}{x_1 \cdots x_n} &= \Gamma(n) \int_0^1 \mathrm{d}\alpha_1 \cdots \int_0^1 \mathrm{d}\alpha_n\, \delta(1 - \sum_i \alpha_i)\, [\alpha_1 x_1 \cdots \alpha_n x_n]^{-n} \\ &= \Gamma(n) \int_0^1 \mathrm{d}\alpha_1 \cdots \int_0^{\alpha_{n-2}} \mathrm{d}\alpha_n\, [\alpha_1(1 - x_1) + \alpha_2(x_1 - x_2) \cdots \alpha_n x_{n-1}]^{-n}\,.\end{aligned} \tag{4.91}$$

In order to derive this formula for $n = 2$, consider

$$\frac{1}{b-a}\int_a^b \frac{\mathrm{d}x}{x^2} = \frac{1}{b-a}\left(\frac{1}{a} - \frac{1}{b}\right) = \frac{1}{ab} \tag{4.92}$$

for $a, b \in \mathbb{C}$. Setting $x = az + b(1-z)$ and changing the integration variable, we obtain

$$\frac{1}{ab} = \int_0^1 \frac{\mathrm{d}z}{[az + b(1-z)]^2}\,. \tag{4.93}$$

The cases $n > 2$ can be derived by induction, rewriting e.g. $1/(abc)$ as $1/(aB)$ with $B = bc$, and using the result for $n-1$. In particular, for $n = 3$ it follows

$$\frac{1}{abc} = 2\int_0^1 \mathrm{d}x \int_0^{1-x} \frac{\mathrm{d}y}{[ax+by+c(1-x-y)]^3}\,. \tag{4.94}$$

Finally, we can generalise these formulae to expressions like $1/(a^n b^m)$ by taking derivatives with respect to a and b.

Evaluation of Feynman integrals We want to calculate integrals of the type

$$I_0(\omega,\alpha) = \int \frac{\mathrm{d}^{2\omega}k}{(2\pi)^{2\omega}} \frac{1}{[k^2-m^2+\mathrm{i}\varepsilon]^\alpha} \tag{4.95}$$

defined in Minkowski space. Performing a Wick rotation to Euclidean space and introducing spherical coordinates results in

$$I_0(\omega,\alpha) = \mathrm{i}(-1)^\alpha \int \frac{\mathrm{d}^{2\omega}k}{(2\pi)^{2\omega}} \frac{1}{[k^2+m^2]^\alpha} = \mathrm{i}\frac{(-1)^\alpha}{(2\pi)^{2\omega}}\Omega_{2\omega}\int_0^\infty \mathrm{d}k\,\frac{k^{2\omega-1}}{[k^2+m^2]^\alpha}\,, \tag{4.96}$$

where we denoted the volume $\mathrm{v}ol(S^{2\omega-1})$ of a unit sphere in 2ω dimensions[4] by $\Omega_{2\omega}$. You are asked in problem 4.?? to show that $\Omega_{2\omega} = 2\pi^\omega/\Gamma(\omega)$ and thus $\Omega_4 = 2\pi^2$. Substituting $k = m\sqrt{x}$ and using the integral representation (A.28) for Euler's Beta function allows us to express the k integral as a product of Gamma functions,

$$\int_0^\infty \mathrm{d}k\,\frac{k^{2\omega-1}}{[k^2+m^2]^\alpha} = \frac{1}{2}m^{2\omega-2\alpha}\int_0^\infty \mathrm{d}x\,\frac{x^{\omega-1}}{[1+x]^\alpha} = \tag{4.97a}$$

$$\frac{1}{2}m^{2\omega-2\alpha}B(\omega,\alpha-\omega) = \frac{1}{2}m^{2\omega-2\alpha}\frac{\Gamma(\omega)\Gamma(\alpha-\omega)}{\Gamma(\alpha)}\,. \tag{4.97b}$$

Combining this result with $\Omega_{2\omega} = 2\pi^\omega/\Gamma(\omega)$ we obtain

$$I_0(\omega,\alpha) = \int \frac{\mathrm{d}^{2\omega}k}{(2\pi)^{2\omega}} \frac{1}{[k^2-m^2+\mathrm{i}\varepsilon]^\alpha} = \mathrm{i}\,\frac{(-1)^\alpha}{(4\pi)^\omega}\,\frac{\Gamma(\alpha-\omega)}{\Gamma(\alpha)}\,[m^2-\mathrm{i}\varepsilon]^{\omega-\alpha}\,. \tag{4.98}$$

Note that m^2 can denote any function of the external momenta and masses, since we required only that it is independent of the loop momentum. We can generate additional formulae by adding first a dependence on a external momentum p^μ, shifting then the integration variable $k \to k+p$,

$$I(\omega,\alpha) = \int \frac{\mathrm{d}^{2\omega}k}{(2\pi)^{2\omega}} \frac{1}{[k^2+2pk-m^2+\mathrm{i}\varepsilon]^\alpha} = \mathrm{i}\,\frac{(-1)^\alpha}{(4\pi)^\omega}\,\frac{\Gamma(\alpha-\omega)}{\Gamma(\alpha)}\,[m^2+p^2-\mathrm{i}\varepsilon]^{\omega-\alpha}\,. \tag{4.99}$$

Taking then derivatives with respect to the external momentum p^μ results in

$$I_\mu(\omega,\alpha) = \int \frac{\mathrm{d}^{2\omega}k}{(2\pi)^{2\omega}} \frac{k_\mu}{[k^2+2pk-m^2+\mathrm{i}\varepsilon]^\alpha} = -p_\mu I(\omega,\alpha) \tag{4.100}$$

and

$$I_{\mu\nu}(\omega,\alpha) = \int \frac{\mathrm{d}^{2\omega}k}{(2\pi)^{2\omega}} \frac{k_\mu k_\nu}{[k^2+2pk-m^2+\mathrm{i}\varepsilon]^\alpha} = \tag{4.101}$$

$$= \mathrm{i}\,\frac{(-\pi)^\omega}{(2\pi)^{2\omega}}\,\frac{\Gamma(\alpha-\omega-1)}{\Gamma(\alpha)}\,\frac{p_\mu p_\nu(\alpha-\omega-1)-\frac{1}{2}\eta_{\mu\nu}(m^2+p^2)}{[m^2+p^2-\mathrm{i}\varepsilon]^{\alpha-\omega}}\,. \tag{4.102}$$

[4] A $n-1$-dimensional sphere $S^{n-1}(R)$ encloses the n-dimensional volume $x_1^2+\ldots+x_n^2 \le R^2$, while its own $n-1$-dimensional volume is given by $x_1^2+\ldots+x_n^2 = R^2$. The volume of a unit 1-sphere is a length, $\mathrm{v}ol(S^1) = 2\pi$, of a unit 2-sphere an area, $\mathrm{v}ol(S^2) = 4\pi$, and of a unit 3-sphere a volume, $\mathrm{v}ol(S^3) = 2\pi^2$. If we say that the volume of a sphere is $4\pi R^3/3$, we mean in fact the volume of the 3-ball $B^3(R)$, $x_1^2+x_2^2+x_3^2 \le R^2$ which is enclosed by the 2-sphere $S^2(R)$.

Contracting both sides with $k_\mu k_\nu$ and using $\eta^{\mu\nu}\eta_{\mu\nu} = 2\omega$ gives

$$I_2(\omega,\alpha) = \int \frac{\mathrm{d}^{2\omega}k}{(2\pi)^{2\omega}} \frac{k^2}{[k^2+2pk-m^2+\mathrm{i}\varepsilon]^\alpha} = \mathrm{i}\,\frac{(-\pi)^\omega}{(2\pi)^{2\omega}} \frac{\Gamma(\alpha-\omega-1)}{\Gamma(\alpha)} \frac{(\alpha-2\omega-1)p^2-\omega m^2}{[m^2+p^2-\mathrm{i}\varepsilon]^{\alpha-\omega}}\,. \tag{4.103}$$

Special cases often needed are

$$I(\omega,2) = \frac{\mathrm{i}}{(4\pi)^\omega}\,\Gamma(2-\omega)\,(m^2+p^2-\mathrm{i}\varepsilon)^{\omega-2}\,, \tag{4.104}$$

$$I_2(\omega,2) = -\frac{\mathrm{i}}{(4\pi)^\omega}\,\omega\Gamma(1-\omega)\,(m^2+p^2-\mathrm{i}\varepsilon)^{\omega-1}\,, \tag{4.105}$$

and

$$I(2,3) = -\frac{\mathrm{i}}{32\pi^2}\,\frac{1}{m^2+p^2-\mathrm{i}\varepsilon}\,. \tag{4.106}$$

Summary

Disconnected n-point Green functions are generated by the functional $Z[J]$, while $\mathrm{i}W[J] = \ln Z[J]$ generates connected Green functions. The three loop diagrams we calculated in the $\lambda\phi^4$ theory were infinite and had to be regularised. Renormalising the three parameters contained in the classical Lagrangian of the $\lambda\phi^4$ theory, ρ_0, m^2, and λ, eliminated the divergences and converted them into "running" quantities.

Further reading

The derivation of the Feynman rules using the graphical approach is discussed extensively in [GR08]. [Sre07] treats the $\lambda\phi^3$ theory which resembles more QED than the $\lambda\phi^4$ theory we discussed.

5. Global symmetries and Noether's theorem

Emmy Noether showed 1917 that any global continuous symmetry of a classical system described by a Lagrangian leads to a locally conserved current. We can divide such symmetries into two classes: Symmetries of space-time and internal symmetries of a group of fields. Prominent examples for the latter are the global symmetries that lead to the conservation laws of electric charge or baryon number, respectively. For a quantum system, we have to study the impact of symmetries on its generating functional. If this functional remains invariant, the symmetry holds also at the quantum level. Then conserved charges exists which commute with the Hamiltonian. Analysing space-time symmetries, we restrict ourselves in this chapter to the case when we can neglect gravity. Then space-time is the familiar Minkowski space characterised by the Poincaré symmetry group, i.e. the product of the translation and the Lorentz group, with its corresponding conservation laws.

5.1. Internal symmetries

We have used up-to now mainly space-time symmetry of Minkowski space, namely the requirement of Lorentz invariance, to deduce possible terms in the action. If we allow for more than one field, e.g. several scalar fields, the new possibility of *internal* symmetries arise. For instance, we can look at a theory of two massive scalar fields with quartic interactions,

$$\mathscr{L} = \frac{1}{2}\left(\partial_\mu \phi_1\right)^2 - \frac{1}{2} m_1^2 \phi_1^2 - \frac{1}{4}\lambda_1 \phi_1^4 + \frac{1}{2}(\partial_\mu \phi_2)^2 - \frac{1}{2} m_2^2 \phi_2^2 - \frac{1}{4}\lambda_2 \phi_2^4 - \frac{1}{2}\lambda_3 \phi_1^2 \phi_2^2 . \tag{5.1}$$

In order to maintain the discrete Z_2 symmetry $\phi_i \to -\phi_i$ of the individual Lagrangians for the fields ϕ_1 and ϕ_2, we have omitted odd terms like $\phi_1 \phi_2^3$. Then the theory contains five arbitrary parameters, two masses m_i and three coupling constants λ_i. For arbitrary values of these parameters, no new additional symmetry results. In nature, we find however often a set of particles with nearly the same mass and (partly) similar couplings. One of the first examples was suggested by Heisenberg after the discovery of the neutron, which has a mass very close to the one of the proton, $m_n \simeq m_p$: With respect to strong interactions, it is useful to view the proton and neutron as two different "isospin" states of the nucleon, similar as an electron has two spin states. An example of an exact symmetry are particles and their antiparticles, as e.g. the charged pions $\pi^\pm$ which can be combined into one complex scalar field.

If we set in our case $m_1 = m_2$ and $\lambda_2 = \lambda_1$, the Lagrangian becomes invariant under the exchange $\phi_1 \leftrightarrow \phi_2$. Adding the further condition that $\lambda_3 = \lambda_1 = \lambda_2$, we arrive at

$$\mathscr{L} = \frac{1}{2}\left[(\partial_\mu \phi_1)^2 + (\partial_\mu \phi_2)^2\right] - \frac{1}{2} m^2 (\phi_1^2 + \phi_2^2) - \frac{\lambda}{4}(\phi_1^2 + \phi_2^2)^2 . \tag{5.2}$$

Now any orthogonal transformation $O \in \mathrm{O}(2)$ in the two-dimensional field space $\{\phi_1, \phi_2\}$ leads to the same Lagrangian $\mathscr{L}$. In particular, the Lagrangian is invariant under a rotation

$R(\alpha) \in SO(2)$ which mixes $\{\phi_1, \phi_2\}$ as

$$\begin{pmatrix} \phi_1' \\ \phi_2' \end{pmatrix} = \begin{pmatrix} \cos\alpha & \sin\alpha \\ -\sin\alpha & \cos\alpha \end{pmatrix} \begin{pmatrix} \phi_1 \\ \phi_2 \end{pmatrix} . \tag{5.3}$$

The fields transform as a vector $\boldsymbol{\phi} = \{\phi_1, \phi_2\}$, and a rotation leaves the length of this vector invariant. Generalising this to n scalar fields, $\boldsymbol{\phi} = \{\phi_1, \dots, \phi_n\}$, we can write down immediately a theory that is invariant under transformations $\phi_a \to R_{ab}\phi_b$ where R is an element of $\mathrm{O}(n)$,

$$\mathscr{L} = \frac{1}{2}(\partial_\mu \boldsymbol{\phi})^2 - \frac{1}{2}m^2\boldsymbol{\phi}^2 - \frac{\lambda}{4}(\boldsymbol{\phi}^2)^2 . \tag{5.4}$$

Note that the Lagrangian is only invariant under global, i.e. space-time independent rotations, since the term $\partial_\mu[R(x)\boldsymbol{\phi}]$ would breaking the invariance for $R = R(x)$.

The free Lagrangian $\mathscr{L}_0$, i.e. the part quadratic in the fields, is diagonal, $\mathscr{L}_0 = \mathscr{L}_0(\phi_1) + \ldots + \mathscr{L}_0(\phi_n)$. Thus the propagator $D_{ab}(x - x')$ is diagonal too, $D_{ab}(x - x') \propto \delta_{ab}$. An interaction vertex at x connects four propagators $D_{ab}(x - x_i)$. As a result of the Z_2 symmetry, an even number of ϕ_1 and ϕ_2 particles are connected at each vertex which has therefore the form $-\mathrm{i}\lambda(\delta_{ab}\delta_{cd} + \delta_{ac}\delta_{bd} + \delta_{ad}\delta_{bc})$.

5.2. Noether's theorem

From our experience in classical and quantum mechanics, we expect that global continuous symmetries lead also in field theory to conservation laws for the generators of the symmetry. In order to derive such a conservation law, we consider an infinitesimal change $\delta\phi_a$ of the fields that keeps by assumption $\mathscr{L}(\phi_a, \partial_\mu\phi_a)$ invariant,

$$0 = \delta\mathscr{L} = \frac{\delta\mathscr{L}}{\delta\phi_a}\,\delta\phi_a + \frac{\delta\mathscr{L}}{\delta\partial_\mu\phi_a}\,\delta\partial_\mu\phi_a . \tag{5.5}$$

Now we exchange $\delta\partial_\mu = \partial_\mu\delta$ in the second term and use then the Lagrange equations, $\delta\mathscr{L}/\delta\phi_a = \partial_\mu(\delta\mathscr{L}/\delta\partial_\mu\phi_a)$, in the first one. Then we can combine the two terms using the product rule,

$$0 = \delta\mathscr{L} = \partial_\mu\left(\frac{\delta\mathscr{L}}{\delta\partial_\mu\phi_a}\right)\delta\phi_a + \frac{\delta\mathscr{L}}{\delta\partial_\mu\phi_a}\,\partial_\mu\delta\phi_a = \partial_\mu\left(\frac{\delta\mathscr{L}}{\delta\partial_\mu\phi_a}\,\delta\phi_a\right) . \tag{5.6}$$

Hence the invariance of $\mathscr{L}$ under the change $\delta\phi_a$ implies the existence of a conserved current, $\partial_\mu j^\mu = 0$, with

$$j^\mu = \frac{\delta\mathscr{L}}{\delta\partial_\mu\phi_a}\,\delta\phi_a . \tag{5.7}$$

If the transformation $\delta\phi_a$ leads to change in $\mathscr{L}$ that is a total four-divergence, $\delta\mathscr{L} = \partial_\mu K^\mu$, and boundary terms can be dropped, then the equations of motion remain invariant too. The conserved current j^μ, also called Noether current, is then changed to

$$j^\mu = \frac{\delta\mathscr{L}}{\delta\partial_\mu\phi_a}\,\delta\phi_a - K^\mu . \tag{5.8}$$

In Minkowski space, we can convert this differential form of a conservation law into a global one using Gauss' theorem: Recall that this theorem allows us to convert a n dimensional volume integral over the divergence of a tensor field into a $n-1$ dimensional surface integral of the tensor field,

$$\int_\Omega \mathrm{d}^4x\, \partial_\mu X^{\mu\nu\cdots} = \int_{\partial\Omega} \mathrm{d}S_\mu X^{\mu\nu\cdots} . \tag{5.9}$$

Applied to the Noether current j^μ, we obtain assuming $|\boldsymbol{j}| \to 0$ for $|\boldsymbol{x}| \to \infty$

$$\int_\Omega \mathrm{d}^4x\, \partial_\mu j^\mu = \int_{V(t_2)} \mathrm{d}^3x\, j^0 - \int_{V(t_1)} \mathrm{d}^3x\, j^0 = 0 . \tag{5.10}$$

Thus the volume integral $Q = \int_V \mathrm{d}^3x\, j^0$ over the charge density j^0 remains constant. Often (but not always) this charge has a profound physical meaning. Finally, we note that the conserved current j^μ is not unique, since we can add the four-divergence $\partial_\mu K^\mu = \delta\mathscr{L}$.

Internal symmetries As an example, we can use the n scalar fields invariant under the group[1] SO(n). We need the infinitesimal generators T_i of rotations,

$$\phi'_a = R_{ab}\phi_b = (1 + \alpha_i T_i + \mathcal{O}(\alpha_i^2))_{ab}\phi_b . \tag{5.11}$$

SO(n) has an antisymmetric Lie algebra with $n(n-1)/2$ generators. Thus a theory invariant under SO(n) has $n(n-1)/2$ conserved currents. The special case $n=2$ has as important application e.g. the charged pions $\pi^\pm$. We combine the two real fields ϕ_1 and ϕ_2 into the complex field $\phi = (\phi_1 + \mathrm{i}\phi_2)/\sqrt{2}$, then the Lagrangian becomes

$$\mathscr{L} = \partial_\mu\phi^\dagger\partial^\mu\phi - m^2\phi^\dagger\phi - \lambda(\phi^\dagger\phi)^2 . \tag{5.12}$$

Now the Lagrangian $\mathscr{L}$ is invariant under the combined phase transformations $\phi \to \mathrm{e}^{-\mathrm{i}\vartheta}\phi$ and $\phi^\dagger \to \mathrm{e}^{\mathrm{i}\vartheta}\phi^\dagger$: Using complex fields, the SO(2) symmetry has become an U(1) symmetry. With $\delta\phi = -\mathrm{i}\phi$, $\delta\phi^\dagger = \mathrm{i}\phi^\dagger$, the conserved current follows as

$$j^\mu = \mathrm{i}\left[\phi^\dagger\partial^\mu\phi - (\partial^\mu\phi^\dagger)\phi\right] . \tag{5.13}$$

The conserved charge $Q = \int \mathrm{d}^3x\, j^0$ can be also negative and thus we cannot interpret j^0 as the probability density to observe a ϕ particle. Instead, we should associate Q with a conserved additive quantum number as e.g. the electric charge.

Note that Noethers theorem requires only the existence of a *global* symmetry. In the cases of the conservation of electric and colour charge, the global symmetry is a consequence of an underlying *local* gauge symmetry which we will study later in chapter 10 in detail. In most other cases however, as e.g. the conservation of baryon or lepton number, the global symmetry can not be generalised to a local one, and one speaks therefore of accidental symmetries. Such symmetries are not protected against quantum corrections and there is no reason to expect them to hold exactly. We will see later that baryon and lepton number are indeed broken.

[1] Although the Lagrangian is invariant under the larger group O(n), we consider only the subgroup SO(n) which is continuously connected to the identity. The additional discrete transformations contained in O(n) can be used to classify solutions of the Lagrangian, but do not lead to additional conservation laws.

Space-time symmetries of Minkowski space The Poincaré group as symmetry group of Minkowski space has ten generators[2]. If the Lagrangian does not depend explicitly on space-time coordinates, i.e. $\mathscr{L} = \mathscr{L}(\phi_a, \partial_\mu \phi_a)$, ten conservation laws for the fields ϕ_a follow. We consider first the behaviour of the fields ϕ_a and the Lagrangian under an infinitesimal translation $x^\mu \to x^\mu + \varepsilon\xi^\mu$. As in the case of internal symmetries, we consider only global transformations and thus ε does not depend on x. From

$$\phi_a(x^\mu) \to \phi_a(x^\mu + \varepsilon\xi^\mu) \approx \phi_a(x^\mu) + \varepsilon\xi^\nu \partial_\nu \phi_a(x^\mu) \tag{5.14}$$

we find the change

$$\delta\phi_a(x) = \xi^\mu \partial_\mu \phi_a(x) = \partial_\mu [\xi^\mu \phi_a(x)] \,.$$

Since the Lagrange density $\mathscr{L}$ contains by assumption no explicit space-time dependence, it will change simply as $\mathscr{L}(x^\mu) \to \mathscr{L}(x^\mu + \varepsilon\xi^\mu)$ or

$$\delta\mathscr{L}(x) = \xi^\mu \partial_\mu \mathscr{L}(x) = \partial_\mu [\xi^\mu \mathscr{L}(x)] \,. \tag{5.15}$$

Thus $K^\mu = \xi^\mu \mathscr{L}(x)$ and inserting both in the Noether current gives

$$j^\mu = \frac{\partial\mathscr{L}}{\partial(\partial_\mu\phi_a)} [\xi^\nu \partial_\nu \phi_a] - \xi^\mu \mathscr{L} = \xi_\nu \left[\frac{\partial\mathscr{L}}{\partial(\partial_\mu\phi_a)} \frac{\partial\phi_a}{\partial x_\nu} - \eta^{\mu\nu} \mathscr{L} \right] \equiv \xi_\nu T^{\mu\nu} \,, \tag{5.16}$$

where the square bracket defines the (energy-momentum) stress tensor $T^{\mu\nu}$ of the fields ϕ_a. The corresponding four conserved Noether charges are the components of the four-momentum

$$p^\nu = \int \mathrm{d}^3 x \, T^{0\nu} \,. \tag{5.17}$$

The conserved tensor defined by Eq. (5.16) is called the *canonical* stress tensor. The definition (5.16) does not guarantee that $T^{\mu\nu}$ is symmetric. A symmetric stress tensor, $T^{\mu\nu} = T^{\nu\mu}$, is however the condition for the conservation of the total angular momentum, as we will show in the next paragraph. Another reason to require a symmetric stress tensor $T^{\mu\nu}$ is that it serves as source term for the symmetric gravitational field. Since the Lagrange density is only determined up to a four-divergence, we can symmetrize always $T^{\mu\nu}$ adding an appropriate divergence of an antisymmetric tensor. Thus in general, the canonical stress tensor has to be symmetrised by hand.

Example 5.1: The general expression (5.16) for the canonical stress tensor becomes for a free complex scalar field

$$T^{\mu\nu} = 2\partial^\mu \phi^\dagger \partial^\nu \phi - \eta^{\mu\nu} \mathscr{L} \,. \tag{5.18}$$

Thus the canonical stress tensor of a scalar field is already symmetric. Its 00 component,

$$T^{00} = \rho = \mathscr{H} = 2|\dot{\phi}|^2 - \mathscr{L} = |\dot{\phi}|^2 + |\boldsymbol{\nabla}\phi|^2 + m^2|\phi|^2 \,, \tag{5.19}$$

agrees with twice the result (3.14) for the energy-density ρ of a single real scalar field. We consider now plane-wave solutions to the Klein-Gordon equation, $\phi = N \exp(\mathrm{i}kx)$. If we insert $\partial_\mu \phi = \mathrm{i}k_\mu \phi$ into $\mathscr{L}$, we find $\mathscr{L} = 0$ and thus

$$T^{00} = 2N^2 k^0 k^0 \,. \tag{5.20}$$

Changing from the continuum normalisation to a box of size $V = L^3$ amounts to replace $(2\pi)^3$ by L^3. Thus the normalisation constant $N^{-2} = (2\pi)^3 2\omega$ becomes for a finite volume $N^{-2} = 2\omega V$. Thence

[2] If this sound unfamiliar, read first the appendices B.3 and B.4 before continuing

the energy-density $T^{00} = \omega/V$ agrees with the expectation for one particle with energy ω per volume V. The remaining components of $T^{\mu\nu}$ are fixed by its tensor structure,

$$T^{\mu\nu} = 2N^2 k^\mu k^\nu = \frac{k^\mu k^\nu}{\omega V}\,. \tag{5.21}$$

Since the stress tensor $T^{\mu\nu}$ is symmetric, we can find a frame in which $T^{\mu\nu}$ is diagonal with $T \propto$ d$iag(\omega, v_x k_x, v_y k_y, v_z k_z)/V$. The spatial part of the stress tensor agrees with the pressure tensor of an ideal fluid, cf. problem 5.??. Thus a scalar field can be viewed as an ideal fluid with energy density ρ and pressure P, or $T^{\mu\nu} =$ d$iag(\rho, P_x, P_y, P_z)$.

Angular momentum If the tensor $T^{\mu\nu}$ is symmetric, we can construct six more conserved quantities. If we define

$$M^{\mu\nu\lambda} = x^\nu\, T^{\mu\lambda} - x^\lambda\, T^{\mu\nu}\,, \tag{5.22}$$

then $M^{\mu\nu\lambda}$ is conserved with respect to the index μ,

$$\partial_\mu M^{\mu\nu\lambda} = \delta^\nu_\mu\, T^{\mu\lambda} - \delta^\lambda_\mu T^{\mu\nu} = T^{\nu\lambda} - T^{\lambda\nu} = 0\,, \tag{5.23}$$

provided that $T^{\nu\lambda} = T^{\lambda\nu}$. In this case,

$$J^{\mu\nu} = \int \mathrm{d}^3x\; M^{0\mu\nu} = \int \mathrm{d}^3x\; \left[x^\mu T^{0\nu} - x^\nu T^{0\mu}\right]\,, \tag{5.24}$$

is a globally conserved tensor. The antisymmetry of $J^{\mu\nu}$ implies that there exist six conserved charges. The three charges

$$J^{ij} = \int \mathrm{d}^3x\; \left[x^i T^{0j} - x^j T^{0i}\right] \tag{5.25}$$

correspond to the conservation of total angular momentum, since T^{0j} is the momentum density. The remaining three charges J^{0i} express the fact that the center-of-mass moves with constant velocity.

While $J^{\mu\nu}$ transforms as expected for a tensor under Lorentz transformations, it is not invariant under translations $x^\mu \to x^\mu + \xi^\mu$. Instead, the angular momentum changes as

$$J^{\mu\nu} \to J^{\mu\nu} + \xi^\mu p^\nu - \xi^\nu p^\mu\,. \tag{5.26}$$

Clearly, this is a consequence of the definition of the orbital angular momentum with respect to the center of rotation. We want therefore to split the total angular momentum $J^{\mu\nu}$ into the orbital angular momentum $L^{\mu\nu}$ and an intrinsic part connected to a non-zero spin of the field. The latter we require to be invariant under translations. We set

$$S_\alpha = \frac{1}{2}\varepsilon_{\alpha\beta\gamma\delta} J^{\beta\gamma} u^\delta\,, \tag{5.27}$$

where u^α is the four-velocity of the center-of-mass system (cms). Because of the antisymmetry in $\beta\gamma$ of $\varepsilon_{\alpha\beta\gamma\delta}$, the change in (5.26) induced by a translation drops out in S_α. In the cms, $u^\alpha = (1,0,0,0)$ and thus $S_0 = 0$ and $S_\alpha u^\alpha = 0$. The other components are $S_1 = J^{23}$, $S_2 = J^{31}$, and $S_3 = J^{12}$. Thus the vector S_α describes as desired the intrinsic angular momentum of a field. It is called the Pauli-Lubanski spin vector.

Remark 5.1: To see that the contraction of a symmetric tensor $S_{\mu\nu}$ with an antisymmetric tensor $A_{\mu\nu}$ gives zero, consider

$$S_{\mu\nu}A^{\mu\nu} = -S_{\mu\nu}A^{\nu\mu} = -S_{\nu\mu}A^{\nu\mu} = -S_{\mu\nu}A^{\mu\nu} \,. \tag{5.28}$$

Here we used first the antisymmetry of $A^{\mu\nu}$, then the symmetry of $S_{\mu\nu}$, and finally exchanged the dummy summation indices. Clearly, this remains true if the tensor expression contains additional indices. Applied to the Pauli-Lubanski spin vector and recalling $p^\gamma = mu^\gamma$, its change contains terms as $\varepsilon_{\alpha\beta\gamma\delta}\xi^\beta u^\gamma u^\delta$ which are zero.

5.3. Quantum symmetries

Conserved currents We have seen that Noether's theorem guaranties on the classical level the conservation of currents generated by global continuous symmetries. In the corresponding quantum theory, we have to study the impact of this symmetry on the generating functional Z. Since we used the equations of motion to derive Noether's theorem, current conservation holds only for classically allowed paths in field space, or in other words for on-shell fields. Thus the action evaluated for off-shell fields is not invariant under global symmetry transformations. In the path integral, the fields are however only integration variables. The generating functional is therefore invariant, if we can find a field transformation $\phi_i \to \tilde{\phi}_i$ which eliminates the change of the action for off-shell fields and keeps the integration measure invariant, $\mathcal{D}\phi_i = \mathcal{D}\tilde{\phi}_i$.

Let us assume that our theory has a global symmetry under which the classical solutions transform as $\phi_a \to \phi'_a = \phi_a + \varepsilon\eta_a$. Here, ε takes the same value at all space-time points and η_a is a function of the original fields, $\eta_a = \eta_a(\phi_a(x))$. The classically forbidden solutions will be transformed into $\tilde{\phi}_a \neq \phi'_a$. We can express $\tilde{\phi}_a$ always as

$$\phi_a \to \tilde{\phi}_a = \phi_a + \varepsilon(x)\eta_a \,, \tag{5.29}$$

promoting thereby $\varepsilon(x)$ to a space-time dependent function. To be concrete, we consider again a global U(1) symmetry for a complex scalar field. Since the transformation (5.29) is *local*, kinetic terms breaks the symmetry: Direct calculation shows that the Lagrangian (5.12) changes as

$$\delta\mathscr{L} = \mathrm{i}\,[\phi^*\partial^\mu\phi - \partial^\mu\phi^*\phi]\,\partial_\mu\varepsilon = j^\mu\,\partial_\mu\varepsilon \,, \tag{5.30}$$

where j^μ is the classical Noether current, cf. problem 5.??. The final result $\delta\mathscr{L} = j^\mu\,\partial_\mu\varepsilon$ holds in general.

Next we have to generalise the generating functional for a single, real scalar field given in Eq. (4.1) to a complex scalar field. We treat ϕ and ϕ^* as the two independent degrees of freedom, and add therefore also two independent sources J and J^*. Coupling them as $\mathscr{L}_\mathrm{s} \equiv J\phi^* + J^*\phi$ to the fields keeps the Lagrangian real. We denote the total Lagrangian as $\mathscr{L}_{eff} = \mathscr{L}_{cl} + \mathscr{L}_\mathrm{s}$, with $\mathscr{L}_{cl} = \mathscr{L}_0 + \mathscr{L}_{int}$ as the Lagrangian used to derive to the classical equation of motions. Thus the generating functional is

$$Z[J, J^*] = \int \mathcal{D}\phi\mathcal{D}\phi^* \exp \mathrm{i} \int \mathrm{d}^4x \; (\mathscr{L}_{cl} + J\phi^* + J^*\phi) \; . \tag{5.31}$$

We want to calculate matrix elements of the operator representing the classical Noether current (5.8). Using the path integral formalism, we can derive the time-ordered vacuum expectation value of a product of fields ϕ and the current operator j^μ by adding a classical external source v_μ coupled to j^μ,

$$Z[J, J^*, v_\mu] = \int \mathcal{D}\phi\mathcal{D}\phi^* \exp \mathrm{i} \int \mathrm{d}^4x \; (\mathscr{L}_{cl} + J\phi^* + J^*\phi + v_\mu j^\mu) \; . \tag{5.32}$$

Then we obtain the vacuum expectation value of the current as

$$\langle j^\mu(x)\rangle \equiv \langle 0|j^\mu(x)|0\rangle = \frac{1}{\mathrm{i}}\frac{\delta}{\delta v_\mu(x)}\mathrm{i}W[J, J^*, v_\mu]\bigg|_{v_\mu=J=0} \; . \tag{5.33}$$

Inverting this relation we find

$$\delta W[J, J^*, v_\mu] = \int \mathrm{d}^4x \,\langle j^\mu(x)\rangle \delta v_\mu(x) \, . \tag{5.34}$$

We are interested how W and Z change under a transformation of the external source v_μ. To deduce their transformation properties, it is sufficient to consider them for zero external sources J and J^*. Setting $Z[0, 0, v_\mu] \equiv Z[v_\mu]$ and choosing $\delta v_\mu(x) = -\partial_\mu\varepsilon(x)$, it follows

$$\delta W[v_\mu] = W[v_\mu - \partial_\mu\varepsilon(x)] - W[v_\mu] = -\int \mathrm{d}^4x \,\langle j^\mu(x)\rangle\partial_\mu\varepsilon(x) = \int \mathrm{d}^4x \,\partial_\mu\langle j^\mu(x)\rangle\varepsilon(x) \, . \tag{5.35}$$

Thus $\delta W[v_\mu] = 0$ guarantees current conservation in the quantum theory, $\partial_\mu\langle j^\mu\rangle = 0$. The corresponding change of $Z[v_\mu]$ under the same transformation is

$$Z[v_\mu - \partial_\mu\varepsilon(x)] = \int \mathcal{D}\phi\mathcal{D}\phi^* \exp \mathrm{i} \int \mathrm{d}^4x \; \{\mathscr{L} + [v_\mu - \partial_\mu\varepsilon(x)]j^\mu\} \; . \tag{5.36}$$

We now assume that the substitution $\phi_a \to \tilde\phi_a = \phi_a + \varepsilon(x)\eta_a$ keeps the integration measure invariant, $\mathcal{D}\phi\mathcal{D}\phi^* = \mathcal{D}\tilde\phi\mathcal{D}\tilde\phi^*$. Recalling then that $\delta\mathscr{L} = j^\mu\,\partial_\mu\varepsilon$, we find that the generating functional is invariant,

$$Z[v_\mu - \partial_\mu\varepsilon(x)] = \int \mathcal{D}\tilde\phi\mathcal{D}\tilde\phi^* \exp \mathrm{i} \int \mathrm{d}^4x \; \left(\mathscr{L}(\tilde\phi, \partial_\mu\tilde\phi) + v_\mu j^\mu\right) = Z[v_\mu] \, . \tag{5.37}$$

In the case of the U(1) transformation, the two phases cancel in the integration measure

$$\mathcal{D}\tilde\phi\mathcal{D}\tilde\phi^* = \prod_x \mathrm{d}\tilde\phi(x)\mathrm{d}\tilde\phi^*(x) = \prod_x \mathrm{d}\phi(x)\mathrm{d}\phi^*(x) \, . \tag{5.38}$$

As a result, the vacuum expectation value of the electromagnetic current is conserved, $\partial_\mu\langle j^\mu\rangle = 0$.

Anomalies The substitution $\phi_a \to \tilde\phi_a = \phi_a + \varepsilon(x)\eta_a$ shifts the center of the integration at each space-time point by the value $\varepsilon(x)\eta_a$. Such a linear shift seems harmless. Therefore it was taken for granted that the path integral remains invariant under this change and, consequently, that this approach predicts that all classical global symmetries hold also on the quantum level. It was only realised by Fujikawa in 1979 that the integration measure in the path integral may transform non-trivially under a symmetry transformation: Since the path integral is divergent, we have to regularise it and this procedure may break the classical symmetry.

If the classical symmetry is broken, one speaks of an “anomaly”. The three most important examples are the trace anomaly, the chiral anomaly, and the breaking of conformal invariance in string theory. We will discuss the first two cases later in some detail. The anomalous term breaking conformal invariance in string theory vanishes for a definite number of space-time dimensions, $D = 10$ or 26, what is the reason for the predictions of extra-dimensions in string theory.

Summary

Noether’s theorem shows that continuous global symmetries lead classically to conservation laws. Such symmetries can be divided into space-time and internal symmetries. Minkowski space-time is invariant under global Poincaré transformations. The corresponding ten Noether charges are the four-momentum p^μ and the total angular momentum $J^{\mu\nu}$. Examples for conserved charges due to internal symmetries are electric and colour charge, as well as baryon or lepton number. In a quantum theory, the vacuum expectation value of a Noether current is conserved, if the symmetry transformation keeps the path integral measure invariant.

Further reading

A more complete discussion of Noether’s theorem can be found in [GR08] and [Hil51].

6. Space-time symmetries

In the previous chapter, we discussed the symmetries of Minkowski space. In this case, we could view the Poincaré group as the group generating global symmetry transformations on Minkowski space and find the resulting conservation laws. Aim of the present chapter is to extend this discussion to the case of a Riemannian manifold, i.e. to a curved space which looks only locally Euclidean. We will show how one can find the symmetries of such manifolds and how they determine conservation laws. Riemannian manifolds arise naturally in classical mechanics using generalised coordinates q^i, since the kinetic energy $T = a_{ik}\dot{q}^i\dot{q}^k$ defines a quadratic form a_{ik} which we can view as metric tensor on the configuration space $\{q^i\}$. However, the for us more important appearance of a (pseudo-) Riemannian manifold is in Einstein's theory of general relativity which replaces Minkowski space by a curved space-time. Most of the mathematical structures we will introduce have also a close analogue in gauge theories which we will use later on to describe the electroweak and strong interactions.

Equivalence principle As a start, we motivate why one can replace the gravitational force by the curvature of space-time discussing the equivalence principle. The idea underlying this principle emerged in the 16th century, when among others Galileo Galilei found experimentally that the acceleration g of a test mass in a gravitational field is universal. Because of this universality, the gravitating mass m_g and the inertial mass m_i are identical in classical mechanics. While $m_i = m_g$ can be achieved for one material always by a convenient choice of units, there should be in general deviations for test bodies with differing compositions. Current limits for departures from universal gravitational attraction for different materials are however very tight, $|\Delta g/g| < 10^{-12}$.

As a result, gravity has compared to the three other known fundamental interactions the unique property that it can be switched-off locally: Inside a freely falling elevator, one does not feel any gravitational effects except tidal forces. The latter arise if the gravitational field is non-uniform and tries to deform the elevator. Inside a sufficiently small freely falling system, also tidal effects plays no role. Einstein promoted the equivalence of inertial and gravitating mass to the postulate of the "strong equivalence principle": In a small enough region around the center of a freely falling coordinate system all physics is described by the laws of special relativity.

In general relativity, the gravitational force of Newton's theory that accelerates particles in an Euclidean space is replaced by a curved space-time in which particles move force-free along geodesic lines. In particular, photons move still as in special relativity along curves satisfying $\mathrm{d}s^2 = 0$, while all effects of gravity are now encoded in the non-Euclidean geometry of space-time which is determined by the line-element $\mathrm{d}s$ or the metric tensor $g_{\mu\nu}$,

$$\mathrm{d}s^2 = g_{\mu\nu}\mathrm{d}x^\mu \mathrm{d}x^\nu \,. \tag{6.1}$$

Switching on a gravitational field, the metric tensor $g_{\mu\nu}$ can be transformed only locally by a coordinate change into the form $\eta_{\mu\nu} = \mathrm{d}iag(1, -1, -1, -1)$. Thus we should develop the tools necessary to do analysis on a curved manifold $\mathscr{M}$ which geometry is described by the metric tensor $g_{\mu\nu}$.

6.1. Manifolds and tensor fields

Manifolds A manifold $\mathscr{M}$ is any set that can be continuously parametrised. The number of independent parameters needed to specify uniquely any point of $\mathscr{M}$ is its dimension n, the parameters $x = \{x^1, \dots, x^n\}$ are called coordinates. Locally, a manifold with dimension n can be approximated by $\mathbb{R}^n$. Examples for manifolds are Lie groups, the configuration space q^i or the phase space (q^i, p_i) of classical mechanics, and space-time in general relativity. We require the manifold to be smooth: The transitions from one set of coordinates to another one, $x^i = f(\tilde{x}^i, \dots, \tilde{x}^n)$, should be C^∞. In general, it is impossible to cover all $\mathscr{M}$ with one coordinate system that is well-defined on all $\mathscr{M}$. An example are spherical coordinate (ϑ, ϕ) on a sphere S^2, where ϕ is ill-defined at the poles. Instead one has to cover the manifold with patches of different coordinates that partially overlap.

Vector fields A vector field $\boldsymbol{V}(x^a)$ on (a subset $\mathscr{S}$ of) $\mathcal{M}$ is a set of vectors associating to each space-time point $x^a \in \mathscr{S}$ exactly one vector. The paradigm for such a vector field is the four-velocity $\boldsymbol{u}(\tau) = \mathrm{d}\boldsymbol{x}/\mathrm{d}\tau$ which is the tangent vector to the world-line $x(\tau)$ of a particle. Since the differential equation $\mathrm{d}\boldsymbol{x}/\mathrm{d}\sigma = \boldsymbol{X}(\sigma)$ has locally always a solution, we can find for any given $\boldsymbol{X}$ a curve $x(\sigma)$ which has $\boldsymbol{X}$ as tangent vector. Although the definition $\boldsymbol{u}(\tau) = \mathrm{d}\boldsymbol{x}/\mathrm{d}\tau$ coincides with the one familiar from Minkowski space, there an important difference: In a general manifold, we can not imagine a vector $\boldsymbol{V}$ as an "arrow" $\overrightarrow{PP'}$ pointing from a certain point P to another point P' of the manifold. Instead, the vectors $\boldsymbol{V}$ generated by all smooth curves through P span a n-dimensional vector space at the point P called tangent space T_P. We can visualise the tangent space for the case of a two-dimensional manifold embedded in $\mathbb{R}^3$: At any point P, the tangent vectors lie in a plane $\mathbb{R}^2$ which we can associate with T_P. In general, $T_P \neq T_{P'}$ and we cannot simply move a vector $\boldsymbol{V}(x^\mu)$ to another point $\tilde{x}^\mu$. This implies in particular that we cannot add the vectors $\boldsymbol{V}(x^\mu)$ and $\boldsymbol{V}(\tilde{x}^\mu)$, if the points x^μ and $\tilde{x}^\mu$ differ. Therefore we cannot differentiate a vector field without introducing an additional mathematical structure which allows us to transport a vector from one tangent space to another.

If we want to decompose the vector $\boldsymbol{V}(x^\mu)$ into components $V^\nu(x^\mu)$, we have to introduce a basis $\boldsymbol{e}_\mu$ in the tangent space. There are two natural choices for such a basis: First, we could use Cartesian basis vectors as in a Cartesian inertial system in Minkowski space. We will follow this approach later, when we discuss gravity as a gauge theory in chapter 18. Now, we will use the more conventional approach and use as basis vectors the tangential vectors along the coordinate lines x^μ in $\mathscr{M}$,

$$\boldsymbol{e}_\mu = \frac{\partial}{\partial x^\mu} \equiv \partial_\mu \,. \tag{6.2}$$

Here the index μ with value i in $\boldsymbol{e}_\mu$ denotes the i.th basis vector $\boldsymbol{e}_\mu = (0, \dots, 1, \dots 0)$, with an one at the i.th position, not a component. Using this basis, a vector can be decomposed as

$$\boldsymbol{V} = V^\mu \boldsymbol{e}_\mu = V^\mu \partial_\mu \,. \tag{6.3}$$

A coordinate change

$$x^\mu = f(\tilde{x}^1, \dots, \tilde{x}^n)\,, \tag{6.4}$$

or more briefly $x^\mu = x^\mu(\tilde{x}^\nu)$, changes the basis vectors as

$$\boldsymbol{e}_\mu = \frac{\partial}{\partial x^\mu} = \frac{\partial \tilde{x}^\nu}{\partial x^\mu}\frac{\partial}{\partial \tilde{x}^\nu} = \frac{\partial \tilde{x}^\nu}{\partial x^\mu}\,\tilde{\boldsymbol{e}}_\nu\,. \tag{6.5}$$

Therefore the vector $\boldsymbol{V}$ will be invariant under general coordinate transformations,

$$\boldsymbol{V} = V^\mu \partial_\mu = \tilde{V}^\mu \tilde{\partial}_\mu = \tilde{\boldsymbol{V}}, \tag{6.6}$$

if its components transform opposite to the basis vectors $\boldsymbol{e}_\mu = \partial_\mu$, or

$$V^\mu = \frac{\partial x^\mu}{\partial \tilde{x}^\nu}\,\tilde{V}^\nu. \tag{6.7}$$

If x^μ and $\tilde{x}^\mu$ are two inertial frames in Minkowski space, we came back to Lorentz transformations $\partial x^\mu/\partial\tilde{x}^\nu = \Lambda^\mu{}_\nu$ as a special case of general coordinate transformations.

Covectors or one-forms In quantum mechanics, we use Dirac's bracket notation to associate to each vector $|a\rangle$ a dual vector $\langle a|$ and to introduce a scalar product $\langle a|\, b\rangle$. If the vectors $|n\rangle$ form a basis, then the dual basis $\langle n|$ is defined by $\langle n|\, n'\rangle = \delta_{nn'}$. Similarly, we define a basis $\boldsymbol{e}^\mu$ dual to the basis $\boldsymbol{e}_\mu$ in T_P by

$$\boldsymbol{e}^\mu(\boldsymbol{e}_\nu) = \delta^\mu_\nu\,. \tag{6.8}$$

This basis can be used to form a new vector space T_P^* called the cotangent space which is dual to T_P. Its elements $\boldsymbol{\omega}$ are called covectors or one-forms,

$$\boldsymbol{\omega} = \omega_\mu \boldsymbol{e}^\mu\,. \tag{6.9}$$

Combining a vector and an one-form, we obtain a map into the real numbers,

$$\boldsymbol{\omega}(\boldsymbol{V}) = \omega_\mu V^\nu \boldsymbol{e}^\mu(\boldsymbol{e}_\nu) = \omega_\mu V^\mu\,. \tag{6.10}$$

The last equality shows that we can calculate $\boldsymbol{\omega}(\boldsymbol{V})$ in component form without reference to the basis vectors. In order to simplify notation, we will use therefore in the future simply $\omega_\mu V^\mu$; we also write $\boldsymbol{e}^\mu\boldsymbol{e}_\nu$ instead of $\boldsymbol{e}^\mu(\boldsymbol{e}_\nu)$.

Using a coordinate basis, the duality condition (6.8) is obviously satisfied, if we choose $\boldsymbol{e}^\mu = \mathrm{d}x^\mu$. Then the one-form becomes

$$\boldsymbol{\omega} = \omega_\mu \mathrm{d}x^\mu\,. \tag{6.11}$$

Thus the familiar "infinitesimals" $\mathrm{d}x^\mu$ are actually the finite basis vectors of the cotangent space T_P^*. We require again that the transformation of the components ω_μ of a covector cancels the transformation of the basis vectors,

$$\omega_\mu = \frac{\partial \tilde{x}^\nu}{\partial x^\mu}\,\tilde{\omega}_\nu\,. \tag{6.12}$$

This condition guarantees that the covector itself is an invariant object, since

$$\boldsymbol{\omega} = \omega_\mu \mathrm{d}x^\mu = \frac{\partial \tilde{x}^\nu}{\partial x^\mu}\,\tilde{\omega}_\nu \frac{\partial x^\mu}{\partial \tilde{x}^\sigma}\,\mathrm{d}\tilde{x}^\sigma = \tilde{\omega}_\mu \mathrm{d}\tilde{x}^\mu = \tilde{\boldsymbol{\omega}}\,. \tag{6.13}$$

Covariant and contravariant tensors Next we generalise the concept of vectors and covectors. We call a vector $\boldsymbol{X}$ also a *contravariant tensor* of rank one, while we call a covector also a *covariant vector* or covariant tensor of rank one. A general tensor of rank (n,m) is a multilinear map

$$\boldsymbol{T} = T^{\mu,\dots,\nu}_{\alpha,\dots,\beta} \underbrace{\partial_\mu \otimes \ldots \otimes \partial_\nu}_{n} \otimes \underbrace{\mathrm{d}x^\alpha \otimes \ldots \otimes \mathrm{d}x^\beta}_{m} \tag{6.14}$$

which components transforms as

$$\tilde{T}^{\mu,\dots,\nu}_{\alpha,\dots,\beta}(\tilde{x}) = \underbrace{\frac{\partial \tilde{x}^\mu}{\partial x^\rho} \cdots \frac{\partial \tilde{x}^\nu}{\partial x^\sigma}}_{n} \underbrace{\frac{\partial x^\gamma}{\partial \tilde{x}^\alpha} \cdots \frac{\partial x^\delta}{\partial \tilde{x}^\beta}}_{m} T^{\rho,\dots,\sigma}_{\gamma,\dots,\delta}(x) \tag{6.15}$$

under a coordinate change.

Metric tensor A (pseudo-) Riemannian manifold is a differentiable manifold containing as additional structure a symmetric tensor field $g_{\mu\nu}$ which allows us to measure distances and angles. We define the scalar product of two vectors $\boldsymbol{a}(x)$ and $\boldsymbol{b}(x)$ which have the coordinates a^μ and b^μ in a certain basis $\boldsymbol{e}_\mu$ as

$$\boldsymbol{a} \cdot \boldsymbol{b} = (a^\mu \boldsymbol{e}_\mu) \cdot (b^\nu \boldsymbol{e}_\nu) = (\boldsymbol{e}_\mu \cdot \boldsymbol{e}_\nu) a^\mu b^\nu = g_{\mu\nu} a^\mu b^\nu \,. \tag{6.16}$$

Thus we can evaluate the scalar product between any two vectors, if we know the symmetric matrix $g_{\mu\nu}$ composed out of the N^2 products of the basis vectors,

$$g_{\mu\nu}(x) = \boldsymbol{e}_\mu(x) \cdot \boldsymbol{e}_\nu(x) \,, \tag{6.17}$$

at any point x of the manifold. This symmetric matrix $g_{\mu\nu}$ is called *metric tensor*. The manifold is called Riemannian, if all eigenvalues of $g_{\mu\nu}$ are positive, and thus the scalar product defined by $g_{\mu\nu}$ is positive-definite. If the scalar product is indefinite, as in the case of general relativity, one calls the manifold pseudo-Riemannian.

In the same way, we define for the dual basis $\boldsymbol{e}^\mu$ the metric $g^{\mu\nu}$ via

$$g^{\mu\nu} = \boldsymbol{e}^\mu \cdot \boldsymbol{e}^\nu \,. \tag{6.18}$$

A comparison with Eq. (6.10) shows that the metric $g^{\mu\nu}$ maps covariant vectors X_μ into contravariant vectors X^μ, while $g_{\mu\nu}$ provides a map into the opposite direction. Similarly, we can use the metric tensor to raise and lower indices of any tensor.

Next we want to determine the relation of $g^{\mu\nu}$ with $g_{\mu\nu}$. We multiply $\boldsymbol{e}^\rho$ with $\boldsymbol{e}_\mu = g_{\mu\nu}\boldsymbol{e}^\nu$, obtaining

$$\delta^\rho_\mu = \boldsymbol{e}^\rho \cdot \boldsymbol{e}_\mu = \boldsymbol{e}^\rho \cdot g_{\mu\nu}\boldsymbol{e}^\nu = g^{\rho\nu} g_{\mu\nu} \tag{6.19}$$

or

$$\delta^\rho_\mu = g_{\mu\nu} g^{\nu\rho} \,. \tag{6.20}$$

Thus the components of the covariant and the contravariant metric tensors, $g_{\mu\nu}$ and $g^{\mu\nu}$, are inverse matrices of each other. Moreover, the mixed metric tensor of rank (1,1) is given by the Kronecker delta, $g^\nu_\mu = \delta^\nu_\mu$. Note that the trace of the metric tensor is therefore not -2, but

$$tr(g_{\mu\nu}) = g^{\mu\mu} g_{\mu\mu} = \delta^\mu_\mu = 4 \,, \tag{6.21}$$

because we have to contract an upper and a lower index.

6.2. Covariant derivative and the geodesic equation

Covariant derivative In an inertial system in Minkowski space, taking the partial derivative ∂_μ maps a tensor of rank (n,m) into a tensor of rank $(n+1,m)$. Additionally, this map obeys linearity and the Leibniz product rule. We will see that in general the partial derivative in a curved space does not satisfy these rules. We therefore introduce a new derivative called covariant derivative modified such that it fulfils these rules.

We start by considering the gradient $\partial_\mu\phi$ of a scalar ϕ. By definition, a scalar quantity does not depend on the coordinate system, $\phi(x)=\tilde{\phi}(\tilde{x})$. Therefore its gradient transforms as

$$\partial_\mu\phi \to \tilde{\partial}_\mu\tilde{\phi} = \frac{\partial x^\nu}{\partial\tilde{x}^\mu}\partial_\nu\phi\,. \tag{6.22}$$

Thus the gradient is a covariant vector. Similarly, the derivative of a vector $\boldsymbol{V}$ transforms as a tensor,

$$\partial_\mu\boldsymbol{V} \to \tilde{\partial}_\mu\tilde{\boldsymbol{V}} = \frac{\partial x^\nu}{\partial\tilde{x}^\mu}\partial_\nu\boldsymbol{V}\,, \tag{6.23}$$

because $\boldsymbol{V}$ is an invariant quantity. If we consider however its components $V^\mu=\boldsymbol{e}^\mu\cdot\boldsymbol{V}$, then the moving coordinate basis in curved space-time, $\partial_\mu\boldsymbol{e}^\nu\neq 0$, leads to an additional term in the derivative,

$$\partial_\mu V^\nu = \boldsymbol{e}^\nu\cdot(\partial_\mu\boldsymbol{V}) + \boldsymbol{V}\cdot(\partial_\mu\boldsymbol{e}^\nu)\,. \tag{6.24}$$

The term $\boldsymbol{e}^\nu\cdot(\partial_\mu\boldsymbol{V})$ transforms as a tensor, since both $\boldsymbol{e}^\nu$ and $\partial_\mu\boldsymbol{V}$ are tensors. This implies that the combination of the two remaining terms has to transform as tensor too, which we define as covariant derivative

$$\nabla_\mu V^\nu \equiv \boldsymbol{e}^\nu\cdot(\partial_\mu\boldsymbol{V}) = \partial_\mu V^\nu - \boldsymbol{V}\cdot(\partial_\mu\boldsymbol{e}^\nu)\,. \tag{6.25}$$

The first equality tells us that we can view the covariant derivative $\nabla_\mu V^\nu$ as the projection of $\partial_\mu\boldsymbol{V}$ onto the direction $\boldsymbol{e}^\nu$.

We expand now the partial derivatives of the basis vectors as a linear combination of the basis vectors,

$$\partial_\rho\boldsymbol{e}^\mu = -\Gamma^\mu{}_{\rho\sigma}\boldsymbol{e}^\sigma\,. \tag{6.26}$$

The n^3 numbers $\Gamma^\mu{}_{\rho\sigma}$ are called (affine) connection coefficients or symbols, in order to stress that they are not the components of a tensor. You are asked to derive their transformation properties in problem 6.??. Introducing this expansion into (6.25) we can rewrite the covariant derivative of a vector field as

$$\nabla_\mu V^\nu = \partial_\mu V^\nu + \Gamma^\nu{}_{\sigma\mu}V^\sigma\,. \tag{6.27}$$

Using $\nabla_\sigma\phi=\partial_\sigma\phi$ and requiring that the usual Leibniz rule is valid for $\phi=X_\mu X^\mu$ leads to

$$\nabla_\sigma X_\mu = \partial_\sigma X_\mu - \Gamma^\nu{}_{\mu\sigma}X_\nu \tag{6.28}$$

and to

$$\partial_\rho\boldsymbol{e}_\mu = \Gamma^\sigma{}_{\mu\rho}\boldsymbol{e}_\sigma\,. \tag{6.29}$$

For a general tensor, the covariant derivative is defined by the same reasoning as

$$\nabla_\sigma T^{\mu\dots}_{\nu\dots} = \partial_\sigma T^{\mu\dots}_{\nu\dots} + \Gamma^\mu{}_{\rho\sigma}T^{\rho\dots}_{\nu\dots} + \dots - \Gamma^\rho{}_{\nu\sigma}T^{\mu\dots}_{\rho\dots} - \dots \tag{6.30}$$

Note that it is the last index of the connection coefficients that is the same as the index of the covariant derivative. The *p*lus sign goes together with upper (su*p*erscripts), the minus with lower indices.

Parallel transport We say a tensor $\boldsymbol{T}$ is parallel transported along the curve $x(\sigma)$, if its components $T^{\mu\cdots}_{\nu\cdots}$ stay constant. In flat space, this means simply

$$\frac{\mathrm{d}}{\mathrm{d}\sigma}T^{\mu\dots}_{\nu\dots} = \frac{\mathrm{d}x^\alpha}{\mathrm{d}\sigma}\partial_\alpha T^{\mu\dots}_{\nu\dots} = 0\,. \tag{6.31}$$

In curved space, we have to replace the normal derivative by a covariant one. We define the directional covariant derivative along $x(\sigma)$ as

$$\frac{D}{\mathrm{d}\sigma} = \frac{\mathrm{d}x^\alpha}{\mathrm{d}\sigma}\nabla_\alpha\,. \tag{6.32}$$

Then a tensor is parallel transported along the curve $x(\sigma)$, if

$$\frac{D}{\mathrm{d}\sigma}T^{\mu\dots}_{\nu\dots} = \frac{\mathrm{d}x^\alpha}{\mathrm{d}\sigma}\nabla_\alpha T^{\mu\dots}_{\nu\dots} = 0\,. \tag{6.33}$$

Metric compatibility Relations like $\mathrm{d}s^2 = g_{\mu\nu}\mathrm{d}x^\mu\mathrm{d}x^\nu$ or $g_{\mu\nu}p^\mu p^\nu = m^2$ become invariant under parallel transport only, if the metric tensor is covariantly constant,

$$\nabla_\sigma g_{\mu\nu} = \nabla_\sigma g^{\mu\nu} = 0\,. \tag{6.34}$$

A connection satisfying Eq. (6.34) is called metric compatible and leaves lengths and angles invariant under parallel transport. This requirement guarantees that we can introduce locally in the whole space-time Cartesian inertial coordinate systems where the laws of special relativity are valid. Moreover, these local inertial systems can be consistently connected by parallel transport using an affine connection satisfying the constraint (6.34).

Note that we have already built in this constraint into our definition of the covariant derivative: If the length of a vector would not be conserved under parallel transport, then we should differentiate in (6.24) also the scalar product in $V^\mu = \boldsymbol{e}^\mu \cdot \boldsymbol{V}$, leading to an additional term in Eq. (6.25).

Geodesic equation The requirement that the affine connection is metric compatible fixes the connection not uniquely, and thus the question arises which connection describes physics on a general space-time? Ultimately, the combined action for gravity and matter should select the correct connection—an approach we resume in Chapter 18. For the moment, we use a simple workaround which does not require the knowledge of the action of gravity: In flat space, we know that the solution to the equation of motions of a free particle is a straight line. Such a path is characterised by two properties: It is the shortest curve between the considered initial and final point, and it is the curve whose tangent vectors remains constant if they are parallel transported along it. Both conditions can be generalised to curved space and the curves satisfying either one of them are called geodesics. Using the definition of a geodesics as the "straightest" line on a manifold requires as mathematical structure only the possibility to parallel transport a tensor and thus the existence of an affine connection. In contrast, the concept of an "extremal" (shortest or longest) line between two points on a manifold relies on the existence of a metric. Requiring that these two definitions agree fixes uniquely the connection to be used in the covariant derivative.

We start by defining geodesics as the "straightest" line or an autoparallel curve on a manifold—the case which is almost trivial: The tangent vector along the path $x(\tau)$ is

$u^\mu = \mathrm{d}x^\mu/\mathrm{d}\tau$. Then the requirement (6.33) of parallel transport for u^μ becomes

$$\frac{D}{\mathrm{d}\tau}\frac{\mathrm{d}x^\mu}{\mathrm{d}\tau} = \frac{\mathrm{d}^2x^\mu}{\mathrm{d}\tau^2} + \Gamma^\mu{}_{\rho\sigma}\frac{\mathrm{d}x^\rho}{\mathrm{d}\tau}\frac{\mathrm{d}x^\sigma}{\mathrm{d}\tau} = 0\,. \tag{6.35}$$

Introducing $\dot{x}^\mu = \mathrm{d}x^\mu/\mathrm{d}\tau$, we obtain the geodesic equation in its standard form,

$$\ddot{x}^\mu + \Gamma^\mu{}_{\rho\sigma}\dot{x}^\rho\dot{x}^\sigma = 0\,. \tag{6.36}$$

Note that a possible antisymmetric part of the connection $\Gamma^\mu{}_{\rho\sigma}$ drops out of the geodesic equation, because $\dot{x}^\rho\dot{x}^\sigma$ is symmetric.

Next we derive the defining equation for a geodesics as the extremal curve between two points on a manifold. The Lagrangian of a free particle in Minkowski space, Eq. (1.53), is generalised to a curved space-time manifold with the metric tensor $g_{\mu\nu}$ by replacing $\eta_{\mu\nu}$ with $g_{\mu\nu}$ (we set also $m = -1$),

$$L = g_{\mu\nu}\dot{x}^\mu\dot{x}^\nu\,. \tag{6.37}$$

The Lagrange equations are

$$\frac{\mathrm{d}}{\mathrm{d}\sigma}\frac{\partial L}{\partial(\dot{x}^\lambda)} - \frac{\partial L}{\partial x^\lambda} = 0\,. \tag{6.38}$$

Only the metric tensor $g_{\mu\nu}$ depends on x^μ and thus $\partial L/\partial x^\lambda = g_{\mu\nu,\lambda}\dot{x}^\mu\dot{x}^\nu$. Here we introduced also the short-hand notation $g_{\mu\nu,\lambda} = \partial_\lambda g_{\mu\nu}$ for partial derivatives. Now we use $\partial\dot{x}^\mu/\partial\dot{x}^\nu = \delta^\mu_\nu$ and apply the chain rule for $g_{\mu\nu}(x(\sigma))$, obtaining first

$$g_{\mu\nu,\lambda}\dot{x}^\mu\dot{x}^\nu = 2\,\frac{\mathrm{d}}{\mathrm{d}\sigma}\,(g_{\mu\lambda}\dot{x}^\mu) = 2(g_{\mu\lambda,\nu}\dot{x}^\mu\dot{x}^\nu + g_{\mu\lambda}\ddot{x}^\mu) \tag{6.39}$$

and then

$$g_{\mu\lambda}\ddot{x}^\mu + \frac{1}{2}(2g_{\mu\lambda,\nu} - g_{\mu\nu,\lambda})\dot{x}^\mu\dot{x}^\nu = 0\,. \tag{6.40}$$

Next we rewrite the second term as

$$2g_{\lambda\mu,\nu}\dot{x}^\mu\dot{x}^\nu = (g_{\lambda\mu,\nu} + g_{\lambda\nu,\mu})\dot{x}^\mu\dot{x}^\nu\,, \tag{6.41}$$

multiply everything by $g^{\kappa\mu}$ and arrive at our desired result,

$$\ddot{x}^\kappa + \frac{1}{2}g^{\kappa\lambda}(g_{\mu\nu,\lambda} + g_{\mu\lambda,\nu} - g_{\mu\nu,\lambda})\dot{x}^\mu\dot{x}^\nu = \ddot{x}^\kappa + \{{}^{\kappa}_{\mu\nu}\}\dot{x}^\mu\dot{x}^\nu = 0\,. \tag{6.42}$$

Here we defined in the last step the Christoffel symbols

$$\{{}^{\mu}_{\nu\lambda}\} = \frac{1}{2}g^{\mu\kappa}(\partial_\nu g_{\kappa\lambda} + \partial_\lambda g_{\nu\kappa} - \partial_\kappa g_{\nu\lambda})\,. \tag{6.43}$$

They are also called Levi-Civita or Riemannian connection. A comparison with Eq. (6.36) shows that our two geodesic equations agree, if we choose as connection the Christoffel symbols. Moreover, the Christoffel symbols are symmetric in their two lower indices and, as we will show next, compatible to the metric tensor. Following standard practise, we will denote them also with $\Gamma^\lambda{}_{\mu\nu}$. In the remainder of this section, we will use always as affine connection the Christoffel symbols.

We define[1]

$$\Gamma_{\mu\nu\lambda} = g_{\mu\kappa}\Gamma^{\kappa}{}_{\nu\lambda}\,. \tag{6.44}$$

Thus $\Gamma_{\mu\nu\lambda}$ is symmetric in the last two indices. Then it follows

$$\Gamma_{\mu\nu\lambda} = \frac{1}{2}(\partial_\nu g_{\mu\lambda} + \partial_\lambda g_{\nu\mu} - \partial_\mu g_{\nu\lambda})\,. \tag{6.45}$$

Adding $2\Gamma_{\mu\nu\lambda}$ and $2\Gamma_{\nu\mu\lambda}$ gives

$$\begin{aligned}2(\Gamma_{\mu\nu\lambda} + \Gamma_{\nu\mu\lambda}) &= \partial_b g_{\mu\lambda} + \partial_\lambda g_{\nu\mu} - \partial_\mu g_{\nu\lambda}\\ &\quad + \partial_\mu g_{\nu\lambda} + \partial_\lambda g_{\mu\nu} - \partial_\nu g_{\mu\lambda} = 2\partial_\lambda g_{\mu\nu}\end{aligned} \tag{6.46}$$

or

$$\partial_\lambda g_{\mu\nu} = \Gamma_{\mu\nu\lambda} + \Gamma_{\nu\mu\lambda}\,. \tag{6.47}$$

Applying the general rule for covariant derivatives, Eq. (6.30), to the metric,

$$\nabla_\lambda g_{\mu\nu} = \partial_\lambda g_{\mu\nu} - \Gamma^{\kappa}{}_{\mu c} g_{\kappa\nu} - \Gamma^{\kappa}{}_{\nu\lambda} g_{\mu\kappa} = \partial_\lambda g_{\mu\nu} - \Gamma_{\mu\lambda\nu} - \Gamma_{\nu\lambda\mu}\,, \tag{6.48}$$

and inserting Eq. (6.47) shows that

$$\nabla_\lambda g_{\mu\nu} = \nabla_\lambda g^{\mu\nu} = 0\,. \tag{6.49}$$

Hence ∇_λ commutes with contracting indices,

$$\nabla_\lambda(X^\mu X_\mu) = \nabla_\lambda(g_{\mu\nu}X^\mu X^\mu) = g_{\mu\nu}\nabla_\lambda(X^\mu X^\nu) \tag{6.50}$$

and conserves the norm of vectors as announced. Thus the Christoffel symbols are symmetric and compatible with the metric. These two properties specify uniquely the connection.

Example 6.1: Calculate the Christoffel symbols of the two-dimensional unit sphere S^2.
The line-element of the two-dimensional unit sphere S^2 is given by $\mathrm{d}s^2 = \mathrm{d}\vartheta^2 + \sin^2\vartheta \mathrm{d}\phi^2$. A faster alternative to the definition (6.43) of the Christoffel coefficients is the use of the geodesic equation: From the Lagrange function $L = g_{ab}\dot{x}^a\dot{x}^b = \dot\vartheta^2 + \sin^2\vartheta\dot\phi^2$ we find

$$\begin{aligned}\frac{\partial L}{\partial\phi} &= 0\,, &\qquad \frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial L}{\partial\dot\phi} &= \frac{\mathrm{d}}{\mathrm{d}t}(2\sin^2\vartheta\dot\phi) = 2\sin^2\vartheta\ddot\phi + 4\cos\vartheta\sin\vartheta\dot\vartheta\dot\phi\\ \frac{\partial L}{\partial\vartheta} &= 2\cos\vartheta\sin\vartheta\dot\phi^2\,, &\qquad \frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial L}{\partial\dot\vartheta} &= \frac{\mathrm{d}}{\mathrm{d}t}(2\dot\vartheta) = 2\ddot\vartheta\end{aligned}$$

and thus the Lagrange equations are

$$\ddot\phi + 2\cot\vartheta\dot\vartheta\dot\phi = 0 \qquad \textit{and} \qquad \ddot\vartheta - \cos\vartheta\sin\vartheta\dot\phi^2 = 0\,.$$

Comparing with the geodesic equation $\ddot{x}^\kappa + \Gamma^\kappa{}_{\mu\nu}\dot{x}^\mu\dot{x}^\nu = 0$, we can read off the non-vanishing Christoffel symbols as $\Gamma^\phi{}_{\vartheta\phi} = \Gamma^\phi{}_{\phi\vartheta} = \cot\vartheta$ and $\Gamma^\vartheta{}_{\phi\phi} = -\cos\vartheta\sin\vartheta$. (Note that $2\cot\vartheta = \Gamma^\phi{}_{\vartheta\phi} + \Gamma^\phi{}_{\phi\vartheta}$.)

We can use also the Hamiltonian formulation for a relativistic particle. From the Lagrangian $L = \frac{1}{2}g_{\mu\nu}\dot{x}^\mu\dot{x}^\nu$ we determine first the conjugated momenta $p_\mu = \partial L/\partial\dot{x}^\mu = \dot{x}_\mu$ and perform then a Legendre transformation,

$$H(x^\mu, p_\mu, \tau) = p_\mu\dot{x}^\mu - L(x^\mu, \dot{x}^\mu, \tau) = \frac{1}{2}g^{\mu\nu}p_\mu p_\nu\,. \tag{6.51}$$

[1] We showed that the metric tensor can be used to raise or to lower tensor indices, but the connection Γ is not a tensor.

Since the Lagrangian of a free particle does not depend explicitly on the evolution parameter σ, there exists at least one conserved quantity. This conservation law, $H = 1/2$, expresses the fact that the tangent vector $\dot{x}^\mu$ has a constant norm. Hamilton equations give then

$$\dot{x}^\mu = \frac{\partial H}{\partial p_\mu} = g^{\mu\nu} p_\nu \tag{6.52}$$

and

$$\dot{p}_\mu = -\frac{\partial H}{\partial x^\mu} = -\frac{1}{2}\frac{\partial g^{\alpha\beta}}{\partial x^\mu} p_\alpha p_\beta \,. \tag{6.53}$$

This is a useful alternative to the standard geodesic equation: First, it makes clear that the momentum component p_μ is conserved, if the metric tensor is independent of the coordinate x^μ. Second, we can calculate $\dot{p}_\mu$ directly from the metric tensor, without knowing the Christoffel symbols. Combining the Eqs. (6.52) and (6.53) one can re-derive the standard form of the geodesic equation, cf. problem 6.??.

6.3. Integration and Gauss' theorem

Having defined the covariant derivative of an arbitrary tensor field, it is natural to ask how the inverse, the integral over a tensor field, can be defined. The short answer is that this is in general impossible: Integrating a tensor field requires to sum tensors at different points in an invariant way, which is only possible for scalars. Restricting ourselves to scalar fields, we should generalise an integral like $I = \int \mathrm{d}^4x\, \phi(x)$ valid in an Cartesian inertial frame x^μ in Minkowski space to a general space-time with coordinates $\tilde{x}$. For a general coordinate transformation $x^\mu \to \tilde{x}^\mu$, we have to take into account that the Jacobi determinant $J = \det(\partial\tilde{x}^\mu/\partial x^\rho)$ of the transformation can deviate from one. We can express this Jacobian by the determinant $g \equiv \det(g_{\mu\nu})$ of the metric tensor as follows: Applying the transformation law of the metric tensor,

$$\tilde{g}^{\mu\nu}(\tilde{x}) = \frac{\partial\tilde{x}^\mu}{\partial x^\rho}\frac{\partial\tilde{x}^\nu}{\partial x^\sigma}\, g^{\rho\sigma}(x) \tag{6.54}$$

to the case where the x^ρ are inertial coordinates, we obtain with $g = \det(\eta_{\mu\nu}) = -1$ that

$$\det(\tilde{g}) = J^2 \det(g) = -J^2 \tag{6.55}$$

or $J = \sqrt{|\tilde{g}|}$. Thus $I = \int \mathrm{d}^4x \sqrt{|g|}\, \phi$ is an invariant definition of an integral over a scalar field which agrees for inertial coordinates with the one known from special relativity. Now we choose as scalar ϕ the divergence of a vector field, $\phi = \nabla_\mu X^\mu$, or

$$I = \int \mathrm{d}^4x \sqrt{|g|}\, \nabla_\mu X^\mu = \int \mathrm{d}^4x \sqrt{|g|}\, \left(\partial_\mu X^\mu + \Gamma^\mu{}_{\lambda\mu} X^\lambda\right) \,. \tag{6.56}$$

Our aim is to generalise Gauss' theorem (5.9). The only way how this theorem may be reconciled with (6.56) is to hope that we can express the covariant divergence as $1/\sqrt{|g|}\partial_\mu(\sqrt{|g|}X^\mu)$. In order to check this possibility, we determine first the partial derivative of the metric determinant g. As preparation, we consider the variation of a general matrix M with elements $m_{ij}(x)$ under an infinitesimal change of the coordinates, $\delta x^\mu = \varepsilon x^\mu$. It is convenient to look at the change of $\ln\det M$,

$$\delta \ln\det M \equiv \ln\det(M + \delta M) - \ln\det(M) \tag{6.57a}$$

$$= \ln\det[M^{-1}(M + \delta M)] = \ln\det[1 + M^{-1}\delta M] = \tag{6.57b}$$

$$= \ln[1 + tr(M^{-1}\delta M)] + \mathcal{O}(\varepsilon^2) = tr(M^{-1}\delta M) + \mathcal{O}(\varepsilon^2) \,. \tag{6.57c}$$

In the last step, we used $\ln(1+\varepsilon) = \varepsilon + \mathcal{O}(\varepsilon^2)$. Expressing now both the LHS and the RHS as $\delta M = \partial_\mu M \delta x^\mu$ and comparing then the coefficients of δx^μ gives

$$\partial_\mu \ln \det M = \mathrm{tr}(M^{-1}\partial_\mu M)\,. \tag{6.58}$$

Applied to derivatives of $\sqrt{|g|}$, we obtain

$$\frac{1}{2}g^{\mu\nu}\partial_\lambda g_{\mu\nu} = \frac{1}{2}\partial_\lambda \ln g = \frac{1}{\sqrt{|g|}}\partial_\lambda(\sqrt{|g|})\,. \tag{6.59}$$

This expression coincides with contracted Christoffel symbols,

$$\Gamma^\mu{}_{\mu\nu} = \frac{1}{2}g^{\mu\kappa}(\partial_\mu g_{\kappa\nu} + \partial_\nu g_{\mu\kappa} - \partial_\kappa g_{\mu\nu}) = \frac{1}{2}g^{\mu\kappa}\partial_\nu g_{\mu\kappa} = \frac{1}{2}\partial_\nu \ln g = \frac{1}{\sqrt{|g|}}\partial_\nu(\sqrt{|g|})\,. \tag{6.60}$$

Now we can express the divergence of a vector field as

$$\nabla_\mu X^\mu = \partial_\mu X^\mu + \Gamma^\mu{}_{\lambda\mu}X^\lambda = \partial_\mu X^\mu + \frac{1}{\sqrt{|g|}}(\partial_\mu\sqrt{|g|})X^\mu = \frac{1}{\sqrt{|g|}}\partial_\mu(\sqrt{|g|}X^\mu)\,. \tag{6.61}$$

Gauss' theorem for the divergence of a vector field follows directly,

$$\int_\Omega \mathrm{d}^4x\sqrt{|g|}\;\nabla_\mu X^\mu = \int_\Omega \mathrm{d}^4x\;\partial_\mu(\sqrt{|g|}X^\mu) = \int_{\partial\Omega} \mathrm{d}S_\mu\sqrt{|g|}\;X^\mu\,. \tag{6.62}$$

This implies in particular that we can drop terms like $\nabla_\mu X^\mu$ in the action, if the vector field X^μ vanishes on the boundary. Similarly, Gauss' theorem allows us to derive global conservation laws from $\nabla_\mu X^\mu = 0$ in the same way as in Minkowski space.

Next we consider the divergence of an antisymmetric tensor of rank 2,

$$\nabla_\mu A^{\mu\nu} = \partial_\mu A^{\mu\nu} + \Gamma^\mu{}_{\lambda\mu}A^{\lambda\nu} + \Gamma^\nu{}_{\lambda\mu}A^{\mu\lambda} = \frac{1}{\sqrt{|g|}}\partial_\mu(\sqrt{|g|}A^{\mu\nu})\,. \tag{6.63}$$

Because of the antisymmetry of $A^{\mu\lambda}$ the term $\Gamma^\nu{}_{\lambda\mu}A^{\mu\lambda}$ vanishes, and we can combine the first two terms as in the vector case. This generalises to completely antisymmetric tensors of all orders. In contrast, we find for a symmetric tensor of rank 2,

$$\nabla_\mu S^{\mu\nu} = \partial_\mu S^{\mu\nu} + \Gamma^\mu{}_{\lambda\mu}S^{\lambda\nu} + \Gamma^\nu{}_{\lambda\mu}S^{\mu\lambda} = \frac{1}{\sqrt{|g|}}\,\partial_\mu(\sqrt{|g|}S^{\mu\nu}) + \Gamma^\nu{}_{\lambda\mu}S^{\mu\lambda}\,. \tag{6.64}$$

Hence the divergence of a symmetric tensor of rank two contains an additional term $(\partial_\nu g_{\mu\lambda})S^{\mu\lambda}$ which prohibits the use of Gauss' theorem.

6.4. Symmetries of a general space-time

In the case of a Riemannian space-time manifold $(\mathscr{M}, \boldsymbol{g})$, we say the space-time possess a symmetry if it looks the same as one moves from a point P along a vector field ξ^μ to a different point $\tilde{P}$. More precisely, we mean with "looking the same" that the metric tensor transported along ξ^μ remains the same.

Such symmetries may be obvious, if one uses coordinates adapted to these symmetries: For instance, the metric may be independent from one or several coordinates. Let us assume that the metric is e.g. independent from the time coordinate x^0. Then x^0 is a cyclic coordinate, $\partial L/\partial x^0 = 0$, of the Lagrangian $L = \mathrm{d}\tau/\mathrm{d}\sigma$ of a free test particle moving in $\mathscr{M}$. With $L = \mathrm{d}\tau/\mathrm{d}\sigma$, the resulting conserved quantity $\partial L/\partial \dot{x}^0 = const.$ can be written as

$$\frac{\partial L}{\partial \dot{x}^0} = g_{0\beta}\frac{\mathrm{d}x^\beta}{L\mathrm{d}\sigma} = g_{0\beta}\frac{\mathrm{d}x^\beta}{\mathrm{d}\tau} = \boldsymbol{\xi}\cdot\boldsymbol{u} \tag{6.65}$$

with $\boldsymbol{\xi} = \boldsymbol{e}_0$ and $\boldsymbol{u}$ as the four-velocity. Hence the quantity $\boldsymbol{\xi}\cdot\boldsymbol{u} = p^0/m$ is conserved along the solutions $x^\alpha(\sigma)$ of the Lagrange equation of a free particle on $\mathscr{M}$, i.e. along geodesics: In other words, the motion of all test particles in the corresponding space-time conserve energy. The vector field $\boldsymbol{\xi}$ that points in the direction in which the metric does not change is called Killing vector field.

Since we allow arbitrary coordinate systems, space-time symmetries are however in general not evident by a simple inspection of the metric tensor. We say the metric is invariant moving along the Killing vector field ξ^μ, when the resulting change $\delta g^{\mu\nu}$ of the metric is zero. In order to use this condition, we have to be able to calculate how the tensor $g^{\mu\nu}$ changes as we transport it along a vector field ξ^μ. Clearly, it is sufficient to consider an infinitesimal distance. Then we can work in the approximation

$$\tilde{x}^\mu = x^\mu + \varepsilon\xi^\mu(x^\nu) + \mathcal{O}(\varepsilon^2)\,, \qquad \varepsilon \ll 1\,, \tag{6.66}$$

and neglect all terms quadratic in ε.

We recall first the transformation law for a rank two tensor as the metric under an arbitrary coordinate transformation,

$$\tilde{g}^{\mu\nu}(\tilde{x}) = \frac{\partial\tilde{x}^\mu}{\partial x^\alpha}\frac{\partial\tilde{x}^\nu}{\partial x^\beta}\, g^{\alpha\beta}(x)\,. \tag{6.67}$$

Applied to the transport along $\boldsymbol{\xi}$ defined in (6.66), we obtain

$$\tilde{g}^{\mu\nu}(\tilde{x}) = \frac{\partial\tilde{x}^\mu}{\partial x^\alpha}\frac{\partial\tilde{x}^\nu}{\partial x^\beta}\, g^{\alpha\beta}(x) = (\delta^\mu_\alpha + \varepsilon\xi^\mu{}_{,\alpha})(\delta^\nu_\beta + \varepsilon\xi^\nu{}_{,\beta})g^{\alpha\beta}(x) \tag{6.68a}$$

$$= g^{\mu\nu}(x) + \varepsilon(\xi^{\mu,\nu} + \xi^{\nu,\mu}) + \mathcal{O}(\varepsilon^2)\,. \tag{6.68b}$$

In order to be able to compare the new $\tilde{g}^{\mu\nu}(\tilde{x})$ with $g^{\mu\nu}(x)$, we have to express $\tilde{g}^{\mu\nu}(\tilde{x})$ as function of x. A Taylor expansion gives

$$\tilde{g}^{\mu\nu}(\tilde{x}) = \tilde{g}^{\mu\nu}(x+\varepsilon\xi) = \tilde{g}^{\mu\nu}(x) + \varepsilon\xi^\alpha\partial_\alpha\tilde{g}^{\mu\nu}(x) + \mathcal{O}(\varepsilon^2)\,. \tag{6.69}$$

Setting equal Eqs. (6.68b) and (6.69), we obtain

$$g^{\mu\nu}(x) + \varepsilon(\xi^{\mu,\nu} + \xi^{\nu,\mu}) = \tilde{g}^{\mu\nu}(x) + \varepsilon\xi^\alpha\partial_\alpha\tilde{g}^{\mu\nu}(x)\,. \tag{6.70}$$

Thus the metric is kept invariant, $\tilde{g}^{\mu\nu}(x) = g^{\mu\nu}(x)$, if the condition

$$\xi^{\mu,\nu} + \xi^{\nu,\mu} - \xi^\rho\partial_\rho g^{\mu\nu} = 0 \tag{6.71}$$

or

$$g^{\mu\rho}\partial_\rho\xi^\nu + g^{\nu\rho}\partial_\rho\xi^\mu - \xi^\rho\partial_\rho g^{\mu\nu} = 0 \tag{6.72}$$

is satisfied. Expressing partial derivatives as covariant ones[2], the terms containing connection coefficients cancel and we obtain the Killing equation

$$\nabla_\mu \xi_\nu + \nabla_\nu \xi_\mu = 0\,. \tag{6.73}$$

Its solutions $\boldsymbol{\xi}$ are the Killing vectors of the metric.

We now check that Eq. (6.73) leads indeed to a conservation law, as required by our initial definition of a Killing vector field. We multiply the equation for geodesic motion,

$$\frac{Du^\mu}{\mathrm{d}\tau} = 0, \tag{6.74}$$

by the Killing vector ξ_μ and use Leibniz's product rule together with the definition of the absolute derivative (6.32),

$$\xi_\mu \frac{Du^\mu}{\mathrm{d}\tau} = \frac{\mathrm{d}}{\mathrm{d}\tau}(\xi_\mu u^\mu) - \nabla_\mu \xi_\nu u^\mu u^\nu = 0. \tag{6.75}$$

The second term vanishes for a Killing vector field ξ^μ, because the Killing equation implies the antisymmetry of $\nabla_\mu \xi_\nu$. Hence the quantity $\xi_\mu u^\mu$ is indeed conserved along any geodesics.

Example 6.2: Find all ten Killing vector fields of Minkowski space and specify the corresponding symmetries and conserved quantities.
The Killing equation $\nabla_\mu \xi_\nu + \nabla_\nu \xi_\mu = 0$ simplifies in Minkowski space to

$$\partial_\mu \xi_\nu + \partial_\nu \xi_\mu = 0\,. \tag{6.76}$$

Taking one more derivative and using the symmetry of partial derivatives, we arrive at

$$\partial_\rho \partial_\mu \xi_\nu + \partial_\rho \partial_\nu \xi_\mu = 2\partial_\mu \partial_\rho \xi_\nu = 0\,. \tag{6.77}$$

Integrating this equation twice, we find

$$\xi^\mu = \omega_\nu{}^\mu x^\nu + a^\mu\,. \tag{6.78}$$

The matrix $\omega^{\mu\nu}$ has to be antisymmetric in order to satisfy Eq. (6.76). Thus the Killing vector fields are determined by ten integration constants. They agree with the infinitesimal generators of Lorentz transformations, cf. appendix B.3.
The four parameters a^μ generate translations, $x^\mu \to x^\mu + a^\mu$, described by four Killing vector fields which can be chosen as the Cartesian basis vectors of Minkowski space,

$$\boldsymbol{T}_0 = \partial_t, \qquad \boldsymbol{T}_1 = \partial_x, \qquad \boldsymbol{T}_2 = \partial_y, \qquad \boldsymbol{T}_3 = \partial_z\,.$$

For a particle with momentum $p^\mu = mu^\mu$ moving along $x^\mu(\lambda)$, the existence of a Killing vector $\boldsymbol{T}^\mu$ implies

$$\frac{\mathrm{d}}{\mathrm{d}\lambda}(\boldsymbol{T}^\mu \cdot \boldsymbol{u}) = \frac{\mathrm{d}}{m\mathrm{d}\lambda}(\boldsymbol{T}^\mu \cdot \boldsymbol{p}) = 0\,,$$

i.e. the conservation of the four-momentum component p_μ.
Consider next the ij (=spatial) components of the Killing equation. Three additional Killing vectors are

$$\boldsymbol{J}_1 = y\partial_z - z\partial_y\,, \qquad \text{and cyclic permutations.} \tag{6.79}$$

[2] Since Eq. (6.73) is tensor equation, the previous Eq. (6.72) is also invariant under arbitrary coordinate transformations, although it contains only partial derivatives. This suggests that one can introduce the derivative of an arbitrary tensor along a vector field, called Lie derivative, without the need for a connection.

The existence of Killing vectors $\boldsymbol{J}_i$ implies that $\boldsymbol{J}_i \cdot \boldsymbol{p}$ is conserved along a geodesics of particle. But

$$\boldsymbol{J}_1 \cdot \boldsymbol{p} = yp_z - zp_y = J_x$$

and thus the angular momentum around the origin of the coordinate system is conserved.
The other three components satisfy the 0α component of the Killing equations (${\omega_1}^0 = {\omega_0}^1$),

$$\boldsymbol{K}_1 = t\partial_z + z\partial_t\,, \qquad \text{and cyclic permutations.} \tag{6.80}$$

The conserved quantity $tp_z - zE = const.$ now depends on time and is therefore not as popular as the previous ones. Its conservation implies that the center of mass of a system of particles moves with constant velocity $v_\alpha = p_\alpha/E$.

Global conservation laws An immediate consequence of Eq. (6.61) is a covariant form of Gauss' theorem for vector fields. In particular, we can conclude from local current conservation, $\nabla_\mu j^\mu = 0$, the existence of a globally conserved charge. If the conserved current j^μ vanishes at infinity, then we obtain also in a general space-time

$$\int_\Omega \mathrm{d}^4x\sqrt{|g|}\,\nabla_\mu j^\mu = \int_\Omega \mathrm{d}^4x\partial_\mu(\sqrt{|g|}j^\mu) = \int_{\partial\Omega} \mathrm{d}S_\mu\sqrt{|g|}\,j^\mu = 0\,. \tag{6.81}$$

Thus the conservation of Noether charges of internal symmetries as the electric charge, baryon number, etc., continues to hold in a curved space-time.

Next we consider the energy-momentum stress tensor as an example for a locally conserved symmetric tensors of rank two. Now, the second term in Eq. (6.64) prevents us to convert the local conservation law into a global one. If the space-time admits however a Killing field ξ_μ, then we can form the vector field $P^\mu = T^{\mu\nu}\xi_\nu$ with

$$\nabla_\mu P^\mu = \nabla_\mu(T^{\mu\nu}\xi_\nu) = \xi_\nu\nabla_\mu T^{\mu\nu} + T^{\mu\nu}\nabla_\mu\xi_\nu = 0\,. \tag{6.82}$$

Here, the first term vanishes since $T^{\mu\nu}$ is conserved and the second because $T^{\mu\nu}$ is symmetric, while $\nabla_\mu\xi_\nu$ is antisymmetric. Therefore the vector field $P^\mu = T^{\mu\nu}\xi_\nu$ is also conserved, $\nabla_\mu P^\mu = 0$, and we obtain thus the conservation of the component of the energy-momentum vector in direction ξ.

Global energy conservation requires thus the existence of a time-like Killing vector field. If the metric is time-dependent, as e.g. in the case of the expanding Universe, a time-like Killing vector field does not exist and the energy contained in a "comoving" volume changes with time.

Summary

In a curved space-time $\mathcal{M}$ we require a connection to compare vectors at different points. The unique connection which is symmetric and compatible with the metric are the Christoffel symbols. The symmetries of a space-time $\mathcal{M}$ are determined by its Killing vector fields ξ^μ. The momentum component parallel to ξ^μ of test particles moving in $\mathcal{M}$ is conserved. Locally conserved currents lead in general only for vector currents to globally conserved charges. In the case of locally conserved tensors (as $\nabla_\mu T^{\mu\nu} = 0$), the global conservation of the corresponding charges requires the existence of Killing vector fields.

Further reading

[LL80] introduces classical field theory including general relativity. [Car03] presents a clear introduction to differential geometry on a level accessible for physicists.

7. Spin-1 and spin-2 fields

We introduced fields transforming as tensors under coordinate general transformations. Such fields have integer spin and obey Bose-Einstein statistics. Therefore they can exist as macroscopic fields and are thus candidates to describe the electric and the gravitational force. Since both the electric and the gravitational potential $V(r)$ follow a $1/r$ law, we expect from our discussion of the Yukawa potential that the two forces are mediated by massless particles. We will find later that no interacting theory of massless particles with spin $s > 2$ exists. Therefore it is sufficient to consider the two cases $s = 1$ and $s = 2$.

7.1. Tensor fields

The momentum modes $\propto \mathrm{e}^{-\mathrm{i}kx}$ of massive fields can be boosted to their rest-frame, where $k^\mu = (m, \mathbf{0})$. In this frame, the total angular momentum reduces to spin, and nonrelativistic quantum mechanics is valid. Thus a field with spin s has $2s + 1$ spin or polarisation states. On the other hand, we can determine the spin[1] s of a field calculating the angle $s\alpha$ it is turned by a rotation $R_{ij}(\alpha)$. If we consider the transformation law of a tensor field of rank n, $\tilde{T}^{\mu_1\cdots\mu_n} = \Lambda^{\mu_1}{}_{\nu_1} \cdots \Lambda^{\mu_n}{}_{\nu_n} T^{\nu_1\cdots\nu_n}$, for the special case of rotations, we see that the n factors $\Lambda(\alpha)$ rotate some components of $T^{\nu_1\cdots\nu_n}$ by the angle $n\alpha$. Therefore a tensor field of rank n has spin $s = n$. This implies that fields with spin $s \geq 1$ contain unphysical degrees of freedom: For instance, a massive spin-1 field has three and a massive spin-2 field has five polarisation states. On the other hand, a vector field A^μ has four components, and a symmetric tensor field $h^{\mu\nu}$ of rank two has ten components in $d = 4$ space-time dimensions. The purpose of a relativistic wave equation is thus to impose the correct relativistic dispersion relation and to select the correct physical polarisation states in the chosen frame.

The first requirement is fulfilled if each component of a free field ϕ_a satisfies the Klein-Gordon equation. Additionally, we have to impose constraints f_i which eliminate the redundant components,

$$\left(\Box + m^2\right) \phi_a(x) = 0\,, \qquad and \quad f_i(\phi_a(x)) = 0\,. \tag{7.1}$$

The reason for this mismatch in the number of degrees of freedom is that in general a tensor of rank n is reducible, i.e. it contains components of rank $< n$. For instance, the trace h^μ_μ of a second rank tensor transforms clearly as a scalar. Therefore we should choose the constraints for massive fields with spin s such that all components with spin $< s$ are eliminated.

Example 7.1: An object which contains invariant subgroups with respect to a symmetry operation is called reducible. As example, consider the reducible subgroups of a symmetric tensor $h^{\mu\nu}$ of rank two with respect to spatial rotations. Since one can boost a massive particle into its rest frame, this the relevant decomposition to find its spin states. We split $h^{\mu\nu}$ into a scalar h^{00}, a vector h^{0i} and a reducible tensor h^{ij},

$$h^{\mu\nu} = \begin{pmatrix} h^{00} & h^{0i} \\ h^{i0} & h^{ij} \end{pmatrix}.$$

[1] See the appendices B.3 and B.4 for a brief discussion.

Then we decompose h^{ij} again into its trace h^{ii} and its traceless part $h^i_j - h\delta^i_j/(d-1)$. The latter has $6-1=5$ degrees of freedom in $d=4$, as required for a massive spin-2 field.

This problem is more severe for massless fields: We know from classical electrodynamics that the photon has only two polarisation states, and in appendix B.4 it is shown that this holds for massless fields with any spin $s>0$. The redundant degrees of freedom of massless fields can be consistently eliminated only, if some redundancy of the field variables exists which in turn leads to a local symmetry of the field. In this chapter, we discuss the consequences of this redundancy called gauge symmetry on the level of the wave equations and their solutions for the photon and the graviton.

Tensor structure of the propagator We can gain some insight into the general tensor structure of the Feynman propagator for fields with spin $s>0$ using the definition of the 2-point Green function as the time-ordered vacuum expectation values of fields. In general, we can express an arbitrary solutions of a free spin $s=0,1$ and 2 field by its Fourier components as

$$\phi(x) = \int \frac{\mathrm{d}^3k}{\sqrt{(2\pi)^3 2\omega_k}} \left[a(\boldsymbol{k})\mathrm{e}^{-\mathrm{i}(\omega_k t - \boldsymbol{kx})} + \mathrm{h.}c.\right] , \tag{7.2}$$

$$A^\mu(x) = \sum_r \int \frac{\mathrm{d}^3k}{\sqrt{(2\pi)^3 2\omega_k}} \left[a_r(\boldsymbol{k})\varepsilon^\mu_r(\boldsymbol{k})\mathrm{e}^{-\mathrm{i}(\omega_k t - \boldsymbol{kx})} + \mathrm{h.}c.\right] , \tag{7.3}$$

$$h^{\mu\nu}(x) = \sum_r \int \frac{\mathrm{d}^3k}{\sqrt{(2\pi)^3 2\omega_k}} \left[a_r(\boldsymbol{k})\varepsilon^{\mu\nu}_r(\boldsymbol{k})\mathrm{e}^{-\mathrm{i}(\omega_k t - \boldsymbol{kx})} + \mathrm{h.}c.\right] , \tag{7.4}$$

where the momentum is on-shell, $k^\mu = (\omega_k, \boldsymbol{k})$ and r labels the spin or polarisation states. The constraints $f_i = 0$ are now conditions on the polarisation vector and tensor, respectively, which depend on $\boldsymbol{k}$. Proceeding as in the scalar case, we expect that e.g. the propagator for a vector field is given by

$$\mathrm{i}D_F^{\mu\nu}(x) = \langle 0|T\{A^\mu(x)A^{\nu *}(0)\}|0\rangle = \tag{7.5}$$

$$= \sum_r \int \frac{\mathrm{d}^3k}{(2\pi)^3 2\omega_k} \left[\varepsilon^\mu_r(\boldsymbol{k})\varepsilon^{\nu *}_r(\boldsymbol{k})\mathrm{e}^{-\mathrm{i}kx}\vartheta(x^0) + \varepsilon^\mu_r(\boldsymbol{k})\varepsilon^{\nu *}_r(\boldsymbol{k})\mathrm{e}^{\mathrm{i}kx}\vartheta(-x^0)\right] \tag{7.6}$$

$$= \int \frac{\mathrm{d}^4k}{(2\pi)^4} \frac{\mathcal{P}^{\mu\nu}(k)\,\mathrm{e}^{-\mathrm{i}kx}}{k^2 - m^2 + \mathrm{i}\varepsilon} . \tag{7.7}$$

The expression (7.6) is in line with the interpretation of the propagator as the probability for the creation of a particle at x with any momentum k and polarisation r, its propagation to x' followed by its annihilation. In the last step, we introduced the tensor $\mathcal{P}^{\mu\nu}(k)$ which corresponds to the sum over the polarisation states $\varepsilon^\mu_r(\boldsymbol{k})\varepsilon^{\nu *}_r(\boldsymbol{k})$. We will show that the polarisation tensors are polynomials in the momentum k, and thus Eq. (7.6) shows that $\mathcal{P}^{\mu\nu}(k)$ is even in the momentum. As a result, our discussion of causality in the scalar case applies for all tensor field, implying that these fields seen as quantum fields commute. Therefore the particles described by these fields satisfy Bose-Einstein statistics.

7.2. Vector fields

Proca and Maxwell equations A massive vector field A^μ has four components in $d = 4$ space-time dimensions, while it has only $2s+1 = 3$ independent spin components. Correspondingly, a four-vector A^μ transforms under a rotation as $(A^0, \boldsymbol{A})$, i.e. it contains a scalar and a three-vector. Therefore we have to add to the four Klein-Gordon equations for A^μ one constraint which eliminates A^0: The only linear, Lorentz invariant possibility is

$$\left(\Box + m^2\right) A^\mu(x) = 0 \qquad \textit{and} \quad \partial_\mu A^\mu = 0\,. \tag{7.8}$$

In momentum space, this translates into $(k^2 - m^2)A^\mu(k) = 0$ and $k_\mu A^\mu(k) = 0$. In the rest frame of the particle, $k^\mu = (m, \mathbf{0})$, and the constraint becomes $A^0 = 0$. Hence a field satisfying (7.8) has only three space-like components as required for a massive $s = 1$ field. We can choose the three polarisation vectors which label the three degrees of freedom in the rest frame e.g. as the Cartesian unit vectors, $\boldsymbol{\varepsilon}_i \propto \mathbf{e}_i$.

The two equations can be combined into one equation called Proca equation,

$$\left(\eta^{\mu\nu}\Box - \partial^\mu\partial^\nu\right) A_\nu + m^2 A^\mu = 0\,. \tag{7.9}$$

To show the equivalence of this equation with (7.8), we act with ∂_μ on it,

$$\left(\partial^\nu\Box - \Box\partial^\nu\right) A_\nu + m^2\partial_\mu A^\mu = m^2\partial_\mu A^\mu = 0\,. \tag{7.10}$$

Hence, a solution of the Proca equation fulfils automatically the constraint $\partial_\mu A^\mu = 0$ for $m^2 > 0$. On the other hand, we can neglect the second term in (7.9) for $\partial_\nu A^\nu = 0$ and obtain the Klein-Gordon equation.

We now go over to the case of a massless spin-1 field which is described by the Maxwell equations. In classical electrodynamics, the field-strength tensor $F_{\mu\nu}$ is an observable quantity, while the potential A_μ is merely a convenient auxiliary quantity. From the definition

$$F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu \tag{7.11}$$

it is clear that $F_{\mu\nu}$ is invariant under the transformations

$$A_\mu(x) \to A'_\mu(x) = A_\mu(x) - \partial_\mu\Lambda(x)\,. \tag{7.12}$$

Thus $A'_\mu(x)$ is for any $\Lambda(x)$ physically equivalent to $A_\mu(x)$, leading to the same field-strength tensor and thus e.g. to the same Lorentz force on a particle. The transformations (7.12) are called gauge transformations. Note that the mass term m^2A^μ in the Proca equation breaks gauge invariance.

If we insert into the Maxwell equation the definition of the potential,

$$\partial_\mu F^{\mu\nu} = \partial_\mu(\partial^\mu A^\nu - \partial^\nu A^\mu) = \Box A^\nu - \partial_\mu\partial^\nu A^\mu = j^\nu\,, \tag{7.13}$$

we see that this expression equals the $m = 0$ limit of the Proca equation. Gauge invariance allows us to choose a potential A^μ such that $\partial_\mu A^\mu = 0$. Such a choice is called fixing the gauge, and the particular case $\partial_\mu A^\mu = 0$ is denoted as Lorenz gauge. In the Lorenz gauge, the wave equation simplifies to

$$\Box A^\mu = j^\mu\,. \tag{7.14}$$

Additionally, we can add to the potential A^μ any function $\partial^\mu\chi$ satisfying $\Box\chi = 0$. We can use this freedom to set $A^0 = 0$. Inserting then a plane-wave $A^\mu \propto \varepsilon^\mu \mathrm{e}^{\mathrm{i}kx}$ into the free wave equation, $\Box A^\nu = 0$, we find that k is a null-vector and that $\varepsilon^\mu k_\mu = -\boldsymbol{\varepsilon} \cdot \boldsymbol{k} = 0$. Thus the photon propagates with the speed of light, is transversely polarised and has two polarisation states as expected for a massless particle.

Closely connected to the gauge invariance of electrodynamics is the fact that its source, the electromagnetic current, is conserved. The antisymmetry of $F^{\mu\nu}$, which is the basis for the symmetry (7.12), leads also to $\partial_\mu \partial_\nu F^{\mu\nu} = 0$. Thus the Maxwell equation $\partial_\mu F^{\mu\nu} = j^\nu$ implies the conservation of the electromagnetic current j^μ,

$$\partial_\mu \partial_\nu F^{\mu\nu} = \partial_\nu j^\nu = 0 . \tag{7.15}$$

Propagator for massive spin-1 fields The propagator $D_{\mu\nu}$ for a massive spin-1 field is determined by

$$\left[\eta^{\mu\nu}(\Box + m^2) - \partial^\mu \partial^\nu\right] D_{\nu\lambda}(x) = \delta^\mu_\lambda \delta(x) . \tag{7.16}$$

Inserting the Fourier transformation of the propagator and of the delta function gives

$$\left[\left(-k^2 + m^2\right) \eta^{\mu\nu} + k^\mu k^\nu\right] D_{\nu\lambda}(k) = \delta^\mu_\lambda . \tag{7.17}$$

We will apply the tensor method to solve this equation: In this approach, we use first all tensors available in the problem to construct the required tensor of rank 2. In the case at hand, we have at our disposal only the momentum k_μ of the particle—which we can combine to $k_\mu k_\nu$—and the metric tensor $\eta_{\mu\nu}$. Thus the tensor structure of $D_{\mu\nu}(k)$ has to be of the form

$$D_{\mu\nu}(k) = A\eta_{\mu\nu} + Bk_\mu k_\nu \tag{7.18}$$

with two unknown scalar functions $A(k^2)$ and $B(k^2)$. Inserting this ansatz into (7.17) and multiplying out, we obtain

$$\left[(-k^2 + m^2)\eta^{\mu\nu} + k^\mu k^\nu\right]\left[A\eta_{\nu\lambda} + Bk_\nu k_\lambda\right] = \delta^\mu_\lambda , \tag{7.19a}$$

$$-Ak^2\delta^\mu_\lambda + Am^2\delta^\mu_\lambda + Ak^\mu k_\lambda + Bm^2 k^\mu k_\lambda = \delta^\mu_\lambda , \tag{7.19b}$$

$$-A(k^2 - m^2)\delta^\mu_\lambda + (A + Bm^2)k^\mu k_\lambda = \delta^\mu_\lambda . \tag{7.19c}$$

In the last step, we regrouped the LHS into the two tensor structures δ^μ_λ and $k^\mu k_\lambda$. A comparison of their coefficients gives then $A = -1/(k^2 - m^2)$ and

$$B = -\frac{A}{m^2} = \frac{1}{m^2(k^2 - m^2)} .$$

Thus the massive spin-1 propagator follows as

$$D_F^{\mu\nu}(k) = \frac{-\eta^{\mu\nu} + k^\mu k^\nu/m^2}{k^2 - m^2 + \mathrm{i}\varepsilon} . \tag{7.20}$$

Note that there is a sign ambiguity, since we could have added a minus sign to the Proca equation.

Next we check this sign and our claim that the propagator $D_F^{ab}(k)$ of spin $s > 0$ fields can be obtained as sum over their polarisation states $\varepsilon_a^{(r)}$ times the scalar propagator $\Delta(k)$. As the theory is Lorentz invariant, we can choose the frame most convenient for this comparison which is the rest-frame of the massive particle. Then $k^\mu = (m, \mathbf{0})$ and the three polarisation vectors can be chosen as the Cartesian basis vectors. Comparing then

$$-\eta^{\mu\nu} + k^\mu k^\nu/m^2 = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} = \sum_r \varepsilon^{\mu(r)}\varepsilon^{\nu(r)}\,, \tag{7.21}$$

shows that both methods agree and can be used to derive the Feynman propagator. In the latter approach, working from the RHS to the LHS of Eq. (7.21), we derive first the expression valid for the Feynman propagator in a specific frame. Then we have to rewrite the expression in an invariant way using the relevant tensors, here $\eta^{\mu\nu}$ and $k^\mu k^\nu$. Moreover, Eq. 7.21 shows that we have chosen the right sign for the propagator (7.20).

Propagator for massless spin-1 fields As we have seen, we can set $m = 0$ in the Proca equation and obtain the Maxwell equation. The corresponding limit of the propagator (7.20) leads however to an ill-defined result. As we know that the number of degrees of freedom differs between the massive and the massless case, this is not too surprising. If we try next the limit $m \to 0$ in Eq. (7.19c), then we find

$$-Ak^2\delta^\mu_\lambda + Ak^\mu k_\lambda = \delta^\mu_\lambda. \tag{7.22}$$

This equation has for arbitrary k with $A = -1/k^2$ and $A = 0$ no solution. Moreover, the function B is undetermined. We can understand this physically, since for a massless field current conservation holds. But $\partial_\mu J^\mu(x) = 0$ implies $k_\mu J^\mu(k) = 0$ and thus the $k^\mu k^\nu$ term does not influence physical quantities: In physical measurable quantities, as e.g. $W[J]$, the propagator is always matched between conserved currents, and the longitudinal part $k^\mu k^\nu$ drops out.

We now try to construct the photon propagator from its sum over polarisation states. First we consider a linearly polarised photon with polarisation vectors $\varepsilon^{(r)}_\mu$ lying in the plane perpendicular to its momentum vector $\boldsymbol{k}$. If we perform a Lorentz boost on $\varepsilon^{(1)}_\mu$, we will find

$$\tilde{\varepsilon}^{(1)}_\mu = \Lambda^\nu{}_\mu \varepsilon^{(1)}_\nu = a_1\varepsilon^{(1)}_\mu + a_2\varepsilon^{(2)}_\mu + a_3 k_\mu\,, \tag{7.23}$$

where the coefficients a_i depend on the direction $\boldsymbol{\beta}$ of the boost. Thus, in general the polarisation vector will not be anymore perpendicular to $\boldsymbol{k}$. Similarly, if we perform a gauge transformation

$$A_\mu(x) \to A'_\mu(x) = A_\mu(x) - \partial_\mu\Lambda(x) \tag{7.24}$$

with

$$\Lambda(x) = -\mathrm{i}\lambda\exp(-\mathrm{i}kx) + \text{h.c.}\,, \tag{7.25}$$

then

$$A'_\mu(x) = (\varepsilon_\mu + \lambda k_\mu)\exp(-\mathrm{i}kx) + \text{h.c.} = \varepsilon'_\mu\exp(-\mathrm{i}kx) + \text{h.c.} \tag{7.26}$$

Choosing e.g. a photon propagating in z direction, $k^\mu = (\omega, 0, 0, \omega)$, we see that the gauge transformation does not affect the transverse components k_x and k_y. Thus only the components of ε^μ transverse to $\boldsymbol{k}$ can have physical significance. On the other hand, the time-like and longitudinal components depend on the arbitrary parameter λ and are therefore unphysical. In particular, they can be set to zero by a gauge transformation: First, $\varepsilon'_\mu k'^\mu = 0$ implies (again for a photon propagating in z direction) $\varepsilon'_0 = -\varepsilon'_3$. From $\varepsilon'_3 = \varepsilon_3 + \lambda\omega$, we see that $\lambda = \varepsilon_3/\omega$ sets $\varepsilon'_3 = -\varepsilon'_0 = 0$. Thus the transformation law (7.23) for the polarisation vector of a massless spin-1 particles requires the existence of the gauge symmetry (7.24). The gauge symmetry in turn implies that the massless spin-1 particle couples only to conserved currents.

We can exploit the transformation law $\varepsilon'_\mu = \varepsilon_\mu + \lambda k_\mu$ as follows: Since the dependence on k_μ is nonphysical, any Feynman amplitude $\mathcal{A} = \varepsilon_\mu A^\mu$ has to vanish, if we replace the polarisation vector ε_μ of an external photon by its four-momentum, $k_\mu A^\mu = 0$. This quantum analogue of classical current conservation $k_\mu J^\mu(k) = 0$ is called "Ward identity." As an example for its application, we derive a convenient expression for the propagator of a massless vector particle. The two polarisation vectors of a photon should satisfy the normalisation $\varepsilon_\mu^{(a)*}\varepsilon^{\mu(b)} = \delta^{ab}$. For a linearly polarised photon propagating in z direction, $k^\mu = (\omega, 0, 0, \omega)$, the polarisation vectors are $\varepsilon_\mu^{(1)} = \delta_\mu^1$ and $\varepsilon_\mu^{(2)} = \delta_\mu^2$. If we perform the sum over the two polarisation states, we find

$$\sum_r \varepsilon_\mu^{(r)*}\varepsilon_\nu^{(r)} = \mathrm{d}iag\{0,1,1,0\}\,. \tag{7.27}$$

If we try to rewrite this expression in an invariant way using $\eta_{\mu\nu}$ and $k_\mu k_\nu/k^2$, we fail: We cannot cancel at the same time $\eta_{00} = +1$ and $\eta_{33} = -1$ by $k_\mu k_\nu/k^2$. We introduce therefore additionally the momentum vector $\tilde{k}^\mu = (\omega, 0, 0, -\omega)$ obtained by a spatial reflection from k^μ. This allows us to write the polarisation sum as an invariant tensor expression,

$$\sum_r \varepsilon_\mu^{(r)*}\varepsilon_\nu^{(r)} = -\eta_{\mu\nu} + \frac{k_\mu\tilde{k}_\nu + \tilde{k}_\mu k_\nu}{k\tilde{k}} \equiv -\pi_{\mu\nu}\,. \tag{7.28}$$

Current conservation, $k_\mu J^\mu(k) = 0$, implies that the second term in the polarisation sum does not contribute to physical observables. For the same reason, we can add an arbitrary term $\xi k_\mu k_\nu$. We use this freedom to eliminate the $\tilde{k}$ dependence and to set

$$J^{\mu*}\left(\sum_r \varepsilon_\mu^{(r)*}\varepsilon_\nu^{(r)}\right) J^\nu = J^{\mu*}\left(-\eta_{\mu\nu} + (1-\xi)\frac{k_\mu k_\nu}{k^2}\right) J^\nu\,. \tag{7.29}$$

Now we can read off the photon propagator as

$$D_F^{\mu\nu}(k) = \frac{-\eta^{\mu\nu} + (1-\xi)k^\mu k^\nu/k^2}{k^2 + \mathrm{i}\varepsilon}\,. \tag{7.30}$$

A specific choice of the parameter ξ called gauge fixing parameter corresponds to the choice of a gauge in Eq. (7.13). In particular, the Feynman gauge $\xi = 1$, which leads to a form of the propagator often most convenient in calculations, correspond to the Lorenz gauge in Eq. (7.13). In this gauge, $\sum_{r=0}^3 \varepsilon_\mu^{(r)*}\varepsilon_\nu^{(r)} = \mathrm{d}iag\{-1,1,1,1\}$: the propagator contains nonphysical degrees of freedom, time-like and longitudinal photons, which contributions cancel however in physical observables. Similarly, for all other values ξ the propagator is explicitly Lorentz invariant but contains unphysical degrees of freedom. We will see later that it is a general feature of gauge theories as electrodynamics that we have to choose between a covariant gauge which introduces unphysical degrees of freedom and a gauge which contains only the transverse degree of freedom but selects a specific frame.

Repulsive Coulomb potential by vector exchange We consider as in the scalar case two static point charges as external sources, but use now a vector current $J^\mu = J_1^\mu(x_1) + J_2^\mu(x_2)$. Since $J^\mu = (\rho, \boldsymbol{j})$, only the zero component, $J_i^\mu = \delta_0^\mu \delta(\boldsymbol{x} - \boldsymbol{x}_i)$, contributes for a static source to $W[J]$. Moreover, we can neglect the longitudinal part $k^\mu k^\nu / m^2$ of the propagator. This is justified, since the concept of a potential energy makes only sense in the non-relativistic limit, i.e. for $V \ll m$ or equivalently $r \gg 1/m$. Hence

$$\begin{aligned} W_{12}[J] &= -\frac{1}{2} \int \mathrm{d}^4x \mathrm{d}^4x' \int \frac{\mathrm{d}^4k}{(2\pi)^4} J_1^\mu(x) \, \frac{-\eta_{\mu\nu} \, \mathrm{e}^{-\mathrm{i}k(x-x')}}{k^2 - m^2 + \mathrm{i}\varepsilon} \, J_2^\nu(x') && (7.31) \\ &= \frac{1}{2} \int \mathrm{d}t\mathrm{d}t' \int \frac{\mathrm{d}^4k}{(2\pi)^4} \frac{\mathrm{e}^{-\mathrm{i}k(x-x')}}{k^2 - m^2 + \mathrm{i}\varepsilon} \, . && (7.32) \end{aligned}$$

Comparing with our earlier result for scalar exchange in Eq. (3.37), it becomes clear without further calculation that spin-1 exchange between equal charges is repulsive. In the limit $m \to 0$, we obtain the Coulomb potential with the correct sign for electromagnetic interactions.

7.3. Gravity

Wave equation From Newton's law we know that gravity is fundamentally attractive and of long range. Thus the gravitational force has to be mediated by a massless particle which can not be a spin $s = 1$ particle. Analog to the electric field $\boldsymbol{E} = -\boldsymbol{\nabla}\phi$ we can introduce a classical gravitational field $\boldsymbol{g}$ as the gradient of the gravitational potential, $\boldsymbol{g} = -\boldsymbol{\nabla}\phi$. We obtain then $\boldsymbol{\nabla} \cdot \boldsymbol{g}(\boldsymbol{x}) = -4\pi G \rho(\boldsymbol{x})$ and as Poisson equation

$$\Delta\phi(\boldsymbol{x}) = 4\pi G \rho(\boldsymbol{x}) \, , \tag{7.33}$$

where ρ is the mass density, $\rho = \mathrm{d}m/\mathrm{d}^3x$.

Special relativity gives us two hints how we should transfer this equation into a relativistic framework: First, the Laplace operator Δ on the LHS is the $c \to \infty$ limit of minus the d'Alembert operator $\Box$. Second, the RHS should be the $v/c \to 0$ limit of something incorporating not only the mass density but all types of energy densities. To proceed, consider first how the mass density ρ transforms under a Lorentz transformation: An observer moving with the speed β relative to the rest frame of the matter distribution ρ measures the energy density $\rho' = \gamma \mathrm{d}m/(\gamma^{-1}\mathrm{d}V) = \gamma^2\rho$, with $\gamma = 1/\sqrt{1-\beta^2}$. This is the transformation law of the 00 component of a tensor of rank two and ρ as 00 component, alas the energy-momentum stress tensor $T^{\mu\nu}$.

Thus the field equation for a purely scalar theory of gravity would be

$$\Box\phi = -4\pi G T_\mu^\mu. \tag{7.34}$$

Such a theory predicts no coupling between photons and gravitation, because the trace of the stress tensor of the electromagnetic field vanishes, $T_\mu^\mu = 0$, and is therefore in contradiction to the observed gravitational lensing of light. A purely vector theory for gravity fails too, since it predicts not attraction but repulsion of two masses. Hence we are forced to consider a symmetric spin-2 field $\bar{h}_{\mu\nu}$ as mediator of the gravitational force; its source is the energy-momentum stress tensor

$$\Box\bar{h}^{\mu\nu} = -2\kappa T^{\mu\nu}. \tag{7.35}$$

The normalisation constant $\kappa \propto G_N$ has to be determined such that in the non-relativistic limit the Poisson equation (7.33) holds.

Let us consider as a warm-up first the case of a massive particle: A symmetric, massive spin-2 field has ten independent components, but only $2s+1=5$ physical spin degrees of freedom. Thus we have to impose five constraints additional to the source-free equation

$$(\Box + m^2)\bar{h}^{\mu\nu} = 0\,. \tag{7.36}$$

Proceeding as in the $s=1$ case, (7.8), we use as constraint $\partial_\mu \bar{h}^{\mu\nu} = 0$ which provides now four conditions. We can use them to set $\bar{h}^{0\mu} = 0$. We obtain the missing fifth constraint subtracting the trace $\bar{h}^{ii}$ which transforms as a scalar from $\bar{h}^{ij}$.

We move now to the massless case considering a plane wave $\bar{h}_{\mu\nu} = \varepsilon_{\mu\nu}\exp(-\mathrm{i}kx)$. In analogy to the photon case, we expect that also the graviton has only two, transverse degrees of freedom. If we choose the plane wave propagating in the z direction, $\boldsymbol{k} = k\boldsymbol{e}_z$, then we expect that the polarisation tensor can be expressed as

$$\varepsilon^{\mu\nu} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & \varepsilon_{11} & \varepsilon_{12} & 0 \\ 0 & \varepsilon_{12} & -\varepsilon_{11} & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}. \tag{7.37}$$

Here we used that the polarisation tensor has to be symmetric and traceless. The choice (7.37) is called the *transverse traceless* (TT) gauge.

Metric perturbations as a tensor field In the case of the photon, we could reduce the degrees of freedom from four to two, because of the redundancy implied by the gauge symmetry of electromagnetism. Moreover, the gauge symmetry lead to the conservation of the electromagnetic current. The two obvious questions to address next are which symmetry and which conservation law are connected to gravitation.

The second question is the simpler one, since we know already that in flat space $\partial_\mu T^{\mu\nu} = 0$ holds. Thus for gravity energy-momentum conservation will play the role of current conservation, implying that a gravitational wave is transverse, $k_\mu T^{\mu\nu}(k) = 0$. In order to answer the first question, we have to consider the properties of $h^{\mu\nu}$. The equivalence principle implies that all test-particles move along the same world-line, if they are released at the same initial point and move only under the influence of the gravitational force. This universality motivated Einstein to describe the effect of gravity by the curvature of space-time. We associate therefore the symmetric tensor field[2] $h_{\mu\nu}$ with small perturbations around the Minkowski metric $\eta_{\mu\nu}$,

$$g_{\mu\nu} = \eta_{\mu\nu} + \varepsilon h_{\mu\nu}\,, \qquad \varepsilon \ll 1\,. \tag{7.38}$$

We choose a Cartesian coordinate system x^μ and ask ourselves which transformations are compatible with the splitting (7.38) of the metric. If we consider global Lorentz transformations $\Lambda^\nu{}_\mu$, then $x'^\nu = \Lambda^\nu{}_\mu x^\mu$, and the metric tensor transforms as

$$g'_{\alpha\beta} = \Lambda^\rho{}_\alpha \Lambda^\sigma{}_\beta g_{\rho\sigma} = \Lambda^\rho{}_\alpha \Lambda^\sigma{}_\beta(\eta_{\rho\sigma} + h_{\rho\sigma}) = \eta_{\alpha\beta} + \Lambda^\rho{}_\alpha \Lambda^\sigma{}_\beta h_{\rho\sigma} = \eta'_{\alpha\beta} + \Lambda^\rho{}_\alpha \Lambda^\sigma{}_\beta h_{\rho\sigma}\,. \tag{7.39}$$

[2] We drop the bar, anticipating that $\bar{h}^{\mu\nu}$ may differ from $h^{\mu\nu}$ in Eq. (7.35). We will derive their relation, $\bar{h}_{\mu\nu} \equiv h_{\mu\nu} - \frac{1}{2}\eta_{\mu\nu}h^\alpha_\alpha$, in section 18.3.

Since $h'_{\alpha\beta} = \Lambda^\rho{}_\alpha \Lambda^\sigma{}_\beta h_{\rho\sigma}$, we see that global Lorentz transformations respect the splitting (7.38). Thus $h_{\mu\nu}$ transforms as a rank-2 tensor under global Lorentz transformations. We can view therefore the perturbation $h_{\mu\nu}$ as a symmetric rank-2 tensor field defined on Minkowski space that satisfies the wave equation (7.35), similar as the photon field is a rank-1 tensor field fulfilling Maxwell's equations.

The splitting (7.38) is however clearly not invariant under general coordinate transformations, as they allow e.g. the finite rescaling $g_{\mu\nu} \to \Omega g_{\mu\nu}$. We restrict therefore ourselves to infinitesimal coordinate transformations,

$$\bar{x}^\mu = x^\mu + \varepsilon\xi^\mu(x^\nu) \tag{7.40}$$

with $\varepsilon \ll 1$. Then the Killing equation (6.72) simplifies to

$$h'_{\mu\nu} = h_{\mu\nu} + \partial_\mu \xi_\nu + \partial_\nu \xi_\mu \,, \tag{7.41}$$

because the term $\xi^\rho \partial_\rho h_{\mu\nu}$ is quadratic in the small quantities $\varepsilon h_{\mu\nu}$ and $\varepsilon\xi_\mu$ and can be neglected. Recall that the $\xi^\rho \partial_\rho h_{\mu\nu}$ term appeared, because we compared the metric tensor at different points. In its absence, it is more fruitful to view Eq. (7.41) not as a coordinate but as a gauge transformation analogous to (7.12): In this interpretation, we stay in Minkowski space and the fields $h'_{\mu\nu}$ and $h_{\mu\nu}$ describe the same physics, since the gravitational field equations do not fix uniquely $h_{\mu\nu}$ for a given source. In momentum space, Eq. (7.41) specifies how the polarisation tensor transforms under gauge transformations,

$$\varepsilon'_{\mu\nu} = \varepsilon_{\mu\nu} + \xi_\mu k_\nu + \xi_\nu k_\mu . \tag{7.42}$$

We can use this gauge freedom to eliminate four components of $\varepsilon_{\mu\nu}$. After that, we can perform another gauge transformation (7.42) using any four functions χ_μ satisfying the wave equation $\Box\chi_\mu = 0$, eliminating thereby four additional components. This justifies the use of the TT gauge.

Graviton propagator We follow the same approach as in the derivation of the photon propagator. For a graviton propagating in z direction, $k^\mu = (\omega, 0, 0, \omega)$, we choose as the two polarisation states $\varepsilon^{(1)}_{\mu\nu}$ setting $\varepsilon_{11} = 1/\sqrt{2}$ and $\varepsilon_{12} = 0$ and $\varepsilon^{(2)}_{\mu\nu}$ setting $\varepsilon_{11} = 0$ and $\varepsilon_{12} = 1/\sqrt{2}$, respectively. They satisfy the normalisation $\varepsilon^{(a)}_{\mu\nu}\varepsilon^{\mu\nu(b)} = \delta^{ab}$. Now we should perform the sum over the two polarisation states, $\sum_r \varepsilon^{(r)}_{\mu\nu}\varepsilon^{(r)}_{\rho\sigma}$, and express the result as a linear combination of $\eta_{\mu\nu}$ and $k_\mu \tilde{k}_\nu + \tilde{k}_\mu k_\nu$. A straightforward way to do this is to combine first the ten independent quantities of the symmetric tensors in ten-dimensional vectors, $\varepsilon_{\mu\nu} \to E_a$, $\eta_{\mu\nu} \to N_a$, and $k_\mu \tilde{k}_\nu + \tilde{k}_\mu k_\nu \to K_a$, to calculate the tensor products of these vectors and to compare then the resulting $10 \otimes 10$ matrices. An alternative, shorter way is to use the requirement that the propagator is transverse in all indices, $k^\mu \sum_r \varepsilon^{(r)}_{\mu\nu}\varepsilon^{(r)}_{\rho\sigma} = \ldots = k^\sigma \sum_r \varepsilon^{(r)}_{\mu\nu}\varepsilon^{(r)}_{\rho\sigma} = 0$, because of energy-momentum conservation, $\partial_\mu T^{\mu\nu}(x) = 0$. This implies that the graviton propagator should be composed of the projection operators $\pi_{\mu\nu}$ used for the photon (cf. Eq. (7.28)) as follows

$$\sum_r \varepsilon^{(r)}_{\mu\nu}\varepsilon^{(r)}_{\rho\sigma} = A\,\pi_{\mu\nu}\pi_{\rho\sigma} + B\left[\pi_{\mu\rho}\pi_{\nu\sigma} + \pi_{\mu\sigma}\pi_{\nu\rho}\right] . \tag{7.43}$$

The last two terms have a common coefficient, since the LHS is invariant under exchanges of $\mu \leftrightarrow \nu$ or $\rho \leftrightarrow \sigma$. We fix A and B by evaluating this expression for two sets of indices.

Since the only non-zero elements of $\pi_{\mu\nu}$ are $\pi_{11} = \pi_{22} = -1$, we obtain choosing e.g. $\{1212\}$ as indices

$$\sum_r \varepsilon_{12}^{(r)} \varepsilon_{12}^{(r)} = \frac{1}{2} = B\pi_{11}\pi_{22}$$

and thus $B = 1/2$. Similarly, it follows $A = -1/2$ choosing e.g. as indices $\{1111\}$. Thus we found—with surprising ease—the polarisation sum required for the graviton propagator,

$$\sum_r \varepsilon_{\mu\nu}^{(r)} \varepsilon_{\rho\sigma}^{(r)} = -\frac{1}{2}\pi_{\mu\nu}\pi_{\rho\sigma} + \frac{1}{2}\pi_{\mu\rho}\pi_{\nu\sigma} + \frac{1}{2}\pi_{\mu\sigma}\pi_{\nu\rho}\,. \tag{7.44}$$

We continue to proceed in the same way as for the photon. Energy-momentum conservation, $k_\mu T^{\mu\nu}(k) = 0$, implies that the $k_\mu \tilde{k}_\nu + \tilde{k}_\mu k_\nu$ term in $\lambda_{\mu\nu}$ does not contribute to physical observables. We drop therefore again all terms proportional to the graviton momentum k_μ,

$$T^{\mu\nu *}\left(\sum_r \varepsilon_{\mu\nu}^{(r)*} \varepsilon_{\rho\sigma}^{(r)}\right) T^{\rho\sigma} = T^{\mu\nu *}\left(-\frac{1}{2}\eta_{\mu\nu}\eta_{\rho\sigma} + \frac{1}{2}\eta_{\mu\rho}\eta_{\nu\sigma} + \frac{1}{2}\eta_{\mu\sigma}\eta_{\nu\rho}\right) T^{\rho\sigma}. \tag{7.45}$$

Thus the graviton propagator in the Feynman gauge is given by

$$D_F^{\mu\nu;\rho\sigma}(k) = \frac{\frac{1}{2}(-\eta^{\mu\nu}\eta^{\rho\sigma} + \eta^{\mu\rho}\eta^{\nu\sigma} + \eta^{\mu\sigma}\eta^{\nu\rho})}{k^2 + \mathrm{i}\varepsilon}\,. \tag{7.46}$$

Other gauges are obtained by the replacement $\eta^{\mu\nu} \to \eta^{\mu\nu} - (1-\xi)k^\mu k^\nu/k^2$. In the case of gravity, the Feynman gauge $\xi = 1$ is most often called harmonic gauge, but also the names Hilbert, Loren(t)z, de Donder and confusingly many others are in use.

Attractive potential by spin-2 exchange We consider now the potential energy created by two point masses as external sources interacting via a tensor current $T^{\mu\nu} = T_1^{\mu\nu}(x_1) + T_2^{\mu\nu}(x_2)$. Specialising to static sources, only the zero-zero component, $T_i^{\mu\nu} = \delta_0^\mu \delta_0^\nu \delta(\boldsymbol{x} - \boldsymbol{x}_i)$, contributes to $W[J]$. Hence

$$W_{12}[J] = -\frac{1}{2}\int \mathrm{d}^4x \mathrm{d}^4x' \int \frac{\mathrm{d}^4k}{(2\pi)^4} T_1^{00}(x)\, D_{F\,00;00}(k)\, \mathrm{e}^{-\mathrm{i}k(x-x')}\, T_2^{00}(x')\,. \tag{7.47}$$

Looking at the numerator of the graviton propagator, we find $-1 + 1 + 1 = 1 > 0$. Thus spin-2 exchange is attractive, as required for the force mediating gravity. Comparing Eq. (7.47) to Newton's gravitational potential, we see that the graviton couples with the strength $(8\pi G)^{1/2} \equiv \kappa^{1/2}$ to the stress tensor $T^{\mu\nu}$ of matter, cf. problem 7.**??**.

Helicity We determine now how a metric perturbation $h_{\mu\nu}$ transforms under a rotation with the angle α. We choose the wave propagating in z direction, $\boldsymbol{k} = k\boldsymbol{e}_z$, the TT gauge, and the rotation in the xy plane. Then the general Lorentz transformation $\Lambda_\mu^{\ \nu}$ becomes

$$\Lambda_\mu^{\ \nu} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & \cos\alpha & \sin\alpha & 0 \\ 0 & -\sin\alpha & \cos\alpha & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}. \tag{7.48}$$

Since $\boldsymbol{k} = k\boldsymbol{e}_z$ and thus $\Lambda_\mu^{\ \nu} k_\nu = k_\mu$, the rotation affects only the polarisation tensor. We rewrite $\varepsilon'_{\mu\nu} = \Lambda_\mu^{\ \rho}\Lambda_\nu^{\ \sigma}\varepsilon_{\rho\sigma}$ in matrix notation, $\varepsilon' = \Lambda\varepsilon\Lambda^T$. It is sufficient to perform the calculation for the xy sub-matrices. The result after introducing circular polarisation states $\varepsilon_\pm = \varepsilon_{11} \pm \mathrm{i}\varepsilon_{12}$ is

$$\varepsilon'^{\mu\nu}_{\pm} = \exp(\mp 2\mathrm{i}\alpha)\varepsilon^{\mu\nu}_{\pm} \,. \tag{7.49}$$

The same calculation for a circularly polarised photon gives $\varepsilon'^{\mu}_{\pm} = \exp(\mp\mathrm{i}\alpha)\varepsilon^{\mu}_{\pm}$. Any plane wave ψ which is transformed into $\psi' = \mathrm{e}^{-\mathrm{i}h\alpha}\psi$ by a rotation of an angle α around its propagation axis is said to have helicity h. Thus if we say that a photon has spin 1 and a graviton has spin 2, we mean more precisely that electromagnetic and gravitational plan waves have helicity one and two, respectively. Doing the same calculation in an arbitrary gauge, one finds that the remaining, unphysical degrees of freedom transform as helicity one and zero (problem 7.??). In general, a tensor field of rank (n,m) contains states with helicity $h = 0,\ldots,n+m$. Thus we can rephrase the statement that tensor fields follow Bose-Einstein statistics as fields with integer helicity (or spin) are bosons.

7.4. Source of gravity

The dynamical energy-momentum tensor If we compare the wave equation for a photon and a graviton, then there is an important difference: The former is in the classical limit exact. The photon carries no charge and does not contribute to its source term. As a result, the wave equation is linear. In contrast, a gravitational wave carries energy-momentum and acts thus as its own source. The LHS of (7.35) should be therefore the limit of a more complicated equation, which we write symbolically as $G_{\mu\nu} = -\kappa T_{\mu\nu}$. The tensor $G_{\mu\nu}$ should be given as the variation of an appropriate action of gravity, called the Einstein-Hilbert action $S_{\mathrm{E}H}$, with respect to the metric tensor $g_{\mu\nu}$. Even without knowing the action $S_{\mathrm{E}H}$, we can derive an important conclusion: If the total action is the sum of $S_{\mathrm{E}H}$ and the action S_{m} including all relevant matter fields,

$$S = \frac{1}{2\kappa}\, S_{\mathrm{E}H} + S_{\mathrm{m}}\,,$$

then the variation of the matter action S_{m} should give the stress tensor as the source of the gravitational field,

$$\frac{2}{\sqrt{|g|}}\,\frac{\delta S_{\mathrm{m}}}{\delta g_{\mu\nu}} = -T^{\mu\nu}\,. \tag{7.50}$$

Here we included a factor $\sqrt{|g|}$ because $T^{\mu\nu}$ is a density, while the factor 2 is required to obtain agreement with the usual definition of $T^{\mu\nu}$. Since the presence of gravity implies a curved space-time, the replacements $\{\partial_\mu, \eta_{\mu\nu}, \mathrm{d}^4x\} \to \{\nabla_\mu, g_{\mu\nu}, \mathrm{d}^4x\sqrt{|g|}\}$ have to performed in S_{m} before the variation is performed. The tensor $T^{\mu\nu}$ defined by this equation is called *dynamical* energy-momentum stress tensor. In order to show that this definition makes sense, we have to prove that the tensor is locally conserved, $\nabla_\mu T^{\mu\nu} = 0$, and we have to convince ourselves that this definition reproduces the standard results we know already.

Conservation of the stress tensor We start by proving that the dynamical energy-momentum tensor defined by Eq. (7.50) is locally conserved. We consider the change of the matter action under a variation of the metric[3],

$$\delta S_{\rm m} = -\frac{1}{2}\int_\Omega \mathrm{d}^4x\sqrt{|g|}\,T^{\mu\nu}\delta g_{\mu\nu} = \frac{1}{2}\int_\Omega \mathrm{d}^4x\sqrt{|g|}\,T_{\mu\nu}\delta g^{\mu\nu}\,. \tag{7.51}$$

We allow infinitesimal but otherwise arbitrary coordinate transformations,

$$\bar{x}^\mu = x^\mu + \xi^\mu(x)\,. \tag{7.52}$$

For the resulting change in the metric $\delta g_{\mu\nu}$ we can use Eqs. (6.68b) and (6.69),

$$\delta g_{\mu\nu} = \nabla_\mu \xi_\nu + \nabla_\nu \xi_\mu\,. \tag{7.53}$$

We use that $T^{\mu\nu}$ is symmetric and that general covariance guarantees that $\delta S_m = 0$ for a coordinate transformation,

$$\delta S_{\rm m} = -\int_\Omega \mathrm{d}^4x\sqrt{|g|}\,T^{\mu\nu}\nabla_\mu\xi_\nu = 0\,. \tag{7.54}$$

Next we apply the product rule,

$$\delta S_{\rm m} = -\int_\Omega \mathrm{d}^4x\sqrt{|g|}\,(\nabla_\mu T^{\mu\nu})\xi_\nu + \int_\Omega \mathrm{d}^4x\sqrt{|g|}\,\nabla_\mu(T^{\mu\nu}\xi_\nu) = 0\,. \tag{7.55}$$

The second term is a four-divergence and thus a boundary term that we can neglect. The remaining first term vanishes for arbitrary ξ_μ only, if the energy-momentum tensor is conserved,

$$\nabla_\mu T^{\mu\nu} = 0\,. \tag{7.56}$$

Hence the local conservation of energy-momentum is a consequence of the general covariance of the gravitational field equations, in the same way as current conservation follows from gauge invariance in electromagnetism. You should convince yourself that the dynamical energy-momentum stress tensor evaluated for the examples of the Klein-Gordon and the Maxwell field agrees with the symmetrised canonical stress tensor, cf. problem 7.??.

7.A. Appendix: Large extra dimensions and massive gravity

Large extra dimensions As mentioned in chapter 5, quantum corrections break the conformal invariance of string theory except we live in a world with $d = 10$ or 26 space-time dimensions. There are two obvious answers to this result: First, one may conclude that string theory is disproven by nature or, second, one may adjust reality. Consistency of the second approach with experimental data could be achieved, if the $d - 4$ dimensions are compactified with a sufficiently small radius R, such that they are not visible in experiments sensible to wave-lengths $\lambda \gg R$.

Let us check what happens to a scalar particle with mass m, if we add a fifth compact dimension y. The Klein-Gordon equation for a scalar field $\phi(x^\mu, y)$ becomes

$$(\Box_5 + m^2)\phi(x^\mu, y) = 0 \tag{7.57}$$

[3] We should view $g_{\mu\nu}$ (and not $g^{\mu\nu}$) as "the" gravitational field: In the Lagrangian of a point particle or the line-element, the coordinates x^μ are contracted with $g_{\mu\nu}$. Having understood this point, we use simply the second relation in (7.51) in the future.

with the five-dimensional d'Alembert operator $\Box_5 = \Box - \partial_y^2$. The equation can be separated, $\phi(x^\mu, y) = \phi(x^\mu)f(y)$, and since the fifth dimension is compact, the spectrum of f is discrete. Assuming periodic boundary conditions, $f(x) = f(x+R)$, gives

$$\phi(x^\mu, y) = \phi(x^\mu)\cos(n\pi y/R)\,. \tag{7.58}$$

The energy eigenvalues of these solutions are $\omega_{\boldsymbol{k},n}^2 = \boldsymbol{k}^2 + m^2 + (n\pi/R)^2$. From a four-dimensional point of view, the term $(n\pi/R)^2$ appears as a mass term, $m_n^2 = m^2 + (n\pi/R)^2$. Since we usually consider states with different masses as different particles, we see the five-dimensional particle as a tower of particles with mass m_n but otherwise identical quantum numbers. Such theories are called Kaluza-Klein theories, and the tower of particles Kaluza-Klein particles. If $R \ll \lambda$, where λ is the length-scale experimentally probed, only the $n = 0$ particle is visible and physics appears to be four-dimensional.

Since string theory includes gravity, one often assumes that the radius R of the extra-dimensions is determined by the Planck length, $R = 1/M_{\mathrm{P}l} = (8\pi G_N)^{1/2} \sim 10^{-34}\,\mathrm{cm}$. In this case it is difficult to imagine any observational consequences of the additional dimensions. More interesting is the possibility that some of the extra dimensions are large,

$$R_{1,\ldots,\delta} \gg R_{\delta+1,\ldots,6} = 1/M_{\mathrm{P}l}\,.$$

Since the $1/r^2$ behaviour of the gravitational force is not tested below $d_* \sim$mm scales, one can imagine that large extra dimensions exists that are only visible to gravity: Relating the $d = 4$ and $d > 4$ Newton's law $F \sim \frac{m_1 m_2}{r^{2+\delta}}$ at the intermediate scale $r = R$, we can derive the "true" value of the Planck scale in this model: Matching of Newton's law in 4 and $4+\delta$ dimensions at $r = R$ gives

$$F(r=R) = G_N \frac{m_1 m_2}{R^2} = \frac{1}{M_D^{2+\delta}} \frac{m_1 m_2}{R^{2+\delta}}\,. \tag{7.59}$$

This equation relates the size R of the large extra dimensions to the true fundamental scale M_D of gravity in this model,

$$G_N^{-1} = 8\pi M_{\mathrm{P}l}^2 = R^\delta M_D^{\delta+2}\,, \tag{7.60}$$

while Newton's constant G_N becomes just an auxiliary quantity useful to describe physics at $r \gtrsim R$. (You may compare this to the case of weak interactions where Fermi's constant $G_F \propto g^2/m_W^2$ is determined by the weak coupling constant g and the mass m_W of the W-boson).

Next we ask, if $M_D \sim$ TeV is possible, i.e. if one may test such theories at accelerators as LHC? Inserting the measured value of G_N and $M_D = 1\,\mathrm{TeV}$ in Eq. (7.60) we find the required value for the size R of the large extra dimension as $10^{13}\,\mathrm{cm}$ and $0.1\,\mathrm{cm}$ for $\delta = 1$ and 2, respectively, Thus the case $\delta = 1$ is excluded by the agreement of the dynamics of the solar system with 4-dimensional Newtonian physics. The cases $\delta \geq 2$ are possible, because Newton's law is experimentally tested only for scales $r \gtrsim 1\,\mathrm{mm}$.

Massive gravity Theories with extra dimensions contain often from our 4-dimensional point of view a Kaluza-Klein tower of massive gravitons. Such modified theories of gravity have found large interest since one may hope to find an alternative explanation for the accelerated expansion of the Universe.

A striking difference between the spin-2 and the spin-0 and 1 cases is that the limit $m \to 0$ of the massive spin-2 propagator and thus of the potential energy V_{12} is not smooth: In problem 7.??, you are asked to derive the massive spin-2 propagator. As result, you should find

$$D_F^{\mu\nu;\rho\sigma}(k) = \frac{1}{2} \frac{-\frac{2}{3}G^{\mu\nu}G^{\rho\sigma} + G^{\mu\rho}G^{\nu\sigma} + G^{\mu\sigma}G^{\nu\rho}}{k^2 - m^2 + \mathrm{i}\varepsilon}\,, \tag{7.61}$$

where

$$G^{\mu\nu}(k) = -\eta^{\mu\nu} + k^\mu k^\nu/m^2 \tag{7.62}$$

is the polarisation tensor for a massive spin-1 particle. Thus the nominator in the massive spin-2 propagator is as in the massless case a linear combination of the tensor products of two spin-1 polarisation tensors. However, the coefficients of the first term differ and thus the $m \to 0$ limit of the massive propagator does not agree with the massless case. In particular, the difference cannot be compensated by a rescaling of the coupling constant $\tilde{G}_N$, because it is not an overall factor: Imagine for instance that we determine the value of $\tilde{G}_N$ by calculating the potential energy of two non-relativistic sources like the Sun and the Earth. This requires in massive gravity—for an arbitrarily small graviton mass—a gravitational coupling constant $\tilde{G}_N$ a factor 3/4 smaller than in the massless case. Having fixed $\tilde{G}_N$, we can predict the deflection of light by the Sun. Since the first term in the propagator couples the traces T^μ_μ of two sources, it does not contribute to the deflection of light. As a result of the reduced coupling strength, the deflection angle of light by the Sun decreases by the same factor and any non-zero graviton mass would be in conflict with observations.

When this result was first derived in 1970, its authors explained this discontinuity by the different number of degrees of freedom in the two theories: Even if the Compton wave-length of a massive graviton is larger than the observable size of the Universe, and thus the Yukawa factor $\exp(-mr)$ indistinguishable from one, the additional spin states of a massive graviton may change physics. Two years later, Vainshtein realised that perturbation theory may break down in massive gravity and thus a calculation using one-graviton exchange is not reliable. More precisely, a theory of massive gravity contains an additional length scale $R_V = (GM/m^4)^{1/5}$ and for distances $r \ll R_V$ the theory has to be solved exactly.

Summary

Tensor fields satisfy second-order differential equations; their propagators are quadratic in p and thus even functions of x. As a result, tensor fields describe bosons, i.e. their field operators are commuting operators. Massless fields have only two, transverse degrees of freedom. A Lorentz invariant description for such fields is only possible, if the remaining number of non-physical degrees of freedom is redundant. This redundancy implies that fields connected by a gauge transformation are equivalent and describe the same physical system. In the case of photons, the gauge symmetry implies that they couple to a conserved current, in the case of gravitons that they couple to the conserved energy-momentum conservation tensor.

Further reading

[Mag07] discusses in detail (massive) gravity as a spin-2 field in Minkowski space. The history of the gauge principle is reviewed by [JO01].

8. Fermions and the Dirac equation

Up to now we have discussed fields which transform as tensors under Lorentz transformations. Such particles have integer spin or helicity. Since we can boost massive particles into their rest-frame, we can use our knowledge of non-relativistic quantum mechanics to anticipate that additional representations of the rotation group SO(3) and thus also of the Lorentz group SO(1, 3) exist which correspond to particles with half-integer spin. Such particles are described by anti-commuting variables what is the fundamental reason for the Pauli principle and thus the stability of matter.

8.1. Spinor representation of the Lorentz group

In order to introduce spinors we have to find the corresponding representation of the Lorentz group. As always it is simpler to work at linear order, which is in this case the Lie algebra. The Lie algebra of the Poincaré group[1] contains ten generators, the three generators $\boldsymbol{J}$ of rotations, the three generators $\boldsymbol{K}$ of Lorentz boosts and the four generators $\boldsymbol{T}$ of translations. The Killing vector fields $\boldsymbol{V}$ of Minkowski space generate these symmetries, and therefore the generators are given by the Killing vector fields. Thus we can use Eqs. (6.79) and (6.80) to calculate their commutation relations as (problem 8.??)

$$[J_i, J_j] = \mathrm{i}\varepsilon_{ijk} J_k\,, \tag{8.1a}$$

$$[J_i, K_j] = \mathrm{i}\varepsilon_{ijk} K_k\,, \tag{8.1b}$$

$$[K_i, K_j] = -\mathrm{i}\varepsilon_{ijk} J_k\,. \tag{8.1c}$$

Here we followed physicist's convention and identified i$\boldsymbol{V}$ as the generators, so that they are Hermitian. Moreover, we restrict our attention to the Lorentz group which is sufficient to derive the concept of a Weyl spinor. Note that the algebra of the boost generators $\boldsymbol{K}$ is not closed. Thus in contrast to rotations, boosts do not form a subgroup of the Lorentz group.

The structure constants in all these commutation relations are $\pm\varepsilon_{ijk}$, suggesting that we can rewrite the Lorentz group as a product of two SU(2) factors. We try to decouple the two sets of generators $\boldsymbol{J}$ and $\boldsymbol{K}$ by introducing two non-Hermitian ladder operators

$$\boldsymbol{J}^{\pm} = \frac{1}{2}(\boldsymbol{J} \pm \mathrm{i}\boldsymbol{K})\,. \tag{8.2}$$

Their commutations relations are

$$[J_i^+, J_j^+] = \mathrm{i}\varepsilon_{ijk} J_k^+\,, \tag{8.3a}$$

$$[J_i^-, J_j^-] = \mathrm{i}\varepsilon_{ijk} J_k^-\,, \tag{8.3b}$$

$$[J_i^+, J_j^-] = 0 \qquad i, j = 1, 2, 3\,. \tag{8.3c}$$

[1] See the Appendices B.3 and B.4 for a brief review of the Poincaré group.

Thus $\boldsymbol{J}^-$ and $\boldsymbol{J}^+$ commute with each other and generate each a SU(2) group. The Lorentz group is[2] therefore $\sim SU(2) \otimes SU(2)$, and states transforming in a well-defined way are labelled by a pair of angular momenta, (j^-, j^+), corresponding to the eigenvalues of J_z^- and J_z^+, respectively. From our knowledge of the angular momentum algebra in nonrelativistic quantum mechanics, we conclude that the dimension of the representation (j^-, j^+) is $(2j^- + 1)(2j^+ + 1)$. Because of $\boldsymbol{J} = \boldsymbol{J}^- + \boldsymbol{J}^+$, the representation (j^-, j^+) contains all possible spins j in integer steps from $|j^- - j^+|$ to $j^- + j^+$.

The representation $(0, 0)$ has dimension one, transforms trivially, $\boldsymbol{J} = \boldsymbol{K} = 0$, and corresponds therefore to the scalar representation. The two smallest non-trivial representations are $\boldsymbol{J}^+ = 0$, i.e. $(j^-, 0)$ with $\boldsymbol{J}^{(1/2)} = -\mathrm{i}\boldsymbol{K}^{(1/2)}$, and $\boldsymbol{J}^- = 0$, i.e. $(0, j^+)$ with $\boldsymbol{J}^{(1/2)} = \mathrm{i}\boldsymbol{K}^{(1/2)}$. Both representations have spin 1/2 and dimension two. We define therefore two types of two-component spinors,

$$\phi_L : \quad (1/2, 0), \quad \boldsymbol{J}^{(1/2)} = \boldsymbol{\sigma}/2, \quad \boldsymbol{K}^{(1/2)} = +\mathrm{i}\boldsymbol{\sigma}/2\,, \tag{8.4a}$$

$$\phi_R : \quad (0, 1/2), \quad \boldsymbol{J}^{(1/2)} = \boldsymbol{\sigma}/2, \quad \boldsymbol{K}^{(1/2)} = -\mathrm{i}\boldsymbol{\sigma}/2\,, \tag{8.4b}$$

which we call left-chiral and right-chiral Weyl spinors. These Weyl spinors form the fundamental representation of the Lorentz group: All higher spin states can be obtained as tensor products involving them. Their transformation properties under an (active) finite Lorentz transformation with parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\eta}$ follow by exponentiating their generators as $\exp(-\mathrm{i}\boldsymbol{J}\boldsymbol{\alpha} + \mathrm{i}\boldsymbol{K}\boldsymbol{\eta})$ (compare to appendix B.3 for our choice of signs),

$$\phi_L \to \phi_L' = \exp\left[-\frac{\mathrm{i}\boldsymbol{\sigma}\boldsymbol{\alpha}}{2} - \frac{\boldsymbol{\sigma}\boldsymbol{\eta}}{2}\right]\phi_L \equiv S_L \phi_L\,, \tag{8.5a}$$

$$\phi_R \to \phi_R' = \exp\left[-\frac{\mathrm{i}\boldsymbol{\sigma}\boldsymbol{\alpha}}{2} + \frac{\boldsymbol{\sigma}\boldsymbol{\eta}}{2}\right]\phi_R \equiv S_R \phi_R\,. \tag{8.5b}$$

While the transformation matrices S_L and S_R agree for rotations, the terms describing Lorentz boosts have opposite signs. Note also that only rotations are described by a unitary transformation, while Lorentz boosts lead to a non-unitary transformation of the Weyl spinors.

We ask now if we can convert a left- into a right-chiral spinor and vice versa. Thus we should find a spinor $\tilde{\phi}_L$ constructed out of ϕ_L which transforms as $S_R\tilde{\phi}_L$. Changing S_L into S_R requires reversing the relative sign between the rotation and the boost term, which we achieve by complex conjugating ϕ_L,

$$\phi_L^{*\prime} = \left[1 + \frac{\mathrm{i}\boldsymbol{\sigma}^*\boldsymbol{\alpha}}{2} - \frac{\boldsymbol{\sigma}^*\boldsymbol{\eta}}{2} + \ldots\right]\phi_L^*\,. \tag{8.6}$$

Because of $\sigma_1^* = \sigma_1$, $\sigma_2^* = -\sigma_2$, $\sigma_3^* = \sigma_3$, and $\sigma_1\sigma_2 = -\sigma_2\sigma_1$, $\sigma_2\sigma_3 = -\sigma_3\sigma_2$, we obtain the desired transformation property multiplying $\phi_L^{*\prime}$ with σ_2,

$$\sigma_2\phi_L^{*\prime} = \sigma_2\left[1 + \frac{\mathrm{i}(\sigma_1, -\sigma_2, \sigma_3)\boldsymbol{\alpha}}{2} - \frac{(\sigma_1, -\sigma_2, \sigma_3)\boldsymbol{\eta}}{2} + \ldots\right]\phi_L^* \tag{8.7a}$$

$$= \left[1 - \frac{\mathrm{i}\boldsymbol{\sigma}\boldsymbol{\alpha}}{2} + \frac{\boldsymbol{\sigma}\boldsymbol{\eta}}{2} + \ldots\right]\sigma_2\phi_L^* = S_R\sigma_2\phi_L^*\,. \tag{8.7b}$$

[2] More precisely, they have the same Lie algebra and are thus locally isomorphic but differ globally.

Thus ϕ_L and ϕ_R are connected by a non-unitary transformation, and therefore ϕ_L and ϕ_R describe different physics. Obviously, we can add to $\sigma_2\phi_L^*$ an arbitrary phase $\mathrm{e}^{\mathrm{i}\delta}$ without changing the transformation properties. We define

$$\phi_R^c \equiv -\mathrm{i}\sigma_2\phi_L^* \quad and \quad \phi_L^c \equiv \mathrm{i}\sigma_2\phi_R^*\,, \,, \tag{8.8}$$

which ensures $(\phi_L^c)^c = \phi_L$ and $(\phi_R^c)^c = \phi_R$. When we discuss later the coupling of a fermion to an external field, we will see that ϕ_L^c is the charge conjugated spinor of ϕ_L.

For the construction of a Lagrangian we need for the mass term scalars and for the kinetic energy vectors built out of the Weyl fields. Both the kinetic and the mass terms should be real to provide a real Lagrangian. In contrast to the real Lorentz transformation $\Lambda^\mu{}_\nu$ acting on tensor fields, the matrices $S_{L/R}$ are however complex and thus the Weyl fields are complex too. This suggests together with the fact that a measurement device should be the same after a rotation by 2π that observables are bilinear quantities in the fermion fields, such that they transform tensorial and their eigenvalues are real.

Out of the two Weyl spinors, we can form four different products $\phi_{L/R}^\dagger\phi_{L/R}$ leading to the combinations $S_L^\dagger S_L$, $S_R^\dagger S_R$, $S_L^\dagger S_R$, and $S_R^\dagger S_L$. The rotation $\mathrm{i}\boldsymbol{\sigma\alpha}/2$ cancels in all four products, since it enters with the same sign in S_L and S_R, and the Pauli matrices are Hermitian, $\boldsymbol{\sigma}^\dagger = \boldsymbol{\sigma}$. By contrast, the cancellation of the boost $\boldsymbol{\sigma\eta}/2$ requires a combination of a left- and right-chiral field,

$$\phi_L'^\dagger\phi_R' = \phi_L^\dagger\left[1+\mathrm{i}\frac{\boldsymbol{\sigma\alpha}}{2}-\frac{\boldsymbol{\sigma\eta}}{2}+\ldots\right]\left[1-\mathrm{i}\frac{\boldsymbol{\sigma\alpha}}{2}+\frac{\boldsymbol{\sigma\eta}}{2}+\ldots\right]\phi_R = \phi_L^\dagger\phi_R\,, \tag{8.9}$$

and similarly for $\phi_R^\dagger\phi_L$. Thus $\phi_L^\dagger\phi_R$ and $\phi_R^\dagger\phi_L$ transform as Lorentz scalars, but not $\phi_L^\dagger\phi_L$ and $\phi_R^\dagger\phi_R$. So what are the transformation properties of the latter two products? Performing an infinitesimal boost along the z axis, we find

$$\phi_R'^\dagger\phi_R' = \phi_R^\dagger\left[1+\frac{\sigma_3\eta}{2}+\ldots\right]\left[1+\frac{\sigma_3\eta}{2}+\ldots\right]\phi_R = \phi_R^\dagger\phi_R + \eta\phi_R^\dagger\sigma_3\phi_R\,. \tag{8.10}$$

This looks like an infinitesimal Lorentz transformation of the time-like component $j^0 = \phi_R^\dagger\phi_R$ of a four-vector j^μ. If this interpretation is correct, we should be able to associate the spatial part $\boldsymbol{j}$ with $\phi_R^\dagger\boldsymbol{\sigma}\phi_R$. Checking thus how $\boldsymbol{j}$ transforms, we find using $\sigma^i\sigma^j = \delta^{ij} + \mathrm{i}\varepsilon^{ijk}\sigma^k$ that j^1 and j^2 are invariant, while j^3 transforms as

$$\phi_R'^\dagger\sigma_3\phi_R' = \phi_R^\dagger\left[1+\frac{\sigma_3\eta}{2}+\ldots\right]\sigma_3\left[1+\frac{\sigma_3\eta}{2}+\ldots\right]\phi_R = \eta\phi_R^\dagger\phi_R + \phi_R^\dagger\sigma_3\phi_R\,. \tag{8.11}$$

Thus $\phi_R^\dagger\sigma^\mu\phi_R$ with $\sigma^\mu \equiv (1,\boldsymbol{\sigma})$ transforms as a four-vector, $j^0 \to j^0 + \eta j^3$ and $j^3 \to \eta j^0 + j^3$. Performing the same calculation for the left-chiral fields reproduces the same result except for an opposite sign of η. We account for this sign change setting now $\bar\sigma^\mu \equiv (1,-\boldsymbol{\sigma})$, so that a four-vector bilinear in ϕ_L is given by $\phi_L^\dagger\bar\sigma^\mu\phi_L$.

The transformations S_L and S_R that belong to the restricted Lorentz group do not mix the left- and right-chiral Weyl spinors. Consider however the effect of a parity transformation, $P\boldsymbol{x} = -\boldsymbol{x}$, on the generators $\boldsymbol{K}$ and $\boldsymbol{J}$. The velocity changes sign, $\boldsymbol{v} \to -\boldsymbol{v}$, i.e. transforms as a polar vector, while the angular momentum $\boldsymbol{J}$ as axial vector remains invariant. Thus parity interchanges $(1/2,0)$ and $(0,1/2)$ and hence ϕ_L and ϕ_R, as one would expect from a left- and right-chiral object. If parity is a symmetry of the theory examined, one can therefore not consider separately the two spinors ϕ_L and ϕ_R. Instead, it proves useful to combine them into a four-spinor called Dirac (or bi-spinor)

$$\psi = \begin{pmatrix} \phi_L \\ \phi_R \end{pmatrix} . \tag{8.12}$$

Another reason to consider Dirac spinors is that the scalar terms $\phi_L^\dagger \phi_R$ and $\phi_R^\dagger \phi_L$ that qualify as mass terms combine a left- and a right-chiral field: Thus the description of a particle with such a mass term seems to require the use of both left- and right-chiral Weyl spinors. Next we will derive a field equation for this type of spinor and discuss its properties.

8.2. Dirac equation

From Weyl spinors to the Dirac equation We can obtain the spinor describing a particle with momentum p by boosting the one describing a particle at rest,

$$\phi_R(p) = \exp\left[\frac{\boldsymbol{\sigma\eta}}{2}\right] \phi_R(0) = \exp\left[\frac{\eta\boldsymbol{\sigma n}}{2}\right] \phi_R(0) = [\cosh(\eta/2) + \boldsymbol{\sigma n}\sinh(\eta/2)]\,\phi_R(0) . \tag{8.13}$$

If we replace the boost parameter η by the Lorentz factor[3] $\gamma = \cosh\eta$ and use the identities $\cosh(\eta/2) = \sqrt{(\cosh\eta + 1)/2}$ and $\sinh(\eta/2) = \sqrt{(\cosh\eta - 1)/2}$, we can express the spinor as

$$\phi_R(p) = \left[\left(\frac{\gamma+1}{2}\right)^{1/2} + \boldsymbol{\sigma}\hat{\boldsymbol{p}}\left(\frac{\gamma-1}{2}\right)^{1/2}\right] \phi_R(0) . \tag{8.14}$$

Here $\hat{\boldsymbol{p}} = \boldsymbol{p}/|\boldsymbol{p}|$ is the unit vector in direction of $\boldsymbol{p}$. Inserting $\gamma = E/m$ and combining the two terms in the angular bracket, we arrive at

$$\phi_R(p) = \frac{E + m + \boldsymbol{\sigma p}}{\sqrt{2m(E+m)}}\,\phi_R(0) . \tag{8.15}$$

Similarly, we find

$$\phi_L(p) = \frac{E + m - \boldsymbol{\sigma p}}{\sqrt{2m(E+m)}}\,\phi_L(0) . \tag{8.16}$$

Thus ϕ_L and ϕ_R differ only by the sign of the operator $\boldsymbol{\sigma p}$ which measures the projection of the spin $\boldsymbol{\sigma}$ on the momentum $\boldsymbol{p}$ of the particle. For a particle at rest, this difference disappears and we set therefore $\phi_L(0) = \phi_R(0)$. This allows us to eliminate the zero momentum spinors, giving

$$\phi_L^R(p) = \frac{E \pm \boldsymbol{\sigma p}}{m}\,\phi_R^L(p) . \tag{8.17}$$

In matrix form, these two equations correspond to

$$\begin{pmatrix} -m & E - \boldsymbol{\sigma p} \\ E + \boldsymbol{\sigma p} & -m \end{pmatrix} \begin{pmatrix} \phi_L(p) \\ \phi_R(p) \end{pmatrix} = \begin{pmatrix} -m & \sigma^\mu p_\mu \\ \bar\sigma^\mu p_\mu & -m \end{pmatrix} \begin{pmatrix} \phi_L(p) \\ \phi_R(p) \end{pmatrix} = 0 . \tag{8.18}$$

We introduce the 4×4 matrices

$$\gamma^\mu = \begin{pmatrix} 0 & \sigma^\mu \\ \bar\sigma^\mu & 0 \end{pmatrix} . \tag{8.19}$$

[3] Recall the relations $E = m\cosh\eta$ and $p = m\sinh\eta$ connecting E, p, and the rapidity η.

Then we arrive with $\gamma^\mu p_\mu = \gamma^0 E - \boldsymbol{\gamma p}$ at the compact expression

$$(\gamma^\mu p_\mu - m)\,\psi(p) = 0\,. \tag{8.20}$$

Setting $p_\mu = \mathrm{i}\partial_\mu$ we obtain the Dirac equation $(\mathrm{i}\partial_\mu\gamma^\mu - m)\psi(x) = 0$ in coordinate space. The representation used for the Dirac spinor and the gamma matrices is called chiral or Weyl representation. Other representations can be obtained performing a unitary transformation, $U\tilde{\gamma}^\mu U^\dagger = \gamma^\mu$ and $U\tilde{\psi} = \psi$.

We can apply the tensor method to derive a definition of the gamma matrices and their properties which is independent of the considered representation. The only invariant tensor at our disposal is the metric tensor $\eta^{\mu\nu}$ and thus the gamma matrices have to satisfy $\{\gamma^\mu, \gamma^\nu\} = A\eta^{\mu\nu}$. Considering $\{\gamma^0, \gamma^0\}$ shows that $A = 2$, or

$$\{\gamma^\mu, \gamma^\nu\} = 2\eta^{\mu\nu}\,. \tag{8.21}$$

These anti-commutation relations define a Clifford algebra, implying

$$\left(\gamma^0\right)^2 = 1\,, \qquad \left(\gamma^i\right)^2 = -1 \qquad and \qquad \gamma^\mu\gamma^\nu = -\gamma^\nu\gamma^\mu \tag{8.22}$$

for $\mu \neq \nu$. The last condition shows that the Clifford algebra cannot be satisfied by normal numbers.

The definition (8.21) implies that we can apply in the usual way the metric tensor to raise or to lower the indices of the gamma matrices, $\gamma_\mu = \eta_{\mu\nu}\gamma^\nu$. Thus we can write $\gamma^\nu\partial_\nu = \gamma_\nu\partial^\nu$. Since the contraction of the gamma matrices γ^μ with a four-vector A^μ will appear frequently, we introduce the so-called Feynman slash,

$$\not{A} \equiv A_\mu\gamma^\mu\,, \tag{8.23}$$

as useful shortcut. This notation also stresses that the gamma matrices γ^μ allow us to map a four-vector A^μ onto an element $\not{A}$ of the Clifford algebra which then can be applied on a spinor ψ. Although we suppress the spinor indices, you should keep in mind that the matrices $(\gamma_{ab})^\mu$ carry both tensor and spinor indices.

Dirac's way towards the Dirac equation The Klein-Gordon equation was historically the first wave equation derived in relativistic quantum mechanics. Applied to the hydrogen atom, it failed to reproduce the correct energy spectrum. Dirac tried to derive as an alternative an equation linear in the derivatives ∂_μ. Since Lorentz invariance requires that ∂_μ has to be contracted with another object carrying the Lorentz index μ, a first order equation has the form

$$(\mathrm{i}\gamma^\mu\partial_\mu - m)\psi(x) = 0\,. \tag{8.24}$$

Main task for Dirac was to uncover the nature of the quantities γ^μ in this equation. They cannot be normal numbers, since then they would form a four-vector, specify one direction in space-time and thus break Lorentz invariance. Multiplying the Dirac equation with $-(\mathrm{i}\gamma^\mu\partial_\mu + m)$ and comparing the result to the Klein-Gordon equation, we find

$$-(\mathrm{i}\gamma^\mu\partial_\mu + m)(\mathrm{i}\gamma^\nu\partial_\nu - m)\psi = (\gamma^\mu\gamma^\nu\partial_\mu\partial_\nu + m^2)\psi = (\Box + m^2)\psi = 0\,. \tag{8.25}$$

Using the symmetry of partial derivatives, we can rewrite

$$\gamma^\mu\gamma^\nu\partial_\mu\partial_\nu = \frac{1}{2}\{\gamma^\mu, \gamma^\nu\}\partial_\mu\partial_\nu\,. \tag{8.26}$$

Remembering next the definition of the d'Alembert operator, $\Box = \eta^{\mu\nu}\partial_\mu\partial_\nu$, we re-derive that the γ^μ form a Clifford algebra.

Lagrange density For a complex scalar field, we could rewrite after a partial integration the Lagrange density as $\mathscr{L} = -\phi^\dagger(\Box + m^2)\phi$. This expression corresponds to the Klein-Gordon operator $-(\Box + m^2)$ sandwiched between the quadratic form $\phi^\dagger\phi$, as the correspondence of the propagator and a two-point Green function requires. This suggests to try for the Dirac field as Lagrangian

$$\mathscr{L} = \psi^\dagger A(\mathrm{i}\gamma^\mu\partial_\mu - m)\psi = \bar{\psi}(\mathrm{i}\gamma^\mu\partial_\mu - m)\psi\,, \tag{8.27}$$

where we have used as quadratic from $\psi^\dagger A\psi$ with a matrix A yet to be determined. In the second step, we defined the *adjoint* spinor $\bar{\psi} \equiv \psi^\dagger A$. Varying then the action $S[\psi, \bar{\psi}]$, we obtain

$$\delta S = \int \mathrm{d}^4 \left\{\delta\bar{\psi}(\mathrm{i}\gamma^\mu\partial_\mu - m)\psi - \bar{\psi}(\mathrm{i}\gamma^\mu\overleftarrow{\partial_\mu} + m)\delta\psi\right\}\,. \tag{8.28}$$

Here we made a partial integration of the $\partial_\mu\delta\psi$ term, and thus the derivative $\overleftarrow{\partial_\mu}$ acts to the left. Since we treat ψ and $\bar{\psi}$ as two independent variables, we obtain from $\delta S = 0$ two equations of motion,

$$\bar{\psi}(\mathrm{i}\gamma^\mu\overleftarrow{\partial_\mu} + m) = 0 \quad and \quad (\mathrm{i}\gamma^\mu\partial_\mu - m)\psi = 0\,. \tag{8.29}$$

Next we determine the unknown matrix A: Taking the Hermitian conjugate of the RHS of (8.29) results in

$$\psi^\dagger(-\mathrm{i}\gamma^{\mu\dagger}\overleftarrow{\partial_\mu} - m) = 0\,. \tag{8.30}$$

This agrees with the LHS of (8.29), if A satisfies

$$A^{-1}\gamma^{\mu\dagger}A = \gamma^\mu\,. \tag{8.31}$$

One can readily check that the γ^μ matrices in the Weyl representation fulfil this relation, if we set $A = \gamma^0$. With $(\gamma^0)^2 = 1$ and $(\gamma^0)^{-1} = (\gamma^0)^\dagger = \gamma^0$, we can express this condition as

$$\gamma^{\mu\dagger} = \gamma^0\gamma^\mu\gamma^0 = \begin{cases} (\gamma^0)^2\gamma^0 = \gamma^0\,, \\ -\gamma^i(\gamma^0)^2 = -\gamma^i\,. \end{cases} \tag{8.32}$$

Thus the action principle implies that γ^0 is Hermitian, while the γ^i are anti-Hermitian matrices.

Using the gamma matrices and the Dirac spinor in the chiral representation it is straightforward to express the Dirac Lagrangian (8.27) by Weyl fields,

$$\mathscr{L} = \mathrm{i}\phi_R^\dagger\sigma^\mu\partial_\mu\phi_R + \mathrm{i}\phi_L^\dagger\bar{\sigma}^\mu\partial_\mu\phi_L - m(\phi_L^\dagger\phi_R + \phi_R^\dagger\phi_L)\,. \tag{8.33}$$

This implies that the Dirac Lagrangian and the Dirac equation are invariant under Lorentz transformations, because we have already checked that all ingredients of (8.33) are invariant. Note also that out of the two possible combinations of the two Lorentz scalars we found, only the one invariant under parity entered the mass term. Moreover, $P(\sigma^\mu\partial_\mu) = \bar{\sigma}^\mu\partial_\mu$, and thus the combination of the kinetic energies of ϕ_L and ϕ_R is also invariant under parity.

Hamiltonian form The Dirac equation can be transformed into Hamiltonian form by multiplying with γ^0,

$$\mathrm{i}\partial_t\psi = H_D\psi = (-\mathrm{i}\gamma^0\gamma^i\partial_i + \gamma^0 m)\psi\,. \tag{8.34}$$

Looking back at the (anti-) Hermiticity properties (8.32) of the γ^μ matrices, we see that they correspond to the one required to make the Dirac Hamiltonian Hermitian. By tradition, one re-writes H_D often with $\beta = \gamma^0$ and $\alpha^i = \gamma^0\gamma^i$ as

$$\mathrm{i}\partial_t\psi = H_D\psi = (\boldsymbol{\alpha}\cdot\boldsymbol{p} + \beta m)\psi\,. \tag{8.35}$$

Considering the semi-classical limit, one sees that the matrix $\boldsymbol{\alpha}$ has the meaning of a velocity operator, see problem 8.**??**.

Clifford algebra and bilinear quantities We now determine the minimal matrix representation for the Clifford algebra defined by Eq. (8.21). We find first the maximal number of independent products that we can form out of the four gamma matrices. Five obvious elements are the unit matrix $1 = (\gamma^0)^2$ and the four gamma matrices γ^μ themselves. Because of $(\gamma^\mu)^2 = \pm 1$, the remaining products should consist of γ^μ matrices with different indices. Thus the only product of four γ^μ matrices that we have to consider is $\gamma^0\gamma^1\gamma^2\gamma^3$. This combination will appear very often and deserves therefore a special name. Including the imaginary unit to make it Hermitian, we define

$$\gamma^5 \equiv \gamma_5 \equiv \mathrm{i}\gamma^0\gamma^1\gamma^2\gamma^3\,. \tag{8.36}$$

Because the four gamma matrices in γ^5 anti-commute, we can rewrite its definition introducing the completely anti-symmetric tensor $\varepsilon_{\alpha\beta\gamma\delta}$ in four dimensions as

$$\gamma^5 = \frac{\mathrm{i}}{24}\,\varepsilon_{\alpha\beta\gamma\delta}\,\gamma^\alpha\gamma^\beta\gamma^\gamma\gamma^\delta\,. \tag{8.37}$$

This suggests that bilinear quantities containing one γ^5 matrix transform as pseudo-tensors, i.e. change sign under a parity transformation $\boldsymbol{x} \to -\boldsymbol{x}$. Two important properties of the γ^5 matrix are $(\gamma^5)^2 = 1$ and $\{\gamma^\mu, \gamma^5\} = 0$.

Next we consider products of three γ^μ matrices. For instance,

$$\gamma^1\gamma^2\gamma^3 = \underbrace{\gamma^0\gamma^0}_{1}\gamma^1\gamma^2\gamma^3 = -\mathrm{i}\gamma^0\gamma^5\,. \tag{8.38}$$

Hence these products are equivalent to $\gamma^\mu\gamma^5$, giving us four more basis elements.

Finally, we are left with products of two different γ^μ matrices. We can associate these six products with the commutator $[\gamma^\mu, \gamma^\nu]$. Adding again for later convenience a factor i/2, we define the anti-symmetric tensor $\sigma^{\mu\nu}$ as

$$\sigma^{\mu\nu} \equiv \frac{\mathrm{i}}{2}[\gamma^\mu, \gamma^\nu]\,. \tag{8.39}$$

The six matrices $\sigma^{\mu\nu}$ are the remaining independent elements we choose as basis for our matrix representation of the Clifford algebra. All together, the basis has dimension 16,

$$\Gamma = \{1, \gamma^5, \gamma^\mu, \gamma^5\gamma^\mu, \sigma^{\mu\nu}\}\,, \tag{8.40}$$

as the 4×4 matrices. Hence an arbitrary 4×4 matrix can be decomposed into a linear combination of these basis elements. Moreover, the smallest matrix representation of the Clifford algebra is given by 4×4 matrices. Some useful properties of gamma matrices are collected in the Appendix A.2.

Knowing the dimension of the γ matrices, we can count the number of degrees of freedom represented by a Dirac spinor ψ. As the γ matrices and the Lorentz transformation acting on spinors are complex, the field ψ is complex too and has thus four complex degrees of freedom. We know already that the Dirac equation describes spin 1/2 particles, which come with $2s+1=2$ spin degrees of freedom for a particle plus 2 for its anti-particle. Thus in this case the number of physical states matches the four components of the fields ψ. Note also the difference to the case of a complex scalar or vector field: There we introduced two complex fields $\phi^{\pm}=(\phi_1\pm \mathrm{i}\phi_2)/\sqrt{2}$, which are connected by $(\phi^{\pm})^*=\phi^{\mp}$. The real fields ϕ_1 and ϕ_2 are not mixed by Lorentz transformations and thus we count them as two real degrees of freedom.

We come now to the construction of bilinear quantities out of the Dirac spinors. Since the Lagrangian is a scalar, we know already that $\bar{\psi}\psi$ transforms as a scalar while $j^\mu=\bar{\psi}\gamma^\mu\psi$ is vector. In general, bilinear quantities are constructed as

$$\psi^\dagger\gamma^0\Gamma\psi\equiv\bar{\psi}\Gamma\psi\,, \tag{8.41}$$

where $\bar{\psi}=\psi^\dagger\gamma^0$ is the adjoint spinor and Γ is any of the 16 basis elements given in Eq. (8.40). In this way, the complex conjugated of a bilinear becomes

$$(\bar{\psi}\Gamma\psi')^*=(\bar{\psi}\Gamma\psi')^\dagger=\psi'^\dagger\Gamma^\dagger\gamma^0\psi=\psi'^\dagger\gamma^0\gamma^0\Gamma^\dagger\gamma^0\psi\equiv\bar{\psi}'\overline{\Gamma}\psi \tag{8.42}$$

with

$$\overline{\Gamma}\equiv\gamma^0\Gamma^\dagger\gamma^0\,. \tag{8.43}$$

For $\psi=\psi'$, these bilinears are real as desired. The analogue $\psi^\dagger\psi$ to the probability density $\psi^*\psi$ of the Schrödinger equation is thus the zero-component of a four-current, $\psi^\dagger\psi=\psi^\dagger\gamma^0\gamma^0\psi=\bar{\psi}\gamma^0\psi=j^0$, as one should expect in a relativistic theory.

Finally, we note that γ^0 and γ^5 are involutory matrices, i.e. they satisfy the relation $A^2=1$. Because of $(1\pm A)^2=2(1\pm A)$, we can construct the projection operators $P_\pm=(1\pm A)/2$, satisfying

$$P_\pm^2=P_\pm,\qquad P_\pm P_\mp=0,\quad and\quad P_++P_-=1\,.$$

Thus we should be able to classify the four independent solutions of the Dirac equation with the help of $(1\pm\gamma^0)/2$ and $(1\pm\gamma^5)/2$, or their suitable covariant generalisations.

Lorentz transformations Our derivation of the Weyl spinors as the fundamental representation of the Lorentz group provided automatically their transformation properties under a finite Lorentz transformation. Using the Weyl representation, the transformation law for a Dirac spinor follows as

$$\psi(x)\to\psi(x')=S(\Lambda)\psi(x)=\begin{pmatrix}\phi_L(x')\\ \phi_R(x')\end{pmatrix}=\begin{pmatrix}S_L & 0\\ 0 & S_R\end{pmatrix}\begin{pmatrix}\phi_L(x)\\ \phi_R(x)\end{pmatrix}. \tag{8.44}$$

We want to express the transformation matrix $S(\Lambda)$ by gamma matrices, such that it is representation independent and manifestly Lorentz invariant. We set

$$S(\Lambda)=\exp\left(-\mathrm{i}\omega_{\mu\nu}J^{\mu\nu}/2\right)\,, \tag{8.45}$$

where the antisymmetric matrix $\omega_{\mu\nu}$ parametrises the Lorentz transformation and the six generators $(J_{ab})_{\mu\nu}$ have to be determined. Since the generators are the covariant generalisation of $\boldsymbol{\sigma}$, we suspect that they are connected to $\sigma_{\mu\nu}$. Using Eq. (8.19), we obtain as explicit expression for the $\sigma^{\mu\nu}$ matrices in the Weyl representation

$$\sigma^{0i} = \mathrm{i}\begin{pmatrix} -\sigma^i & 0 \\ 0 & \sigma^i \end{pmatrix}, \quad and \quad \sigma^{ij} = \varepsilon_{ijk}\begin{pmatrix} \sigma^k & 0 \\ 0 & \sigma^k \end{pmatrix}. \tag{8.46}$$

We split $J^{\mu\nu}$ into boosts and rotations,

$$\frac{1}{2}\omega_{\mu\nu}J^{\mu\nu} = \omega_{i0}J^{i0} + \omega_{12}J^{12} + \omega_{13}J^{13} + \omega_{23}J^{23}\,. \tag{8.47}$$

Identifying $\eta_i = \omega_{i0}$ and $\alpha_i = (1/2)\varepsilon_{ijk}\omega_{jk}$, we obtain $J^{\mu\nu} = \sigma^{\mu\nu}/2$. In contrast to (8.44), the expression $S(\Lambda) = \exp(-\mathrm{i}\sigma_{\mu\nu}\omega^{\mu\nu}/4)$ is valid for any representation of the gamma matrices.

Solutions We search for plane wave solutions $u\mathrm{e}^{-\mathrm{i}px}$ and $v\mathrm{e}^{+\mathrm{i}px}$ of the Dirac equation with $m > 0$ and $E = p^0 = |\boldsymbol{p}| > 0$. The algebra is simplified, if we construct the solutions first in the rest frame of the particle. Then $\not{p} = m\gamma^0$, and thus the use of the Dirac representation,

$$\gamma^0 = 1 \otimes \tau_3 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, \quad and \quad \gamma^i = \sigma^i \otimes \mathrm{i}\tau_2 = \begin{pmatrix} 0 & \sigma^i \\ -\sigma^i & 0 \end{pmatrix}, \tag{8.48}$$

where γ^0 is diagonal is most convenient. Here σ_i and τ_i are the Pauli matrices, $\otimes$ denotes the tensor product, 0 and 1 are 2×2 matrices. In the Dirac representation, the γ^5 matrix is off-diagonal,

$$\gamma^5 = 1 \otimes \tau_1 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}. \tag{8.49}$$

The Dirac equation becomes

$$(\not{p} - m)u = m(\gamma^0 - 1)u = 0 \tag{8.50a}$$

$$(\not{p} + m)v = m(\gamma^0 + 1)v = 0\,. \tag{8.50b}$$

The RHS shows that $(1 \pm \gamma^0)/2$ project a general spinor at rest on the subspaces of solutions with positive or negative energy. Inserting the explicit form of γ^0 into (8.50), the four solutions in the rest frame of the particle follow as

$$u(m,+) \propto \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \quad u(m,-) \propto \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \end{pmatrix}, \quad v(m,-) \propto \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \end{pmatrix}, \quad v(m,+) \propto \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \end{pmatrix}. \tag{8.51}$$

The additional ± label should be the quantum number of a suitable operator labelling the two spin states of a Dirac particle. Note the opposite order of the spin label in the v spinor compared to u. We will see later that this choice is required by the structure of the relativistic spin operator s^μ. As an intuitive argument, we add that this labelling corresponds to our interpretation of antiparticles as particles moving backwards in time: The spinor v describes two states with negative energy, negative 3-momentum $\boldsymbol{p}$ and negative spin s relative to u.

The solutions are orthogonal,

$$\bar{u}(p,s)u(p,s') = \mathcal{N}^2\delta_{s,s'} \quad and \quad \bar{v}(p,s)v(p,s') = -\mathcal{N}^2\delta_{s,s'}\,, \tag{8.52}$$

but not normalised to one. Note also the minus sign introduced in $\bar{v}v$ by the corresponding minus in the (3,4) corner of γ^0. Since we know that $\psi^\dagger\psi$ is the zero component of a four-vector, the normalisation of the corresponding spinor products is

$$u^\dagger(p,s)u(p,s') = \mathcal{N}^2\frac{E_p}{m}\delta_{s,s'} \quad and \quad v^\dagger(p,s)v(p,s') = -\mathcal{N}^2\frac{E_p}{m}\delta_{s,s'}\,. \tag{8.53}$$

Summing over spins, we obtain in the rest frame

$$\sum_s u_a(m,s)\bar{u}_b(m,s) = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}_{ab}\mathcal{N}^2 = \frac{1}{2}(\gamma^0+1)_{ab}\mathcal{N}^2\,, \tag{8.54}$$

$$\sum_s v_a(m,s)\bar{v}_b(m,s) = \begin{pmatrix} 0 & 0 \\ 0 & -1 \end{pmatrix}_{ab}\mathcal{N}^2 = \frac{1}{2}(\gamma^0-1)_{ab}\mathcal{N}^2\,. \tag{8.55}$$

We saw that $\gamma^0 \pm 1$ corresponds in an arbitrary frame to $(\not{p} \pm m)/m$. Thus in general these relations become

$$\Lambda_+ \equiv \sum_s u_a(p,s)\bar{u}_b(p,s) = \mathcal{N}^2\left(\frac{\not{p}+m}{2m}\right)_{ab}\,, \tag{8.56}$$

$$\Lambda_- \equiv -\sum_s v_a(p,s)\bar{v}_b(p,s) = \mathcal{N}^2\left(\frac{-\not{p}+m}{2m}\right)_{ab}\,, \tag{8.57}$$

where we defined $\Lambda_\pm$ as the projection operator on states with positive and negative energy, respectively.

The two most common normalisation conventions for the Dirac spinors are $\mathcal{N} = \sqrt{2m}$ and $\mathcal{N} = 1$. We will use the former, $\mathcal{N} = \sqrt{2m}$, which has three advantages: First, the expressions for $\Lambda_\pm$ which appear frequently become more compact. Second, spurious singularities in the limit $m \to 0$ disappear. Finally, the normalisation of fermion states and thus also the phase space volume becomes identical to the one of bosons.

The solutions of the Dirac equation for an arbitrary frame can be simplest obtained remembering $(\not{p}-m)(\not{p}+m) = p^2 - m^2 = 0$, i.e.

$$u(p,\pm) = \frac{\not{p}+m}{\sqrt{m+E}}\,u(0,\pm) \quad and \quad v(p,\pm) = \frac{-\not{p}+m}{\sqrt{m+E}}\,v(0,\pm)\,. \tag{8.58}$$

Here, the normalisation was fixed using (8.52).

Spin We have seen that the Dirac equation describes a particle with helicity one-half. Thus the ± degeneracy of the u and v spinors should correspond to the different helicity or spin states of a Dirac particle. We introduce the spin operator

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\sigma} & 0 \\ 0 & \boldsymbol{\sigma} \end{pmatrix}\,, \tag{8.59}$$

as an obvious generalisation of the non-relativistic spin matrices. This operator has the eigenvalues $\Sigma_z u(m,\pm) = \pm u(m,\pm)$ and $\Sigma_z v(m,\pm) = \mp v(m,\pm)$ and can therefore be used to classify the spin states of a Dirac particle in the rest frame, where $[H_D,\Sigma_z] \propto [\gamma^0,\Sigma_z] = 0$. Note however that $[H_D,\Sigma_z] \neq 0$ for $p^2 \neq m^2$, and thus the eigenvalue of Σ_z is not conserved for a moving particle. This comes not as a surprise, because the total angular momentum $\boldsymbol{L}+\boldsymbol{s}$ and not only the spin $\boldsymbol{s}$ should be conserved.

We are looking now for the relativistic generalisation of the three-dimensional spin operator $\boldsymbol{\Sigma}$. It should be a product of gamma matrices which contains in the rest frame of the particle $\boldsymbol{\sigma}$ in the diagonal. We note first that $\gamma^5\boldsymbol{\gamma} = \left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)\left(\begin{smallmatrix} 0 & \boldsymbol{\sigma} \\ -\boldsymbol{\sigma} & 0 \end{smallmatrix}\right)$ has the required structure. Then we define the spin vector s^μ with the properties $s^2 = -1$, $s^\mu = (0, \boldsymbol{s})|_{p=m}$ and thus $s \cdot p = 0$. Since

$$\gamma^5 \not{s}|_{p=m} = -\gamma^5 \boldsymbol{s\gamma} = \begin{pmatrix} \boldsymbol{\sigma s} & 0 \\ 0 & -\boldsymbol{\sigma s} \end{pmatrix}, \tag{8.60}$$

we see that $\gamma^5\not{s}$ measures in the rest frame the projection of the spin along the chosen axis $\boldsymbol{s}$. Moreover, $\gamma^5\not{s}$ commutes with the Dirac Hamiltonian, $[\gamma^5\not{s}, \not{p}] = 0$, and has because of $(\gamma^5\not{s})^2 = 1$ as eigenvalues ±1. If we apply $\gamma^5\not{s}$ on the spinor $v(p,s)$—what is easiest done in the rest frame—

$$\gamma^5\not{s}\; v(m,s) = \begin{pmatrix} \boldsymbol{\sigma s} & 0 \\ 0 & -\boldsymbol{\sigma s} \end{pmatrix} \begin{pmatrix} 0 \\ \chi \end{pmatrix} = \begin{pmatrix} 0 \\ -\boldsymbol{\sigma s}\chi \end{pmatrix} = s_j \begin{pmatrix} 0 \\ \chi \end{pmatrix}, \tag{8.61}$$

we see that χ_1 has the eigenvalue $s_1 = -1$, while χ_2 has the eigenvalue $s_2 = +1$. This explains the "wrong" order of the two spin states of $v(p,s)$ in (8.51). Finally, we can define a projection operator on a definite spin state by

$$\Lambda_s = \frac{1}{2}(1 + \gamma^5\not{s}). \tag{8.62}$$

Thus we can obtain from an arbitrary Dirac spinor ψ a state with definite sign of the energy and spin by applying the two projection operators $\Lambda_\pm$ and Λ_s.

Helicity An important special case of the spin operator $\gamma^5\not{s}$ is the helicity operator $h \equiv \boldsymbol{sp}/|\boldsymbol{p}|$ which measures the projection of the spin $\boldsymbol{s} = \boldsymbol{\Sigma}/2$ on the momentum $\boldsymbol{p}$ of a particle,

$$\frac{\boldsymbol{\Sigma}\cdot\boldsymbol{p}}{2|\boldsymbol{p}|}\,\psi = h\psi\,. \tag{8.63}$$

The helicity operator and the Dirac Hamiltonian commute, $[H_D, \boldsymbol{\Sigma p}] = 0$, because there is no orbital angular momentum in the direction of $\boldsymbol{p}$. Therefore common eigenfunctions of H_D and h called helicity states can be constructed, cf. problem 8.**??**. Positive helicity particles are called right-handed, negative helicity particles left-handed. For a massive particle, helicity is a frame-dependent quantity: If we choose e.g. a frame with $\boldsymbol{\beta}||\boldsymbol{p}$ and $\beta > p$, then the particles moves in the opposite direction and h changes sign. Since we cannot "overtake" a massless particle, helicity becomes in this case a Lorentz invariant quantity.

Axial and vector U(1) symmetries Out of the 16 bilinear forms, two transform as vectors under proper Lorentz transformations, $j^\mu = \bar\psi\gamma^\mu\psi$ and $j^\mu_5 = \bar\psi\gamma^5\gamma^\mu\psi$. We now want to check if these two currents are conserved. Inspection of the Lagrange density shows immediately that global U(1) transformations,

$$\psi(x) \to \psi'(x) = \mathrm{e}^{\mathrm{i}\phi}\psi(x) \qquad and \qquad \bar\psi(x) \to \bar\psi'(x) = \mathrm{e}^{-\mathrm{i}\phi}\bar\psi(x)\,, \tag{8.64}$$

keep the Lagrangian invariant, $\delta\mathscr{L}=0$. Noether's theorem leads then to the conserved vector current $j^\mu=\bar\psi\gamma^\mu\psi$. In the second case, the underlying symmetry is using $\{\gamma^5,\gamma^\mu\}=0$,

$$\psi'(x)\to \mathrm{e}^{\mathrm{i}\phi\gamma^5}\psi(x) \quad and \quad \bar\psi(x)\to\bar\psi'(x)=(\mathrm{e}^{\mathrm{i}\phi\gamma^5}\psi(x))^\dagger\gamma^0=\bar\psi(x)\mathrm{e}^{\mathrm{i}\phi\gamma^5}\,. \tag{8.65}$$

The resulting (infinitesimal) change is

$$\delta\mathscr{L}=2m\mathrm{i}\bar\psi\gamma^5\psi\,. \tag{8.66}$$

Thus the axial or chiral symmetry $\mathrm{U_A}(1)$ is broken by the mass term, leading to the non-conservation of the axial current j_5^μ for a massive fermion.

Chirality To understand this better we re-express the Dirac Lagrangian using eigenfunctions of γ^5. We can split any solution ψ of the Dirac equation into

$$\psi_L=\frac{1}{2}(1-\gamma^5)\psi\equiv P_L\psi \quad and \quad \psi_R=\frac{1}{2}(1+\gamma^5)\psi\equiv P_R\psi\,. \tag{8.67}$$

Since $\gamma^5\psi_L=-\psi_L$ and $\gamma^5\psi_R=\psi_R$, $\psi_{L,R}$ are eigenfunctions of γ^5 with eigenvalue ±1. Expressing the mass term through these fields as

$$\bar\psi\psi=\bar\psi\left(P_L^2+P_R^2\right)\psi=\psi^\dagger\left(P_R\gamma^0P_L+P_L\gamma^0P_R\right)\psi=\bar\psi_R\psi_L+\bar\psi_L\psi_R \tag{8.68}$$

and similarly for the kinetic term,

$$\bar\psi\partial\!\!\!/\psi=\bar\psi\left(P_L^2+P_R^2\right)\partial\!\!\!/\psi=\psi^\dagger\left(P_R\gamma^0\gamma^\mu P_R+P_L\gamma^0\gamma^\mu P_L\right)\partial_\mu\psi=\bar\psi_L\partial\!\!\!/\psi_L+\bar\psi_R\partial\!\!\!/\psi_R\,, \tag{8.69}$$

the Dirac Lagrange density becomes

$$\mathscr{L}=\mathrm{i}\bar\psi_L\partial\!\!\!/\psi_L+\mathrm{i}\bar\psi_R\partial\!\!\!/\psi_R-m(\bar\psi_L\psi_R+\bar\psi_R\psi_L)\,. \tag{8.70}$$

Comparing this expression to (8.33), we see that we can identify the Dirac fields $\psi_{L/R}$ in the chiral representation with the Weyl fields $\phi_{L/R}$ as follows,

$$\psi_L=\begin{pmatrix}\phi_L\\0\end{pmatrix} \quad and \quad \psi_R=\begin{pmatrix}0\\\phi_R\end{pmatrix}\,. \tag{8.71}$$

Thus the projection operators (8.67) allows us to define the left- and right-chiral components of a Dirac field in an arbitrary representation. If the mass or interaction terms treat ψ_L and ψ_R not symmetrically, one calls them chiral fermions.

The two kinetic terms which are invariant under chiral transformations connect left- to left-chiral and right- to right-chiral fields, while the mass term mixes left- and right-chiral fields. Such a mass term is called Dirac mass. The distinction between left- and right-chiral fields is Lorentz invariant: In terms of Weyl spinors, we saw that the Lorentz transformations S_L and S_R do not mix ϕ_L and ϕ_R—which qualified them to form the irreducible representation of the Lorentz group. In terms of Dirac spinors, the relation $[\gamma^5,\sigma^{\mu\nu}]=0$ guarantees that left and right chiral fields transform separately under a Lorentz transformation, $\psi'_{L/R}=S(\Lambda)\psi_{L/R}$. However, the mass term of a massive Dirac particle will mix left- and right-chiral fields as they evolve in time.

Helicity and chirality eigenstates can be seen as complimentary states. The former one is a conserved, frame-dependent quantum number, while the latter is frame-independent, but not conserved. Thus helicity states are e.g. useful to describe scattering processes where the detector measures spin in a definite frame. If on the other hand the interactions of a fermion are spin-dependent, then one should choose chiral fields, since the Lagrangian should be Lorentz invariant.

Charge conjugation From $\mathscr{L} = \bar\psi(\mathrm{i}\gamma^\mu\partial_\mu - m)\psi$ and $[\mathscr{L}] = m^4$ in four dimensions, we see that the dimension of a fermion field in four dimension is $[\psi] = m^{3/2}$. Thus we can order possible couplings of a fermion to spin-1 particles according to their dimension as

$$\mathscr{L}_I = c_1 A_\mu \bar\psi\gamma^\mu\psi + c_2 A_\mu \bar\psi\gamma^5\gamma^\mu\psi + \frac{c_3}{M} F_{\mu\nu}\bar\psi\sigma^{\mu\nu}\psi + \dots\,, \tag{8.72}$$

where the coupling constants c_i are dimensionless and we introduced the mass scale M. The only coupling to the photon field with a dimensionless coupling that also respects parity is $\mathscr{L}_I = -qj^\mu A_\mu = -q\bar\psi\gamma^\mu\psi A_\mu$. Solving the Lagrange equations for $\mathscr{L}_0 + \mathscr{L}_I$ gives the Dirac equation including a coupling to the electromagnetic field as

$$[\mathrm{i}\gamma^\mu(\partial_\mu + \mathrm{i}qA_\mu) - m]\psi(x) = 0\,. \tag{8.73}$$

This corresponds to the "minimal coupling" prescription known from quantum mechanics.

Having defined the coupling to an external electromagnetic field, we can ask ourselves how the Dirac equation for a charged conjugated field ψ_c should look like. In the case of a scalar particle, complex conjugation transformed a positively charged particle into a negative one and vice versa. We try the same for the Dirac equation,

$$[-\mathrm{i}\gamma^{\mu *}(\partial_\mu - \mathrm{i}qA_\mu) - m]\psi^*(x) = 0\,. \tag{8.74}$$

The matrix $\gamma^{\mu *}$ satisfies also the Clifford algebra. Hence we should find the unitary transformation $U^{-1}\gamma^\mu U = -\gamma^{\mu *}$ or setting $U \equiv C\gamma^0$

$$(C\gamma^0)^{-1}\gamma^\mu C\gamma^0 = -\gamma^{\mu *}\,. \tag{8.75}$$

If it exists, then the charge-conjugated field $\psi^c \equiv C\gamma^0\psi^*$ satisfies the Dirac equation with $-q$,

$$[\mathrm{i}\gamma^\mu(\partial_\mu - \mathrm{i}qA_\mu) - m]C\gamma^0\psi^*(x) = 0\,. \tag{8.76}$$

Explicit calculation shows that we may choose $C = \mathrm{i}\gamma^2\gamma^0$, see problem 8.??. In the chiral representation, $\psi^c = C\gamma^0\psi^* = \mathrm{i}\gamma^2\psi^*$ becomes

$$\psi^c = \begin{pmatrix} 0 & \mathrm{i}\sigma^2 \\ -\mathrm{i}\sigma^2 & 0 \end{pmatrix}\begin{pmatrix} \phi_L^* \\ \phi_R^* \end{pmatrix} = \begin{pmatrix} \mathrm{i}\sigma^2\phi_R^* \\ -\mathrm{i}\sigma^2\phi_L^* \end{pmatrix}, \tag{8.77}$$

which is in agreement with $\phi_L^c = \mathrm{i}\sigma^2\phi_R^*$ and $\phi_R^c = -\mathrm{i}\sigma^2\phi_L^*$ found earlier.

Example 8.1: Since the γ^2 matrix has the same form in the Dirac and the chiral representation, we find applying C on the spinors $u(p,\pm)$ and $v(p,\pm)$ immediately that

$$u^c(p,s) = C\gamma^0 u^*(p,s) = v(p,s) \quad \textit{and} \quad v^c(p,s) = C\gamma^0 v^*(p,s) = u(p,s)\,.$$

Inserted into Eq. (8.85) this implies that $S_F^T(x) = CS_F(-x)C^{-1}$.

Feynman propagator The Green functions of the free Dirac equation are defined by

$$(\mathrm{i}\partial\!\!\!/ - m)S(x,x') = \delta(x-x')\,, \tag{8.78}$$

where we omit on the RHS a unit matrix in spinor space. Translation invariance implies $S(x,x') = S(x-x')$ and, performing a Fourier transformation, the Fourier components $S(p)$ have to obey

$$(p\!\!\!/ - m)S(p) = 1\,. \tag{8.79}$$

After multiplication with $p\!\!\!/ + m$ and use of $a\!\!\!/^2 = \frac{1}{2}\{\gamma^\mu,\gamma^\nu\}a_\mu a_\nu = a^2$, we can solve for the propagator in momentum space,

$$\mathrm{i}S_F(p) = \mathrm{i}\frac{p\!\!\!/ + m}{p^2 - m^2 + \mathrm{i}\varepsilon} = \frac{\mathrm{i}}{p\!\!\!/ - m + \mathrm{i}\varepsilon}\,, \tag{8.80}$$

where the last step is only meant as a symbolical short-cut. Here, we chose again with the $-\mathrm{i}\varepsilon$ prescription the causal or Stückelberg-Feynman propagator for the electron and, more generally, for spin 1/2 particles. Note also the connection to the scalar propagator Δ_F,

$$\mathrm{i}S_F(x) = -(\mathrm{i}\partial\!\!\!/ + m)\mathrm{i}\Delta_F(x)\,. \tag{8.81}$$

Example 8.2: Express the Feynman propagator as sum over the solutions $u(p,s)$ and $v(p,s)$: We follow the steps from (3.25a) to (3.28) in the scalar case, finding now

$$S_F(x) = \int \frac{\mathrm{d}^3p}{(2\pi)^3}\int\frac{\mathrm{d}p_0}{2\pi}\frac{(p\!\!\!/ + m)\mathrm{e}^{-\mathrm{i}p_0 t}\mathrm{e}^{\mathrm{i}\boldsymbol{px}}}{(p_0 - E_p + \mathrm{i}\varepsilon)(p_0 + E_p - \mathrm{i}\varepsilon)} \tag{8.82}$$

$$= \int \frac{\mathrm{d}^3p}{(2\pi)^3}\left[-\mathrm{i}\,\frac{p\!\!\!/ + m}{2E_p}\mathrm{e}^{-\mathrm{i}E_p t}\vartheta(x^0) + \mathrm{i}\,\frac{-E_p\gamma^0 - \boldsymbol{p}\boldsymbol{\gamma} + m}{-2E_p}\mathrm{e}^{\mathrm{i}E_p t}\vartheta(-x^0)\right]\mathrm{e}^{\mathrm{i}\boldsymbol{px}}\,. \tag{8.83}$$

Next we change as in the bosonic case the integration variable as $\boldsymbol{p}\to-\boldsymbol{p}$ in the second term,

$$\mathrm{i}S_F(x) = \int\frac{\mathrm{d}^3p}{2E_p(2\pi)^3}\left[(p\!\!\!/ + m)\mathrm{e}^{-\mathrm{i}(E_p t - \boldsymbol{px})}\vartheta(t) + (-p\!\!\!/ + m)\mathrm{e}^{\mathrm{i}(E_p t - \boldsymbol{px})}\vartheta(-t)\right]\,. \tag{8.84}$$

Using finally (8.56), we arrive at

$$\mathrm{i}S_F(x) = \int\frac{\mathrm{d}^3p}{2E_p(2\pi)^3}\sum_s\left[u(p,s)\bar u(p,s)\mathrm{e}^{-\mathrm{i}(E_p t - \boldsymbol{px})}\vartheta(x^0) - v(p,s)\bar v(p,s)\mathrm{e}^{\mathrm{i}(E_p t - \boldsymbol{px})}\vartheta(-x^0)\right]\,. \tag{8.85}$$

Thus the phase space volume of fermionic states is the same as the one of bosons for the normalisation of the Dirac spinors chosen by us. The minus sign between the positive energy solution propagating forward in time and the negative energy solution propagating backward in time is a direct consequence of our $-\mathrm{i}\varepsilon$ prescription. It implies that fermionic fields anti-commute,

$$\mathrm{i}S_{F,ab}(x) = \langle 0|T\{\psi_a(x)\bar\psi_b(0)\}|0\rangle = \langle 0|\psi_a(x)\bar\psi_b(0)|0\rangle\vartheta(t) - \langle 0|\bar\psi_b(0)\psi_a(x)|0\rangle\vartheta(-t)\,, \tag{8.86}$$

(we have added for clarity the spinor indices) and explains thereby the Pauli exclusion principle and thus the stability of matter.

Let us have a look back to understand why the sign appears. To simplify the discussion, we neglect the inessential mass term. In the positive frequency term $(p^0\gamma^0 - \boldsymbol{p}\boldsymbol{\gamma})\mathrm{e}^{\mathrm{i}px}$, we pick up an additional minus relative to the bosonic case from the variable change $\boldsymbol{p}\to-\boldsymbol{p}$, resulting in $p\!\!\!/ \to -p\!\!\!/$ and

$$S_F(x) \propto p\!\!\!/\,\mathrm{e}^{-\mathrm{i}px}\vartheta(t) - p\!\!\!/\,\mathrm{e}^{\mathrm{i}px}\vartheta(-t)\,.$$

Thus the relative minus sign has its origin in the fact that the fermion propagator $S_F(p)$ is odd in the momentum, while a bosonic propagator is even. In turn, the fermion propagator is linear in the momentum, because the fermion wave equation is a first-order equation.

8.3. Quantizing Dirac fermions

Spin-statistic connection We have noted that fermionic fields should anti-commute examining the Feynman propagator in example 8.2. In a relativistic quantum field theory the spin and the statistics of a field is connected:

- The wave equations of bosons are second-order differential equations. Therefore the propagators of bosonic fields are even in the momentum p. As a result, bosonic fields commute and satisfy Bose-Einstein statistics.

- In contrast, fermions satisfy first-order differential equations and therefore the fermion propagator $S_F(p)$ is odd in p. This implies that fermions are described by anticommuting classical spinors or operators, and satisfy Fermi-Dirac statistics.

This leads to a practical and a principal question: First, the practical one: How do we implement that classical functions which enter the path integral do anti-commute? And second, does the anticommutation of fermionic variables lead to a consistent picture? In particular is the Hamiltonian of such a theory bounded from below?

We will start to address the latter question calculating the energy density $\rho = \mathscr{H}$ of the Dirac field assuming that the classical spinors u and v *do* commute. We determine first the canonically conjugated momenta as

$$\pi = \frac{\partial \mathscr{L}}{\partial \dot{\psi}} = \mathrm{i}\bar{\psi}\gamma^0 = \mathrm{i}\psi^\dagger \tag{8.87}$$

and $\bar{\pi} = 0$. Thus the Hamilton density is

$$\mathscr{H} = \pi\dot{\psi} - \mathscr{L} = \mathrm{i}\psi^\dagger \partial_t \psi - \bar{\psi}(\mathrm{i}\gamma^\mu \partial_\mu - m)\psi = \mathrm{i}\psi^\dagger \partial_t \psi \,, \tag{8.88}$$

where we used the Dirac equation in the last step. To make this expression more explicit, we express now ψ by plane wave solutions,

$$\psi(x) = \sum_s \int \frac{\mathrm{d}^3 p}{\sqrt{(2\pi)^3 2E_p}} \left[b_s(p) u_s(p) e^{-\mathrm{i}px} + d_s^\dagger(p) v_s(p) e^{+\mathrm{i}px} \right] , \tag{8.89a}$$

$$\psi^\dagger(x) = \sum_s \int \frac{\mathrm{d}^3 p}{\sqrt{(2\pi)^3 2E_p}} \left[b_s^\dagger(p) u_s^\dagger(p) e^{+\mathrm{i}px} + d_s(p) v_s^\dagger(p) e^{-\mathrm{i}px} \right] . \tag{8.89b}$$

Inserting these expressions into (8.88) gives schematically $(b^\dagger + d)(b - d^\dagger)$, where the relative minus sign comes from ∂_t acting on ψ. Since the spinors u and v are orthonormal, cf. (8.53), only the diagonal terms survive, $(b^\dagger + d)(b - d^\dagger) \to b^\dagger b - dd^\dagger$. Hence the energy of a Dirac field is given by

$$H = \int \mathrm{d}^3 x \, \mathscr{H} = \sum_s \int \mathrm{d}^3 p \, E_p \left[b_s^\dagger(p) b_s(p) - d_s(p) d_s^\dagger(p) \right] . \tag{8.90}$$

If d and $d^\dagger$ would be normal Fourier coefficients of an expansion into plan waves, the second term would be negative and the energy density of a fermion field could be made arbitrarily negative.

This conclusion is avoided if the fermion fields anticommute: In canonical quantisation, one promotes the Fourier coefficients to operators. Requiring then anti-commutation relations between the creation and annihilation operators for particles and anti-particles,

$$\{b_s(\boldsymbol{p}), b^\dagger_{s'}(\boldsymbol{p}')\} = \delta_{s,s'}\delta(\boldsymbol{p}-\boldsymbol{p}') \quad and \quad \{d_s(\boldsymbol{p}), d^\dagger_{s'}(\boldsymbol{p}')\} = \delta_{s,s'}\delta(\boldsymbol{p}-\boldsymbol{p}'), \tag{8.91}$$

compensates the sign in the second term. Repeating the discussion in section 3.3 one can show that these anti-commutation relations implement also correctly causality for fermionic fields. If we restore units, then we have to add a factor $\hbar$ on the RHS of the anti-commutation relations (8.91). Since the RHS vanishes in the classical limit $\hbar \to 0$, classical spinors should anticommute. Thus we should perform the path integral of fermionic fields over anti-commuting numbers which are called Graßmann numbers.

Graßmann variables We now proceed to the question how we can implement the analogue of the anticommutation relations for operators into the path integral formalism. We define a Graßmann algebra $\mathcal{G}$ requiring that for $a, b \in \mathcal{G}$ the anticommutation rules

$$\{a,a\} = \{a,b\} = \{b,b\} = 0 \tag{8.92}$$

and thus $a^2 = b^2 = 0$ are valid. Then any smooth function f of a and b can be expanded into a power-series as

$$\begin{aligned} f(a,b) &= f_0 + f_1 a + \tilde{f}_1 b + f_2 ab \\ &= f_0 + f_1 a + \tilde{f}_1 b - f_2 ba\,. \end{aligned} \tag{8.93}$$

Defining the derivative as acting to the right, $\partial \equiv \overrightarrow{\partial}$, we find

$$\frac{\partial f}{\partial a} = f_1 + f_2 b\,, \qquad \frac{\partial f}{\partial b} = \tilde{f}_1 - f_2 a\,, \tag{8.94}$$

and

$$\frac{\partial^2 f}{\partial a \partial b} = -\frac{\partial^2 f}{\partial b \partial a} = -f_2\,. \tag{8.95}$$

As integration rules for Graßmann variables, we require linearity and that the infinitesimals $\mathrm{d}a$, $\mathrm{d}b$ are also Graßmann variables,

$$\{a, \mathrm{d}a\} = \{b, \mathrm{d}b\} = \{a, \mathrm{d}b\} = \{\mathrm{d}a, b\} = \{\mathrm{d}a, \mathrm{d}b\} = 0\,. \tag{8.96}$$

Multiple integrals are iterated,

$$\int \mathrm{d}a\mathrm{d}b f(a,b) = \int \mathrm{d}a \left(\int \mathrm{d}b f(a,b) \right)\,. \tag{8.97}$$

We have to determine the value of $\int \mathrm{d}a$ and $\int \mathrm{d}a a$. For the first, we write

$$\left(\int \mathrm{d}a\right)^2 = \left(\int \mathrm{d}a\right)\left(\int \mathrm{d}b\right) = \int \mathrm{d}a\mathrm{d}b = -\int \mathrm{d}b\mathrm{d}a = -\left(\int \mathrm{d}a\right)^2 \tag{8.98}$$

and find thus $\int \mathrm{d}a = 0$. We are left with $\int \mathrm{d}a a$: Since there is no intrinsic scale—states are empty or occupied—we are free to set

$$\int \mathrm{d}a a = 1\,. \tag{8.99}$$

This implies also that there is no difference between definite and indefinite integrals for Graßmann variables. Moreover, differentiation and integration are equivalent for Graßmann variables.

Assume now that η_1 and η_2 are real Graßmann variables and $A \in \mathbb{R}$. Then

$$\int \mathrm{d}\eta_1 \mathrm{d}\eta_2 \ \mathrm{e}^{\eta_2 A \eta_1} = \int \mathrm{d}\eta_1 \mathrm{d}\eta_2 \ (1 + \eta_2 A \eta_1) = \int \mathrm{d}\eta_1 \mathrm{d}\eta_2 \ \eta_2 A \eta_1 = \int \mathrm{d}\eta_1 \ A \eta_1 = A \,. \tag{8.100}$$

Next we consider a two-dimensional integral with an anti-symmetric matrix A and $\eta = (\eta_1, \eta_2)$. Then

$$A = \begin{pmatrix} 0 & a \\ -a & 0 \end{pmatrix} \tag{8.101}$$

and $\eta^T A \eta = 2a\eta_1\eta_2$. Using an arbitrary matrix would lead to the same result, since its symmetric part cancels. Expanding again the exponential gives

$$\int \mathrm{d}^2\eta \ \exp\left(\frac{1}{2}\eta^T A \eta\right) = a = (\det(A))^{1/2}. \tag{8.102}$$

An arbitrary antisymmetric matrix can be transformed into block diagonal form, where the diagonal is composed of matrices of the type (8.101). Thus the last formula holds for arbitrary n.

Finally, we introduce complex Graßmann variables $\eta = (\eta_1, \ldots, \eta_n)$ and their complex conjugates $\eta^* = (\eta_1^*, \ldots, \eta_n^*)$. For any complex matrix A,

$$\int \mathrm{d}^n\eta \mathrm{d}^n\eta^* \ \exp\left(\eta^\dagger A \eta\right) = \prod_{i=1}^n a_i = \det(A)\,. \tag{8.103}$$

We can compare this to the result over commuting complex variables, $z_i = (x_i + \mathrm{i}y_i)/\sqrt{2}$ and $\bar{z}_i = (x_i - \mathrm{i}y_i)/\sqrt{2}$, with $\mathrm{d}x\mathrm{d}y = \mathrm{d}z\mathrm{d}z^*$ and

$$\int \mathrm{d}^n z \mathrm{d}^n z^* \ \exp\left(-z^\dagger A z\right) = \frac{(2\pi)^n}{\det(A)}\,. \tag{8.104}$$

Thus for Graßmann variables the determinant appearing in the evaluation of a Gaussian integral is in the numerator, while it is in the denominator for real or complex valued functions.

Path integral for fermions In the bosonic case, the action $S[\phi, \pi]$ is quadratic in the canonically conjugated momenta π. The path integral over the momenta can thus be performed and we started directly with the path integral in configuration space. For a fermion, $\pi = \mathrm{i}\psi^\dagger$, and thus the path integral in phase space is

$$Z[0] = \int \mathcal{D}\psi\mathcal{D}\bar{\psi} \ \mathrm{e}^{\mathrm{i}S[\psi,\bar{\psi}]} = \int \mathcal{D}\psi\mathcal{D}\bar{\psi} \ \mathrm{e}^{\mathrm{i}\int \mathrm{d}^4x\, \bar{\psi}(\mathrm{i}\partial\!\!\!/ - m)\psi}\,, \tag{8.105}$$

where we changed to $\bar{\psi}$ as integration variable. For its evaluation, we use (8.103) in the limit $n \to \infty$. Since the action is quadratic in the fields, we can perform the path integral formally,

$$Z[0] = \mathrm{D}et(\mathrm{i}\partial\!\!\!/ - m) = \exp \mathrm{T}r \ln(\mathrm{i}\partial\!\!\!/ - m)\,. \tag{8.106}$$

Using the cyclic property of the trace, we write

$$\mathrm{T}r\ln(\mathrm{i}\partial\!\!\!/ - m) = \mathrm{T}r\ln\gamma^5(\mathrm{i}\partial\!\!\!/ - m)\gamma^5 = \mathrm{T}r\ln(-\mathrm{i}\partial\!\!\!/ - m) = \tag{8.107a}$$

$$= \frac{1}{2}[\mathrm{T}r\ln(\mathrm{i}\partial\!\!\!/ - m) + \mathrm{T}r\ln(-\mathrm{i}\partial\!\!\!/ - m)] = \frac{1}{2}[\mathrm{T}r\ln(\Box + m^2)]\,. \tag{8.107b}$$

Thus $Z[0] = \exp[+\mathrm{T}r\ln(\Box + m^2)/2]$: We have found the remarkable result that the zero-point energy of fermions has the opposite sign compared to the one of bosons. We arrive at the same conclusion, using anti-commutation relations $\{d_s(\boldsymbol{p}), d^\dagger_{s'}(\boldsymbol{p}')\} = \delta_{s,s'}\delta(\boldsymbol{p} - \boldsymbol{p}')$ in the Hamiltonian (8.90),

$$H = \sum_s \int \mathrm{d}^3p\, E_p \left[b^\dagger b + d^\dagger d - \delta^{(3)}(0)\right]\,. \tag{8.108}$$

With $\delta^{(3)}(0) = \int \mathrm{d}^3x/(2\pi)^3$ we see that the last term corresponds to the negative zero-point energies of a fermion.

Note that this opens the possibility that the zero-point energies of (groups of) bosons and fermions cancel exactly, provided that the degrees of freedom of fermions and bosons agree. For instance, the trace in Eq. (8.107b) includes the trace over the 4×4 matrix in spinor space, leading to a factor four larger results than for a single scalar. Moreover, their masses have to be the same, $m_f = m_b$. And finally their interactions have to match, so that also higher-order corrections are identical for fermions and bosons. The corresponding symmetry that guarantees automatically that the conditions i)-iii) are satisfied is called "supersymmetry". As result, the vacuum energy would be zero in an unbroken supersymmetric theory, Clearly, the second condition is most problematic, since e.g. no bosonic partner of the electron has been found (yet). Hence supersymmetry must be a broken symmetry, but as long as the mass spitting $m_f^2 - m_b^2$ between fermions and bosons is not too large, it might be still "useful".

Feynman rules Next we add Graßmannian sources η and $\bar\eta$ to the action, $S[\psi,\bar\psi] + \bar\eta\psi + \bar\psi\eta$. Then we complete the square,

$$\bar\psi A\psi + \bar\eta\psi + \bar\psi\eta = (\bar\psi + \bar\eta A^{-1})A(\psi + A^{-1}\eta) - \bar\eta A^{-1}\eta\,, \tag{8.109}$$

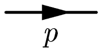
obtaining

$$Z[\eta,\bar\eta] = \int \mathcal{D}\psi\mathcal{D}\bar\psi\; \mathrm{e}^{\mathrm{i}\int \mathrm{d}^4x\bar\psi A\psi + \bar\eta\psi + \bar\psi\eta} = Z[0]\,\mathrm{e}^{-\mathrm{i}\bar\eta A^{-1}\eta} \tag{8.110a}$$

$$= Z[0]\,\exp\left(-\mathrm{i}\int \mathrm{d}^4x\,\mathrm{d}^4x'\;\bar\eta(x)S_F(x-x')\eta(x')\right)\,. \tag{8.110b}$$

Here, $A^{-1}(x,x') = S_F(x-x') = -S_F^T(x'-x)$ which corresponds to the fact that the matrix A is antisymmetric.

The propagator of a Dirac fermion is a line with an arrow representing the flow of the conserved charge which distinguishes particles and antiparticles. Thus a fermion line cannot split, and the arrow cannot change direction,

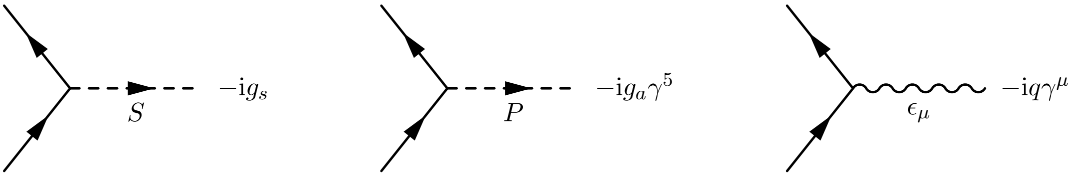


$$= \mathrm{i}S_F(p) = \frac{\mathrm{i}}{p\!\!\!/ - m + \mathrm{i}\varepsilon}\,.$$

Next we look at possible interaction terms of a Dirac fermion with scalars and photons, restricting ourselves to dimensionless coupling constants,

$$\mathscr{L}_I = -g_s S\bar{\psi}\psi - g_a P\bar{\psi}\gamma^5\psi - e\bar{\psi}\gamma^\mu\psi A_\mu\,. \tag{8.111}$$

Both interaction terms of the fermion with the scalars respect parity, if the field S is a true scalar and the field P a pseudo-scalar. Analogous to the $-\mathrm{i}\lambda$ coupling in the case of a scalar self-interactions, we read off from the Lagrangian the following interaction vertices in momentum space,



Fermion loops A closed fermion loop with n propagators corresponds to

$$\underbrace{\overline{\psi}(x_1)\underbrace{\psi(x_1)\overline{\psi}(x_2)}\psi(x_2)\cdots\overline{\psi}(x_n)\psi(x_n)}\;.$$

In order to combine $\bar{\psi}(x_1)$ and $\psi(x_n)$ into $T\{\psi(x_n)\bar{\psi}(x_1)\}$, we have to anticommute $\bar{\psi}(x_1)$ with the $2n-1$ fields $\psi(x_1)\cdots\psi(x_n)$, generating a minus sign. Thus we have to add to our set of Feynman rules that each fermion loop generates a minus sign. Another way to understand the minus sign of fermion loops is to look at the generating functional for connected graphs setting the sources to zero, $\mathrm{i}W[0] = \ln Z[0] = \ln\det A$. The generated graphs are single-closed loops with n Feynman propagators. The change from $1/\det A$ in Z for bosonic fields to $\det A$ in Z for fermionic fields implies an additional minus sign for closed fermion loops. Similarly, diagrams contributing to the same process which differ only by the exchange of two identical external fermion lines carry a relative minus sign. This applies also the exchange of a particle and anti-particle in the initial and final state (cf. also the discussion of crossing symmetry in section 9.3.1).

Furry's theorem What is the relation between diagrams containing fermion loops with opposite orientation in QED? A fermion loop with n external photons attached corresponds to a trace over n fermion propagators separated by gamma matrices,

$$G_1 = \mathrm{tr}[\gamma_{\mu_1}S_F(y_1,y_n)\gamma_{\mu_n}S_F(y_n,y_{n-1})\cdots\gamma_{\mu_2}S_F(y_2,y_1)]\,. \tag{8.112}$$

If we insert $CC^{-1} = 1$ between all factors in the trace, use $C\gamma^\mu C^{-1} = -\gamma^{\mu T}$ and $CS_F(-x)C^{-1} = S_F^T(x)$, then we find

$$\begin{aligned} G_1 &= (-1)^n\mathrm{tr}[\gamma_{\mu_1}^T S_F^T(y_n,y_1)\gamma_{\mu_n}^T S_F^T(y_{n-1},y_n)\cdots\gamma_{\mu_2}^T S_F^T(y_1,y_2)] && (8.113)\\ &= (-1)^n\mathrm{tr}[\gamma_{\mu_1}S_F(y_1,y_2)\cdots\gamma_{\mu_n}S_F(y_n,y_1)] = (-1)^n G_2\,. && (8.114)\end{aligned}$$

Here we used $B^TA^T = (AB)^T$ in the last step. Except for the factor $(-1)^n$, the last expression corresponds to the loop G_2 with opposite orientation. Hence for an odd number of propagators, the two contributions cancel, while they are equal for an even number of propagators.

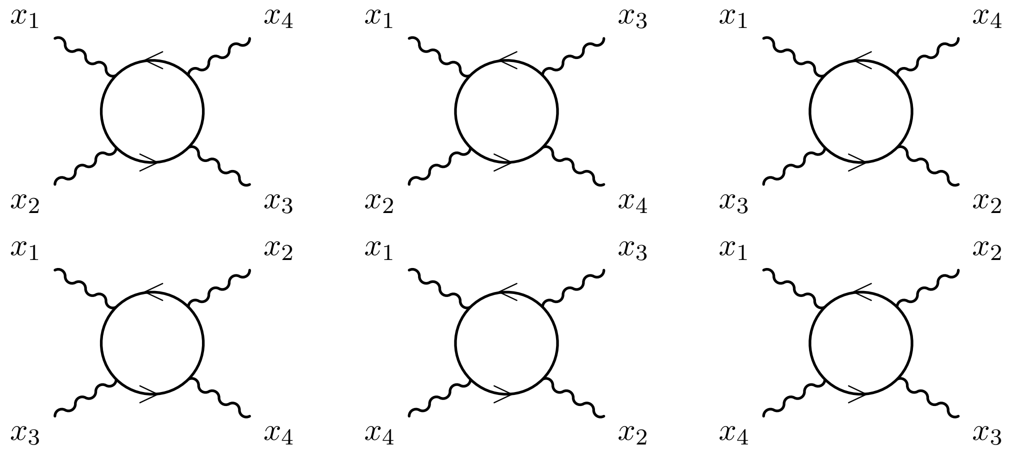
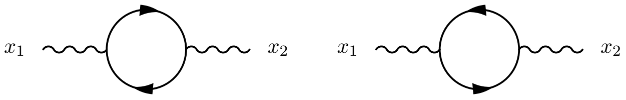


Figure 8.1.: The six topologically distinct diagrams contributing to the photon four-point function.

Symmetry factors in QED The last issue we want to address in this section is the question if the interactions in (8.111) lead to symmetry factors. We recall that drawing Feynman diagrams we should include only diagrams which are topologically distinct after the integration over internal coordinates. For instance, the two diagrams



are not topologically distinct, because a rotation around the x_1–x_2 axis interchanges them. Therefore, the corresponding two-point function $G(x_1,x_2)$ has to be invariant under a change of the orientation of the fermion loop – as it is guarantied by the Furry theorem. The two-point function $G(x_1,x_2)$ consists of two identical diagrams obtained by exchanging the integration variables y_1 and y_2. Thus the factor 2! compensates the 1/2! from the Taylor expansion of $\exp(\mathrm{i}\mathscr{L}_{int})$, if we draw only one diagram, and its symmetry factor is one.

Next we consider the four-point function $G(x_1,x_2,x_3,x_4)$ which describes photon-photon scattering. The four-point function $G(x_1,x_2,x_3,x_4)$ contains $4! \times 3!$ diagrams, obtained by permutating y_1,y_2,y_3,y_4 and x_2,x_3,x_4. After integration over the free y_i variables, the factor 4! compensates the 1/4! from the Taylor expansion of $\exp(\mathrm{i}\mathscr{L}_{int})$. In configuration space, the $3! = 6$ topologically distinct diagrams shown in Fig. 8.1 remain which carry no additional symmetry factor. Thus the resulting rule for QED is very simple: We do not need symmetry factors, if we draw all diagrams which are topologically distinct after the integration over internal coordinates. Fermion loops with an odd number of fermions are zero and can be omitted. Independent of the type of interaction, any fermion loop leads to an additional minus sign.

8.4. Weyl and Majorana fermions

Up to now we have discussed the Dirac equation, having in mind a massive particle carrying a conserved U(1) charge that allows us to distinguish particles and anti-particles. We call such particles Dirac fermions. In the SM, all particles except neutrinos carry a non-zero electric charge, are massive and are therefore Dirac fermions. In this section, we consider the case where one of these two conditions is not fulfilled.

Weyl fermions, $q \neq 0$ and $m = 0$: The Dirac equation (8.18) in the chiral representation decouples for $m = 0$ into two equations called Weyl equations,

$$(E + \boldsymbol{\sigma p})\phi_L(p) = 0 \quad and \quad (E - \boldsymbol{\sigma p})\phi_R(p) = 0\,. \tag{8.115}$$

A fermion described by the Weyl equations is called a Weyl fermion. The correct dispersion relation, $E = |\boldsymbol{p}|$, requires that ϕ_L is an eigenstate of the helicity operator $h = \boldsymbol{\sigma p}/(2|\boldsymbol{p}|)$ with eigenvalue $h = -1/2$, while ϕ_R has the eigenvalue $h = +1/2$. Recall also that helicity is frame independent for a massless particle; in this case positive helicity agrees with right chirality[4]. Until the 1990's, the experimental data on neutrino masses were consistent with zero and neutrinos were incorporated into the SM as Weyl fermions. Since only left-chiral particles and right-chiral antiparticles participate in weak interactions, one set $\nu = \binom{\phi_L}{0}$, while antineutrinos were described by the CP transformed state. The lepton number L_α of the three flavours $\alpha = \{e, \mu, \tau\}$ of leptons played the role of the conserved U(1) charge that distinguishes neutrinos and antineutrinos. As result, the difference in the number of leptons and antileptons of each individual flavour was conserved. Neutrino oscillations that occur if neutrinos are massive conserve the total lepton number $L = \sum_\alpha L_\alpha$ but interchange the individual neutrino flavours L_α. Thus the observation of neutrino oscillations showed that neutrinos are not Weyl fermions.

Majorana fermions, $m > 0$ and $q = 0$: The Dirac field ψ_D has to be complex, because it transforms under the complex representation $S(\Lambda)$ of the Lorentz group. In the case of a neutral fermion, where we cannot distinguish particles and antiparticles, we should have only half of the degrees of freedom of a charged Dirac field. By analogy with the scalar case, we expect that we can halve the number of degrees of the complex Dirac field by imposing a reality condition, $\psi_M = \psi_M^*$. But this condition can be Lorentz invariant only in a special representation of the gamma matrices where $\sigma_{\mu\nu}^* = -\sigma_{\mu\nu}$ and thus $S(\Lambda)$ is real. This condition defines the Majorana representation of the γ matrices in which all γ^μ and thus $\sigma_{\mu\nu}$ are imaginary, and the charge conjugation matrix C is the unity matrix, $C = 1$. Then also the Dirac equation becomes real and thus the time-evolution preserves the reality condition. Since the spinors are real in this representation, no phase invariance $\psi(x) \to \psi'(x) = \exp(\mathrm{i}\phi)\psi(x)$ as in (8.64) can be implemented for a Majorana fermion[5] and thus they cannot carry any conserved U(1) charge.

[4] Most authors call $\psi_{L/R}$ and $\phi_{L/R}$ not left and right-chiral but left and right-handed, although this identification holds only for massless particles.

[5] This argument does not forbid that a Majorana fermion carries conserved charges which transform under a real representation of a symmetry group: An example are gluinos, the suggested supersymmetric partners of the gluons, which are Majorana fermions and transform under the adjoint representation of SU(3).

We can halve the number of degrees of freedom of a Dirac fermion in a representation independent way by using a self-conjugated field $\psi^c = \psi$. A fermion described by a self-conjugated field ψ_M is called a Majorana fermion and the corresponding spinor a Majorana spinor. The field operator of a Majorana field contains only one type of annihilation and creation operators,

$$\psi_M(x) = \sum_s \int \frac{\mathrm{d}^3p}{\sqrt{(2\pi)^3 2E_p}} \left[a_s(p)u_s(p)e^{-\mathrm{i}px} + a_s^\dagger(p)v_s(p)e^{+\mathrm{i}px}\right]. \tag{8.116}$$

Using then

$$\psi_M^c(x) = C\gamma^0\psi_M^*(x) = \sum_s \int \frac{\mathrm{d}^3p}{\sqrt{(2\pi)^3 2E_p}} \left[a_s^\dagger(p)C\gamma^0u_s^*(p)e^{\mathrm{i}px} + a_s(p)C\gamma^0v_s^*(p)e^{-\mathrm{i}px}\right], \tag{8.117}$$

we confirm immediately the Majorana property $\psi_M^c(x) = \psi_M(x)$. Expressed by Weyl spinors, a Majorana spinor becomes

$$\psi^c = \psi = \begin{pmatrix} \phi_L \\ -\mathrm{i}\sigma^2\phi_L^* \end{pmatrix} = \begin{pmatrix} \mathrm{i}\sigma^2\phi_R^* \\ \phi_R \end{pmatrix}. \tag{8.118}$$

Thus a Majorana fermion ($m > 0$, $q = 0$) contains two degrees of freedom, which we may choose either as a left-chiral or or a right-chiral two-spinor with both helicities.

We can replace any Dirac field ψ_D by a pair of self-conjugated fields,

$$\psi_{M,1} = \frac{1}{\sqrt{2}}\left(\psi_D + \psi_D^c\right), \tag{8.119a}$$

$$\psi_{M,2} = \frac{1}{\sqrt{2}}\left(\psi_D - \psi_D^c\right). \tag{8.119b}$$

and vice versa inverting these relations. Thus it is only a question of taste, if one describes fermions by Dirac, Weyl or Majorana spinors.

Dirac versus Majorana mass terms Charge conjugated Dirac spinors were defined by

$$\psi^c = C\gamma^0\psi^* = C\bar\psi^t, \qquad \bar\psi^c = \psi^t C\,.$$

We define also

$$\psi_L^c \equiv (\psi_L)^c = \frac{1}{2}(1+\gamma^5)\psi^c = (\psi^c)_R\,, \tag{8.120}$$

which is consistent with our previous definition for Weyl spinors. As we saw, a Dirac mass term connects the left- and right-chiral components of the same field and $\psi = \psi_L + \psi_R$ is a mass eigenstate. We now use the observation that $(\psi_L)^c = (\psi^c)_R$ allows us to obtain new mass terms[6] called Majorana mass terms,

$$-\mathscr{L}_L = m_L(\bar\psi_L^c\psi_L + \bar\psi_L\psi_L^c) \tag{8.121}$$

$$-\mathscr{L}_R = m_R(\bar\psi_R^c\psi_R + \bar\psi_R\psi_R^c) \tag{8.122}$$

[6] Note that terms like $\bar\psi_L^c\psi_L = \psi_L^t C\psi_L$ vanish because of $C^T = -C$, if one does not already assumes on the classical level that fields are anticommuting Graßmann variables.

which connect the left- and right-chiral components of charge-conjugated fields. The corresponding mass eigenstates are the self-conjugated fields

$$\chi = \psi_L + \psi_L^c = \chi^c \quad \textit{and} \quad \omega = \psi_R + \psi_R^c = \omega^c \tag{8.123}$$

with $\mathscr{L}_L = -m_L\bar{\chi}\chi$ and $\mathscr{L}_R = -m_R\bar{\omega}\omega$. In the general case, both Dirac and Majorana mass terms may be present,

$$\begin{aligned} -\mathscr{L}_{DM} &= m_D\bar{\psi}_L\psi_R + m_L\bar{\psi}_L^c\psi_L + m_R\bar{\psi}_R^c\psi_R + \text{h.}c. && \text{(8.124a)} \\ &= \frac{1}{2}m_D(\bar{\chi}\omega + \bar{\omega}\chi) + m_L\bar{\chi}\chi + m_R\bar{\omega}\omega\,, && \text{(8.124b)} \end{aligned}$$

or in matrix form

$$-\mathscr{L}_{DM} = (\bar{\chi}, \bar{\omega}) \begin{pmatrix} m_L & m_D/2 \\ m_D/2 & m_R \end{pmatrix} \begin{pmatrix} \chi \\ \omega \end{pmatrix}. \tag{8.125}$$

Physical states have a definite mass and thus we have to diagonalise the mass matrix. Its eigenvalues are

$$m_{1,2} = \frac{1}{2}\left\{(m_L + m_R) \pm \sqrt{(m_L - m_R)^2 + m_D^2}\right\} \tag{8.126}$$

and its eigenvectors

$$\begin{aligned} \eta_1 &= \cos\vartheta\chi - \sin\vartheta\omega && \text{(8.127a)} \\ \eta_2 &= \sin\vartheta\chi + \cos\vartheta\omega && \text{(8.127b)} \end{aligned}$$

with $\tan 2\vartheta = m_D/(m_L - m_R)$. Thus the most general mass term $\mathscr{L}_{DM}$ for a four-component fermion spinor corresponds to two Majorana particles with different masses. Therefore we can view a Dirac particle as a special case of two Majorana particles with identical masses and interactions.

The *seesaw model* tries to explain why neutrinos have much smaller masses than all other particles in the standard model. Let us assume that there exist both left- and right-chiral neutrinos and that they obtain Dirac masses as the other fermions, say of order $m_D \sim 100\,\text{GeV}$. The right-chiral ν_R does not participate in any SM interaction and suffers the same fate as a scalar particle: Its mass will be driven by quantum corrections to a value close to the cutoff scale used, and so we expect $m_R \gg m_D$. Moreover, in many models it is $m_L = 0$. Expanding then

$$m_{1,2} \approx \frac{1}{2}\left\{m_R \pm m_R\sqrt{1 + m_D^2/m_R^2}\right\}, \tag{8.128}$$

the two eigenvalues are $m_1 \approx m_D^2/(4m_R)$ and $m_2 \approx m_R$. For $m_R \sim 10^{14}\,\text{GeV}$, the light neutrino mass is in the eV or sub-eV range as required by experimental data.

Summary

The fundamental representation of the proper Lorentz group for massive particles is given by left and right-chiral Weyl spinors. These two-spinors are mixed by parity and thus one combines them into a Dirac four-spinor for parity conserving theories like electromagnetic and strong interactions.

Fermions satisfy first-order differential equations and have mass dimension 3/2. Therefore the fermion propagator $S_F(p)$ is linear in p and thus $S_F(x)$ is an antisymmetric function in x. As a result, fermions satisfy Fermion-Dirac statistics and are described either by Graßmann variables or by anticommuting operators.

A Weyl fermion has $m = 0$ and $q \neq 0$ and satisfies the Weyl equation; its solution has two degrees of freedom, a left-chiral field with negative helicity and a right-chiral field with positive helicity. A Majorana fermion is a self-conjugated field with $m \neq 0$ and $q = 0$ which has therefore also only two degrees of freedom. It is described either by a left-chiral or a right-chiral 2-spinor with both helicities. In the Majorana representation, this spinor can be chosen to be real.

Further reading

The symmetries of the Dirac equation as well as of other relativistic wave equations are extensively discussed by [Gre00]. More details on two-component Weyl and Majorana spinor can be found e.g. in [Sre07].

9. Scattering processes

Most information about the properties of fundamental interactions and particles is obtained from scattering experiments. In a scattering process, the initial and final state contain widely separated particles which can be approximated as free, real particles which are on mass-shell. By contrast, n-point Green functions describe the propagation of virtual particles. In order to make contact with experiments, we have to find therefore the link between Green functions and experimental results from scattering experiments. The latter can be predicted knowing the scattering matrix S which is an unitary operator mapping an initial state at $t=-\infty$ on a final state at $t=+\infty$. We introduce first the S-matrix and show then that its unitarity restricts the analytic structure of Feynman amplitudes; in particular it implies the optical theorem. Then we derive the connection between n-point Green functions and scattering amplitudes, before we perform some explicit calculations of few tree-level processes. Finally, we consider the special case when in a scattering event additional soft particles are emitted. The relation between Feynman amplitudes and cross sections or decay widths which is essentially the same as in non-relativistic quantum mechanics is reviewed in the appendix of this chapter.

9.1. Unitarity of the S-matrix and its consequences

A scattering process is fully described in the Schrödinger picture by the knowledge how initial states $|i,t\rangle$ at $t\to-\infty$ are transformed into final states $|f,t\rangle$ at $t\to\infty$. This knowledge is encoded in the S-matrix elements

$$|f,t=\infty\rangle = S_{fi}\,|i,t=-\infty\rangle\ . \tag{9.1}$$

An intuitive, but mathematically delicate definition of the scattering operator S is the $t\to\infty$ limit of the time-evolution operator $U(t,-t)$,

$$S=\lim_{t\to\infty}U(t,-t)\,. \tag{9.2}$$

Thus the scattering operator S evolves an eigenstate $|n,t\rangle$ of the Hamiltonian from $t=-\infty$ to $t=+\infty$,

$$S\,|n,-\infty\rangle = |n,\infty\rangle\ . \tag{9.3}$$

The unitarity of the scattering operator, $S^\dagger S = SS^\dagger = 1$, expresses the fact that we (should) use a complete set of states for the initial and final states in a scattering process,

$$1=\sum_n |n,+\infty\rangle\,\langle n,+\infty| = \sum_n S\,|n,-\infty\rangle\,\langle n,-\infty|\,S^\dagger = SS^\dagger\,. \tag{9.4}$$

Optical theorem We split the scattering operator S into a diagonal part and the transition operator T, $S = 1 + \mathrm{i}T$, and thus

$$1 = (1 + \mathrm{i}T)(1 - \mathrm{i}T^\dagger) = 1 + \mathrm{i}(T - T^\dagger) + TT^\dagger \tag{9.5}$$

or

$$\mathrm{i}TT^\dagger = T - T^\dagger \,. \tag{9.6}$$

Note that in perturbation theory the LHS is $\mathcal{O}(g^{2n})$, while the RHS is $\mathcal{O}(g^n)$. Hence this equation implies a non-linear relation between the transition operator evaluated at different orders. At lowest order perturbation theory, the LHS vanishes and T is real, $T = T^\dagger$.

We consider now matrix elements between the initial and final state,

$$\langle f|\, T - T^\dagger \,|i\rangle = T_{fi} - T^*_{if} = \mathrm{i}\,\langle f|\, TT^\dagger \,|i\rangle = \mathrm{i}\sum_n T_{fn}T^*_{in} \,. \tag{9.7}$$

If we set $|i\rangle = |f\rangle$, we obtain a connection between the forward scattering amplitude T_{ii} and the total cross section $\sigma_{\mathrm{t}ot}$ called the optical theorem,

$$2\Im T_{ii} = \sum_n |T_{in}|^2 \,. \tag{9.8}$$

The optical theorem relates the attenuation of a beam of particles in the state i, $\mathrm{d}N_i \propto -|\Im T_{ii}|^2 N_i$, to the probability that they scatter into all possible states n. Its RHS is given by the total cross section σ_{tot} up to a factor depending on the flux of initial particles and possible symmetry factors. For the case of two particles in the initial state, comparison with Eqs. (9.149) and (9.153) from the appendix shows that

$$\Im\mathcal{A}_{ii} = 2p_{cms}\sqrt{s}\,\sigma_{tot} \,. \tag{9.9}$$

Note also that the forward scattering amplitude T_{ii} means *scattering* without change in any conserved quantum number, since we extracted already the identity part, $T_{ii} = (S_{ii} - 1)/\mathrm{i}$.

Imaginary part of the amplitude Let us consider the Feynman amplitude $\mathcal{A}$ as a complex function of the squared center-of-mass (c.m.) energy s. The threshold energy $\sqrt{s_0}$ in the c.m. system equals the minimal energy for which the reaction is kinematically allowed. The optical theorem implies that $\mathcal{A}$ is real for $s < s_0$ and $s \in \mathbb{R}$. Thence $s = s^*$ and $\mathcal{A}(s) = [\mathcal{A}(s)]^*$ and therefore

$$\mathcal{A}(s) = [\mathcal{A}(s^*)]^* \qquad \mathrm{f}or \quad s < s_0 \,. \tag{9.10}$$

If $\mathcal{A}(s)$ is an analytic function, then also $[\mathcal{A}(s^*)]^*$ is analytic and we can continue this relation into the complex s plane. In particular, along the real axis we have for $s > s_0$

$$\Re\mathcal{A}(s + \mathrm{i}\varepsilon) = \Re\mathcal{A}(s - \mathrm{i}\varepsilon) \quad \mathrm{a}nd \quad \Im\mathcal{A}(s + \mathrm{i}\varepsilon) = -\Im\mathcal{A}(s - \mathrm{i}\varepsilon) \,. \tag{9.11}$$

Thus starting from s_0, the amplitude $\mathcal{A}$ has a discontinuity along the real s axis. Since the amplitude $\mathcal{A}$ should be single-valued, it has to contain a branch cut along the real s axis starting at s_0. Feynman's $m^2 - \mathrm{i}\varepsilon$ prescription tells us then which side of the cut we should pick out as the "physical" one.

The second relation in (9.11) allows us to obtain the imaginary part of a Feynman amplitude from its discontinuity,

$$\mathrm{d}isc(\mathcal{A}) \equiv \mathcal{A}(s+\mathrm{i}\varepsilon) - \mathcal{A}(s-\mathrm{i}\varepsilon) = 2\mathrm{i}\Im\mathcal{A}(s+\mathrm{i}\varepsilon)\,. \tag{9.12}$$

The prototype of a function having a discontinuity and a branch cut (along $\mathbb{R}^-$) is the logarithm,

$$\mathrm{L}n(z) = \mathrm{L}n(r\mathrm{e}^{\mathrm{i}\vartheta}) = \ln(r\mathrm{e}^{\mathrm{i}\vartheta}) + 2k\pi\mathrm{i} = \ln(r) + (\vartheta + 2k\pi)\mathrm{i} \tag{9.13}$$

with $\Im\ln(x+\mathrm{i}\varepsilon) = \pi$. How does an imaginary part in a Feynman diagram arise? Comparing the relation

$$\frac{1}{x\pm\mathrm{i}\varepsilon} = P\left(\frac{1}{x}\right) \mp \mathrm{i}\pi\delta(x) \tag{9.14}$$

with the propagator of a virtual particle, we see that virtual particles which propagate on-shell lead to poles and to imaginary terms in the amplitude.

Example 9.1: Verify the optical theorem for $\phi\phi \to \phi\phi$ scattering in the $\lambda\phi^4$ theory at $\mathcal{O}(\lambda^2)$:
The logarithmic terms in the scattering amplitude (4.74b) for $\phi\phi \to \phi\phi$ scattering at one-loop have the form

$$F(q^2,m) = \int_0^1 \mathrm{d}z\, \ln\left[m^2 - q^2 z(1-z)\right] \tag{9.15}$$

with $q^2 = \{s,t,u\}$. In the physical region, the relation $q^2 > 4m^2$ holds only for the s channel diagram. The argument of the logarithm becomes negative for

$$z_{1/2} = \frac{1}{2}\left[1 \pm \sqrt{1-4m^2/s^2}\right] = \frac{1}{2} \pm \frac{1}{2}\beta \tag{9.16}$$

with $\beta = \sqrt{1-4m^2/s^2}$ as the velocity of the ϕ particles in the center of mass system. Using now $\Im[\ln(-q^2-\mathrm{i}\varepsilon)] = -\pi$, the imaginary part follows as

$$\Im(A) = \pi\frac{\lambda^2}{32\pi^2}\int_{\frac{1}{2}-\frac{1}{2}\beta}^{\frac{1}{2}+\frac{1}{2}\beta} \mathrm{d}z = \frac{\lambda^2}{32\pi}\beta\,. \tag{9.17}$$

The optical theorem implies thus that the total cross section $\sigma_{tot}(\phi\phi \to \text{all})$ at $\mathcal{O}(\lambda^2)$ equals

$$\sigma_{\mathrm{tot}} = \frac{\Im\mathcal{A}_{ii}}{2p_{cms}\sqrt{s}} = \frac{\lambda^2}{32\pi s}\,, \tag{9.18}$$

where we used $2p_{cms} = \sqrt{s}\beta$. On the other hand, the Feynman amplitude at tree level is simply $\mathcal{A} = -\lambda$ and thus the elastic cross section for $\phi\phi \to \phi\phi$ scattering follows as $\sigma_{el} = \lambda^2/(32\pi s)$. At $\mathcal{O}(\lambda^2)$, the only reaction contributing to the total cross section is elastic scattering, and thus the two cross sections agree. Note also the treatment of the symmetry factors: In the loop diagram, the symmetry factor $S = 1/2!$ is already included, while the corresponding factor for the two identical particles in the final state is added only integrating the cross section.

9.2. LSZ reduction formula

We defined the generating functional $Z[J] = \langle 0,\infty|0,-\infty\rangle_J$ as the vacuum-vacuum transition amplitude in the presence of a classical source J. Thus the generating functional contains the boundary condition $\phi(x) \to 0$ for $t \to \pm\infty$. We have two options to find a bridge between

S-matrix elements and the formalism we have derived up to now: One possibility is to find the connection between the Green functions derived from $Z[J]$ and S-matrix elements. Another one is to define first a new functional $Z'[J]$ with the correct boundary conditions, and to establish then the connection between $Z[J]$ and $Z'[J]$. We choose the first way, restricting ourselves for simplicity to the case of a real scalar field.

Let us start with the case of a $2 \to 2$ scattering process. We can generate a free two-particle[1] state composed of plane-waves by applying two creation operators on the vacuum,

$$|\boldsymbol{k}_1, \boldsymbol{k}_2\rangle = a^\dagger(\boldsymbol{k}_1) a^\dagger(\boldsymbol{k}_2) |0\rangle \ . \tag{9.19}$$

We obtain localised wave packets defining new creation operators

$$a_i^\dagger = \int \mathrm{d}^3 k \, f_i(\boldsymbol{k}) a^\dagger(\boldsymbol{k}) \,, \tag{9.20}$$

where $f_i(\boldsymbol{k})$ is e.g. a Gaussian centered around $\boldsymbol{k}_i$,

$$f_i(\boldsymbol{k}) \propto \exp[-(\boldsymbol{k} - \boldsymbol{k}_i)^2/(2\sigma^2)] \,. \tag{9.21}$$

We assume that the initial state of the scattering process at $t = -\infty$ can be described by freely propagating wave packets,

$$|i\rangle = \lim_{t\to-\infty} a_1^\dagger(t) a_2^\dagger(t) |0\rangle = |\boldsymbol{k}_1, \boldsymbol{k}_2; -\infty\rangle \ , \tag{9.22}$$

and similarly the final state as

$$|f\rangle = \lim_{t\to\infty} a_{1'}^\dagger(t) a_{2'}^\dagger(t) |0\rangle = |\boldsymbol{k}_{1'}, \boldsymbol{k}_{2'}; +\infty\rangle \ . \tag{9.23}$$

Here we changed to the Heisenberg picture, since our Green functions are time-dependent.

Our task is to connect this transition amplitude $\langle f | i\rangle$ to the corresponding four-point Green function. The latter is the time-ordered vacuum expectation value of field operators. The first property, time-ordering, is automatically satisfied for the transition amplitude $\langle f | i\rangle$, since we can write

$$\langle f | i\rangle = \lim_{t\to\infty} \langle 0| a_{1'}(t) a_{2'}(t) a_1^\dagger(-t) a_2^\dagger(-t) |0\rangle \tag{9.24a}$$

$$= \lim_{t\to\infty} \langle 0| \mathrm{T}\{a_{1'}(t) a_{2'}(t) a_1^\dagger(-t) a_2^\dagger(-t)\} |0\rangle \ . \tag{9.24b}$$

Thus we only have to re-express the creation and annihilation operators as (projected) field operators. We define a scalar product for solutions of the Klein-Gordon equation as follows,

$$(\phi, \chi) = \mathrm{i} \int \mathrm{d}^3 x \, \phi^*(x) \overleftrightarrow{\partial_0} \chi(x) \equiv \mathrm{i} \int \mathrm{d}^3 x \left[\phi^*(x) \frac{\partial \chi(x)}{\partial t} - \frac{\partial \phi^*(x)}{\partial t} \chi(x) \right] \ . \tag{9.25}$$

Comparing this definition to Eq. (5.13), we see that the scalar product is the zero component of the conserved current j^μ. Thus the value of the scalar product (ϕ, χ) is time-independent and corresponds to the number of particles minus the number of anti-particles.

[1] To reduce clutter, we assume $\boldsymbol{k}_1 \neq \boldsymbol{k}_2$.

For plane-wave components with definite momentum,

$$\phi_{\boldsymbol{k}}(x) = \frac{1}{\sqrt{(2\pi)^3 2\omega_k}}\,\mathrm{e}^{-\mathrm{i}kx} = N_{\boldsymbol{k}}\,\mathrm{e}^{-\mathrm{i}kx}\,, \tag{9.26}$$

the scalar product is given by

$$(\phi_{\boldsymbol{k}}, \phi_{\boldsymbol{k}'}) = \mathrm{i}N_{\boldsymbol{k}}N_{\boldsymbol{k}'}\int \mathrm{d}^3x\left[\mathrm{e}^{\mathrm{i}k'x}(-\mathrm{i}\omega_{\boldsymbol{k}})\mathrm{e}^{-\mathrm{i}kx} - \mathrm{i}\omega_{\boldsymbol{k}'}\mathrm{e}^{\mathrm{i}k'x}\mathrm{e}^{-\mathrm{i}kx}\right] \tag{9.27a}$$

$$= N_{\boldsymbol{k}}^2(2\pi)^3\delta(\boldsymbol{k}-\boldsymbol{k}')2\omega_{\boldsymbol{k}}\mathrm{e}^{\mathrm{i}(\omega_{\boldsymbol{k}}-\omega_{\boldsymbol{k}})t} = \delta(\boldsymbol{k}-\boldsymbol{k}')\,. \tag{9.27b}$$

Similarly it follows $(\phi^*_{\boldsymbol{k}}, \phi^*_{\boldsymbol{k}'}) = -\delta(\boldsymbol{k}-\boldsymbol{k}')$, while the two terms in the scalar product cancel otherwise,

$$(\phi_{\boldsymbol{k}}, \phi^*_{\boldsymbol{k}'}) = (\phi^*_{\boldsymbol{k}}, \phi_{\boldsymbol{k}'}) = 0.$$

Thus we can invert the free field operator

$$\phi(x) = \int \mathrm{d}^3k\left[a(\boldsymbol{k})\phi_{\boldsymbol{k}}(x) + a^\dagger(\boldsymbol{k})\phi^*_{\boldsymbol{k}}(x)\right] = \int \frac{\mathrm{d}^3k}{\sqrt{(2\pi)^3 2\omega_k}}\left[a(\boldsymbol{k})\mathrm{e}^{-\mathrm{i}kx} + a^\dagger(\boldsymbol{k})\mathrm{e}^{+\mathrm{i}kx}\right] \tag{9.28}$$

to obtain

$$a^\dagger(\boldsymbol{k}) = -(\phi^*_{\boldsymbol{k}}, \phi) = -\mathrm{i}N_{\boldsymbol{k}}\int \mathrm{d}^3x\;\mathrm{e}^{-\mathrm{i}kx}\overleftrightarrow{\partial_t}\,\phi(x)\,. \tag{9.29}$$

Next we want to rewrite this expression in a way that shows explicitly its Lorentz invariance. Using the identity

$$a^\dagger(\boldsymbol{k},\infty) - a^\dagger(\boldsymbol{k},-\infty) = \int_{-\infty}^{\infty}\mathrm{d}t\,\frac{\partial}{\partial t}\,a^\dagger(\boldsymbol{k},t) \tag{9.30}$$

we insert first (9.29) assuming a wave-package localised around $\boldsymbol{k}_1$ and perform then the time differentiation,

$$\begin{aligned} a^\dagger(\boldsymbol{k}_1,\infty) - a^\dagger(\boldsymbol{k}_1,-\infty) &= -\mathrm{i}\int \mathrm{d}^3k f_1(\boldsymbol{k})\int \mathrm{d}^4x\;\partial_t\left(\mathrm{e}^{-\mathrm{i}kx}\overleftrightarrow{\partial_t}\,\phi(x)\right) \\ &= -\mathrm{i}\int \mathrm{d}^3k f_1(\boldsymbol{k})\int \mathrm{d}^4x\;\left(\mathrm{e}^{-\mathrm{i}kx}\partial_t^2\phi(x) - \phi(x)\partial_t^2\mathrm{e}^{-\mathrm{i}kx}\right)\,, \end{aligned} \tag{9.31}$$

where the two terms linear in ∂_t cancelled. Then we use that the field is on-shell, $k^2 = m^2$, for the replacement

$$\partial_t^2\mathrm{e}^{-\mathrm{i}kx} = (\boldsymbol{\nabla}^2 - m^2)\mathrm{e}^{-\mathrm{i}kx}.$$

Since the field is localised in space, we can perform two partial integrations moving thereby $\boldsymbol{\nabla}^2$ to the left, obtaining

$$a^\dagger(\boldsymbol{k}_1,\infty) - a^\dagger(\boldsymbol{k}_1,-\infty) = -\mathrm{i}\int \mathrm{d}^3k f_1(\boldsymbol{k})\int \mathrm{d}^4x\;\mathrm{e}^{-\mathrm{i}kx}\left(\Box + m^2\right)\phi(x)\,. \tag{9.32}$$

In a free theory, $\phi(x)$ satisfies the Klein-Gordon equation and the RHS would vanish. In an interacting theory with e.g. $\mathscr{L}_I = -\lambda\phi^4/4!$, the RHS is however proportional to $\left(\Box + m^2\right)\phi(x) = \lambda\phi^3/3! \neq 0$.

Having performed the partial integrations, we can forget the wave-packets, $\sigma \to 0$, and write simply

$$a^\dagger(\boldsymbol{k},-\infty) = a^\dagger(\boldsymbol{k},\infty) + \mathrm{i}N_{\boldsymbol{k}}\int \mathrm{d}^4x\;\mathrm{e}^{-\mathrm{i}kx}\left(\Box + m^2\right)\phi(x)\,. \tag{9.33}$$

Taking the Hermitian conjugate, we obtain for the annihilation operator

$$a(\boldsymbol{k},\infty)=a(\boldsymbol{k},-\infty)+\mathrm{i}N_{\boldsymbol{k}}\int \mathrm{d}^4x\ \mathrm{e}^{\mathrm{i}kx}\left(\Box+m^2\right)\phi(x)\,. \tag{9.34}$$

When we insert these expressions into $\langle f\,|i\rangle$, we obtain a four-point function combining the second terms from the RHS of (9.33) and (9.34). Including the terms $a^\dagger(\boldsymbol{k},\infty)$ and $a(\boldsymbol{k},-\infty)$ generates particles propagating from $t=-\infty$ to $t=+\infty$ with momenta unchanged, i.e. to terms corresponding to disconnected graphs. Hence we do not need to consider these contributions, restricting ourselves to connected Green functions. For n particles in the initial and m particles in the final state, we obtain

$$\begin{aligned}\langle \boldsymbol{k}_1',\ldots,\boldsymbol{k}_m';+\infty\,|\boldsymbol{k}_1,\ldots,\boldsymbol{k}_n;-\infty\rangle &= \mathrm{i}^{n+m}\prod_{i=1}^{n}\int \mathrm{d}^4x_i\ N_{\boldsymbol{k}_i}\,\mathrm{e}^{-\mathrm{i}k_ix_i}\left(\Box_{x_i}+m^2\right)\\ \times\prod_{j=1}^{m}\int \mathrm{d}^4y_j\ N_{\boldsymbol{k}_j}\,\mathrm{e}^{\mathrm{i}k_j'y_j}&\left(\Box_{y_j}+m^2\right)\langle 0|\,\mathrm{T}\{\phi(x_1)\cdots\phi(y_m)\,|0\rangle\ .\end{aligned} \tag{9.35}$$

This is the reduction formula of Lehmann, Symanzik and Zimmermann (LSZ): For each external particle we obtained the corresponding plan wave component and a Klein-Gordon operator. Since the latter is the inverse of the free 2-point function, we can rephrase the content of the LZS formula simply as follows: Replace the 2-point functions of external lines by appropriate wave functions, e.g. $\phi_{\boldsymbol{k}}(x)$ for scalar particles in the initial state and $\phi^*_{\boldsymbol{k}}(x)$ for scalar particles in the final state.

Since we started from field operators in the Heisenberg picture, the matrix element is in the Heisenberg picture, too. In the Schrödinger picture, it is

$$\langle \boldsymbol{k}_1',\ldots,\boldsymbol{k}_m'|\mathrm{i}T\,|\boldsymbol{k}_1,\ldots,\boldsymbol{k}_n\rangle\equiv \mathrm{i}T_{fi}\,,$$

where we used also $S=1+\mathrm{i}T$ and the fact that we neglected disconnected parts. Finally, we define the Fourier transformed n-point function as

$$G(x_1,\ldots,x_n)=\int\prod_{i=1}^{n}\frac{\mathrm{d}^4k_i}{(2\pi)^4}\ \exp\left(-\mathrm{i}\sum_i k_ix_i\right)G(k_1,\ldots,k_n)\,. \tag{9.36}$$

Then we obtain the LSZ reduction formula in momentum space,

$$\begin{aligned}\mathrm{i}T_{fi}&=\mathrm{i}^{n+m}N_{\boldsymbol{k}_1}\cdots N_{\boldsymbol{k}_n}N_{\boldsymbol{k}_1'}\cdots N_{\boldsymbol{k}_m'}(k_1^2-m^2)\cdots(k_n^2-m^2)(k_1'^2-m^2)\cdots(k_m'^2-m^2)\\ &\times G(k_1,\cdots,k_n,-k_1',\ldots,-k_m')\,.\end{aligned} \tag{9.37}$$

The Green function $G(k_1,\cdots,k_n,-k_1',\ldots,-k_m')$ is multiplied by zeros, since the external particles satisfy $k^2=m^2$. Thus T_{fi} vanishes, except when poles $1/(k^2-m^2)$ of $G(k_1,\cdots,k_n,-k_1',\ldots,-k_m')$ cancel these zeros. In the case of external scalar particles, only their normalisation factors are left. As they are not essential for the calculation of the transition amplitudes, one include these normalisation factors into the phase space of final state particle and in the flux factor of initial particles. This explains our Feynman rule to replace the scalar propagator by one for amplitudes in momentum space.

The derivation of the LSZ formula for particles with spin $s>0$ proceeds in the same way. Their wave-functions contain additionally polarisation vectors $\varepsilon^\mu(k)$, tensors $\varepsilon^{\mu\nu}(k)$, or spinors $u(p)$ and $\bar{u}(p)$ and their charge conjugated states. In the case of a photon (graviton), we have to add $\varepsilon^\mu(k)$ ($\varepsilon^{\mu\nu}(k)$) in the initial state and the complex conjugated $\varepsilon^{*\mu}(k)$ ($\varepsilon^{*\mu\nu}(k)$) in the final state. In the case of Dirac fermions, we have to assign four different spinors to the four possible combinations of particle and anti-particles in the initial and final state. Having chosen $u(p)$ as particle state with an arrow along the direction of time, the simple rule that a fermion line corresponds to a complex number $\bar{\psi}\cdots\psi$ fixes the designation of the other spinors as shown in Fig. 9.1: Connecting the upper fermion lines, $\bar{u}(p')\cdots u(p)$, corresponds to the scattering $e^-(p) + X \to e^-(p') + X'$, while $\bar{v}(q)\cdots u(p)$ describes the annihilation process $e^+(q) + e^-(p) \to X + X'$. Connecting the lower fermion lines, $\bar{v}(q)\cdots v(q)$, corresponds to the scattering $e^+(q) + X \to e^+(q') + X'$, while $\bar{u}(p')\cdots v(q)$ describes the pair creation process $X + X' \to e^+(q') + e^-(p')$. Recall that the Feynman amplitude $\mathcal{A}$ is defined omitting the normalisation factor $N_p = [2\omega_p(2\pi)^3]^{-1/2}$ from all wave-functions—the splitting of S-matrix elements into Feynman amplitudes and phase space is discussed in appendix 9.A.

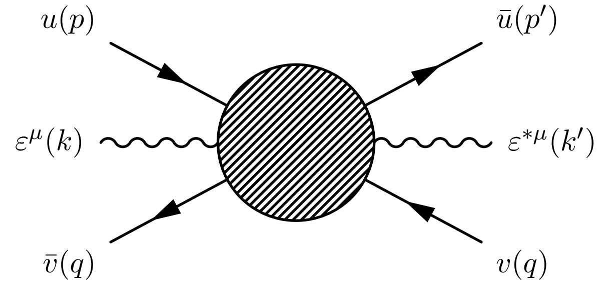


Figure 9.1.: Feynman rules for external particles in momentum space; initial state on the left, final state on the right.

Wave function renormalisation Up to now we have pretended that we can describe the fields in the initial and final state as free particles. Although e.g. Yukawa interactions between two, by assumption, widely separated particles at $t = \pm\infty$ are negligibly small, self-interactions persist. These interactions lead to a renormalisation of the external wave-functions.

We can rephrase the problem as follows: If the creation operator corresponds to the one of a free theory, $a_0^\dagger(\boldsymbol{k}, -\infty)$, then it can only connect one-particle states with the vacuum. In contrast, the interacting field can also connect many-particle states to the vacuum and therefore its overlap with single-particle states is reduced,

$$a^\dagger(\boldsymbol{k}, -\infty)\,|0\rangle = \sqrt{Z}\,|\boldsymbol{k}\rangle + \sqrt{1-Z}\left\{|\boldsymbol{k}', \boldsymbol{k}'', \boldsymbol{k}'''\rangle + \ldots\right\} \tag{9.38a}$$

$$= \sqrt{Z} a_0^\dagger(\boldsymbol{k}_1)\,|0\rangle + \sqrt{1-Z}\left\{|\boldsymbol{k}', \boldsymbol{k}'', \boldsymbol{k}'''\rangle + \ldots\right\} \,.. \tag{9.38b}$$

Therefore the free and the interacting fields are connected by

$$\phi(x) \to \sqrt{Z}\phi_0(x) \tag{9.39}$$

for $t \to \pm\infty$, where we call the factor Z the wave-function (or the field-strength) renormalisation constant. We will show in section 11.4.2 that this factor can be extracted from the self-energy diagrams of the corresponding field. More precisely, including a factor $\sqrt{Z_k}$ for each external line of type k takes into account self-energy corrections in the external lines. Therefore it is enough to calculate the self-energy and to extract $\sqrt{Z_k}$ once; after that we can omit self-energy corrections in the external lines adding simply factors $\sqrt{Z_k}$. Finally, note that we can set $Z = 1$ in tree-level processes, since in perturbation theory $Z = 1 + \mathcal{O}(g)$ holds.

9.3. Specific processes

We consider now in detail a few specific processes. First, we derive the Klein-Nishina formula for the Compton scattering cross section using the standard "trace method". Then we calculate polarised $e^+e^- \to \mu^+\mu^-$ and $e^+e^- \to \gamma\gamma$ scattering applying helicity methods.

9.3.1. Trace method and Compton scattering

Matrix element The Feynman amplitude $\mathcal{A}$ of Compton scattering $e^-(p) + \gamma(k) \to e^-(p') + \gamma(k')$ at $\mathcal{O}(e^2)$ consists of the two diagrams shown in Fig. 9.2 and is given by

$$\mathrm{i}\mathcal{A} = -\mathrm{i}e^2 \bar{u}(p') \left[\not{\varepsilon}^{*\prime} \, \frac{\not{p} + \not{k} + m}{(p+k)^2 - m^2} \, \not{\varepsilon} + \not{\varepsilon} \, \frac{\not{p} - \not{k}' + m}{(p-k')^2 - m^2} \, \not{\varepsilon}^{*\prime} \right] u(p) \,. \tag{9.40}$$

Since the denominator is a non-zero light-like vector, we have omitted the iε. Note that the two amplitudes can be transformed into each other replacing $\varepsilon \leftrightarrow \varepsilon^{*\prime}$ and $k \leftrightarrow -k'$. This symmetry called crossing symmetry relates processes where a particle is replaced by an anti-particle with negative momentum on the other side of the reaction.

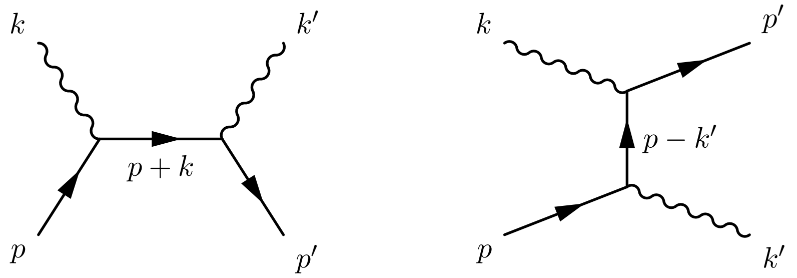


Figure 9.2.: The two Feynman diagrams contributing to Compton scattering at $\mathcal{O}(e^2)$.

We evaluate the process in the rest-frame of the initial electron. Then $p^\mu = (m, \mathbf{0})$ and choosing $\varepsilon^\mu = (0, \boldsymbol{\varepsilon})$ as well as $\varepsilon'^\mu = (0, \boldsymbol{\varepsilon}')$, it follows

$$p \cdot \varepsilon = p \cdot \varepsilon' = 0 \,. \tag{9.41}$$

Moreover, the photons are transversely polarised,

$$k \cdot \varepsilon = k' \cdot \varepsilon' = 0 \,, \tag{9.42}$$

and we choose real polarisation vectors. We anti-commute $\not{p}$ in the numerator to the right, $\not{p}\not{\varepsilon}' = 2p \cdot \varepsilon' - \not{\varepsilon}'\not{p} = -\not{\varepsilon}'\not{p}$ and use the Dirac equation, $\not{p}u(p) = mu(p)$. Then we simplify also the denominator using $p^2 = m^2$ and obtain

$$\mathcal{A} = -e^2 \bar{u}(p') \left[\frac{\not{\varepsilon}'\not{k}\not{\varepsilon}}{2p\cdot k} + \frac{\not{\varepsilon}\not{k}'\not{\varepsilon}'}{2p\cdot k'} \right] u(p)\,. \tag{9.43}$$

Typically the electron target is not polarised, and the spin of the final electron is not measured. Thus we sum the squared matrix element over the final and average over the initial electron spin,

$$|\overline{\mathcal{A}}|^2 = \frac{1}{2}\sum_{s,s'} |\mathcal{A}|^2 = \frac{e^4}{2}\sum_{s,s'} \left| \bar{u}(p') \left(\frac{\not{\varepsilon}'\not{k}\not{\varepsilon}}{2p\cdot k} + \frac{\not{\varepsilon}\not{k}'\not{\varepsilon}'}{2p\cdot k'} \right) u(p) \right|^2 . \tag{9.44}$$

Calculating $|\mathcal{A}|^2 = \mathcal{A}\mathcal{A}^*$ requires the knowledge of $\mathcal{A}^* = \mathcal{A}^\dagger$. Recall that we defined $\bar{\psi} = \psi^\dagger\gamma^0$ such that a general amplitude $\mathcal{A}$ composed of spinors,

$$\mathcal{A} = \bar{\psi}(p')\Gamma\psi(p) = \psi^\dagger(p')\gamma^0\Gamma\psi(p) \tag{9.45}$$

with Γ denoting a product of the basis elements given in Eq. (8.40), becomes

$$\mathcal{A}^* = \bar{\psi}(p)\gamma^0\Gamma^\dagger\gamma^0\psi(p') \equiv \bar{\psi}(p)\overline{\Gamma}\psi(p')\,. \tag{9.46}$$

Important special cases worth to memorise are $\overline{\gamma^\mu} = \gamma^\mu$, $\overline{\gamma^5} = -\gamma^5$, and $\overline{\not{a}\not{b}\cdots\not{c}} = \not{c}\cdots\not{b}\not{a}$. We now write out the spinor indices,

$$|\overline{\mathcal{A}}|^2 = \frac{e^4}{2}\sum_{s,s'} \bar{u}_a(p') \left[\frac{\not{\varepsilon}'\not{k}\not{\varepsilon}}{2p\cdot k} + \frac{\not{\varepsilon}\not{k}'\not{\varepsilon}'}{2p\cdot k'} \right]_{ab} u_b(p)\bar{u}_c(p) \left[\frac{\not{\varepsilon}\not{k}\not{\varepsilon}'}{2p\cdot k} + \frac{\not{\varepsilon}'\not{k}'\not{\varepsilon}}{2p\cdot k'} \right]_{cd} u_d(p')\,. \tag{9.47}$$

Using the property (8.56) of the Dirac spinors, $\sum_s u_a(p,s)\bar{u}_b(p,s') = (\not{p}+m)_{ab}$, we obtain

$$|\overline{\mathcal{A}}|^2 = \frac{e^4}{2}\,[\not{p}'+m]_{da} \left[\frac{\not{\varepsilon}'\not{k}\not{\varepsilon}}{2p\cdot k} + \frac{\not{\varepsilon}\not{k}'\not{\varepsilon}'}{2p\cdot k'} \right]_{ab} [\not{p}+m]_{bc} \left[\frac{\not{\varepsilon}\not{k}\not{\varepsilon}'}{2p\cdot k} + \frac{\not{\varepsilon}'\not{k}'\not{\varepsilon}}{2p\cdot k'} \right]_{cd} . \tag{9.48}$$

Since $\not{p}+m$ combines one spinor in $\mathcal{A}$ and one in $\mathcal{A}^*$, the result is a trace over gamma matrices,

$$|\overline{\mathcal{A}}|^2 = \frac{e^4}{8}\mathrm{t}r \left\{ (\not{p}'+m) \left[\frac{\not{\varepsilon}'\not{k}\not{\varepsilon}}{p\cdot k} + \frac{\not{\varepsilon}\not{k}'\not{\varepsilon}'}{p\cdot k'} \right] (\not{p}+m) \left[\frac{\not{\varepsilon}\not{k}\not{\varepsilon}'}{p\cdot k} + \frac{\not{\varepsilon}'\not{k}'\not{\varepsilon}}{p\cdot k'} \right] \right\} . \tag{9.49}$$

Working out some more examples of this type (e.g. in problem 9.??), you should convince yourself that each fermion line in $\mathcal{A}$ is converted into a trace in $|\overline{\mathcal{A}}|^2$. Useful identities for the evaluation of such traces are given in the appendix A.2.

We simplify this trace by anti-commuting identical variables, such that they become neighbours. Then we can use $\not{a}\not{a} = a^2$ and reduce thereby the number of gamma matrices in each step by two. Multiplying out the terms in the trace, we obtain three contributions that we denote by

$$\mathrm{t}r\,\{\quad\} = \frac{S_1}{(p\cdot k)^2} + \frac{S_2}{(p\cdot k')^2} + \frac{2S_3}{(p\cdot k)\,(p\cdot k')}\,. \tag{9.50}$$

We consider only the first term S_1 in detail. Starting from

$$\begin{aligned} S_1 &= \mathrm{t}r\left\{ \left(\not{p}'+m\right) \not{\varepsilon}'\not{\varepsilon}\not{k}\,(\not{p}+m)\,\not{k}\not{\varepsilon}\not{\varepsilon}' \right\} = \mathrm{t}r\{\not{p}'\not{\varepsilon}'\not{\varepsilon}\, \underbrace{\not{k}\not{p}}_{2kp-\not{p}\not{k}}\, \not{k}\not{\varepsilon}\not{\varepsilon}'\} + m^2\mathrm{t}r\{\not{\varepsilon}'\not{\varepsilon}\, \underbrace{\not{k}\not{k}}_{k^2=0}\, \not{\varepsilon}\varepsilon'\} = \\ &= 2\,(k\cdot p)\,\mathrm{t}r\ \{\not{p}'\not{\varepsilon}'\not{\varepsilon}\not{k}\not{\varepsilon}\not{\varepsilon}'\} - \mathrm{t}r\ \{\not{p}'\not{\varepsilon}'\not{\varepsilon}\not{p}\not{k}\not{k}\not{\varepsilon}\not{\varepsilon}'\} \end{aligned} \tag{9.51}$$

we arrive at an expression with only six gamma matrices. We continue the work,

$$S_1 = 2\,(k\cdot p)\,\mathrm{t}r\ \left\{\not{p}'\not{\varepsilon}'\not{\varepsilon}\not{k}\not{\varepsilon}\not{\varepsilon}'\right\} = -2\,(k\cdot p)\,\mathrm{t}r\left\{\not{p}'\not{\varepsilon}'\not{k}\not{\varepsilon}\not{\varepsilon}\not{\varepsilon}'\right\} = \tag{9.52a}$$

$$= 2\,(k\cdot p)\,\mathrm{t}r\left\{p'\not{\varepsilon}'\not{k}\not{\varepsilon}'\right\} = 2\,(k\cdot p)\left[2\left(k\cdot\varepsilon'\right)\mathrm{t}r\ \left(\not{p}'\not{\varepsilon}'\right) - \mathrm{t}r\ \left\{\not{p}'\not{\varepsilon}'\not{\varepsilon}'\not{k}\right\}\right] = \tag{9.52b}$$

$$= 8\,(k\cdot p)\left[2\left(k\cdot\varepsilon'\right)\left(p'\cdot\varepsilon'\right) + \left(p'\cdot k\right)\right]\,. \tag{9.52c}$$

where we have used $\not{\varepsilon}\not{\varepsilon} = \not{\varepsilon}'\not{\varepsilon}' = -1$. We want to eliminate as next step the two scalar products that include p'. Four-momentum conservation implies $(p'-k)^2 = (p-k')^2$ and thus

$$p'\cdot k = p\cdot k'\,. \tag{9.53}$$

Multiplying the four-momentum conservation equation by ε', it follows moreover

$$p+k = p'+k' \quad\Rightarrow\quad \underbrace{\varepsilon'\cdot p}_{0} + \varepsilon'\cdot k = \varepsilon'\cdot p' + \underbrace{\varepsilon'\cdot k'}_{0}\,. \tag{9.54}$$

Thus our final result for S_1 is

$$S_1 = 8\,(k\cdot p)\left[2\left(k\cdot\varepsilon'\right)^2 + k'\cdot p\right]\,. \tag{9.55}$$

S_2 can be obtained observing the crossing symmetry of the amplitude by the replacements $\varepsilon \leftrightarrow \varepsilon'$ and $k \leftrightarrow -k'$. The cross term S_3 has to be calculated and we give here only the final result for the combination of the three terms, where some terms cancel

$$\left|\overline{\mathcal{A}}\right|^2 = e^4\left[\frac{\omega'}{\omega} + \frac{\omega}{\omega'} + 4(\varepsilon\cdot\varepsilon')^2 - 2\right]\,. \tag{9.56}$$

Cross section To obtain the cross section, we have to calculate the flux factor and to perform the integration over the phase space of the final state,

$$\mathrm{d}\sigma = \frac{1}{4I}\,(2\pi)^4\,\delta^{(4)}(P_i - P_f)|\mathcal{A}_{fi}|^2\prod_{f=1}^{n}\frac{\mathrm{d}^3p_f}{2E_f(2\pi)^3} = \frac{1}{4I}\,|\mathcal{A}_{fi}|^2\,\mathrm{d}\Phi^{(n)}\,, \tag{9.57}$$

with the final state phase space $\mathrm{d}\Phi^{(n)}$. The flux factor I in the rest system of the electron is simply

$$I \equiv v_{\mathrm{rel}}\,p_1\cdot p_2 = m\omega\,. \tag{9.58}$$

Using Eq. (**??**),

$$\frac{\mathrm{d}^3p'}{2E'} = \mathrm{d}^4p'\,\delta^{(4)}(p'^2 - m^2)\vartheta(p_0')\,, \tag{9.59}$$

the phase space integration becomes

$$\mathrm{d}\Phi^{(2)} = \frac{1}{(2\pi)^2}\int\mathrm{d}\Omega_{k'}\frac{|\boldsymbol{k'}|^2\mathrm{d}k'}{2|\boldsymbol{k'}|}\int\frac{\mathrm{d}^3p'}{2E'}\,\delta^{(4)}\left(p'+k'-p-k\right) = \tag{9.60}$$

$$= \frac{1}{8\pi^2}\int\mathrm{d}\Omega_{k'}|\boldsymbol{k'}|\mathrm{d}k'\,\delta\left(\left(p+k-k'\right)^2 - m^2\right)\,. \tag{9.61}$$

The argument of the delta function is

$$\left(p+k-k'\right)^2 - m^2 = m^2 + 2p\cdot k - 2p\cdot k' - 2k\cdot k' - m^2 \tag{9.62a}$$

$$= 2m\left(\omega - \omega'\right) - 2\omega\omega'\left(1-\cos\vartheta\right) \equiv f(\omega')\,. \tag{9.62b}$$

In order to evaluate the delta function we have to determine the derivative $f'(\omega')$,

$$f'(\omega') = -2m - 2\omega\,(1-\cos\vartheta)\ , \tag{9.63}$$

and the zeros of $f(\omega')$,

$$0 = 2m\left(\omega - \omega'\right) - 2\omega\omega'\,(1-\cos\vartheta)\ . \tag{9.64}$$

Solving for ω' gives $\omega'\left[\omega\,(1-\cos\vartheta) + m\right] = m\omega$ and

$$\omega' = \frac{\omega}{1+\frac{\omega}{m}\,(1-\cos\vartheta)}\ . \tag{9.65}$$

This is the famous relation for the frequency shift of a photon found first experimentally in the scattering of X-rays on electrons by Compton 1921. The observed energy change of photons was crucial in accepting the quantum nature ("particle-wave duality") of photons. Combining everything, we obtain

$$\mathrm{d}\Phi^{(2)} = \frac{1}{8\pi^2}\int \mathrm{d}\Omega_{k'}|\omega'|\mathrm{d}\omega'\;\delta\left(2m\left(\omega-\omega'\right) - 2\omega\omega'(1-\cos\vartheta)\right) \tag{9.66a}$$

$$= \frac{1}{8\pi^2}\int \mathrm{d}\Omega_{k'}\,\frac{\omega'}{2m\left[1+\frac{\omega}{m}\,(1-\cos\vartheta)\right]} = \frac{1}{16\pi^2\,m\omega}\int \mathrm{d}\Omega_{k'}\,\omega'^2 \tag{9.66b}$$

and thus as differential Klein-Nishina cross section

$$\frac{\mathrm{d}\sigma}{\mathrm{d}\Omega} = \frac{1}{4m\omega}\,\frac{\omega'^2}{16\pi^2\,m\omega}\,\left|\overline{\mathcal{A}}\right|^2 = \frac{\alpha^2}{4m^2}\frac{\omega'^2}{\omega^2}\left[\frac{\omega'}{\omega} + \frac{\omega}{\omega'} + 4(\varepsilon\cdot\varepsilon')^2 - 2\right] \tag{9.67}$$

with $\alpha \equiv e^2/(4\pi)$. For scatterings in the forward direction, $\vartheta \to 0$ and thus $\omega' \to \omega$, the scattered photon retains (in the lab frame) its energy even in the ultra-relativistic limit $\omega \gg m$. The same holds in the classical limit, $\omega \ll m$, but now for all directions. Thus we obtain as classical limit of the Klein-Nishina formula the polarised Thomson cross section

$$\frac{\mathrm{d}\sigma}{\mathrm{d}\Omega} \approx \frac{\alpha^2}{m^2}\left(\varepsilon\cdot\varepsilon'\right)^2 = r_0^2\left(\varepsilon\cdot\varepsilon'\right)^2\ , \tag{9.68}$$

with $r_0 = \alpha/m$ as the classical electron radius.

Averaging and summing over the photon polarisation vectors is simplest, if we choose the angle between $\boldsymbol{\varepsilon}$ and $\boldsymbol{\varepsilon}'$ as ϑ. Then

$$\sum_{r,r'}\left(\varepsilon\cdot\varepsilon'\right)^2 = 1+\cos^2\vartheta\ . \tag{9.69}$$

The integration over the scattering angle ϑ can be done analytically. We use $x = \cos\vartheta$ and set $\tilde\omega \equiv \omega/m$,

$$\sigma = \frac{\pi\alpha^2}{m^2}\int_{-1}^{1}\mathrm{d}x\left[\frac{1}{[1+\tilde\omega(1-x)]^3} + \frac{1}{1+\tilde\omega(1-x)} - \frac{1-x^2}{[1+\tilde\omega(1-x)]^2}\right] \tag{9.70a}$$

$$= \frac{\pi\alpha^2}{2m^2}\left\{\frac{1+\tilde\omega}{\tilde\omega^3}\left[\frac{2\tilde\omega(1+\tilde\omega)}{(1+2\tilde\omega} - \ln(1+2\tilde\omega)\right] + \frac{\ln(1+2\tilde\omega)}{2\tilde\omega} - \frac{1+3\tilde\omega}{(1+2\tilde\omega)^2}\right\}\ . \tag{9.70b}$$

Since in the electron rest frame $s = (p+k)^2 = m^2 + 2m\omega = m^2(1+2\tilde\omega)$, we can use $\tilde\omega = (s/m^2-1)/2$ to express σ in an explicit Lorentz invariant form.

Approximations for the non-relativistic and the ultra-relativistic limit are

$$\sigma = \sigma_{\mathrm{T}h} \times \begin{cases} 1 - 2\tilde{\omega} + \mathcal{O}(\tilde{\omega}^2) & \text{for } \tilde{\omega} \ll 1\,, \\ \frac{3}{8\tilde{\omega}}\left(\ln(2\tilde{\omega}) + \frac{1}{2}\right) + \mathcal{O}(\tilde{\omega}^{-2}) & \text{for } \tilde{\omega} \gg 1\,, \end{cases} \tag{9.71}$$

where the Thomson cross section is given by $\sigma_{\mathrm{T}h} = 8\pi\alpha^2/(3m^2)$. These approximations are shown together with the exact result in the left panel of Fig. 9.3. In the ultra-relativistic limit $s \gg m^2$, the total cross section for Compton scattering decreases as $\sigma \propto 1/s$. On the other hand, the differential cross section in the forward direction is constant. As a result, the relative importance of the forward region $\vartheta \sim 0$ increases for increasing s: While $\mathrm{d}\sigma/\mathrm{d}x$ is symmetric around $x = 0$ in the classical limit $\omega \to 0$, it becomes more and more asymmetric with a a shrinking peak around the forward region at $\vartheta \sim 0$, cf. the right panel of Fig. 9.3.

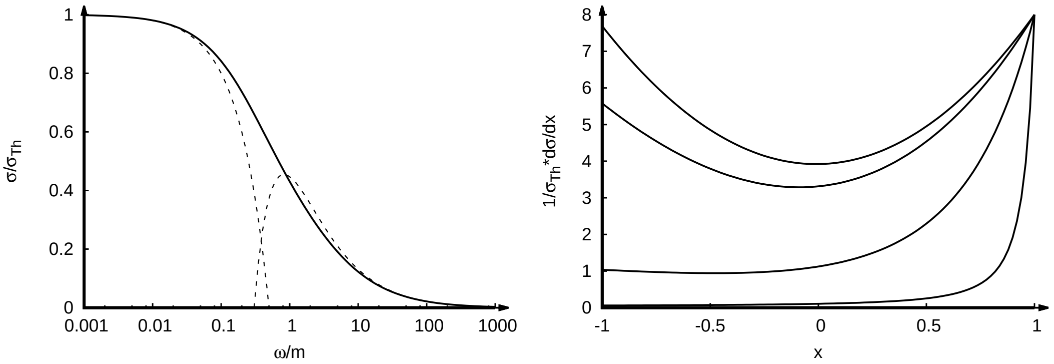


Figure 9.3.: Left: The total cross section $\sigma/\sigma_{\mathrm{T}h}$ as function of $\tilde{\omega}$ together with the classical and ultra-relativistic limits given in Eq. (9.71). Right: The normalised differential cross section $\sigma_{\mathrm{T}h}^{-1}\,\mathrm{d}\sigma/\mathrm{d}x$ as function of $x = \cos\vartheta$ for $\tilde{\omega} = 0.01$, 0.1, 1, and 10 (from top to down).

Crossing symmetry We noticed that the two amplitudes in Compton scattering can be transformed into each other replacing $\varepsilon \leftrightarrow \varepsilon^{*\prime}$ and $k \leftrightarrow -k'$. This is an example of a general symmetry of relativistic quantum field theories called crossing symmetry. Using the Feynman rules for in and out particles, it follows that matrix elements where an in-going particle is replaced by an out-going anti-particle or vice versa are related by the following substitutions,

- exchange the momentum $k \leftrightarrow -k'$;
- exchange particle and anti-particle wave functions; thus in momentum space, $1 \leftrightarrow 1$ for spinless particles, $\varepsilon \leftrightarrow \varepsilon^{*\prime}$ for spin-1 and $u \leftrightarrow v$ for fermions.
- multiply by -1 for each exchanged fermion pair.

The additional minus for fermions is required, because the spin sums of fermions and antifermions are related by

$$\sum_s u(p,s)\bar{u}(p,s) = (\not{p} + m) = -\left(\not{p}' - m\right) = -\sum_s v(p',s)\bar{v}(p',s)\,. \tag{9.72}$$

Note that this symmetry allows us to obtain the matrix elements of different processes: For instance, we can relate the processes $e^-e^+ \to \mu^-\mu^+$ with $e^-\mu^- \to e^-\mu^-$ and $\mu^-\mu^+ \to e^-e^+$.

Using a more formal approach, one can derive the crossing symmetry not relying on perturbation theory and the Feynman rules, but using the analytical properties of S-matrix elements: The LSZ reduction formula distinguishes in and out particles only by the sign of the momenta used in the Fourier transformation. If one can analytically continue the residue of a pole in an S-matrix element from p^0 to $-p^0$, then one converts the S-matrix for a particle with $\phi(p)$ into the one for an anti-particle with $\phi^*(-p)$. Remarkably, the basic properties of a relativistic quantum field theory, locality and causality, are sufficient to proof that this analytical continuation is possible.

Finally, note that the factor -1 for each exchanged external fermion pair implies a relative minus sign for diagrams connected by crossing which contribute to the same process. Thus there is a relative minus sign between e.g. the t and the u channel diagrams for $e^-e^- \to e^-e^-$ scattering.

9.3.2. Helicity method and polarised QED processes

Using the trace method, the number of terms that have to be calculated grows as $\sim n^2$ with the number n of diagrams. For large n, it should be therefore favourable to calculate the amplitude $\mathcal{A}(s_1, \dots, s_f)$ for fixed polarisations s_i of the external particles: The amplitude is a complex number and can be trivially squared. An efficient way to calculate polarised amplitudes uses helicity spinors, an approach used also in most modern computer programs for the calculation of scattering processes.

Massless fermions We restrict our short introduction into helicity methods to massless particles. In the case of fermions, we know that then the use of Weyl spinors in the chiral representation is most convenient,

$$u_L(p) = \begin{pmatrix} \phi_L(p) \\ 0 \end{pmatrix} \quad \text{and} \quad u_R(p) = \begin{pmatrix} 0 \\ \phi_R(p) \end{pmatrix} . \tag{9.73}$$

We do not need to consider $v_{L,R}(p)$, since they correspond to particle spinors of opposite helicity, $u_{R,L}(p)$. Moreover, two out of the fours possible scalar products involving $u_{L,R}$ are zero for massless fermions,

$$\bar{u}_L(p)u_L(q) = \bar{u}_R(p)u_R(q) = 0 . \tag{9.74}$$

This motivates us to introduce a bracket notation for the helicity spinors as follows

$$\bar{u}_L(p) = \langle p \;, \quad \bar{u}_R(p) = [p \;, \quad u_L(p) = p] \;, \quad u_R(p) = p\rangle . \tag{9.75}$$

We call the quantities on the RHS *angle* and *square brackets*. The only non-zero Lorentz-invariant spinor products are given by a pair of brackets of the same type,

$$\bar{u}_L(p)u_R(q) = \langle pq\rangle \qquad \text{and} \qquad \bar{u}_R(p)u_L(q) = [pq] . \tag{9.76}$$

Next we consider the tensor product of the spinors,

$$p\rangle[p = u_R(p)\bar{u}_R(p) = P_R\not{p} \;, \quad \text{and} \quad p]\langle p = u_L(p)\bar{u}_L(p) = P_L\not{p} . \tag{9.77}$$

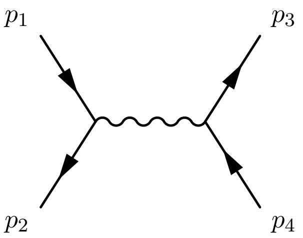


Figure 9.4.: Feynman diagrams for the process $e^-e^+ \to \mu^+\mu^-$.

These identities connect the massless spinors $p\rangle$ and $[p$ to the light-like four-vector p^μ.

We are now in position to derive some basic properties of the brackets. First, we can connect the two types of spinor products as

$$\langle pq\rangle = \bar{u}_L(p)u_R(q) = [\bar{u}_R(q)u_L(p)]^* = [qp]^* \,. \tag{9.78}$$

Multiplying then $p\rangle[p$ and $q]\langle q$ and taking the trace gives

$$\langle pq\rangle[qp] = \mathrm{tr}\{P_R\not{q}P_L\not{p}\} = \mathrm{tr}\{P_R\not{q}\not{p}\} = 2p\cdot q\,, \tag{9.79}$$

so that

$$|\langle pq\rangle|^2 = |[qp]|^2 = 2p\cdot q\ . \tag{9.80}$$

Next, we express the spinor products through Weyl spinors and use $u_R(p) = \mathrm{i}\sigma^2 u_L^*(p)$,

$$\langle pq\rangle = \phi_L^\dagger(p)\phi_R(q) = \phi_{La}^*(p)(\mathrm{i}\sigma^2)_{ab}\phi_{Lb}^*(q)\,. \tag{9.81}$$

Then the antisymmetry of $(\mathrm{i}\sigma^2)_{ab} = \varepsilon_{ab}$ implies

$$\langle pq\rangle = -\langle qp\rangle \quad \text{and} \quad [pq] = -[qp]\,. \tag{9.82}$$

Thus the brackets are square roots of the corresponding Lorentz vector products which are antisymmetric in their two arguments. Finally, we note that the Fierz identity applied to the sigma matrices (cf. with problem 8.??),

$$(\bar{\sigma}^\mu)_{ab}(\bar{\sigma}_\mu)_{cd} = 2(\mathrm{i}\sigma^2)_{ac}(\mathrm{i}\sigma^2)_{bd}\,, \tag{9.83}$$

allows the simplification of contracted spinor expressions,

$$\langle p\gamma^\mu q]\langle k\gamma_\mu \ell] = 2\langle pk\rangle[\ell q]\,, \qquad \langle p\gamma^\mu q][k\gamma_\mu \ell\rangle = 2\langle p\ell\rangle[kq]\,. \tag{9.84}$$

$e^-e^+ \to \mu^-\mu^+$ scattering It is now time to apply this new “bracket” formalism. We consider the tree-level amplitude for, e.g., $e_L^-(1)e_R^+(2) \to \mu_L^-(3)\mu_R^+(4)$ in QED, given by the single diagram shown in Fig. 9.4. As it is standard using this formalism, we consider all momenta as outgoing. Then the amplitude is

$$\mathrm{i}\mathcal{A} = (-\mathrm{i}e)^2\frac{-\mathrm{i}}{q^2}\,\bar{u}_L(3)\gamma^\mu u_L(4)\,\bar{u}_L(2)\gamma_\mu u_L(1) \tag{9.85a}$$

$$= \frac{\mathrm{i}e^2}{q^2}\langle 3\gamma^\mu 4]\langle 2\gamma_\mu 1] = \frac{2\mathrm{i}e^2}{q^2}\langle 32\rangle[14]\,, \tag{9.85b}$$

where we have employed the Fierz identity (9.83) in the last step. Since $\langle 32\rangle$ and $[14]$ are both square roots of

$$u = (k_2 + k_3)^2 = (k_1 + k_4)^2\,, \tag{9.86}$$

we can replace them by the Mandelstam invariant u. We consider the process in the cm frame of the e^+e^- pair. With $u = -2E^2(1 + \cos\vartheta)$, and $q^2 = s = 4E^2$, the amplitude becomes

$$|\mathrm{i}\mathcal{A}|^2 = e^4(1 + \cos\vartheta)^2\,. \tag{9.87}$$

You should re-derive this result using the more familiar trace formalism and compare the amount of algebra required in the two approaches (problem 9.??).

Massless gauge bosons In the next step, we incorporate massless gauge bosons as the photon into this framework. We claim that the polarisation vectors of a massless vector boson in the final-state can be represented as

$$\epsilon_R^{*\mu}(k) = \frac{1}{\sqrt{2}}\frac{\langle r\gamma^\mu k]}{\langle rk\rangle}\,, \qquad \epsilon_L^{*\mu}(k) = -\frac{1}{\sqrt{2}}\frac{[r\gamma^\mu k\rangle}{[rk]}\,. \tag{9.88}$$

Here, k is the momentum of the vector boson, and r is a fixed light-like 4-vector, called the reference vector, which is assumed to be not collinear with k.

Now we show that this definition makes sense: First, we note that the vectors satisfy $[\varepsilon_R^*(k)]^* = \varepsilon_L^*(k)$. One can also check that the polarisation vectors are correctly normalised. Moreover, the Dirac equation, $\not{k}\, k] = 0$, guaranties that the polarisation vectors (9.88) are transverse,

$$k_\mu \varepsilon_{R,L}^{*\mu}(k) = 0\,. \tag{9.89}$$

Finally, we have to show that a change from one reference vector r to another light-like vector s corresponds to a gauge transformation and thus does not affect physics. The change of a polarisation vector under a change of reference vector $r \to s$ is

$$\varepsilon_R^{*\mu}(k;r) - \varepsilon_R^{*\mu}(k;s) = \frac{1}{\sqrt{2}}\left(\frac{\langle r\gamma^\mu k]}{\langle rk\rangle} - \frac{\langle s\gamma^\mu k]}{\langle sk\rangle}\right) \tag{9.90a}$$

$$= \frac{1}{\sqrt{2}}\frac{1}{\langle rk\rangle\langle sk\rangle}\big\{ - \langle r\gamma^\mu k]\langle ks\rangle + \langle s\gamma^\mu k]\langle kr\rangle\big\}\,. \tag{9.90b}$$

Now we use first the tensor products (9.77), and then the antisymmetry of the brackets,

$$\varepsilon_R^{*\mu}(k;r) - \varepsilon_R^{*\mu}(k;s) = \frac{1}{\sqrt{2}}\frac{1}{\langle rk\rangle\langle sk\rangle}\big\{ - \langle r\gamma^\mu \not{k} s\rangle + \langle s\gamma^\mu \not{k} r\rangle\big\} \tag{9.91a}$$

$$= \frac{1}{\sqrt{2}}\frac{1}{\langle rk\rangle\langle sk\rangle}\big\{\langle s(\not{k}\gamma^\mu + \gamma^\mu\not{k})r\rangle\big\} \tag{9.91b}$$

$$= \frac{1}{\sqrt{2}}\frac{1}{\langle rk\rangle\langle sk\rangle}\langle sr\rangle\; 2k^\mu\,. \tag{9.91c}$$

In the last line, we have applied the Clifford algebra of Dirac matrices. Thus the difference of the polarisation vectors induced by a change of the reference vector is a function proportional to the photon momentum,

$$\varepsilon_R^{*\mu}(k;r) - \varepsilon_R^{*\mu}(k;s) = f(r,s)k^\mu\,. \tag{9.92}$$

Contracted into an on-shell amplitude, $\mathcal{A} = \varepsilon^\mu\mathcal{A}_\mu$, current conservation implies that this expression vanishes. Thus we can use the most convenient reference vector s which can be chosen differently in any gauge-invariant subset of Feynman diagrams.

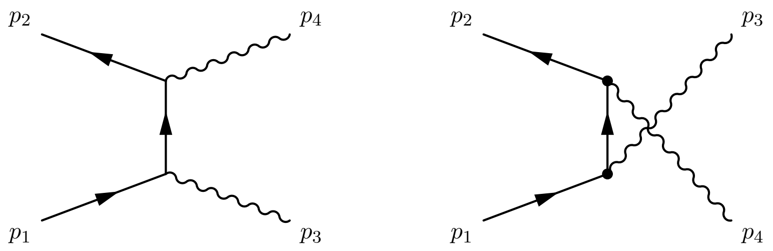


Figure 9.5.: Feynman diagrams for the process $e^-e^+ \to \gamma\gamma$.

$e^-e^+ \to \gamma\gamma$ scattering As second example, we consider now a scattering process with photons as external particles, $e^-e^+ \to \gamma\gamma$, as illustration for the use of the polarisation vectors. We label the momenta as in Fig. 9.5, taking all momenta as outgoing. Then the amplitude for this process is

$$\mathrm{i}\mathcal{A} = (-\mathrm{i}e)^2\langle 2 \left(\not{\varepsilon}(4)\frac{\mathrm{i}(\not{2}+\not{4})}{s_{24}}\not{\varepsilon}(3) + \not{\varepsilon}(3)\frac{\mathrm{i}(\not{2}+\not{3})}{s_{23}}\not{\varepsilon}(4) \right) 1] , \tag{9.93}$$

where we use the shorthand $\not{2}+\not{4}$ for $\not{k}_2+\not{k}_4$ and define $s_{ij}=(i+j)^2$.

There are four possible choices for the photon polarisations. Exchanging the momenta 3 and 4 relates the cases $\gamma_R\gamma_L$ and $\gamma_L\gamma_R$, while parity connects $\gamma_R\gamma_R$ and $\gamma_L\gamma_L$. We start showing that the latter two amplitudes are zero in the massless limit we consider. Considering $\gamma_R\gamma_R$, we choose as reference vector $r=2$ for both polarisation vectors,

$$\varepsilon^\mu(3) = \frac{1}{\sqrt{2}}\frac{[2\gamma^\mu 3\rangle}{[23]} , \qquad \varepsilon^\mu(4) = \frac{1}{\sqrt{2}}\frac{[2\gamma^\mu 4\rangle}{[24]} . \tag{9.94}$$

Inserting the polarisation vectors into Eq. (9.93), we obtain using the Fierz identity (9.83)

$$\langle 2\gamma^\mu\varepsilon_\mu(4) \propto \langle 2\gamma^\mu[2\gamma_\mu 4\rangle = 2\langle 22\rangle[4 = 0 . \tag{9.95}$$

In the last step, we used the antisymmetry of the brackets, $\langle 22\rangle = 0$. A similar cancellation occurs with $\varepsilon(3)$ and hence the entire matrix element vanishes. Parity implies then that the amplitude $\mathcal{A}(\gamma_L\gamma_L)$vanishes too. Alternatively, we can show the same cancellation using $r=1$ in both polarisation vectors.

Next we compute the amplitude for the case $\gamma_R\gamma_L$, choosing

$$\varepsilon^\mu(3) = \frac{1}{\sqrt{2}}\frac{[2\gamma^\mu 3\rangle}{\langle 23\rangle} \quad \text{and} \quad \varepsilon^\mu(4) = -\frac{1}{\sqrt{2}}\frac{[1\gamma^\mu 4\rangle}{[14]} . \tag{9.96}$$

Then the second diagram in Fig. 9.5 vanishes because of (9.95). Using the Fierz identity, the first diagram results in

$$\mathrm{i}\mathcal{A} = \frac{-\mathrm{i}e^2}{s_{24}}\frac{2\cdot 2}{(-2)\langle 23\rangle[14]}\langle 24\rangle[1(\not{2}+\not{4})2\rangle[31] . \tag{9.97}$$

Now we use the Dirac equation, $\not{2}2\rangle = 0$, and replace the vector $\not{4}_L$ by an angle bracket,

$$\mathrm{i}\mathcal{A} = \frac{2\mathrm{i}e^2}{s_{13}\langle 23\rangle[14]}\langle 24\rangle[14]\langle 42\rangle[31] = \frac{2\mathrm{i}e^2}{\langle 13\rangle[31]\langle 23\rangle}\langle 24\rangle\langle 42\rangle[31] = 2\mathrm{i}e^2\frac{\langle 24\rangle^2}{\langle 23\rangle^2} . \tag{9.98}$$

Finally, we introduce Mandelstam variables, $s_{23} = u$, $s_{13} = s_{24} = t$, reproducing the standard result

$$|\mathrm{i}\mathcal{A}|^2 = 4e^4 \, \frac{t}{u} = 4e^4 \, \frac{1-\cos\vartheta}{1+\cos\vartheta} \,. \tag{9.99}$$

This short introduction into helicity methods convinced you hopefully of their efficiency. The advantage of this method over the traditional trace method increases with the number of diagrams involved, since the step $\mathcal{A} \to |\mathcal{A}|^2$ is trivial in this approach. Massive fermions can be treated using the helicity states (A.17), and efficient extensions to massive gauge bosons exist.

9.4. Soft photons and gravitons

The addition of a vertex introduces typically a factor $\alpha/\pi \sim 0.2\%$ into a QED cross section. Thus one may hope that perturbation theory in QED converges, at least initially, reasonably fast. An exception to this rule is the emission of an additional soft or collinear photon from an external line shown in Fig. 9.6. The denominator of the additional propagator goes for $k \to 0$ to

$$\frac{1}{(p+k)^2 - m^2} \to \frac{1}{2p\cdot k} \sim \frac{1}{2E\omega(1-\cos\vartheta)} \,, \tag{9.100}$$

where we assumed $|\boldsymbol{p}| \gg m$ in the last step. Hence the denominator can blow up in two different limits: Firstly, in case of emission of soft photons, $\omega \to 0$. Secondly, in case of collinear emission of photons, $\vartheta \to 0$, if the mass of the emitting particle can be neglected. We have seen in the example of Compton scattering that both soft and collinear emission correspond to the classical limit.

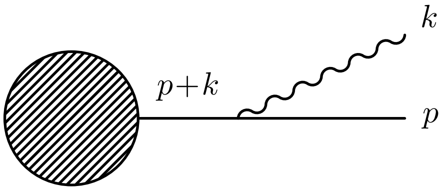


Figure 9.6.: Emission of an additional soft or collinear photon from an external line in the final state.

Universality and factorisation The fact that a photon sees in the soft limit $k \to 0$ a classical current should lead to considerable simplifications: In particular, interference effects should disappear and the amplitude $\mathcal{A}_{n+1}$ for the emission of an additional soft photon should factories into an universal factor $\varepsilon^\mu S_\mu$ and the amplitude $\mathcal{A}_n$ for the original process.

Let us start considering the emission of a soft photon by a spinless particle. If a scalar in the final state with momentum p and charge q emits a photon with momentum $k \to 0$, then[2]

$$\mathcal{A}_{n+1} = q \, \frac{\varepsilon_\mu(2p^\mu + k^\mu)}{(p-k)^2 - m^2 + \mathrm{i}\varepsilon} \, \mathcal{A}_n \to -q \, \frac{\varepsilon \cdot p}{p \cdot k - \mathrm{i}\varepsilon} \, \mathcal{A}_n \,. \tag{9.101}$$

[2] We use the Feynman rule for a $\phi\phi A^\mu$ vertex derived in problem 7.??.

For the emission of a soft photon from an initial state particle, the corresponding factor is $+q\,\varepsilon \cdot p/(p\cdot k + \mathrm{i}\varepsilon)$. In the case of an internal line, in general no factor $(p\cdot k)^{-1}$ appears for $k \to 0$, since the virtual particle is off-shell.

For a spin-1/2 particle in the initial state, the emission of a soft photon adds the factor

$$q\,\frac{\varepsilon_\mu \bar{u}(p,s)\gamma^\mu(\not{p}+\not{k}+m)}{(p-k)^2-m^2+\mathrm{i}\varepsilon} \to -q\,\frac{\varepsilon_\mu \bar{u}(p,s)\gamma^\mu(\not{p}+m)}{-2p\cdot k+\mathrm{i}\varepsilon} \tag{9.102}$$

to the amplitude $\mathcal{A}_n$. Now we replace $\not{p}+m$ by the spin sum $\sum_s u(p,s)\bar{u}(p,s)$, and use

$$\bar{u}(p,s)\gamma^\mu u(p,s') = 2E_p\,\frac{p^\mu}{p^0}\delta_{s,s'} = 2p^\mu\delta_{s,s'}\,. \tag{9.103}$$

This relation can be checked by direct calculation, or by noting that the current $j^\mu = \bar{u}(p,s)\gamma^\mu u(p,s)$ should become $j^\mu = (\rho, \rho\boldsymbol{v})$ in the classical limit $k \to 0$. Thus we obtain the same universal factor describing the emission of a soft photon,

$$S^\mu = -q\,\frac{p^\mu}{p\cdot k - \mathrm{i}\varepsilon}\,, \tag{9.104}$$

as in the case of a scalar. Moreover, we confirmed that the amplitude indeed factorises, $\mathcal{A}_{n+1} = \varepsilon_\mu S^\mu \mathcal{A}_n \equiv \varepsilon_\mu A^\mu_{n+1}$. If we allow for the emission of m soft photons from external particles with charge q_i, then

$$\mathcal{A}^{\mu_1\cdots\mu_m}_{n+m} \to \sum_{i=1}^m \frac{s_i q_i\,p^{\mu_m}}{p\cdot k + \mathrm{i}s_i\varepsilon}\,\mathcal{A}_n\,, \tag{9.105}$$

where the signs are $s_i = -1$ for an initial and $s_i = +1$ for a final state particle.

We have seen that the polarisation vector $\varepsilon_\mu(k)$ of a photon does not transform as a four-vector, cf. Eq. (7.23), but acquires a term proportional to k^μ. As we exploited already at various places, amplitudes containing polarisation vectors $\varepsilon_\mu(k)$ of external photons have to vanish therefore when contracted with k^μ. Thus Eq. (9.105) implies in the limit $k \to 0$

$$k_{\mu_j}\mathcal{A}^{\mu_1\cdots\mu_m}_{n+m} \to \sum_{i=1}^m s_i q_i\,\mathcal{A}_n = 0\,. \tag{9.106}$$

The prefactor of $\mathcal{A}_n$ is the total charge in the final state minus the total charge in the initial state. In order to obtain a Lorentz invariant matrix element for the soft emission of massless spin-1 particles, we have therefore to require that such particles couple to a conserved charge,

$$\sum_i q_i = \sum_f q_f\,.$$

Thus Lorentz invariance is sufficient to guaranty the conservation of the electromagnetic current in the low-energy limit. While this argument does not rely on gauge invariance, it tells us nothing about the behaviour of “hard” photons.

Spin $s > 1$ We can apply the same line of arguments to the emission of massless particles with spin $s > 1$. In the case of gravitons, $s = 2$, one finds that the universal factor becomes

$$S^{\mu\nu} = -f\,\frac{p^\mu p^\nu}{p\cdot k - \mathrm{i}s\varepsilon}\,, \tag{9.107}$$

where f denotes the coupling to the graviton. Requiring again that an amplitude containing the polarisation tensor $\varepsilon_{\mu\nu}(k)$ of external gravitons vanishes when contracted with k^μ gives the constraint

$$k_{\mu_j}\mathcal{A}^{\mu_1\cdots\mu_m}_{n+m} \to \sum_{i=1}^{m} s_i f_i p_i^\nu\,\mathcal{A}_n = 0\,. \tag{9.108}$$

Now the sum $\sum_i f_i p_i^\nu$ is conserved. For $f_i \neq f_j$, a linear combination of the individual four-momenta other than the total four-momentum would be conserved in the scattering process—a condition which is not possible to satisfy in a non-trivial scattering process. Thus we have to conclude that any massless $s = 2$ particle has a *universal* coupling to all types of particle. Recalling from (5.21) the form of the stress tensor, $T^{\mu\nu} = 2N^2p^\mu p^\nu$, we see moreover that a massless spin-2 particles couples with universal strength to the stress tensor, $S^{\mu\nu} \propto T^{\mu\nu}$. This result can be viewed as the basis of the weak equivalence principle. Going further to $s = 3$, the universal factor becomes $S^{\mu\nu\lambda} \propto p^\mu p^\nu p^\lambda$, requiring that sums quadratic in the momenta, $\sum_i \tilde{f}_i p_i^\nu p_i^\mu$, are conserved. This is not possible for any scattering angles except $\vartheta = 0$ and $180°$, and thus no consistent theory of interacting massless particles with spin $s \geq 2$ is possible.

Bremsstrahlung We discuss now as a concrete example the case of bremsstrahlung, i.e. the emission of a real photon in the scattering of a charged particle in the Coulomb field $A^0 = -Ze/(4\pi|\boldsymbol{x}|)$ of a static nuclei with charge Ze. The S-matrix element of this process is

$$\mathrm{i}S_{fi} = 2\pi\delta(E' + \omega - E')\,\frac{-Ze^3}{|\boldsymbol{q}|^2}\bar{u}(p')\left[\not{\varepsilon}\,\frac{\not{p}' + \not{k} + m}{2p'\cdot k}\,\gamma^0 + \gamma^0\,\frac{\not{p} - \not{k} + m}{-2p\cdot k}\not{\varepsilon}\right]u(p)\,, \tag{9.109}$$

where $1/|\boldsymbol{q}|^2$ is the Fourier transform of A^0. Note that the external field breaks translation invariance and the momentum is not conserved. We commute now $\not{p}$ and $\not{p}'$,

$$\mathrm{i}S_{fi} \propto e^2\bar{u}(p')\left[\frac{2\varepsilon\cdot p' - (\not{p}' - m)\not{\varepsilon} + \not{k}\not{\varepsilon}}{2p'\cdot k}\,\gamma^0 + \gamma^0\,\frac{2\varepsilon\cdot p - \not{\varepsilon}(\not{p} - m) + \not{k}\not{\varepsilon}}{-2p\cdot k}\right]u(p)\,, \tag{9.110}$$

such that we can use in the next step the Dirac equation. Neglecting additionally in the soft limit the $\not{k}$ term in the numerator, we find

$$\mathrm{i}S_{fi} \propto e^2\bar{u}(p')\gamma^0u(p)\,\left[\frac{\varepsilon\cdot p'}{p'\cdot k} - \frac{\varepsilon\cdot p}{p\cdot k}\right]\,. \tag{9.111}$$

As we have shown in the previous paragraph in general, the amplitude factorises into the amplitude describing the "hard" process and the universal correction term. The latter consists of the two terms expected for the emission of a soft photon from an initial line with momentum p and a final line with momentum p'. The probability $\mathcal{P}$ for the emission of an additional soft photon is given integrating the square bracket over the phase space,

$$\mathrm{d}\mathcal{P}_{n+1} = \frac{\mathrm{d}\sigma_{n+1}}{\mathrm{d}\sigma_n} = \left[\frac{\varepsilon\cdot p'}{p'\cdot k} - \frac{\varepsilon\cdot p}{p\cdot k}\right]^2\frac{\mathrm{d}^3k}{(2\pi)^3 2\omega_k} \propto \frac{\mathrm{d}\omega_k}{\omega_k}\,. \tag{9.112}$$

This probability diverges for $\omega \to 0$ and therefore the process is called infrared (IR) divergent.

The resolution to this IR problem lies in the fact that soft photons with energy below the energy resolution E_{th} of the used detector are not detectable. Therefore the emission of n real soft photons with $E < E_{th}$ is indistinguishable from the scattering process including virtual photons and thus these cross sections should be added. The IR divergences in the real and virtual corrections cancel, leading to a finite result for the combined cross section. We will discuss a detailed example for how this cancellation works in chapter 17.3.

9.A. Appendix: Decay rates and cross sections

We establish first the connection between the normalised transition matrix element $\mathcal{M}$ and the Feynman amplitude $\mathcal{A}$, where the normalisation factors of external particles are omitted. Then we derive decay rates and cross sections describing $1 \to n$ and $2 \to n$ processes.

Normalisation We have split the scattering operator S into a diagonal part and the transition operator T, $S = 1 + \mathrm{i}T$. Taking matrix elements, we obtain

$$S_{fi} = \delta_{fi} + (2\pi)^4\, \delta^{(4)}(P_i - P_f)\mathrm{i}\mathcal{M}_{fi} \tag{9.113}$$

where we set also $T_{fi} = (2\pi)^4\, \delta^{(4)}(P_i - P_f)\mathcal{M}_{fi}$.

The Feynman amplitude $\mathcal{A}$ neglects all normalisation factors of external particles, while the matrix element T_{fi} defined by (9.37) and thus $\mathcal{M}_{fi}$ contains a factor $N_{\boldsymbol{k}}$ for each external particle. Thus the transition between the matrix element $\mathcal{M}_{fi}$ and the Feynman amplitude $\mathcal{A}$ for a process with n particles in the initial and m in the final states is given by

$$\mathcal{M}_{fi} = \prod_{i=1}^{n}(2E_iV)^{-1/2}\prod_{f=1}^{m}(2E_fV)^{-1/2}\mathcal{A}_{fi}\,. \tag{9.114}$$

Here we changed also to a finite normalisation volume, $2E_{\boldsymbol{p}}(2\pi)^3 \to 2E_{\boldsymbol{p}}V$ what makes defining decay rates and cross sections easier.

9.A.1. Decay rate

We consider the decay of a particle into n particles in the final state. Squaring the scattering amplitude S_{fi} for $i \neq f$ using $(2\pi)^4\, \delta^{(4)}(0) = VT$ gives as differential transition probability

$$\mathrm{d}W_{fi} = (2\pi)^4\, \delta^{(4)}(P_i - P_f)VT|\mathcal{M}_{fi}|^2\prod_{f=1}^{n}\frac{V\mathrm{d}^3p_f}{(2\pi)^3}\,. \tag{9.115}$$

The decay rate or decay width $\mathrm{d}\Gamma$ is the transition probability per time,

$$\mathrm{d}\Gamma_{fi} = \lim_{T\to\infty}\frac{\mathrm{d}W_{fi}}{T} = (2\pi)^4\, \delta^{(4)}(P_i - P_f)V|\mathcal{M}_{fi}|^2\prod_{f=1}^{n}\frac{V\mathrm{d}^3p_f}{(2\pi)^3}\,. \tag{9.116}$$

Going over to the Feynman amplitude $\mathcal{A}$ eliminates the volume factors V,

$$\mathrm{d}\Gamma_{fi} = (2\pi)^4\delta^{(4)}(P_i - P_f)\,\frac{1}{2E_i}\,|\mathcal{A}_{fi}|^2\prod_{f=1}^{n}\frac{\mathrm{d}^3p_f}{2E_f(2\pi)^3}\,. \tag{9.117}$$

Moreover, the phase space integrals in the final state are now Lorentz invariant, $\mathrm{d}^3p_f/(2E_f)$. Introducing the n-particle phase space volume

$$\mathrm{d}\Phi^{(n)} = (2\pi)^4\,\delta^{(4)}(P_i - P_f)\prod_{f=1}^{n}\frac{\mathrm{d}^3p_f}{2E_f(2\pi)^3}\,, \tag{9.118}$$

the decay rate becomes

$$\mathrm{d}\Gamma_{fi} = \frac{1}{2E_i}\,|\mathcal{A}_{fi}|^2\,\mathrm{d}\Phi^{(n)}\,. \tag{9.119}$$

Since both $|\mathcal{A}_{fi}|^2$ and the phase space $\mathrm{d}\Phi^{(n)}$ are Lorentz invariant, the decay rate $\Gamma \propto 1/E_i = 1/(\gamma_i m_i)$ shows explicitly the time dilation effect for a moving particle. Finally, we note that a symmetry factor $S = 1/n!$ has to be added to the total decay width or cross section, if there are n identical particles in the final state.

Two-particle decays We evaluate the two particle phase space $\mathrm{d}\Phi^{(2)}$ in the rest frame of the decaying particle,

$$\mathrm{d}\Phi^{(2)} = (2\pi)^4\,\delta(M - E_1 - E_2)\,\delta^{(3)}(\boldsymbol{p}_1 + \boldsymbol{p}_2)\,\frac{\mathrm{d}^3p_1}{2E_1(2\pi)^3}\,\frac{\mathrm{d}^3p_2}{2E_2(2\pi)^3} \tag{9.120}$$

We perform the integration over d^3p_1 using the momentum delta function. In the resulting expression,

$$\mathrm{d}\Phi^{(2)} = \frac{1}{(2\pi)^2}\,\frac{1}{4E_1E_2}\,\delta(M - E_1 - E_2)\,\mathrm{d}^3p_2\,, \tag{9.121}$$

E_1 is now a function of p_2, $E_1^2 = p_2^2 + m_1^2$. Introducing spherical coordinates, $d^3p_2 = \mathrm{d}\Omega p_2^2 \mathrm{d}p_2$,

$$\mathrm{d}\Phi^{(2)} = \frac{1}{(2\pi)^2}\mathrm{d}\Omega\int_0^\infty \delta(M - E_1 - E_2)\,\frac{p_2^2\mathrm{d}p_2}{4E_1E_2}\,, \tag{9.122}$$

and evaluating the delta function with $M - E_1 - E_2 = M - x$ and $\mathrm{d}p_2/\mathrm{d}x = p_2x/(E_1E_2)$ gives

$$\mathrm{d}\Phi^{(2)} = \frac{|p'_{cms}|}{4\pi^2}\mathrm{d}\Omega\,, \tag{9.123}$$

where

$$p_{cms}^2 = \frac{\lambda(s, m_1^2, m_2^2)}{4s} = \frac{1}{4M^2}\left[M^2 - (m_1 + m_2)^2\right]\left[M^2 - (m_1 - m_2)^2\right] \tag{9.124}$$

is the cms momentum of the two final state particles. The Kibble function $\lambda(x, y, z)$ satisfies

$$\begin{aligned}\lambda(x,y,z) &= \left[(x^2 + y^2 + z^2) - 2xy - 2yz - 2xz\right]^{1/2} && (9.125a)\\ &= \left[x^2 - (\sqrt{y} + \sqrt{z})^2\right]^{1/2}\left[x^2 - (\sqrt{y} - \sqrt{z})^2\right]^{1/2}\,. && (9.125b)\end{aligned}$$

Three-particle decays The three particle phase space $\mathrm{d}\Phi^{(3)}$ is in the rest-frame of the decaying particle given by

$$\mathrm{d}\Phi^{(3)} = (2\pi)^4\,\delta(M - E_1 - E_2 - E_3)\,\delta^{(3)}(\boldsymbol{p}_1 + \boldsymbol{p}_2 + \boldsymbol{p}_3)\,\frac{\mathrm{d}^3p_1}{2E_1(2\pi)^3}\,\frac{\mathrm{d}^3p_2}{2E_2(2\pi)^3}\,\frac{\mathrm{d}^3p_3}{2E_3(2\pi)^3}\,. \tag{9.126}$$

We can use again the momentum delta function to perform the integration over d^3p_3,

$$\mathrm{d}\Phi^{(3)} = \frac{1}{(2\pi)^5}\delta(M - E_1 - E_2 - E_3)\,\frac{\mathrm{d}^3p_1\mathrm{d}^3p_2}{8E_1E_2E_3}\,, \tag{9.127}$$

To proceed we have to know the dependence of the matrix element on the integration variables. If there is no preferred direction (either for scalar particles or after averaging over spins), we obtain

$$\begin{aligned} \mathrm{d}\Phi^{(3)} &= \frac{1}{8(2\pi)^5}\,\frac{4\pi p_1^2\mathrm{d}p_1\,2\pi\mathrm{d}\cos\vartheta\mathrm{d}p_2}{E_1E_2E_3}\,\delta(M-E_1-E_2-E_3) && (9.128a)\\ &= \frac{1}{32\pi^3}\,\frac{p_1\mathrm{d}p_1\,(p_1p_2\mathrm{d}\cos\vartheta)(p_2\mathrm{d}p_2)}{E_1E_2E_3}\,\delta(M-E_1-E_2-E_3)\,. && (9.128b)\end{aligned}$$

We rewrite next the momentum integrals as energy integrals. The energy-momentum relation $E_i^2 = m_i^2 + p_i^2$ gives $E_i\mathrm{d}E_i = p_i\mathrm{d}p_i$ for $i = 1, 2$. Furthermore,

$$E_3^2 = (\boldsymbol{p}_1 + \boldsymbol{p}_2)^2 + m_3^2 = p_1^2 + p_2^2 + 2p_1p_2\cos\vartheta + m_3^2 \tag{9.129}$$

and thus $E_3\mathrm{d}E_3 = p_1p_2\mathrm{d}\cos\vartheta$ for fixed $\boldsymbol{p}_1, \boldsymbol{p}_2$. Performing the angular integral, we obtain

$$\mathrm{d}\Phi^{(3)} = \frac{1}{32\pi^3}\,\mathrm{d}E_1\mathrm{d}E_2\mathrm{d}E_3\delta(M-E_1-E_2-E_3)\,, \tag{9.130}$$

and finally

$$\mathrm{d}\Phi^{(3)} = \frac{1}{32\pi^3}\,\mathrm{d}E_1\mathrm{d}E_2\,. \tag{9.131}$$

The last step is only valid, if the argument of the delta function is non-zero. Thus the remaining task is to determine the boundary of the integration domain. Let us choose E_1 as the outer integration variable. Then we have to determine the allowed range of E_2 for a given value of E_1. Inserting energy and momentum conservation into $E_3^2 = p_3^2 + m_3^2$, we obtain

$$(M - E_1 - E_2)^2 = m_3^2 + \boldsymbol{p}_1^2 + \boldsymbol{p}_2^2 + 2\boldsymbol{p}_1\cdot\boldsymbol{p}_2\,. \tag{9.132}$$

The extrema correspond to

$$\boldsymbol{p}_1\cdot\boldsymbol{p}_2 = \pm|\boldsymbol{p}_1||\boldsymbol{p}_2| = \pm\sqrt{(E_1^2-m_1^2)(E_2^2-m_2^2)}\,. \tag{9.133}$$

Inserting them into Eq. (9.132), we obtain the curve defining the boundary of the integration area as function of E_1 and E_2,

$$M^2 - 2M(E_1+E_2) + 2E_1E_2 + m_1^2 + m_2^2 - m_3^2 = \pm\sqrt{(E_1^2-m_1^2)(E_2^2-m_2^2)}\,. \tag{9.134}$$

In order to visualise the integration area easier, we set first $m_1 = m_2 = 0$. Then the equation with the plus sign becomes $E_2 = M/2 + m_3^2/(4E_1 - 2M)$, while the equation with the minus sign simplifies to a straight line $E_2 = -E_1 + M/(2 - 2m_3^2/M^2)$. The resulting integration area is shown in Fig. 9.7 for $m_3/M = 0.1$. Setting also $m_3 = 0$, the integration area is a triangle in the E_1–E_2 plane.

If one prefers Lorentz invariant integration variables, one can introduce the invariant mass of the pair (i, j)

$$\begin{aligned} m_{23}^2 &= (p-p_1)^2 = (p_2+p_3)^2 = M^2 - 2ME_1 + m_1^2 && (9.135a)\\ m_{13}^2 &= (p-p_2)^2 = (p_1+p_3)^2 = M^2 - 2ME_2 + m_2^2 && (9.135b)\\ m_{12}^2 &= (p-p_3)^2 = (p_1+p_2)^2 = M^2 - 2ME_3 + m_3^2\,, && (9.135c)\end{aligned}$$

Because of $m_{23}^2 + m_{13}^2 + m_{12}^2 = M^2 + m_2^2 + m_3^2$, only two out of the three variables are independent. They can be used to to replace E_1 and E_2 in Eq. (9.131).

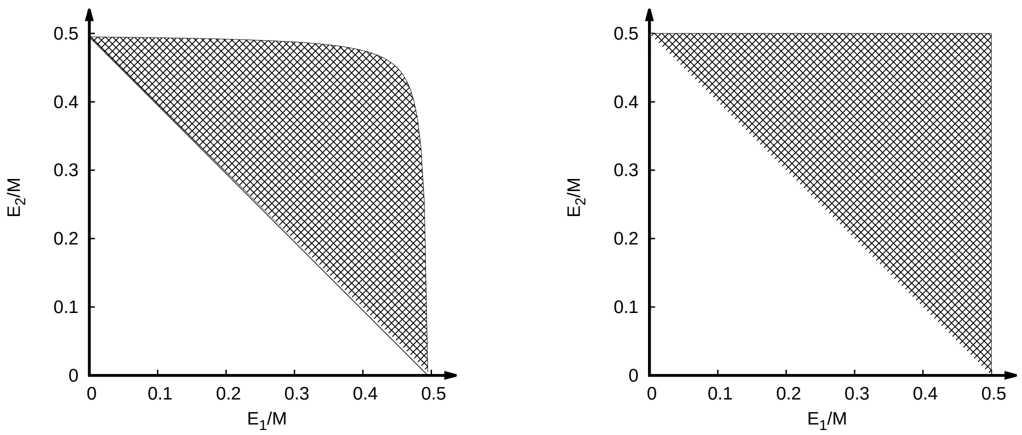


Figure 9.7.: Phase-space boundary for $m_3/M = 0.1$ (left). and $m_3 = 0$ (right); in both cases we set $m_1 = m_2 = 0$.

9.A.2. Cross sections

We consider now the interaction of two particles in the rest system of either particle 1 or 2. For simplicity, we consider two uniform beams. They may produce n final state particles. The total number of such scatterings is

$$\mathrm{d}N \propto v_{\mathrm{M}\o l} n_1 n_2 \mathrm{d}V\mathrm{d}t \tag{9.136}$$

where n_i is the density of particle type $i = 1, 2$. The Møller velocity $v_{\mathrm{M}\o l}$ is a quantity which coincides in the rest frame of particle 1 or 2 with $|\boldsymbol{v}_2|$ and $|\boldsymbol{v}_1|$, respectively. Therefore it is often denoted simply as relative velocity $\boldsymbol{v}_{\mathrm{r}el}$. The proportionality constant in (9.136) has the dimension of an area and is called cross section σ. We define in the rest system of either particle 1 or 2

$$\mathrm{d}N = \sigma v_{\mathrm{M}\o l}\, n_1 n_2\, \mathrm{d}V\mathrm{d}t\,, \tag{9.137}$$

while we set in an arbitrary frame

$$\mathrm{d}N = A n_1 n_2 \mathrm{d}V\mathrm{d}t\,. \tag{9.138}$$

We determine now A. Since both $\mathrm{d}N$ and $\mathrm{d}V\mathrm{d}t = \mathrm{d}^4x$ are Lorentz invariant, the expression An_1n_2 has to be Lorentz invariant too. The densities transform as

$$n_i = n_{i,0}\gamma = n_{i,0}\frac{E_i}{m_i}\,, \tag{9.139}$$

and thus the expression

$$A\,\frac{E_1E_2}{p_1\cdot p_2} \tag{9.140}$$

is also Lorentz invariant. In the rest system of particle 1, it becomes

$$A\,\frac{E_1E_2}{E_1E_2 - \boldsymbol{p}_1\boldsymbol{p}_2} = A = \sigma v_{\mathrm{M}\o l}\,. \tag{9.141}$$

Thus we found that A in an arbitrary frame is given by

$$A = \sigma v_{\mathrm{M}\o l}\,\frac{p_1\cdot p_2}{E_1E_2}\,. \tag{9.142}$$

We still have to determine $v_{\mathrm{M}\o l}$: In the rest frame 1, we have

$$p_1\cdot p_2 = m_1E_2 = m_1\,\frac{m_2}{\sqrt{1-v_{\mathrm{M}\o l}^2}}\,. \tag{9.143}$$

Thus the Møller velocity is given in general by

$$v_{\mathrm{M}\o l} = \sqrt{1 - \frac{m_1^2 m_2^2}{(p_1 \cdot p_2)^2}}\,. \tag{9.144}$$

Since this expression is Lorentz invariant, we see that the notion of the Møller velocity as relative velocity is misleading.

Next we define the flux factor

$$I \equiv v_{\mathrm{M}\o l}\, p_1 \cdot p_2 = \sqrt{(p_1 \cdot p_2)^2 - m_1^2 m_2^2}\,. \tag{9.145}$$

Inserting (9.142) for A together with the definition of I into (9.138), we obtain

$$\mathrm{d}N = \sigma \,\frac{I}{E_1 E_2 V}\,(n_1 V)(n_2 \mathrm{d}V)\mathrm{d}t\,. \tag{9.146}$$

Here, we re-grouped the terms to make clear that after integration the total number N of scattering events is proportional to the number $N_1 = n_1 V$ and $N_2 = \int n_2 \mathrm{d}V$ of particles of type 1 and 2, respectively. The number N of scattering events per time and per particles 1 and 2 is however simply the transition probability per time,

$$\frac{\mathrm{d}W_{fi}}{T} = \frac{\mathrm{d}N}{N_1 N_2 T} = \mathrm{d}\sigma\,\frac{I}{E_1 E_2 V}\,. \tag{9.147}$$

Inserting the expression (9.116) for $\mathrm{d}W_{fi}$, we find

$$\mathrm{d}\sigma = \frac{E_1 E_2 V^2}{I}\,(2\pi)^4\,\delta^{(4)}(P_i - P_f)|\mathcal{M}_{fi}|^2 \prod_{f=1}^{n} \frac{V \mathrm{d}^3 p_f}{(2\pi)^3}\,. \tag{9.148}$$

Changing from the normalised matrix element $\mathcal{M}$ to the Feynman amplitude $\mathcal{A}$ introduces a factor $(2E_1 V)^{-1}(2E_2 V)^{-1}$ for the initial state and $\prod_f (2E_f V)^{-1}$ for the final state. Thus the arbitrary normalisation volume V cancels and we obtain

$$\mathrm{d}\sigma = \frac{1}{4I}\,(2\pi)^4\,\delta^{(4)}(P_i - P_f)|\mathcal{A}_{fi}|^2 \prod_{f=1}^{n} \frac{\mathrm{d}^3 p_f}{2E_f (2\pi)^3} = \frac{1}{4I}\,|\mathcal{A}_{fi}|^2\,\mathrm{d}\Phi^{(n)} \tag{9.149}$$

with the final state phase space $\mathrm{d}\Phi^{(n)}$. The three pieces composing the differential cross section, the flux factor I, the Feynman amplitude $\mathcal{A}$, and the final state phase space $\mathrm{d}\Phi^{(n)}$, are each Lorentz invariant.

2–2 scattering For a $1+2 \to 3+4$ scattering process, it is useful to introduce Mandelstam variables s, t, and u as

$$s = (p_1 + p_2)^2 = (p_2 + p_4)^2\,, \tag{9.150}$$

$$t = (p_1 - p_3)^2 = (p_2 - p_4)^2\,, \tag{9.151}$$

$$u = (p_1 - p_4)^2 = (p_2 - p_3)^2\,. \tag{9.152}$$

Since $s + t + u = \sum_{i=1}^4 m_i^2$, the scattering amplitude $\mathcal{A}$ depends only on two variables, e.g $\mathcal{A}(s,t)$. In the cms, the flux factor becomes

$$I^2 = (p_1 \cdot p_2)^2 - m_1^2 m_2^2 = p_{cms}^2 (E_1 + E_2)^2 = p_{cms}\sqrt{s}\,. \tag{9.153}$$

Adding the expression for the 2-particle phase space gives

$$\frac{\mathrm{d}\sigma}{\mathrm{d}\Omega} = \frac{1}{64\pi^2 s}\,\frac{p'_{cms}}{p_{cms}}\,|\mathcal{A}_{fi}|^2\,. \tag{9.154}$$

Here the cm momentum of the initial state is $p_{cms}^2 = \lambda(s, m_1^2, m_2^2)/(4s)$, while $p_{cms}'^2 = \lambda(s, m_3^2, m_4^2)/(4s)$ is the one of the final state. Using as variable the momentum transfer t, the differential cross section becomes

$$\frac{\mathrm{d}\sigma}{\mathrm{d}t} = \frac{1}{64\pi s p_{cms}^2} \left|\mathcal{A}_{fi}\right|^2 , \tag{9.155}$$

where the allowed range of t has to be determined from Eq. (9.151) and $-1 \leq \cos\vartheta \leq 1$.

The optical theorem connects the imaginary part of the forward amplitude $\Im T_{ii}$ with the total cross section as

$$\sigma_{\mathrm{tot}} = \frac{\Im T_{ii}}{\lambda^{1/2}(s, m_1^2, m_2^2)} . \tag{9.156}$$

Summary

The LSZ reduction formula shows that S-matrix elements are obtained from connected Green functions by a replacement of the propagators on external lines with the corresponding wavefunctions times the wavefunction renormalisation constant $\sqrt{Z}$. Cross sections are calculated from the squared Feynman amplitude $\mathcal{A}$, the final state phase space $\mathrm{d}\Phi^{(n)}$ and the flux factor I, which are all three Lorentz-invariant. Squared Feynman amplitudes can be obtained using "Casimir's trick". If the number of diagrams increases, it is more convenient to calculate directly the amplitude using helicity methods.

The amplitude for the emission of additional soft particles factorises in the amplitude of the hard process and an universal factor. Lorentz invariance requires that a massless spin-1 particle couples in the low-energy limit to a conserved charge, while a massless spin-2 particle has to couples with universal strength to the energy-momentum stress tensor.

Further reading

[Ste93] discusses the optical theorem and its connection to cut diagrams in more detail. The LSZ formula for particles with spin $s > 0$ is presented e.g. in [GR08], while [BL93] derive S-matrix elements defining a new functional $Z'[J]$ with the correct boundary conditions. For additional information about the helicity formalisms see [Hab94] and [Pes11] from which our examples are taken. [A$^+$12] provide a tutorial for several software tools useful for the calculation of scattering processes. The discussion of soft photon emission follows closely the original discussion of [Wei65], for an introduction see [Whi15].

10. Gauge theories

We discuss in this chapter field theories in which the Lagrangian is invariant under a continuous group of *local* transformations in internal field space. The symmetry group of these transformations is called the gauge group and the vector fields associated to the generators of the group the gauge fields. We introduce as a first step unbroken gauge theories, i.e. theories with massless gauge bosons, and defer the more complex case of broken gauge symmetries to the chapters 13 and **??**. The Standard Model (SM) of particle physics contains with quantum electrodynamics (QED) and quantum chromodynamics (QCD) two examples for unbroken gauge theories. While QED is an abelian gauge theory based on the gauge group U(1), QCD which describes the strong interactions is an non-abelian gauge theory with group SU(3). Non-abelian gauge theories were first studied by Yang and Mills and are therefore also often called Yang-Mills theories. The structure of Yang-Mills theories has many similarities with gravity. We use this property to introduce the curvature of a space-time as the analogue of the field-strength in the Yang-Mills case.

10.1. Electrodynamics as abelian gauge theory

In classical electrodynamics, the field-strength tensor $F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu$ is an observable quantity, while the potential A_μ is merely a convenient auxiliary quantity. From its definition as an antisymmetric tensor, it is clear that $F_{\mu\nu}$ is invariant under local gauge transformations

$$A_\mu(x) \to A'_\mu(x) = A_\mu(x) - \partial_\mu \Lambda(x) \tag{10.1}$$

of the potentials. Thus $A'_\mu(x)$ is for any smooth $\Lambda(x)$ physically equivalent to $A_\mu(x)$, leading to the same field-strength tensor and thus e.g. to the same Lorentz force on a particle.

Consider now e.g. a free Dirac field $\psi(x)$ with electric charge q. We saw already that this field is invariant under *global* phase transformations $\exp[\mathrm{i}q\Lambda] \in \mathrm{U}(1)$, implying a conserved current $j^\mu = \bar\psi\gamma^\mu\psi$ via Noether's theorem. Can we promote this global U(1) symmetry to a local one,

$$\psi(x) \to \psi'(x) = U(x)\psi(x) = \exp[\mathrm{i}q\Lambda(x)]\psi(x)\,, \tag{10.2}$$

by making the phase U space-time dependent as in (10.1)? The partial derivatives in the Dirac Lagrangian will lead to an additional term $\propto \partial_\mu U(x)$, destroying the invariance of the free Lagrangian. However, if we add a field $A_\mu(x)$ which transforms as defined in (10.1) and couples to the Noether current j^μ of the complex field as $\mathscr{L}_I = -q j^\mu A_\mu$, the two gauge-dependent terms will cancel. Thus local U(1) gauge invariance of the Dirac field requires the existence of a massless gauge boson and fixes its interaction with matter: The coupling of matter to photons is obtained by replacing the normal derivative by the covariant derivative,

$$\partial_\mu \to D_\mu = \partial_\mu + \mathrm{i}qA_\mu\,, \tag{10.3}$$

which transforms as the matter fields,

$$D_\mu\psi(x) \to D'_\mu\psi'(x) = \{\partial_\mu + \mathrm{i}q[A_\mu(x) - \partial_\mu\Lambda(x)]\}\exp[\mathrm{i}q\Lambda(x)]\psi(x) = \tag{10.4}$$

$$= \exp[\mathrm{i}q\Lambda(x)]\{\partial_\mu + \mathrm{i}qA_\mu(x)]\}\psi(x) = U(x)D_\mu\psi(x)\,. \tag{10.5}$$

We can rewrite the gauge transformation of A_μ as

$$A_\mu(x) \to A'_\mu(x) = A_\mu(x) - \partial_\mu\Lambda(x) = A_\mu(x) - \frac{\mathrm{i}}{q}U(x)\partial_\mu U^\dagger(x)\,, \tag{10.6}$$

expressing the change $\delta A_\mu(x)$ through the group elements $U(x)$. Finally, we note that we can connect the field-strength tensor to the commutator of covariant derivatives,

$$[D_\mu, D_\nu]\psi = \mathrm{i}q([\partial_\mu, A_\nu] - [\partial_\nu, A_\mu])\psi = \mathrm{i}q\psi(\partial_\mu A_\nu - \partial_\nu A_\mu) = \mathrm{i}q\psi F_{\mu\nu}\,. \tag{10.7}$$

To summarise: The invariance of complex (scalar or Dirac) fields under *global* phase transformations $\exp[\mathrm{i}q\Lambda] \in \mathrm{U}(1)$ implies a conserved current, promoting it to a *local* U(1) symmetry requires the existence of a massless U(1) gauge boson coupled via gauge-invariant derivatives to these fields.

10.2. Non-abelian gauge theories

10.2.1. Gauge invariant interactions

We want to generalise now electrodynamics, using as symmetry group instead of the abelian group U(1) larger groups like SO(n) or SU(n). A group like SU(n) will describe the interactions of $n^2 - 1$ gauge bosons with matter, using as a single parameter the gauge coupling g. The gauge transformations will moreover mix fermions living in the same representation of the group, requiring that these fermions have the same interactions and the same mass if the symmetry is unbroken. In this way, non-abelian gauge theories lead to a partial *unification* of matter fields and interactions. Note the difference to an abelian symmetry: The emission of a photon does not change any quantum number (apart from the momentum) and thus does not "mix" different particles. Therefore there is also no connection between the electric charge of different particles.

The two non-abelian groups used in the SM are SU(2) for weak and SU(3) for strong interactions. A matrix representation for the fundamental representation of these two groups are the Pauli matrices, $T^a = \sigma^a/2$, and the Gell-Mann matrices, $T^a = \lambda^a/2$, respectively. Under the fundamental representation the fermions transform as doublets for SU(2), as triplets for SU(3), etc. Since the number of generators is $m = n^2 - 1$ for SU(n), the groups SU(2) contains three gauge bosons, while SU(3) contains eight bosons carrying strong interactions. The most important difference of these non-abelian groups compared to U(1) is that the generators $T^a \equiv T^a_{ij}$ of such groups do not commute with each other. As a result, we may expect that both the expression for the field-strength tensor, Eq. (7.11), and the transformation law for the gauge field, Eq. (7.12), becomes more complicated. In contrast, we postulate that the replacement $\partial_\mu \to D_\mu$ and the interaction law $j_\mu A^\mu$ of gauge bosons with matter remain valid, with the sole difference that now $A_\mu = A^a_\mu T^a$. Thus A_μ is a Lorentz vector with values in the Lie algebra of the gauge group.

We derive now the transformation laws and structure of the gauge sector, requiring that the transformation of the fermions and their interaction with the gauge field are locally invariant. A local gauge transformation

$$U(x)=\exp[\mathrm{i}g\textstyle\sum_{a=1}^{m}\vartheta^a(x)T^a]\equiv\exp[\mathrm{i}g\vartheta(x)] \tag{10.8}$$

changes a vector of fermion fields $\boldsymbol{\psi}$ with components $\{\psi_1,\dots,\psi_n\}$ as[1]

$$\psi(x)\to\psi'(x)=U(x)\psi(x)\,. \tag{10.9}$$

Already global gauge invariance of the fermion mass term requires $m_1=m_2=\ldots=m_n$ and for simplicity we set $m_i=0$. We can implement local gauge invariance, if derivatives transform in the same way as ψ. Hence we define a new covariant derivative D_μ requiring

$$D_\mu\psi(x)\to[D_\mu\psi(x)]'=U(x)[D_\mu\psi(x)]\,. \tag{10.10}$$

The gauge field should compensate the difference between the normal and the covariant derivative,

$$D_\mu\psi(x)=[\partial_\mu+\mathrm{i}gA_\mu(x)]\psi(x)\,. \tag{10.11}$$

In the non-abelian case, the gauge field A_μ is a matrix that is connected to its component fields by

$$A_\mu=A_\mu^aT^a\,. \tag{10.12}$$

We now determine the transformation properties of D_μ and A_μ demanding that (10.9) and (10.10) hold. Combining both requirements gives

$$D_\mu\psi(x)\to[D_\mu\psi]'=UD_\mu\psi=UD_\mu U^{-1}U\psi=UD_\mu U^{-1}\psi'\,, \tag{10.13}$$

and thus the covariant derivative transforms as $D'_\mu=UD_\mu U^{-1}$. Using its definition (10.11), we find

$$[D_\mu\psi]'=[\partial_\mu+\mathrm{i}gA'_\mu]U\psi=UD_\mu\psi=U[\partial_\mu+\mathrm{i}gA_\mu]\psi\,. \tag{10.14}$$

We compare now the second and the fourth term, after having performed the differentiation $\partial_\mu(U\psi)$. The result

$$[(\partial_\mu U)+\mathrm{i}gA'_\mu U]\psi=\mathrm{i}gUA_\mu\psi \tag{10.15}$$

should be valid for arbitrary ψ and hence we arrive after multiplying from the right with U^{-1} at

$$A_\mu\to A'_\mu=UA_\mu U^{-1}+\frac{\mathrm{i}}{g}(\partial_\mu U)U^{-1}=UA_\mu U^{-1}-\frac{\mathrm{i}}{g}U\partial_\mu U^{-1}\,. \tag{10.16}$$

Here we used also $\partial_\mu(UU^{-1})=0$. In most cases, the gauge transformation U is an unitary transformation and one sets $U^{-1}=U^\dagger$. A term changing as $U(x)D_\mu(x)U^\dagger(x)$ is called to transform homogeneously, while the potential A_μ is said to transforms inhomogeneously.

[1] We suppress in the following most indices; writing them out gives e.g. $\psi'_i(x)=U_{ij}(x)\psi_j(x)$ with $U_{ij}(x)=\exp[\mathrm{i}g\sum_{a=1}^{m}\vartheta^a(x)T^a_{ij}]$.

Example 10.1: We can determine the transformation properties of A_μ also by demanding that (10.11) defines the interaction term in a gauge invariant way. Replacing $\partial_\mu \to D_\mu$ in the free Lagrange density of fermions and inserting then $U^{-1}U = 1$ gives

$$\begin{aligned}\mathscr{L}_f + \mathscr{L}_I &= \mathrm{i}\bar\psi\gamma^\mu D_\mu\psi = \mathrm{i}\bar\psi\gamma^\mu\partial_\mu\psi - g\bar\psi\gamma^\mu A_\mu\psi = \\ &= \mathrm{i}\bar\psi U^{-1}U\gamma^\mu\partial_\mu U^{-1}U\psi - g\bar\psi U^{-1}U\gamma^\mu A_\mu U^{-1}U\psi\,. \end{aligned} \tag{10.17}$$

Using then $\psi' = U\psi$, we obtain

$$\begin{aligned}\mathscr{L}_f + \mathscr{L}_I &= \mathrm{i}\bar\psi'\gamma^\mu U\partial_\mu U^{-1}\psi' - g\bar\psi'\gamma^\mu U A_\mu U^{-1}\psi' \\ &= \mathrm{i}\bar\psi'\gamma^\mu\partial_\mu\psi' - g\bar\psi'\gamma^\mu\left\{UA_\mu U^{-1} - \frac{\mathrm{i}}{g}U(\partial_\mu U^{-1})\right\}\psi'\,. \end{aligned} \tag{10.18}$$

The Lagrange density $\mathscr{L}_f + \mathscr{L}_I$ is thus invariant, if the gauge field transforms as in Eq. (10.16).

Specialising to infinitesimal transformations,

$$U(x) = \exp(\mathrm{i}g\vartheta^a(x)T^a) = 1 + \mathrm{i}g\vartheta(x) + \mathcal{O}(\vartheta^2)\,, \tag{10.19}$$

it follows

$$A_\mu(x) \to A'_\mu(x) = A_\mu(x) - \mathrm{i}g[A_\mu(x), \vartheta(x)] - \partial_\mu\vartheta(x)\,. \tag{10.20}$$

In the abelian U(1) case, the commutator term is not present and the transformation law reduces to the known $A_\mu \to A_\mu - \partial_\mu\vartheta$. For a (semi-simple) Lie group one defines

$$[T^a, T^b] = \mathrm{i}f^{abc}T^c \tag{10.21}$$

with structure constants f^{abc} that can be chosen to be completely antisymmetric. Thus

$$A^a_\mu(x) \to A^{a\prime}_\mu(x) = A^a_\mu(x) + gf^{abc}A^b_\mu(x)\vartheta^c(x) - \partial_\mu\vartheta^a(x) \tag{10.22a}$$

$$= A^a_\mu(x) - [\delta^{ac}\partial_\mu - gf^{abc}A^b_\mu(x)]\vartheta^c(x) \tag{10.22b}$$

$$\equiv A^a_\mu(x) - D^{ac}_\mu\vartheta^c(x)\,, \tag{10.22c}$$

where the last line defines how the covariant derivative acts on the gauge fields: Comparing this expression to the general definition $D_\mu = \partial_\mu + \mathrm{i}gA^a_\mu T^a$, we see that the gauge fields live in the adjoint representation of the gauge group[2], cf. problem **??**. The infinitesimal change of the gauge fields A^a_μ is given by the covariant derivative acting on the parameters ϑ^a of the gauge transformation.

Finally, we have to derive the field strength tensor $F_{\mu\nu} = F^a_{\mu\nu}T^a$ and the Lagrange density $\mathscr{L}_{YM}$ of the gauge field. The quantity F^2 requires now additionally a summation over the group index a,

$$\mathscr{L}_{\mathrm{YM}} = -\frac{1}{4}F^a_{\mu\nu}F^{a\mu\nu} = -\frac{1}{2}\,\mathrm{tr}\,F_{\mu\nu}F^{\mu\nu}\,, \tag{10.23}$$

where we assumed in the second step that the standard normalisation $\mathrm{tr}\,T^aT^b = \delta^{ab}/2$ for the group generators T^a holds. The last equation shows that it is sufficient for the gauge invariance of the action that the field-strength tensor transforms homogeneously,

$$F_{\mu\nu}(x) \to F'_{\mu\nu}(x) = U(x)F_{\mu\nu}(x)U^\dagger(x)\,. \tag{10.24}$$

[2] The n complex fermion and n^2-1 real gauge fields of SU(n) live in different representations of the group, as already the mismatch of their number indicates, see also Appendix B. Note also that the gauge transformations of the gauge fields have to be real, in contrast to the ones of the fermion fields.

There are several ways to derive the relation between $F_{\mu\nu}$ and A_μ. The field-strength tensor should be antisymmetric. Thus we should construct it out of the commutator of gauge invariant quantities that in turn should contain A_μ. An obvious try is $\mathrm{i}gF_{\mu\nu} = [D_\mu, D_\nu]$ that worked in the abelian case. Now, additionally the non-zero commutator of the gauge fields contributes,

$$F_{\mu\nu} = F^a_{\mu\nu}T^a = \frac{1}{\mathrm{i}g}\,[D_\mu, D_\nu] = \partial_\mu A_\nu - \partial_\nu A_\mu + \mathrm{i}g[A_\mu, A_\nu]\,. \tag{10.25}$$

In components, this equation reads explicitly

$$F^a_{\mu\nu} = \partial_\mu A^a_\nu - \partial_\nu A^a_\mu - gf^{abc}A^b_\mu A^c_\nu\,. \tag{10.26}$$

Remark 10.1: Antisymmetric tensors of rank n can also be seen as differential forms. We know already functions as forms of order $n = 0$ and co-vectors as forms of order $n = 1$. Since differentials $\mathrm{d}f = \partial_i f\,\mathrm{d}x^i$ of functions are forms of order $n = 1$, the $\mathrm{d}x^i$ form a basis, and one can write in general $\mathbf{A} = A_i\,\mathrm{d}x^i$. For $n > 1$, the basis has to be antisymmetrised. Hence, a two-form as the field-strength tensor is given by

$$\mathbf{F} = \frac{1}{2}\,F_{\mu\nu}\mathrm{d}x^\mu \wedge \mathrm{d}x^\nu \tag{10.27}$$

with $\mathrm{d}x^\mu \wedge \mathrm{d}x^\nu = -\mathrm{d}x^\nu \wedge \mathrm{d}x^\mu$. Looking at $\mathrm{d}f$ suggestions to define the differentiation of a form ω with coefficients w and degree n as an operation that increases its degree by one to $n+1$,

$$\mathrm{d}\omega = \frac{1}{n!}\,(\partial_\beta w_{\alpha_1\ldots,\alpha_n})\mathrm{d}x^\beta \wedge \mathrm{d}x^{\alpha_1} \wedge \ldots \wedge \mathrm{d}x^{\alpha_n}\,. \tag{10.28}$$

Thus we have $\mathbf{F} = \mathrm{d}\mathbf{A}$. Moreover, it follows $\mathrm{d}^2\omega = 0$ for all forms. Hence we can write an abelian gauge transformation as $\mathbf{F}' = \mathrm{d}(\mathbf{A} - \mathrm{d}\Lambda) = \mathbf{F}$.

10.2.2. Gauge fields as connection

There is a close analogy between the covariant derivative ∇_μ introduced for a space-time containing a gravitational field and the gauge invariant derivative D_μ required for a space-time containing a gauge field. In the former case, the moving coordinate basis in curved space-time, $\partial_\mu \boldsymbol{e}^\nu \neq 0$, introduces an additional term in the derivative of vector components $V^\mu = \boldsymbol{e}^\mu \cdot \boldsymbol{V}$. Analogously, a non-zero gauge field A^μ leads to a rotation of the basis vectors $\boldsymbol{e}_i$ in group space which in turn produces an additional term $\boldsymbol{\psi} \cdot (\partial_\mu \boldsymbol{e}_i)$ performing the derivative of a $\psi_i = \boldsymbol{\psi} \cdot \boldsymbol{e}_i$.

Let us rewrite our formulas such that the analogy between the covariant gauge derivative D_μ and the covariant space-time derivative ∇_μ becomes obvious. The vector $\boldsymbol{\psi}$ of fermion fields with components $\{\psi_1, \ldots, \psi_n\}$ transforming under a representation of a gauge group can be written as

$$\boldsymbol{\psi}(x) = \psi_i(x)\boldsymbol{e}_i(x)\,. \tag{10.29}$$

We can pick out the component ψ_j by multiplying with the corresponding basis vector $\boldsymbol{e}_j$,

$$\psi_j = \boldsymbol{\psi} \cdot \boldsymbol{e}_j(x)\,. \tag{10.30}$$

If the coordinate basis in group space depends on x^μ, then the partial derivative of ψ_i acquires a second term,

$$\partial_\mu \psi_i = (\partial_\mu \boldsymbol{\psi}) \cdot \boldsymbol{e}_i + \boldsymbol{\psi} \cdot (\partial_\mu \boldsymbol{e}_i)\,. \tag{10.31}$$

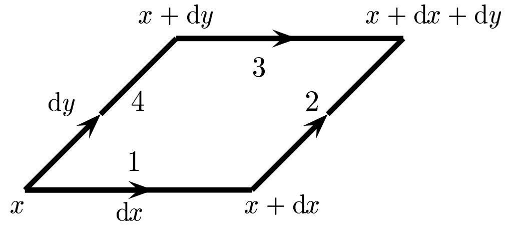


Figure 10.1.: Parallelogram used to calculate the rotation of a test field ψ_i moved along a closed loop in the presence of a non-zero gauge field A^μ.

We can argue as in section 6.2 that $(\partial_\mu \boldsymbol{\psi}) \cdot \boldsymbol{e}_i$ is an invariant quantity, defining therefore as gauge invariant derivative

$$D_\mu \psi_i = (\partial_\mu \boldsymbol{\psi}) \cdot \boldsymbol{e}_i = \partial_\mu \psi_i - \boldsymbol{\psi} \cdot (\partial_\mu \boldsymbol{e}_i) \,. \tag{10.32}$$

The change $\partial_\mu \boldsymbol{e}_i$ of the basis vector in group space should be proportional to gA_μ. Setting

$$\partial_\mu \boldsymbol{e}_i = -\mathrm{i}g(A_\mu)_{ij}\boldsymbol{e}_j \tag{10.33}$$

we are back to our old notation.

Gauge loops The correspondence between the derivatives ∇_μ and D_μ suggests that we can use the gauge field A_μ to transport fields along a curve $x^\mu(\sigma)$. In empty space, we can use the partial derivative $\partial_\mu \psi(x)$ to compare fields at different points,

$$\partial_\mu \psi(x) \propto \psi(x + \mathrm{d}x^\mu) - \psi(x) \,. \tag{10.34}$$

If there is an external gauge field present, the field ψ is additionally rotated in group space moving it from x to $x + \mathrm{d}x$,

$$\begin{aligned} \tilde\psi(x+\mathrm{d}x) &= \psi(x+\mathrm{d}x) + \mathrm{i}gA_\mu(x)\psi(x)\mathrm{d}x^\mu & (10.35\mathrm{a}) \\ &= \psi(x) + \partial_\mu\psi(x)\mathrm{d}x^\mu + \mathrm{i}gA_\mu(x)\psi(x)\mathrm{d}x^\mu \,. & (10.35\mathrm{b}) \end{aligned}$$

Then the total change is

$$\tilde\psi(x+\mathrm{d}x) - \psi(x) = [\partial_\mu + \mathrm{i}gA_\mu(x)]\psi(x)\mathrm{d}x^\mu = D_\mu\psi(x)\mathrm{d}x^\mu \,. \tag{10.36}$$

Thus we can view[3]

$$P_{\mathrm{d}x}(x) = 1 - \mathrm{i}gA_\mu(x)\mathrm{d}x^\mu \tag{10.37}$$

as an operator which allows us to transport a gauge-dependent field the infinitesimal distance from x to $x + \mathrm{d}x$.

We ask now what happens to a field $\psi_i(x)$, if we transport it along an infinitesimal parallelogram, as shown in Fig. 10.1. Calculating the path 2, we find

$$\begin{aligned} P_{\mathrm{d}y}(x+\mathrm{d}x) &= 1 - \mathrm{i}gA_\nu(x+\mathrm{d}x)\mathrm{d}y^\nu \\ &= 1 - \mathrm{i}gA_\nu(x)\mathrm{d}y^\nu - \mathrm{i}g\partial_\mu A_\nu(x)\mathrm{d}x^\mu\mathrm{d}y^\nu \,, \end{aligned} \tag{10.38}$$

[3] Note the sign change compared to the covariant derivative: there we pull-back the field from $x + \mathrm{d}x$ to x.

where we Taylor expanded $A_\nu(x+\mathrm{d}x)$. Combining the paths 1 and 2, we arrive at

$$\begin{aligned} P_{\mathrm{d}y}(x+\mathrm{d}x)P_{\mathrm{d}x}(x) &= [1-\mathrm{i}gA_\nu(x)\mathrm{d}y^\nu - \mathrm{i}g\partial_\mu A_\nu(x)\mathrm{d}x^\mu\mathrm{d}y^\nu][1-\mathrm{i}gA_\mu(x)\mathrm{d}x^\mu] \\ &= 1-\mathrm{i}gA_\mu(x)\mathrm{d}x^\mu - \mathrm{i}gA_\nu(x)\mathrm{d}y^\nu - \mathrm{i}g\partial_\mu A_\nu(x)\mathrm{d}x^\mu\mathrm{d}y^\nu \\ &- g^2A_\nu(x)A_\mu(x)\mathrm{d}y^\nu\mathrm{d}x^\mu + \mathcal{O}(\mathrm{d}x^3)\,. \end{aligned} \tag{10.39}$$

Instead of performing the calculation for a round trip $1\to2\to3\to4$, we evaluate next $4\to3$ which we then subtract from $1\to2$. In this way, we can re-use our result for $1\to2$ after exchanging labels, $A_\mu\mathrm{d}x^\mu \leftrightarrow A_\nu\mathrm{d}y^\nu$, obtaining

$$\begin{aligned} P_{\mathrm{d}x}(x+\mathrm{d}y)P_{\mathrm{d}y}(x) &= 1-\mathrm{i}gA_\nu(x)\mathrm{d}y^\nu - \mathrm{i}gA_\mu(x)\mathrm{d}x^\mu - \mathrm{i}g\partial_\nu A_\mu(x)\mathrm{d}x^\mu\mathrm{d}y^\nu \\ &- g^2A_\mu(x)A_\nu(x)\mathrm{d}x^\mu\mathrm{d}y^\nu + \mathcal{O}(\mathrm{d}x^3)\,. \end{aligned} \tag{10.40}$$

The first three terms on the RHS's of (10.39) and (10.40) cancel in the result $P(\Box)$ for the round-trip, leaving us with

$$\begin{aligned} P(\Box) \equiv P_{\mathrm{d}y}(x+dx)P_{\mathrm{d}x}(x) - P_{\mathrm{d}x}(x+dy)P_{\mathrm{d}y}(x) = \\ -\,\mathrm{i}g\left\{\partial_\mu A_\nu - \partial_\nu A_\mu + \mathrm{i}g[A_\mu, A_\nu]\right\}\mathrm{d}x^\mu\mathrm{d}y^\nu\,. \end{aligned} \tag{10.41}$$

Maxwell's equations inform us that the line integral of the vector potential equals the enclosed flux: The area of the parallelogram corresponds to $\mathrm{d}x^\mu\mathrm{d}y^\nu$, and the prefactor has to be therefore the field-strength tensor. If the enclosed flux is non-zero, then $P(\Box)\psi_i \neq \psi_i$ and thus the field is rotated.

10.2.3. Curvature of space-time

Curvature and the Riemann tensor We continue to work out the analogy between Yang-Mills theories and gravity. Both the gauge field A_μ and the connection $\Gamma^\mu{}_{\kappa\rho}$ transform homogeneously. Therefore we can not use them to judge if a gauge or gravitational field is present. In the gauge case, we introduced therefore the field-strength $F_{\mu\nu}$: It transforms homogeneously and thus the statement $F_{\mu\nu}(x)=0$ holds in any gauge. This suggests to transform (10.25) into a definition for a tensor measuring a non-zero curvature of space-time,

$$(\nabla_\alpha\nabla_\beta - \nabla_\beta\nabla_\alpha)T^{\mu\dots}_{\nu\dots} = [\nabla_\alpha, \nabla_\beta]T^{\mu\dots}_{\nu\dots} \neq 0\,. \tag{10.42}$$

Thus the curvature of space-time should be proportional to the area of a loop and the amount a tensor is rotated.

For the special case of a vector V^α we obtain with

$$\nabla_\rho V^\alpha = \partial_\rho V^\alpha + \Gamma^\alpha{}_{\beta\rho}V^\beta \tag{10.43}$$

first

$$\nabla_\sigma\nabla_\rho V^\alpha = \partial_\sigma(\partial_\rho V^\alpha + \Gamma^\alpha{}_{\beta\rho}V^\beta) + \Gamma^\alpha{}_{\kappa\sigma}(\partial_\rho V^\kappa + \Gamma^\kappa{}_{\beta\rho}V^\beta) - \Gamma^\kappa{}_{\rho\sigma}(\partial_\kappa V^\alpha + \Gamma^\alpha{}_{\beta\kappa}V^\beta)\,. \tag{10.44}$$

The second part of the commutator follows from the simple relabelling $\sigma \leftrightarrow \rho$ as

$$\nabla_\rho\nabla_\sigma V^\alpha = \partial_\rho(\partial_d V^\alpha + \Gamma^a{}_{\beta\sigma}V^\beta) + \Gamma^\alpha{}_{\kappa\rho}(\partial_\sigma V^\kappa + \Gamma^\kappa{}_{b\sigma}V^\beta) - \Gamma^\kappa{}_{\sigma\rho}(\partial_\kappa V^\alpha + \Gamma^\alpha{}_{\beta\kappa}V^\beta)\,. \tag{10.45}$$

Now we subtract the two equations using that $\partial_\rho\partial_\sigma = \partial_\sigma\partial_\rho$ and $\Gamma^\alpha{}_{\beta\rho} = \Gamma^\alpha{}_{\rho\beta}$,

$$[\nabla_\rho, \nabla_\sigma]V^\alpha = \left[\partial_\rho\Gamma^\alpha{}_{\beta\sigma} - \partial_\sigma\Gamma^\alpha{}_{\beta\rho} + \Gamma^\alpha{}_{\kappa\rho}\Gamma^\kappa{}_{\beta\sigma} - \Gamma^\alpha{}_{\kappa\sigma}\Gamma^\kappa{}_{\beta\rho}\right] V^\beta \equiv R^\alpha{}_{\beta\rho\sigma}V^\beta \,. \tag{10.46}$$

The tensor $R^\alpha{}_{\beta\rho\sigma}$ is called *Riemann* or *curvature tensor.* In problem **??**, you are asked to show that the tensor $R_{\alpha\beta\rho\sigma} = g_{\alpha\gamma}R^\gamma{}_{\beta\rho\sigma}$ is antisymmetric in the indices $\rho \leftrightarrow \sigma$, antisymmetric in $\alpha \leftrightarrow \beta$ and symmetric against an exchange of the index pairs $(\alpha\beta) \leftrightarrow (\rho\sigma)$. Therefore, we can construct out of the Riemann tensor only one non-zero tensor of rank two, contracting α either with the third or fourth index, $R^\rho{}_{\alpha\rho\beta} = -R^\rho{}_{\alpha\beta\rho}$. We define the Ricci tensor by

$$R_{\alpha\beta} = R^\rho{}_{\alpha\rho\beta} = -R^\rho{}_{\alpha\beta\rho} = \partial_\rho\Gamma^\rho{}_{\alpha\beta} - \partial_\beta\Gamma^\rho{}_{\alpha\rho} + \Gamma^\rho{}_{\alpha\beta}\Gamma^\sigma{}_{\rho\sigma} - \Gamma^\sigma{}_{\beta\rho}\Gamma^\rho{}_{\alpha\sigma} \,. \tag{10.47}$$

A further contraction gives the curvature scalar,

$$R = R_{\alpha\beta}g^{\alpha\beta} \,. \tag{10.48}$$

Example 10.2: Calculate the Ricci tensor R_{ab} and the scalar curvature R of the two-dimensional unit sphere S^2.
We have already determined the non-vanishing Christoffel symbols of the sphere S^2 as $\Gamma^\phi{}_{\vartheta\phi} = \Gamma^\phi{}_{\phi\vartheta} = \cot\vartheta$ and $\Gamma^\vartheta{}_{\phi\phi} = -\cos\vartheta\sin\vartheta$. We will show later that the Ricci tensor of a maximally symmetric space as a sphere satisfies $R_{ab} = Kg_{ab}$. Since the metric is diagonal, the non-diagonal elements of the Ricci tensor are zero too, $R_{\phi\vartheta} = R_{\vartheta\phi} = 0$. We calculate with

$$R_{ab} = R^c{}_{acb} = \partial_c\Gamma^c{}_{ab} - \partial_b\Gamma^c{}_{ac} + \Gamma^c{}_{ab}\Gamma^d{}_{cd} - \Gamma^d{}_{bc}\Gamma^c{}_{ad}$$

the $\vartheta\vartheta$ component, obtaining

$$\begin{aligned} R_{\vartheta\vartheta} &= 0 - \partial_\vartheta(\Gamma^\phi{}_{\vartheta\phi} + \Gamma^\vartheta{}_{\vartheta\vartheta}) + 0 - \Gamma^d{}_{\vartheta c}\Gamma^c{}_{\vartheta d} = 0 + \partial_\vartheta\cot\vartheta - \Gamma^\phi{}_{\vartheta\phi}\Gamma^\phi{}_{\vartheta\phi} \\ &= 0 - \partial_\vartheta\cot\vartheta - \cot^2\vartheta = 1 \,. \end{aligned}$$

From $R_{ab} = Kg_{ab}$, we find $R_{\vartheta\vartheta} = Kg_{\vartheta\vartheta}$ and thus $K = 1$. Hence $R_{\phi\phi} = g_{\phi\phi} = \sin^2\vartheta$.
The scalar curvature is (diagonal metric with $g^{\phi\phi} = 1/\sin^2\vartheta$ and $g^{\vartheta\vartheta} = 1$)

$$R = g^{ab}R_{ab} = g^{\phi\phi}R_{\phi\phi} + g^{\vartheta\vartheta}R_{\vartheta\vartheta} = \frac{1}{\sin^2\vartheta}\,\sin^2\vartheta + 1 \times 1 = 2 \,.$$

We can push the analogy further by remembering that the field-strength defined in Eq. (10.25) is a matrix. Writing out the implicit matrix indices of $F_{\mu\nu}$ in Eq. (10.25) gives

$$(F_{\mu\nu})_{ij} = \partial_\mu(A_\nu)_{ij} - \partial_\nu(A_\mu)_{ij} + \mathrm{i}g\left\{(A_\mu)_{ik}(A_\nu)_{kj} - (A_\nu)_{ik}(A_\mu)_{kj}\right\} \,. \tag{10.49}$$

Comparing this expression to

$$R^\alpha{}_{\beta\mu\nu} = \partial_\mu\Gamma^\alpha{}_{\beta\nu} - \partial_\nu\Gamma^\alpha{}_{\beta\mu} + \Gamma^\alpha{}_{\rho\mu}\Gamma^\rho{}_{\beta\nu} - \Gamma^\alpha{}_{\rho\nu}\Gamma^\rho{}_{\beta\mu} \tag{10.50}$$

we see that the first two indices of the Riemann tensor, α and β, correspond to the group indices ij in the field-strength tensor. This is in line with the relation of the potential $(A_\mu)_{ij}$ and the connection $\Gamma^\alpha{}_{\beta\nu}$ implied by (10.33).

10.3. Quantisation of gauge theories

10.3.1. Abelian case

We discussed already in Sec. 7.2 that we can derive the photon propagator only fixing a gauge. Now we reconsider this problem, and ask how we should modify the Lagrange density in order to be able to obtain the photon propagator. The Lagrange density that leads to the Maxwell equation is

$$\begin{aligned}\mathscr{L} &= -\frac{1}{4}\,F_{\mu\nu}F^{\mu\nu} = -\frac{1}{2}\left(\partial_\mu A_\nu \partial^\mu A^\nu - \partial_\mu A_\nu \partial^\nu A^\mu\right)\\ &= \frac{1}{2}\left(A_\nu \partial_\mu \partial^\mu A^\nu - A_\nu \partial_\mu \partial^\nu A^\mu\right) = \frac{1}{2} A_\mu \left[\eta^{\mu\nu}\Box - \partial^\mu \partial^\nu\right] A_\nu = \frac{1}{2}\,A^\mu D^{-1}_{\mu\nu} A^\nu\,,\end{aligned} \tag{10.51}$$

where we made a partial integration dropping as usual the surface term. Deriving the photon propagator requires to invert the term in the square bracket. Performing a Fourier transformation, we see that we should find the inverse of the operator

$$D^{-1}_{\mu\nu}(k) = k^2 P^{\mu\nu}_T(k) = k^2\left(\eta^{\mu\nu} - k^\mu k^\nu/k^2\right)\,. \tag{10.52}$$

We have already seen that this operator projects any four-vector on the three-dimensional subspace orthogonal to k. More formally, we see that $P^{\mu\nu}_T(k)$ is a projection operator,

$$P^{\mu\nu}_T P_{T\nu}{}^{\lambda} = P^{\mu\lambda}_T\,, \tag{10.53}$$

and has thus as only eigenvalues 0 and 1. Since $P^{\mu\nu}_T(k)$ is not the unit operator, it has at least one zero eigenvalue and is thus not invertible. More precisely, its trace is

$$P^{\mu}_{T\mu} = \eta_{\mu\nu}P^{\mu\nu}_T = \delta^\mu_\mu - 1 = 3\,, \tag{10.54}$$

and thus three eigenvalues are one and one eigenvalue is zero. The latter eigenvalue corresponds to $k_\mu P^{\mu\nu}_T = 0$, as required for a projection operator on the three-dimensional subspace orthogonal to k. The orthogonal part $\delta^\mu_\nu - P_{T\nu}{}^{\mu}$ is given by the longitudinal projection operator $P^{\mu\nu}_L = k^\mu k^\nu/k^2$.

We can invert $D^{-1}_{\mu\nu}$, if we choose a gauge such that the subspace parallel to k is included. The simplest choice is the Lorenz gauge. Imposing this gauge on the level of the Lagrangian means adding

$$\mathscr{L} \to \mathscr{L}_{\mathrm{e}ff} = \mathscr{L} + \mathscr{L}_{\mathrm{g}f} = \mathscr{L} - \frac{1}{2}(\partial^\mu A_\mu)^2\,. \tag{10.55}$$

More generally, we can add the term

$$\mathscr{L}_{\mathrm{g}f} = -\frac{1}{2\xi}(\partial^\mu A_\mu)^2 \tag{10.56}$$

that depends on the arbitrary parameter ξ. This group of gauges is employed in the proof of the renormalisability of gauge theories and is therefore called R_ξ gauge. The combined effective Lagrange density is thus

$$\mathscr{L}_{\mathrm{e}ff} = -\frac{1}{4}\,F_{\mu\nu}F^{\mu\nu} - \frac{1}{2\xi}(\partial^\mu A_\mu)^2 = \frac{1}{2}\,A_\nu\left[\eta^{\mu\nu}\Box - \left(1 - \frac{1}{\xi}\right)\partial^\mu\partial^\nu\right]A_\mu\,. \tag{10.57}$$

Fourier transforming the term in the square brackets, we obtain

$$P^{\mu\nu} = -k^2\eta^{\mu\nu} + (1-\xi^{-1})k^\mu k^\nu \,. \tag{10.58}$$

Now we split this expression into its transverse and longitudinal parts,

$$\begin{aligned} P^{\mu\nu} &= -\,k^2\left(P_T^{\mu\nu} + \frac{k^\mu k^\nu}{k^2}\right) + (1-\xi^{-1})k^\mu k^\nu & (10.59\text{a}) \\ &= -\,k^2 P_T^{\mu\nu} - \xi^{-1}k^2 P_L^{\mu\nu} \,. & (10.59\text{b}) \end{aligned}$$

Since $P_T^{\mu\nu}$ and $P_L^{\mu\nu}$ project on orthogonal subspaces, we obtain the inverse $P_{\mu\nu}^{-1}$ simply by inverting their prefactors, cf. problem 8.??. Thus the photon propagator in R_ξ gauge is given by

$$\mathrm{i}D_F^{\mu\nu}(k^2) = \frac{-\mathrm{i}P_T^{\mu\nu}}{k^2+\mathrm{i}\varepsilon} + \frac{-\mathrm{i}\xi P_L^{\mu\nu}}{k^2+\mathrm{i}\varepsilon} = \frac{-\mathrm{i}}{k^2+\mathrm{i}\varepsilon}\Big[\eta^{\mu\nu} - (1-\xi)\frac{k^\mu k^\nu}{k^2}\Big] \,. \tag{10.60}$$

Special cases are the Feynman gauge $\xi = 1$, the Landau gauge $\xi = 0$, while $\xi \to \infty$ corresponds to the unitary gauge. The arbitrary, ξ dependent part of the photon propagator vanishes in physical quantities, where it is matched between conserved currents with $\partial_\mu J^\mu(x) = 0$ or $k_\mu J^\mu(k) = 0$.

10.3.2. Non-abelian case

An important conceptional difference between abelian and non-abelian theories is that in the latter case the conserved Noether current is not gauge invariant, cf. problem 10.??. Moreover, the non-abelian gauge transformation (10.22c) adds not only a term $\partial_\mu\vartheta$ but mixes also the fields via the term $f^{abc}A_\mu^b\vartheta^c$. Therefore it is in general not guaranteed that the gauge-dependent unphysical degrees of freedom contained e.g. in the propagator (10.60) decouple and the quantisation of non-abelian theories becomes more challenging.

We consider first as a toy model for the generating functional of a Yang-Mills theory the two-dimensional integral

$$Z \propto \int \mathrm{d}x\mathrm{d}y\; \mathrm{e}^{\mathrm{i}S(x)} \,. \tag{10.61}$$

Since the integration extends from $-\infty$ to ∞, the y integration does not merely change the normalisation of Z but makes the integral ill-defined. We can eliminate the dangerous y integration by introducing a delta function,

$$Z \propto \int \mathrm{d}x\mathrm{d}y\; \delta(y)\mathrm{e}^{\mathrm{i}S(x)} \,. \tag{10.62}$$

Since the value of y in the delta function plays no role, we can replace $\delta(y)$ by $\delta(y-f(x))$ with an arbitrary function $f(x)$. If $y = f(x)$ is the solution of $g(x,y) = 0$, we obtain with

$$\delta(g(x,y)) = \frac{\delta(y-f(x))}{|\partial g/\partial y|} \tag{10.63}$$

assuming that $\partial g/\partial y > 0$

$$Z \propto \int \mathrm{d}x\mathrm{d}y\; \frac{\partial g}{\partial y}\delta(g)\mathrm{e}^{\mathrm{i}S(x)} \,. \tag{10.64}$$

Generalising this to n dimensions, we need n delta functions and have to include the Jacobian,

$$Z \propto \int \mathrm{d}^n x \mathrm{d}^n y \ \det\left(\frac{\partial g_i}{\partial y_j}\right) \prod_i \delta(g_i) \mathrm{e}^{\mathrm{i}S(x)} . \tag{10.65}$$

We now translate this toy example to the Yang-Mills case. The functions g are the gauge fixing conditions that we choose as

$$g^a(x) = \partial^\mu A^a_\mu(x) - \omega^a(x) , \tag{10.66}$$

where the $\omega^a(x)$ are arbitrary functions. The discrete index i corresponds to $\{x, a\}$, explaining why the gauge freedom results in an infinity: Although the integration measure of a compact gauge group is finite, the summation over $\mathbb{R}^4$ gives an infinite answer. Finally, we see that from the transformation law $A^a_\mu(x) \to A^{a\prime}_\mu(x) = A^a_\mu(x) - D^{ac}_\mu \vartheta^c(x)$ that the parameters ϑ^a correspond to the redundant coordinates y_i.

The generating functional for a Yang-Mills theory is thus with $\mathcal{D}A \equiv \prod_{\mu=0}^3 \prod_{a=1}^m \mathcal{D}A^a_\mu$ as short-cut given by

$$Z[0] \propto \int \mathcal{D}A \ \mathrm{Det}\left(\frac{\delta g^a}{\delta \vartheta^b}\right) \prod_{x,a} \delta(g^a) \mathrm{e}^{\mathrm{i}S_{YM}} , \tag{10.67}$$

where we set for the moment the sources to zero. Our task is to evaluate first $\delta g^a/\delta\vartheta^b$ and then to transform the determinant into the Lagrangian of new, auxiliary fields such that we can use the language of Feynman diagrams to perform perturbative calculations in the usual way. Inserting into the gauge fixing condition (10.66) the infinitesimal gauge transformation (10.22c), we obtain as change

$$\delta g^a(x) = -\partial^\mu D^{ab}_\mu \vartheta^b(x) . \tag{10.68}$$

Thus the required functional derivative is

$$\frac{\delta g^a(x)}{\delta \vartheta^b(y)} = -\partial^\mu D^{ab}_\mu \delta(x-y) . \tag{10.69}$$

We can eliminate the determinant remembering $\int \mathrm{d}\eta \mathrm{d}\bar\eta \ \mathrm{e}^{\bar\eta A \eta} = \det A$ from Eq. (8.103), expressing the Jacobian as a path integral over Graßmann variables c^a and $\bar c^a$,

$$\mathrm{Det}\left[\frac{\delta g^a(x)}{\delta \vartheta^b(y)}\right] \propto \int \mathcal{D}c \mathcal{D}\bar c \ \mathrm{e}^{\mathrm{i}S_{FP}} . \tag{10.70}$$

The corresponding Lagrangian is

$$\mathscr{L}_{FP} = -\bar c^a \partial^\mu D^{ab}_\mu c^b = (\partial^\mu \bar c^a)(D^{ab}_\mu c^b) = \partial^\mu \bar c^a \partial_\mu c^a + g f^{abc} \partial^\mu \bar c^a c^b A^c_\mu , \tag{10.71}$$

where we made a partial integration and inserted the definition of the covariant derivatives acting on the gauge field, Eq. (10.22c). As a result, we have recast the determinant as the kinetic energy of complex scalar fields c^a that interact with the gauge fields. Since we had to use for the scalar fields Graßmann variables c^a, their statistics is fermionic. Clearly, such fields should be seen as a purely mathematical construct and they are therefore called Faddeev-Popov ghosts. In an abelian theory as U(1), the interaction term in Eq. (10.71) is absent and ghost fields decouple. Since they change then only the normalisation of the path integral, they can be omitted in QED.

Next we have to eliminate the $\delta(g^a(x))$. They contain the arbitrary functions $\omega^a(x)$, but the path integral does not dependent on them. Thus we have the freedom to multiply with a chosen function $f(\omega^a)$, thereby changing only the normalisation. Our aim is to generate after integrating over the delta functions a term $\exp(\mathrm{i}S_{\mathrm{g}f})$, as in the case of QED. Choosing

$$Z \to \exp\left(-\frac{\mathrm{i}}{2\xi}\int \mathrm{d}^4x\, \omega^a(x)\omega^a(x)\right) Z\,, \tag{10.72}$$

integrating $\prod_{x,a}\delta(g^a)\exp\left(-\frac{\mathrm{i}}{2\xi}\int \mathrm{d}^4x\, \omega^a(x)\omega^a(x)\right)$ with the help of $\delta(g^a)$ and (10.66), we obtain as gauge-fixing term the desired

$$\mathscr{L}_{\mathrm{g}f} = -\frac{1}{2\xi}\partial^\mu A^a_\mu \partial^\nu A^a_\nu\,. \tag{10.73}$$

The complete Lagrange density $\mathscr{L}_{\mathrm{e}ff}$ of a non-abelian gauge theory consists thus of four parts,

$$\mathscr{L}_{\mathrm{e}ff} = \mathscr{L}_{\mathrm{Y}M} + \mathscr{L}_{\mathrm{g}f} + \mathscr{L}_{\mathrm{F}P} + \mathscr{L}_{\mathrm{s}}\,, \tag{10.74}$$

where the last one couples sources linearly to the fields,

$$\mathscr{L}_{\mathrm{s}} = J^\mu A_\mu + \bar{\eta}c + \bar{c}\eta\,. \tag{10.75}$$

We break both $\mathscr{L}_{YM}$ and $\mathscr{L}_{\mathrm{F}P}$ into a piece of $\mathcal{O}(g^0)$ defining the free propagator, and pieces of $\mathcal{O}(g)$ corresponding to a three gluon and a two ghost-gluon vertex, respectively, and a four gluon vertex of $\mathcal{O}(g^2)$. After a partial integration of the free part, we obtain

$$\begin{aligned}\mathscr{L}_{YM} + \mathscr{L}_{\mathrm{F}P} = {}& \frac{1}{2}A^a_\nu\left(\eta^{\mu\nu}\Box - \partial^\mu\partial^\nu\right)A^a_\mu - gf^{abc}A^a_\mu A^b_\nu \partial^\mu A^{c\nu} - \frac{g^2}{4}f^{abe}f^{cde}A^a_\mu A^b_\nu A^{c\mu}A^{d\nu} \\ & - \bar{c}^a\Box c^a + gf^{abc}\partial^\mu\bar{c}^a c^b A^c_\mu.\end{aligned} \tag{10.76}$$

The Feynman rules can now be read off after Fourier transforming into momentum space, cf. problem 10.??. Combining the resulting expression with $\mathscr{L}_{\mathrm{g}f}$, we see that the gluon propagator is diagonal in the group indices and otherwise identical to the photon propagator in R_ξ gauge. The ghost propagator is the one of a massless scalar particle,

$$\Delta_{ab}(k) = \frac{\delta_{ab}}{k^2 + \mathrm{i}\varepsilon}\,. \tag{10.77}$$

Being a fermion, a closed ghost loop introduces however a minus sign.

Non-covariant gauges The introduction of ghost fields can be avoided, if one uses non-covariant gauges which depend on an arbitrary vector n^μ. An example used often in QED is the Coulomb or radiation gauge,

$$\partial_\mu A^\mu - (n_\mu \partial^\mu)^2 = 0 \tag{10.78}$$

with $n_\mu = (1, 0, 0, 0)$. In QCD, one employs often the set of gauges

$$n_\mu A^\mu_a = 0, \quad a = 1, \ldots, 8 \tag{10.79}$$

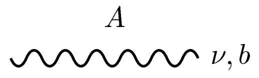
with the constant vector n^μ. More specifically, one calls the case $n^2 = 0$ light-cone, $n^2 < 0$ axial and $n^2 > 0$ temporal gauge. They have in common that the Faddeev-Popov determinant does not depend on A_a^μ, and can be absorbed in the normalisation of the path integral, exercise 10.**??**. While non-covariant gauges thus bypass the introduction of unphysical particles in loop graphs, the resulting propagators are unhandy. Moreover, they contain spurious singularities which require care. Therefore in practical all applications the use of the R_ξ gauge is advantageous.

Let us finally comment on the case of external gluons. In the case of photons, we can sum their polarisation states using $\sum_{r=0}^{3} \varepsilon_\mu^{(r)*} \varepsilon_\nu^{(r)} = -\eta_{\mu\nu}$, since the nonphysical degrees of freedom cancel in physical observables. In the non-abelian case, we can use this "trick" only in the case of a single external gluon. For two or more external gluons, we have to employ the polarisation sum derived in problem 7.**??**, since the non-abelian vertices mix physical and non-physical degrees of freedom. As a result, the conserved Noether current is not gauge invariant, and we cannot use the argument of section 7.2. Alternatively, we can use the R_ξ-gauge if we include Faddeev-Popov ghosts also as *external* particles. In order to subtract correctly the unphysical contributions to the squared matrix elements, one has to add the factor $(-1)^n$ to a term $\mathcal{A}_i \mathcal{A}_j^*$ with $2n$ Faddeev-Popov ghosts [Nac90].

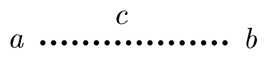
10.A. Appendix: Feynman rules for an unbroken gauge theory

The Feynman rules for a non-broken Yang-Mills theory as QCD are given in the R_ξ gauge; for the abelian case of QED set the structure constants $f_{abc} = 0$, $T = 1$ and replace $g_s \to eq_f$, where q_f is the electric charge of the fermion in units of the elementary charge $e > 0$. The momentum flow is indicated by the thin arrow: For instance, all momenta are chosen as in-going in the triple gauge vertex (10.82).

Propagators

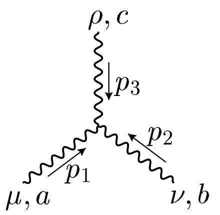


$$-\,\mathrm{i}\delta_{ab}\left[\frac{\eta_{\mu\nu}}{k^2+\mathrm{i}\epsilon} - (1-\xi)\frac{k_\mu k_\nu}{(k^2)^2}\right] \tag{10.80}$$



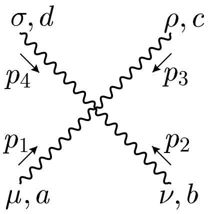
$$\delta_{ab}\,\frac{\mathrm{i}}{k^2+\mathrm{i}\epsilon} \tag{10.81}$$

Triple Gauge Interactions



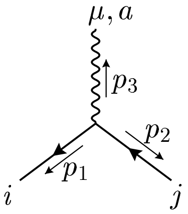
$$\begin{aligned} -g_s f^{abc}[\;& \eta^{\mu\nu}(p_1-p_2)^\rho + \eta^{\nu\rho}(p_2-p_3)^\mu \\ & +\eta^{\rho\mu}(p_3-p_1)^\nu] \\ & p_1+p_2+p_3=0 \end{aligned} \tag{10.82}$$

Quartic Gauge Interactions



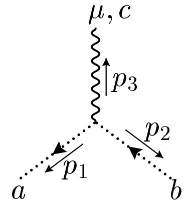
$$\begin{aligned} -\mathrm{i}g_s^2\Big[\; & f_{eab}f_{ecd}(\eta_{\mu\rho}\eta_{\nu\sigma} - \eta_{\mu\sigma}\eta_{\nu\rho}) \\ & + f_{eac}f_{edb}(\eta_{\mu\sigma}\eta_{\rho\nu} - \eta_{\mu\nu}\eta_{\rho\sigma}) \\ & + f_{ead}f_{ebc}(\eta_{\mu\nu}\eta_{\rho\sigma} - \eta_{\mu\rho}\eta_{\nu\sigma})\Big] \\ & p_1 + p_2 + p_3 + p_4 = 0 \end{aligned} \tag{10.83}$$

Fermion Gauge Interactions



$$-\,\mathrm{i}\,g_s\gamma^\mu T^a_{ij} \tag{10.84}$$

Ghost Interactions



$$\begin{aligned} & g_s\; f^{abc}p_1^\mu \\ & p_1 \quad + p_2 + p_3 = 0 \end{aligned} \tag{10.85}$$

Summary

Requiring local symmetry under a gauge group as SU(n) or SO(n) specifies the self-interactions of massless gauge boson as well as their couplings to fermions and scalars. The presence of self-interactions implies that a pure Yang-Mills theory is non-linear. The gauge invariant derivative D_μ is the analogon to the covariant derivative ∇_μ of gravity, while the field-strength corresponds to the Riemann tensor: Both measure the rotation of a vector which is parallel-transported along a closed loop. The quantisation of Yang-Mills theories in the covariant R_ξ gauge leads to ghost particles: These fermionic scalars compensate the unphysical degrees of freedom still contained in the gauge fields A_μ using a covariant gauge fixing condition as $\partial_\mu A^\mu = 0$.

Note also the interplay between local and global symmetries: A global symmetry transformation U maps a physical state onto a different physical state with the same properties, implying via Noether's theorem a conserved current. A local symmetry transformation $U(x)$ maps a physical state on itself, implying a redundancy in our description of the system. Since local symmetries contain global transformations as a subgroup, they imply always also the conservation of global charges via Noether's theorem.

Further reading

The Feynman rules in the appendix are taken from [RS12]. This article contains all Feynman rules for the SM in a convention independent notation which allows an easy comparison of references with differing conventions. Current conservation in non-abelian theories is discussed e.g. in [LP13]. The extension of the helicity formalism to QCD, where it leads to both phenomenological useful and theoretically interesting results, is discussed by [Pes11], [Sch13], and [Wei16].

11. Renormalisation I: Perturbation theory

We encountered three examples of divergent loop integrals discussing the $\lambda\phi^4$ theory. In these cases, it was possible to subtract the infinities in such a way that we obtained finite observables which depend only on the experimentally measured values of m, λ and ρ. The aim of this and the following chapter is to obtain a better understanding of this renormalisation procedure. We will see that the $\lambda\phi^4$ theory as well as the electroweak and strong interactions of the SM are examples for renormalisable theories: For such theories, the renormalisation of the finite number of parameters contained in the classical Lagrangian is sufficient to make all observables finite at any order perturbation theory.

11.1. Overview

Why renormalisation at all? We are using perturbation theory with the free, non-interacting Lagrangian as starting point to evaluate non-linear quantum field theories. Interactions change however the parameters of the free theory, as we know already both from classical electrodynamics and quantum mechanics. In the former case, Lorentz studied 1904 the connection between the measured electron mass m_{phy}, its mechanical or inertial mass m_0 and its electromagnetic self-energy m_{el} in a toy model. He described the electron as a spherically symmetric uniform charge distribution with radius $r_{\rm e}$, obtaining

$$m_{\rm phy} = m_0 + m_{\rm el} = m_0 + \frac{4e^2}{5r_{\rm e}} \,. \tag{11.1}$$

Special relativity forces us to describe the electron as a point particle: Taking thus the limit $r_{\rm e} \to 0$, classical electrodynamics implies an infinite "renormalisation" of the "bare" electron mass m_0 by its electromagnetic self-energy m_{el}.

Another familiar example for renormalisation appears in quantum mechanics. Perturbation theory is possible, if the Hamilton operator H can be split into a solvable part $H^{(0)}$ and an interaction λV,

$$H = H^{(0)} + \lambda V \,, \tag{11.2}$$

and the parameter λ is small. Using then as starting point the normalised solutions $|n^{(0)}\rangle$ of $H^{(0)}$,

$$H^{(0)}|n^{(0)}\rangle = E_n^{(0)}|n^{(0)}\rangle \,, \tag{11.3}$$

we can find the eigenstates $|n\rangle$ of the complete Hamiltonian H as a power-series in λ,

$$|n\rangle = |n^{(0)}\rangle + \lambda|n^{(1)}\rangle + \lambda^2|n^{(2)}\rangle + \ldots \tag{11.4}$$

Since we started with normalised states, $\langle n^{(0)}|n^{(0)}\rangle = 1$, the new states $|n\rangle$ are not longer correctly normalised. Thus going from free (or "bare") to interacting states requires to renormalise the states,

$${}_R\langle n|\, n\rangle_R = 1 \quad \Rightarrow \quad |n\rangle_R \equiv Z^{1/2}\,|n\rangle \,. \tag{11.5}$$

A very similar problem we encountered introducing the LSZ formalism. In the parlance of field theory, we continue often to call this procedure wave-function renormalisation, although Z renormalises field operators.

Why regularisation at all? The familiar process of renormalisation becomes more obscure by the fact that the renormalisation constants are infinite in most quantum field theories. Mathematical manipulations as shifting the integration variable in a divergent loop integral are only well-defined, if we convert first these integrals into convergent ones. Thus we have to regularise as first step, i.e. employing a method which makes our expressions finite, so that our mathematical manipulations are well-defined and we can perform the renormalisation. You should keep in mind that the two operations, regularisation and renormalisation, are logically independent: Renormalisation of the parameters in the free theory is necessary because they are changed by interactions. This change may be finite, as the change of the photon mass in a plasma, and no regularisation is necessary.

The second question to ask is why the renormalisation constants are infinite, or in other words why do we have regularise at all? There are (at least) two possible answers to this question: Either we use a bad theory as starting point, i.e. the full quantum theory defined non-perturbatively by its generating functional $Z[J]$ is ill-defined. Or we employ a bad expansion scheme evaluating $Z[J]$ in perturbation theory.

Example 11.1: An example for a bad expansion is the following toy model for the $\lambda\phi^4$ interaction,

$$Z(\lambda) = \int_{-\infty}^{\infty} \mathrm{d}x\, \mathrm{e}^{-\frac{1}{2}x^2-\lambda x^4} \overset{?}{=} \int_{-\infty}^{\infty} \mathrm{d}x \left[1 - \lambda x^4 + \frac{(\lambda x^4)^2}{2!} - \ldots\right] \mathrm{e}^{-\frac{1}{2}x^2} . \tag{11.6}$$

The LHS is well-defined for $\lambda > 0$. Doing perturbation theory and summing up the first N terms of the expansion on the RHS results in an alternating series,

$$Z_N(\lambda) = \sum_{n=0}^{N} \frac{(-\lambda)^n}{n!} \int_{-\infty}^{\infty} \mathrm{d}x\, x^{4n} \mathrm{e}^{-x^2/2} = \sum_{n=0}^{N} a_n \lambda^n$$

with

$$a_n = \frac{(-1)^n}{n!}\, 2^{2n+1/2}\, \Gamma(2n+1/2) .$$

The coefficients a_n of this series grow like a factorial and thus the convergence radius of the expansion is zero. Plotting $Z_N(\lambda)/Z(0)$ for the first few N as function of λ, problem 11.??, you see firstly that adding more terms makes the expansion worse beyond at certain value $\lambda_{\max}(N)$, and secondly that $\lambda_{\max}(N) \to 0$ for $N \to \infty$, see [FHS12] for more details.

It should be not too surprising that the expansion (11.6) has a zero convergence radius: Moving from $\lambda > 0$ to $\lambda < 0$ changes fundamentally physics, since the vacuum is unstable for arbitrarily small negative λ. An interesting consequence of the failure of perturbation theory is that the complete theory may contain additional non-perturbative physics. Next we look at an example where we start from a bad theory.

Example 11.2: We discussed in problem 2.?? the scattering on a short-range potential in $d = 1$, and found that no consistent solution exists for an odd potential. We rephrase this problem now in a language close to the one used in QFT. The perturbative expansion of the S-matrix in quantum mechanics is given by

$$\langle p_f | S | p_i \rangle = 2\pi\delta(E_i - E_f) \left[\langle p_f | V | p_i \rangle + \int \mathrm{d}p \, \langle p_f | V | p \rangle \frac{\mathrm{i}}{E - p^2/2 + \mathrm{i}\varepsilon} \langle p | V | p_i \rangle + \cdots \right]$$

for $p_i \neq p_f$. Recoiling on the infinitely heavy static source, the (virtual) particle in the intermediate states can have any momentum p while its energy is conserved. With $V_0(x) = c_0\delta(x)$, it follows $\langle p_f | V_0 | p \rangle = c_0/(2\pi)$ and then the momentum integral in the 2.nd order correction becomes

$$\left(\frac{c_0}{2\pi}\right)^2 \int \mathrm{d}p \frac{\mathrm{i}}{E - p^2/2 + \mathrm{i}\varepsilon} \,.$$

Thus this momentum integral, and similarly those at higher orders, are well-defined. Next we set $c_0 = 0$. Using then $\delta' f = -\delta f'$, we obtain $\langle p_f | V_1 | p \rangle = \mathrm{i}c_1(p_f - p)/(2\pi)$ and thus the momentum integral

$$\left(\frac{c_1}{2\pi}\right)^2 \int \mathrm{d}p \frac{\mathrm{i}(p - p_i)(p - p_f)}{E - p^2/2 + \mathrm{i}\varepsilon}$$

is linearly divergent. The divergence means that the scattering probability is sensitive to arbitrarily high momentum modes. We can understand this behaviour looking at the wave-function $\psi(x)$: Because the potential is odd, also $\psi(x)$ is odd and thence has to change rapidly within $|x| < a$. As result, its Fourier transform $\psi(k)$ necessarily contains also high-frequency modes.

In this simple toy-model, the natural way to solve the UV divergence problem is to replace the mathematical idealisation of a delta-function like potential by the true, smooth potential. If we either do not know the "true" potential or if we insist that a delta-function like potential captures all the physics contained in a scattering process at a short-range potential, then we have to regularise the potential, replacing $V(x) = c_1\delta'(x)$ e.g. by

$$V(x) = c_1 \frac{\delta(x + a) - \delta(x - a)}{2a} \,.$$

In this way, we eliminate high-frequency modes with $p \gg 1/a$. Repeating the computation of the transmission amplitude, we find $T \simeq \mathrm{i}ap/c_1^2$. Hence $c_R \equiv c_1^2/a$ plays the role of an effective coupling constant in the regularised theory. Physical observables like the transmission amplitude depend only on the single parameter c_R, if we rescale $c_1(a) \propto a^{-1/2}$. Thus this simple example from quantum mechanics exhibits the key features of a UV divergent QFT: We regularise the theory, cutting off UV modes. Requiring the independence of physical observables from the cutoff scale, we obtain running parameters.

It is very likely that our favourite $\lambda\phi^4$ interaction suffers from both diseases: First, the expansion in λ is not convergent but results in an asymptotic series. Second, the full theory contains only the trivial $\lambda = 0$ case as consistent solution. Even if the interacting theory may be mathematically inconsistent, it can however be used as an effective model describing physics up to a finite energy scale.

Regularisation methods We have already seen that the regularisation of divergent loop integrals can be done in various ways. In general, one reparametrises the integral in terms of a parameter Λ (or ε) called regulator such that the integral becomes finite for a finite value of the regulator, while the limit $\Lambda \to \infty$ (or $\varepsilon \to 0$) returns the original integral.

- We can avoid UV divergences evaluating loop-integrals introducing an (Euclidean) momentum cutoff Λ. Somewhat more sophisticated, we could introduce instead of a hard cutoff a smooth function which suppresses large momenta. Using Schwinger's proper-time representation (4.90) we can cut-off large momenta setting

$$\frac{1}{p^2+m^2} \to \frac{\mathrm{e}^{-(p^2+m^2)/\Lambda^2}}{p^2+m^2} = \int_{\Lambda^{-2}}^{\infty} \mathrm{d}s\, \mathrm{e}^{-s(p^2+m^2)} \,. \tag{11.7}$$

Although conceptual easy, both regularisation schemes violate generically all symmetries of our theory. This is not a principal flaw, since we should be able to recover these symmetries in the limit $\Lambda \to \infty$. However, this "recovery process" may be non-trivial to perform. Moreover, intermediate calculations become much more transparent if we can use the symmetries of the theory, and therefore these schemes are in practise not used except for the simplest cases.

- Pauli-Villars regularisation is a scheme where one adds heavy particles having the same quantum numbers and couplings as the originals ones. Thus the propagator of a massless scalar particle is changed to

  $$\frac{1}{k^2+\mathrm{i}\varepsilon} \to \frac{1}{k^2+\mathrm{i}\varepsilon} + \sum_i \frac{a_i}{k^2-M_i^2+\mathrm{i}\varepsilon} \,.$$

  For $k^2 \ll M_i^2$, physics is unchanged, while for $k^2 \gg M_i^2$ and $a_i < 0$ the combined propagator scales as M_i^2/k^4 and the convergence of loop integrals improves. Since the heavy particles enter with the wrong sign, they are unphysical ghosts and serve only as a mathematical tool to regularise loop diagrams. Pauli-Villars regularisation respects the gauge invariance of QED, if the heavy particles are coupled gauge invariantly to the photon.

- Lattice regularisation replaces the continuous space-time by a discrete lattice. The finite lattice spacing a introduces a momentum cutoff, eliminating all UV divergences. Moreover the (Euclidean) path-integral becomes well-defined and can be calculated numerically without the need to do perturbation theory. Thus this approach is particularly useful in the strong-coupling regime of QCD where it has been used to calculate static quantities as e.g. the hadron mass spectrum. Note that lattice regularisation for finite a respects gauge symmetries, but spoils the translation and Lorentz symmetry of the underlying QFT. Nevertheless, one recovers in the limit $a \to 0$ a relativistic QFT. A longstanding problem of lattice theory was how to implement correctly chiral fermions. This question was solved around the year 2000 and thus the SM can be defined now in a mathematically consistent, non-perturbative way as a lattice theory.

- Dimensional regularisation (DR) is the method we applied in the calculations of the one-loop diagrams of the $\lambda\phi^4$ theory. While DR has the important virtue of preserving Lorentz and gauge invariance, it is one of the least intuitive regularisation methods. We will show later that an integral without mass scale is zero in DR, e.g. $\int \mathrm{d}^d k k^{-2} = 0$. This example shows that the integration measure we implement using physical requirements with DR is not positive—as a mathematician would require. In problem 11.**??**, we examine how DR modifies the range of momentum values contributing to Feynman integrals.

  Using DR with fermions, we have to extend the Clifford algebra to d dimensions. A natural choice is $\mathrm{tr}(\gamma^\mu\gamma^\nu) = d\eta^{\mu\nu}$ and $\mathrm{tr}(1) = 4$. Problematic is however the treatment of $\gamma^5 \equiv \mathrm{i}\gamma^0\gamma^1\gamma^2\gamma^3$ relying heavily on $d = 4$.

- Various other regularisation methods as e.g. zeta function regularisation or point splitting methods exists.

Even fixing a regularisation method as e.g. DR, we can choose various renormalisation *schemes*. Four popular choices are

- on-shell renormalisation. In this scheme, we choose the subtraction such that the on-shell masses and couplings coincide with the corresponding values measured in processes with zero momentum transfer q. For instance, we define the renormalised electric charge via the Thomson limit of the Compton scattering amplitude. While this choice is very intuitive, it is not practical for QCD: We will see soon that in this theory scattering amplitudes calculated in perturbation theory become ill-defined in the limit $q^2 \to 0$.

- The momentum subtraction (MoM) scheme is a generalisation of the on-shell scheme which can be applied also to QCD. Here we subtract from the Green functions counter-terms such that the corrections are zero at a fixed space-like four-momentum $p^2 = -\mu^2$. In this way, divergences in the limit $q^2 \to 0$ are avoided.

- In the minimal subtraction (MS) scheme, we subtract only the divergent $1/\varepsilon$ poles.

- In the modified minimal subtraction $\overline{\mathrm{M}S}$ (read em-es-bar) scheme, we subtract also the $\ln(4\pi) - \gamma$ term appearing frequently. This scheme gives more compact expressions than the others and is most often used in theoretical calculations.

The main advantage of the MS and $\overline{\mathrm{M}S}$ schemes is that they are mass independent, i.e. that the subtraction terms do not depend on the particle masses. This independence simplifies the derivation of “running couplings” (cf. with the calculation of $\lambda(\mu)$ in section 4.3.4). As a drawback of the MS and $\overline{\mathrm{M}S}$ schemes, quantities like the electron mass calculated in these schemes, m_e^{MS} or $m_e^{\overline{MS}}$, have to be translated into the physical mass m_e.

At a fixed order perturbation theory, predictions and reliability of different schemes vary for given external parameters: A simple example is the change from the MS to the $\overline{\mathrm{M}S}$ scheme which are connected by $\tilde{\mu}^2 = 4\pi\mu^2 \mathrm{e}^{-\gamma}$. Thus this transition is equivalent to a change of the renormalisation scale, altering thereby the size of the $\ln(\mu^2)$ term and thus the strength of the running coupling. More drastic changes result moving from a mass independent to a mass dependent scheme, or comparing DR with other schemes. As a result, running couplings which are small enough to allow perturbation theory in one scheme may be prohibitive large in other schemes.

11.2. Anomalous magnetic moment of the electron

After this overview, let us move on to the calculation of the magnetic moment of the electron which is shifted by loop corrections from the tree-level value $g = 2$ you derived in problem 8.**??**. Apart from being the first successful loop calculation in the history of QFT, this process illustrates also several generic properties of loop graphs in renormalisable theories like QED.

Vertex function The tree-level interaction $\mathscr{L}_{int} = -e\bar{\psi}\gamma^\mu\psi A_\mu$ between an electron and a photon corresponds in momentum space to $e\bar{u}(p')\gamma^\mu u(p)\varepsilon_\mu(q)$. Since loop integrals depend generally on the external momenta, the tree-level vertex γ^μ is modified by loop graphs as

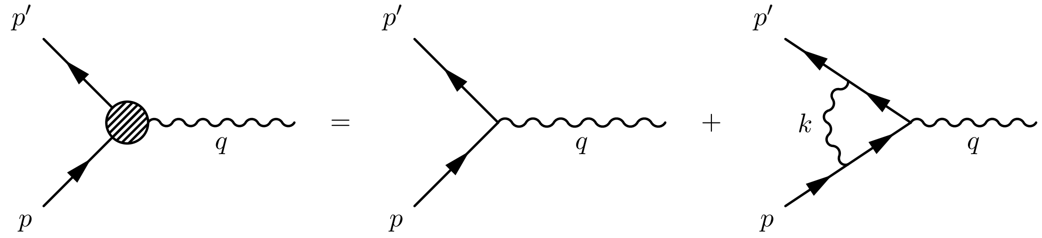


Figure 11.1.: The general vertex for the interaction of a fermion with a photon and its perturbative expansion within QED, corrections in external lines are omitted.

the one shown in Fig. 11.1 and becomes a function of the momenta, $\Lambda^\mu(p,p',q)$. We want to write down the most general form of the vertex function Λ^μ for the coupling between an external electromagnetic field and an on-shell Dirac fermion, consistent with the symmetries of the problem. It is as usually convenient to apply the tensor method, i.e. to express Λ^μ as a sum of linearly independent rank-1 tensors multiplied by scalar functions.

Translation invariance implies $q = p' - p$ and thus Λ^μ is only a function of two momenta which we choose as p and p'. Since $p^2 = p'^2 = m^2$, the only non-trivial scalar variable in the problem is $p \cdot p'$. We choose to use the equivalent quantity $q^2 = (p'-p)^2$ as the variable on which the arbitrary scalar functions in our ansatz for Λ^μ depend. Next we have to form all possible vectors out of the momenta p_μ and p'_μ and the 16 basis elements (8.40) of the Clifford algebra. Restricting ourselves to QED, we have to impose additionally parity conservation what forbids the use of γ^5. Hence the most general ansatz compatible with Poincaré invariance and parity is

$$\Lambda^\mu(p,p') = A(q^2)\gamma^\mu + B(q^2)p^\mu + C(q^2)p'^\mu + D(q^2)\sigma^{\mu\nu}p_\nu + E(q^2)\sigma^{\mu\nu}p'_\nu\,. \tag{11.8}$$

Current conservation requires $q_\mu\Lambda^\mu(p,p') = 0$ and leads to $C = B$ and $E = -D$. Hence

$$\Lambda^\mu(p,p') = A(q^2)\gamma^\mu + B(q^2)(p^\mu + p'^\mu) + D(q^2)\sigma^{\mu\nu}q_\nu\,. \tag{11.9}$$

Hermiticity finally implies that A, B are real and D is purely imaginary.

Gordon decomposition We derive now an identity that allows us to eliminate one of the three terms in Eq. (11.9), if we sandwich Λ^μ between two spinors which are on-shell. We evaluate

$$F^\mu = \bar{u}(p')\left[\not{p}'\gamma^\mu + \gamma^\mu\not{p}\right]u(p) \tag{11.10}$$

first using the Dirac equation for the two on-shell spinors, finding

$$F^\mu = 2m\bar{u}(p')\gamma^\mu u(p)\,. \tag{11.11}$$

Secondly, we can use $\gamma^\mu\gamma^\nu = \eta^{\mu\nu} - \mathrm{i}\sigma^{\mu\nu}$, obtaining

$$F^\mu = \bar{u}(p')\left[(p'+p)^\mu + \mathrm{i}\sigma^{\mu\nu}(p'-p)_\nu\right]u(p)\,. \tag{11.12}$$

Equating (11.11) and (11.12) gives the Gordon identity: It allows us to separate the Dirac current into a part proportional to $(p+p')^\mu$, i.e. with the same structure as a scalar current, and a part vanishing for $q = p' - p \to 0$ which couples to the spin of the fermion,

$$\bar{u}(p')\gamma^\mu u(p) = \bar{u}(p')\left[\frac{(p'+p)^\mu}{2m} + \frac{\mathrm{i}\sigma^{\mu\nu}(p'-p)_\nu}{2m}\right]u(p)\,. \tag{11.13}$$

Using the results from problem 8.**??**, we can identify in the non-relativistic limit the second term as contribution to the magnetic moment of the fermion.

The Gordon identity shows that the three terms in Eq. (11.9) are not independent. Depending on the context, we can eliminate therefore the most annoying term in the vertex function. We follow conventions and introduce the (real) form-factors $F_1(q^2)$ and $F_2(q^2)$ by

$$\Lambda^\mu(p,p') = F_1(q^2)\gamma^\mu + F_2(q^2)\frac{\mathrm{i}\sigma^{\mu\nu}q_\nu}{2m} = \tag{11.14a}$$

$$= F_1(q^2)\frac{(p'+p)^\mu}{2m} + [F_1(q^2) + F_2(q^2)]\frac{\mathrm{i}\sigma^{\mu\nu}q_\nu}{2m}\,. \tag{11.14b}$$

The form-factor F_1 is the coefficient of the electric charge, $eF_1(q^2)\gamma^\mu$, and should thus go to one for small momentum transfer, $F_1(0) = 1$. Therefore the magnetic moment of an electron is shifted by $1 + F_2(0)$ from the tree-level value $g = 2$. The deviation $a \equiv (g-2)/2$ is called anomalous magnetic moment, the two form-factors are often called electric and magnetic form-factors.

Note the usefulness of the procedure to express the vertex function using only general symmetry requirements but not a specific theory for the interaction: Equation (11.14a) allows experimentalists to present their measurements using only two scalar functions which in turn can be easily compared to predictions of specific theories.

Anomalous magnetic moment After having discussed the general structure of the electromagnetic vertex function, we turn now to its calculation in perturbation theory for the case of QED. The Feynman diagrams contributing to the matrix element at $\mathcal{O}(e^3)$ with wave-functions as external lines are shown in Fig. 11.1, where we omit self-energy corrections in the external lines: As we will see soon, the later do not contribute to the anomalous magnetic moment. We separate the matrix element into the tree-level part and the one-loop correction, $-\mathrm{i}e\bar{u}(p')\left[\gamma^\mu + \Gamma^\mu\right]u(p)$. Using the Feynman gauge for the photon propagator, we obtain

$$\Gamma^\mu(p,p') = \int \frac{\mathrm{d}^4k}{(2\pi)^4}\frac{-\mathrm{i}}{k^2+\mathrm{i}\varepsilon}(-\mathrm{i}e\gamma^\nu)\frac{\mathrm{i}}{\not{p}'+\not{k}-m+\mathrm{i}\varepsilon}\gamma^\mu\frac{\mathrm{i}}{\not{p}+\not{k}-m+\mathrm{i}\varepsilon}(-\mathrm{i}e\gamma_\nu)\,. \tag{11.15}$$

This integral is logarithmically divergent for large k,

$$\int^\Lambda \mathrm{d}k\,\frac{k^3}{k^2k^2} \propto \ln\Lambda\,. \tag{11.16}$$

Before we perform the explicit calculation, we want to understand if this divergence is connected to a specific kinematical configuration of the momenta. We split therefore the vertex correction into an on-shell and an off-shell part,

$$\Gamma^\mu(p,p') = \Gamma^\mu(p,p) + [\Gamma^\mu(p,p') - \Gamma^\mu(p,p)] \equiv \Gamma^\mu(p,p) + \Gamma^\mu_{off}(p,p')\,. \tag{11.17}$$

Next we rewrite the first fermion propagator using the identity $(A+B)^{-1} = B^{-1} - B^{-1}AB^{-1} + \ldots$ for small $A = p' - p$ as

$$\frac{1}{\not p' + \not k - m} = \frac{1}{\not p + \not k - m + (\not p' - \not p)} = \tag{11.18a}$$

$$= \frac{1}{\not p + \not k - m} - \frac{1}{\not p + \not k - m}(\not p' - \not p)\frac{1}{\not p + \not k - m} + \ldots \tag{11.18b}$$

The first term of this expansion leads to the logarithmic divergence of the loop integral for large k. In contrast, the remainder of the expansion that vanishes for $p' - p = q \to 0$ contains additional powers of $1/k$ and is thus convergent. Hence the UV divergence is contained solely in the on-shell part of the vertex correction, while the function $\Gamma^\mu_{off}(p,p') = \Gamma^\mu(p,p') - \Gamma^\mu(p,p)$ is well-behaved. Moreover, we learn from Eq. (11.14a) that the divergence is confined to $F_1(0)$, while $F_2(0)$ is finite. This is good news: The divergence is only connected to a quantity already present in the classical Lagrangian, the electric charge. Thus we can predict the function $\Gamma^\mu(p,p')$ for all values $p' \neq p$, after we have renormalised the electric charge in the limit of zero momentum transfer.

We now calculate the vertex function (11.15) explicitly. We set

$$\Gamma^\mu(q) = -\mathrm{i}e^2 \int \frac{\mathrm{d}^4k}{(2\pi)^4} \frac{\mathcal{N}^\mu(k)}{[(p'+k)^2 - m^2]\,[(p+k)^2 - m^2]\; k^2} \tag{11.19}$$

with

$$\mathcal{N}^\mu = \gamma^\nu(\not p' + \not k + m)\gamma^\mu(\not p + \not k + m)\gamma_\nu\,. \tag{11.20}$$

Then we combine the propagators introducing as Feynman parameter integrals

$$\frac{1}{xyz} = 2\int_0^1 \mathrm{d}\alpha \int_0^{1-\alpha} \mathrm{d}\beta\, \frac{1}{[z + \alpha(x-z) + \beta(y-z)]^3} = 2\int_0^1 \mathrm{d}\alpha \int_0^{1-\alpha} \mathrm{d}\beta\, \frac{1}{D}\,. \tag{11.21}$$

Setting $z = k^2$, we obtain

$$D = \{k^2 + \alpha[(p'+k)^2 - m^2 - k^2] + \beta[(p+k)^2 - m^2 - k^2]\}^3\,. \tag{11.22}$$

The complete calculation of the vertex function (11.15) for arbitrary off-shell momenta is already quite cumbersome. In order to shorten the calculation, we restrict ourselves therefore to the part contributing to the magnetic form factor $F_2(0)$. Because of

$$\Lambda^\mu(p,p') = \left[F_1(q^2) + F_2(q^2)\right]\gamma^\mu - F_2(q^2)\frac{(p'+p)^\mu}{2m} \tag{11.23}$$

we can simplify the calculation of $\mathcal{N}^\mu(k)$, throwing away all terms proportional to γ^μ which do not contribute to the magnetic moment. This justifies also why we can neglect diagrams with self-energy corrections in the external lines. Moreover, we can consider the limit that the electrons are on-shell and the momentum transfer to the photon vanishes.

Using the on-shell condition, $p^2 = p'^2 = m^2$, the two square brackets in D simplify to $2p'\cdot k$ and $2p\cdot k$, respectively,

$$D = \left\{k^2 + 2k\cdot(\alpha p' + \beta p)\right\}^3\,. \tag{11.24}$$

Next we eliminate the term linear in k completing the square,

$$D = \left\{\underbrace{(k + \alpha p' + \beta p)}_{\ell}{}^2 - (\alpha p' + \beta p)^2\right\}^3 = \left\{[\ell^2 - (\alpha^2 m^2 + \beta^2 m^2 + 2\alpha\beta p'\cdot p)\right\}^3\,. \tag{11.25}$$

Since the momentum transfer to the photon vanishes, $q^2 = 2m^2 - 2p' \cdot p \to 0$, we can replace $p' \cdot p \to m^2$ and obtain as final result for the denominator

$$D = \left\{\ell^2 - (\alpha+\beta)^2 m^2\right\}^3 . \tag{11.26}$$

Now we move on to the evaluation of the numerator $\mathcal{N}^\mu(k)$. Performing the change of our integration variable from $k = \ell - (\alpha p' + \beta p)$ to ℓ, the numerator becomes

$$\mathcal{N}^\mu(\ell) = \gamma^\nu(\not{P}' + \not{\ell} + m)\gamma^\mu(\not{P} + \not{\ell} + m)\gamma_\nu \tag{11.27}$$

with $\not{P}' \equiv (1-\alpha)\not{p}' - \beta\not{p}$ and $\not{P} \equiv (1-\beta)\not{p} - \alpha\not{p}'$. Multiplying out the two brackets and ordering the result according to powers of m, we observe first that the term $\propto m^2$ leads to $\propto \gamma^\mu$ and thus does not contribute to $F_2(0)$. Next we split further the term linear in m according to powers of ℓ: The term linear in ℓ vanishes after integration, while the term $m\not{\ell}^0$ results in

$$m(\gamma^\nu\not{P}'\gamma^\mu\gamma_\nu + \gamma^\nu\gamma^\mu\not{P}\gamma_\nu) = 4m(P'^\mu + P^\mu) = 4m[(1-2\alpha)p'^\mu + (1-2\beta)p^\mu] . \tag{11.28}$$

Using the symmetry in the integration variables α and β, we can rewrite this expression as

$$\to 4m[(1-\alpha-\beta)(p'^\mu + p^\mu)] . \tag{11.29}$$

We split the m^0 term in the same way according to the powers of $\not{\ell}$. The $m^0\not{\ell}^2$ term gives a γ^μ term, the $m^0\not{\ell}$ vanishes after integration, and the $m^0\not{\ell}^0$ gives after some work

$$\gamma^\nu\not{P}'\gamma^\mu\not{P}\gamma_\nu \to 2m[\alpha(1-\alpha) + \beta(1-\beta)](p'+p)^\mu . \tag{11.30}$$

Finally, the m^0 term contributes to the anomalous magnetic moment

$$\to -2m(p'+p)^\mu[2(1-\alpha)(1-\beta)] . \tag{11.31}$$

Combining all terms, we find

$$\begin{aligned}\mathcal{N}^\mu &= 4m(1-\alpha-\beta)(p'+p)^\mu + 2m[\alpha(1-\alpha)+\beta(1-\beta)](p'+p)^\mu \\ &\quad - 4m(1-\alpha)(1-\beta)(p'+p)^\mu = \\ &= 2m[(1-\alpha-\beta)(\alpha+\beta)](p'+p)^\mu . \end{aligned} \tag{11.32}$$

Thus

$$\Gamma_2^\mu(0) = -2\mathrm{i}e^2 \int \mathrm{d}\alpha \mathrm{d}\beta \int \frac{\mathrm{d}^{2\omega}\ell}{(2\pi)^{2\omega}} \frac{\mathcal{N}^\mu}{[\ell^2 - (\alpha+\beta)^2 m^2]^3} , \tag{11.33}$$

where the subscript 2 indicates that we account only for the contribution to the anomalous magnetic moment. We expressed also the loop integral in 2ω dimensions, such that we can apply the general formula derived in the appendix 4.A. Using Eq. (4.106) for $I(\omega, a)$ with $\omega = 2$ and $a = 3$,

$$I(2,3) = -\frac{\mathrm{i}}{32\pi^2} \frac{1}{(\alpha+\beta)^2 m^2 + \mathrm{i}\varepsilon} , \tag{11.34}$$

we obtain as expected a finite result. As last step, we perform the integrals over the Feynman parameters α and β,

$$\int_0^1 \mathrm{d}\alpha \int_0^{1-\alpha} \mathrm{d}\beta \, \frac{1-\alpha-\beta}{\alpha+\beta} = \frac{1}{2} , \tag{11.35}$$

and find thus

$$\Gamma_2^\mu(0) = -\frac{e^2}{8\pi^2}\,\frac{1}{2m}\,(p'+p)^\mu\,. \tag{11.36}$$

Recalling Eq. (11.23), we can identify the factor $e^2/(8\pi^2)$ with the magnetic form factor $F_2(0)$. We have thus reproduced the result of the first successful calculation of a loop correction in a QFT, performed by Schwinger, and independently by Feynman and Tomonaga, in 1948, $F_2(0) = \alpha/(2\pi)$. Together with Bethe's previous estimate of the Lamb shift in the hydrogen energy spectrum, this stimulated the view that a consistent renormalisation of QED is possible.

The currently most precise experimental value for the electron anomalous magnetic moment $a_e \equiv F_2(0)$ is

$$a_e^{exp} = 0.001\,159\,652\,180\,73 \pm 2.4\times 10^{-10}\,. \tag{11.37}$$

The calculation of the universal (i.e. common to all charged leptons) QED contribution has been completed up to fourth order. There exists also an estimate of the dominant fifth order contribution,

$$\begin{aligned} a_\ell^{uni} &= 0.5\,\left(\frac{\alpha}{\pi}\right) - 0.328\,478\,965\,579\,193\,78\ldots\,\left(\frac{\alpha}{\pi}\right)^2 \\ &+ 1.181\,241\,456\,587\ldots\,\left(\frac{\alpha}{\pi}\right)^3 - 1.9144(35)\,\left(\frac{\alpha}{\pi}\right)^4 + 0.0(4.6)\,\left(\frac{\alpha}{\pi}\right)^5 \\ &= 0.001\,159\,652\,176\,30(43)(10)(31)\cdots \end{aligned} \tag{11.38}$$

The three errors given in round brackets are the error from the uncertainty in α, the numerical uncertainty of the α^4 coefficient and the error estimated for the missing higher order terms [Jeg07]. Comparing the measured value and the prediction using QED, we find an extremely good agreement. First of all, this is strong support that the methods of perturbative QFT we developed so far can be successfully applied to weakly coupled theories as QED. Second, it means that additional contributions to the anomalous magnetic moment of the electron have to be tiny.

Electroweak and other corrections The lowest order electroweak corrections to the anomalous magnetic moment contain in the loop virtual gauge bosons ($W^\pm$, Z) or a Higgs boson h and are shown in Fig. 11.2. We will consider the electroweak theory describing these diagrams only later; for the present discussion it is sufficient to know that the weak coupling constant is $g \simeq 0.6$ and that the scalar and weak gauge bosons are much heavier than leptons, $M \gg m$.

The first diagram corresponds schematically to the expression

$$\sim g^2 \int \frac{\mathrm{d}^4k}{(2\pi)^4}\,\frac{1}{k^2-M^2}\,\frac{A(m^2,k)}{[(p-k)^2-m^2]^2}\,. \tag{11.39}$$

As in QED, this integral has to be finite and we expect that it is dominated by momenta up to the mass M of the gauge bosons, $k \lesssim M$. Therefore its value should be proportional to g^2m^2/M^2 (times a possible logarithm $\ln(M^2/m^2)$) and electroweak corrections to the anomalous magnetic moment of the electron are suppressed by a factor $(m/M)^2 \sim 10^{-10}$ compared to the QED contribution. The property that the contribution of virtual heavy particles to loop processes is suppressed in the limit $|q^2| \ll M^2$ is called "decoupling". Note the difference to the case of the mass of a scalar particle or the cosmological constant: In these examples, the loop corrections are infinite and we cannot predict these quantities. In

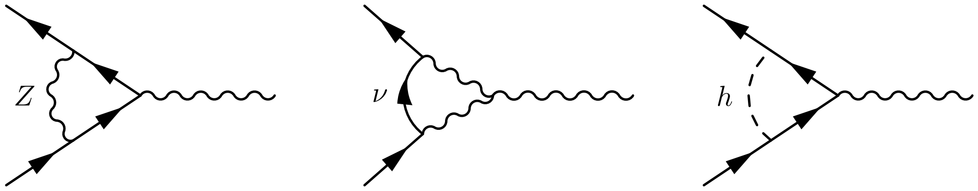


Figure 11.2.: Lowest order electroweak corrections to the anomalous magnetic moment of fermions.

contrast, the anomalous magnetic moment is finite but, as we include loop momenta up to infinity, depends in principal on all particles coupling to the electron, even if they are arbitrarily heavy. Only if these heavy particles "decouple", we can calculate a_e without knowing e.g. the physics at the Planck scale. Thus the decoupling property is a necessary ingredient of any reasonable theory of physics, otherwise no predictions would be possible before knowing the "theory of everything".

Clearly, the contribution of heavy particles (either electroweak gauge and Higgs bosons or other not yet discovered particles) is more visible in the anomalous magnetic moment of the muon than of the electron. Moreover, a relativistic muon lives long enough that a measurement of its magnetic moment is feasible. This is one example how radiative corrections (here evaluated at $q^2 = 0$) are sensitive to physics at higher scales M: If an observable can be measured and calculated with high enough precision, one can be sensitive to suppressed corrections of order g^2m^2/M^2. Other examples are rare processes like $\mu \to e + \gamma$ or $B_s \to \mu^+\mu^-$: These processes are suppressed by a specific property of the SM which one does not expect to hold in general. The achieved precision in measuring and calculating such processes is high enough to probe generically scales of $M \sim 100\,\mathrm{TeV}$, i.e. much higher than the mass scales that can be probed directly at current accelerators as LHC.

Finite versus divergent parts of loop corrections We found that the vertex correction could be split into two parts

$$\Lambda^\mu(p,p') = F_1(q^2)\gamma^\mu + F_2(q^2)\frac{\mathrm{i}\sigma^{\mu\nu}q_\nu}{2m}\,, \tag{11.40}$$

where the form factor $F_2(q^2)$ is finite for all q^2, while the form factor $F_1(q^2)$ diverges for $q^2 \to 0$. The important observation is that $F_2(q^2)$ corresponds to a Lorentz structure that is not present in the original Lagrangian of QED. This suggests that we can require from a "nice" theory that

- all UV divergences are connected to structures contained in the original Lagrangian, all new structures are finite. The basic divergent structures are also called "primitive divergent graphs".

- If there are no anomalies, then loop corrections respect the original (classical) symmetries. Thus, e.g., the photon propagator should be at all orders transverse, respecting gauge invariance. We will see that as consequence the high-energy behaviour of the theory improves.

In such a case, we are able to hide all UV divergences in a renormalisation of the original parameters of the Lagrange density.

11.3. Power counting and renormalisability

We try to make the requirements on a "nice" theory a bit more precise. Let us consider the set of $\lambda\phi^n$ theories in $d = 4$ space-time dimensions and check which graphs are divergent. We define the superficial degree D of divergence of a Feynman graph as the difference between the number of loop momenta in the numerator and denominator of a Feynman graph. We can restrict our analysis to those diagrams called 1P irreducible (1PI) which cannot be disconnected by cutting an internal line: All 1P reducible diagrams can be decomposed into 1PI diagrams which do not contain common loop integrals and can be therefore analysed separately. Moreover, we are only interested in the loop integration and define therefore the *1PI Green functions*[1] as graphs where the propagators on the external lines were stripped off. In $d = 4$ space-time dimensions, the superficial degree D of divergence of a 1PI Feynman graph is thus

$$D = 4L - 2I\,, \tag{11.41}$$

where L is the number of independent loop momenta and I the number of internal lines. The former contributes a factor d^4p, while the latter corresponds to a scalar propagator with $1/(p^2 - m^2) \sim 1/p^2$ for $p \to \infty$.

Momentum conservation at each vertex leads for an 1PI-diagram to

$$L = I - (V - 1)\,, \tag{11.42}$$

where V is the number of vertices and the -1 takes into account the delta function leading to overall momentum conservation: The latter constrains only the external not the loop momenta. Thus

$$D = 2I - 4V + 4\,. \tag{11.43}$$

Each vertex connects n lines and any internal line reduces the number of external lines by two. Therefore the number E of external lines is given by

$$E = nV - 2I\,. \tag{11.44}$$

As result, we can express the superficial degree D by the order of perturbation theory (V), the number of external lines E and the degree n of the interaction polynomial ϕ^n,

$$D = (n - 4)V + 4 - E\,. \tag{11.45}$$

From this expression, we see that

- for $n < 4$, the coefficient of V is negative. Therefore only a finite number of terms in the perturbative expansion are infinite. Such a theory is called super-renormalisable, the corresponding terms in the Hamiltonian are also called relevant.

[1] Similar to their relatives, the (dis-) connected Green functions, also the 1PI Green functions can be derived from a generating functional which we will introduce in the next chapter.

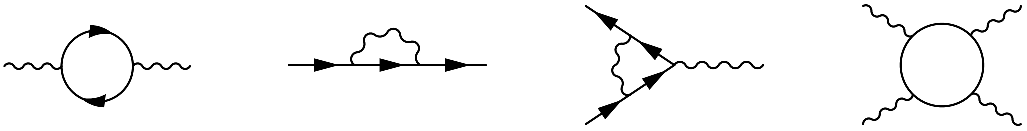


Figure 11.3.: Primitive divergent diagrams in QED (without vacuum diagrams).

- For $n = 4$, we find $D = 4 - E$. Thus the degree of divergence is independent of the order of perturbation theory being only determined by the number of external lines. Such theories contain an infinite number of divergent graphs, but they all correspond to a finite number of divergent structures—the so-called primitive divergent graphs. These interactions are also called marginal and are candidates for a renormalisable theory.
- Finally, for $n > 4$ the degree of divergence increases with the order of perturbation theory. As result, there exists an infinite number of divergent structures, and increasing the order of perturbation theory requires more and more input parameter to be determined experimentally. Such a theory is called non-renormalisable, the interaction irrelevant.

In particular, the $\lambda\phi^4$ theory as an example for a renormalisable theory has only three divergent structures: i) the case $E = 0$ and $D = 4$ contributes to the vacuum energy, ii) the case $E = 2$ and $D = 2$ corresponding to the self-energy, and iii) the case $E = 4$ and $D = 0$, i.e. logarithmically divergence, to the four-point function. As we saw in chapter 3, the three primitive divergent diagrams of the $\lambda\phi^4$ theory correspond to the following physical effects: Vacuum bubbles renormalise the cosmological constant. The effect of self-energy insertions is twofold: Inserted in external lines it renormalises the field, while self-energy corrections in internal propagators lead to a renormalisation of its mass. The vertex correction finally renormalises the coupling strength λ.

Let us move to the case of QED. Repeating the discussion, we obtain the analogue to Eq. (11.45), but accounting now for the different dimension of fermionic and bosonic fields,

$$D = 4 - B - \frac{3}{2}F\,, \tag{11.46}$$

where B and F count the number of external bosonic and fermionic lines, respectively. There are six different superficially divergent primitive graphs in QED shown in Fig. 11.3: The photon and the fermionic contribution to the cosmological constant ($D = 4$), the vacuum polarisation ($D = 2$), the fermion self-energy ($D = 1$), the vertex correction ($D = 0$) and light-by-light scattering ($D = 0$). Recall that Furry's theorem implies that loops with a an odd number of fermion propagator vanish in QED. Therefore we have not included in our list of primitive divergent graphs of QED the tadpole ($B = 1$ and $D = 3$) and the "photon splitting" graph ($B = 3$ and $D = 1$).

In a theory with symmetries such as a gauge theory, the true degree of divergence can be smaller than the superficial one. For instance, light-by-light scattering corresponds to a term $\mathscr{L} \sim A^4$ that violates gauge invariance. Thus either the gauge symmetry is violated by quantum corrections or such a term is finite.

Because of the correspondence of the dimension of a field and the power of its propagator, we can connect the superficial degree of divergence of a graph to the dimension of the coupling constants at its vertices. The superficial degree $D(G)$ of divergence of a graph is connected to the one of its vertices D_v by

$$D(G) - 4 = \sum_v (D_v - 4) \tag{11.47}$$

which in turn depends as

$$D_v = \delta_v + \frac{3}{2} f_v + b_v = 4 - [g_v] \tag{11.48}$$

on the dimension of the coupling constant g at the vertex v. Here, f_v and b_v are the number of fermion and boson fields at the vertex, while δ_v counts the number of derivatives. Thus the dimension of the coupling constant plays a crucial role deciding if a certain theory is "nice" in the naive sense defined above. Clearly $D = 0$ or $[g] = m^0$ is the border-line case:

- If at least one coupling constant has a negative mass dimension, $[g] < 0$ and $D_v > 4$, the theory is non-renormalisable. Examples are the Fermi theory of weak interactions, $[G_F] = m^{-2}$, and gravitation, $[G_N] = m^{-2}$.
- If all coupling constants have positive mass dimension, $[g] > 0$ and $D_v < 4$, the theory is super-renormalisable. An example is the $\lambda\phi^3$ theory in $D = 4$ with $[\lambda] = m^0$.
- The remaining cases, with all $[g_i] = 0$, are candidates for renormalisable theories. Examples are Yukawa interactions, $\lambda\phi^4$, Yang-Mills theories that are unbroken (QED and QCD) or broken by the Higgs mechanism (electroweak interactions).

Theories with massive bosons We have assumed that bosonic propagators behave as $\propto 1/k^2$ for large (Euclidean) momenta k. This is true both for massive and massless scalars, while it holds only for massless particles with spin $s \geq 1$: As we have seen, the massless spin-1 and spin-2 propagators in the R_ξ gauge decrease like $\propto 1/k^2$ for large k. In contrast, the massive spin-1 propagator behaves as $D_F^{\mu\nu}(k) \propto const$. Thus the divergences in loop diagrams are more severe for massive vector particles than for massless ones. For a massive bosonic field of spin s, the polarisation tensors contains s tensor products of $k_\mu k_\nu$ and therefore its propagator scales as $D_{\mu_1,\cdots,\mu_s;\nu_1,\cdots,\nu_s}(k) \propto k^{2s-2}$. This implies that the divergences of loop diagrams aggravate for higher spin fields. In particular, inserting additional massive propagators into a loop graph does not improve its convergence and thus a theory with massive $s > 0$ particles contains an infinite number of divergent diagrams at each loop order. Including an explicit mass term for gauge bosons leads therefore to a non-renormalisable theory. A solution to this problem is the introduction of gauge boson masses via the Higgs mechanism, which we will introduce in chapter 13.3. Combined with our finding that interacting theories of massless bosons are only possible for $s \leq 2$, we can conclude that elementary particles should have spin $s \leq 2$.

11.4. Renormalisation of the $\lambda\phi^4$ theory

We have argued that a theory with dimensionless coupling constant is renormalisable. In this case a multiplicative shift of the parameters contained in the classical Lagrangian is sufficient to obtain finite Green functions. The simplest theory of this type in $d = 4$ is the $\lambda\phi^4$ theory for which we will discuss now the renormalisation procedure at one loop level. As starter, we examine the general structure of the UV divergences.

11.4.1. Structure of the divergences

We learnt that the degree of divergence decreases increasing the number of external lines, since the number of propagators increases. The same effect has taking derivatives w.r.t. external momenta p,

$$\frac{\partial}{\partial \not p}\int \frac{d^4k}{(2\pi)^4}\frac{1}{\not k+\not p-m} = -\int \frac{d^4k}{(2\pi)^4}\frac{1}{(\not k+\not p-m)^2}\,.$$

This means that

1. we can Taylor expand loop integrals, confining the divergences in the lowest order terms. Choosing e.g. $p=0$ as expansion point in the fermion self-energy,

   $$\Sigma(p) = A_0 + A_1\not p + A_2 p^2 + \ldots \qquad \text{with} \qquad A_n = \frac{1}{n!}\frac{\partial^n}{\partial \not p^n}\Sigma(p)\,,$$

   we know that A_0 is (superficially) linear divergent. Thus A_1 can be maximally logarithmically divergent, while all other coefficients A_n are finite.

2. We could choose a different expansion point, leading to different renormalisation conditions (within the same regularisation scheme).

3. The divergences can be subtracted by local operators, i.e. by polynomials of the fields and their derivatives. These terms called counter-terms have for a renormalisable theory the same structure as the terms present in the classical Lagrangian. For instance, the linear divergent term A_0 can be associated to a counter-term $\delta A_0\bar\psi\psi$, while $A_1\not p$ corresponds to the counter-term $\delta A_1\bar\psi\not\partial\psi$.

It is easy to show that the counter-terms are local operators at the one-loop level, where diagrams contain only one integration variable. Any loop integral $I(p)$ with superficial degree of divergence $n-1$ becomes finite after taking n derivatives w.r.t. an external momentum p. Using a cutoff Λ as regulator, this implies that in

$$\frac{\partial^n}{\partial p^n}I(p) = f(p) + \mathcal{O}(p/\Lambda) \tag{11.49}$$

the function $f(p)$ is finite and independent of Λ, while the remainder vanishes in the limit $\Lambda\to\infty$. Integrating this expression n times, we obtain

$$I(p) = F(p) + P_n(p) + \mathcal{O}(p/\Lambda)\,, \tag{11.50}$$

where $F(p)$ is also finite and independent of Λ. The function $P_n(p)$ is a n.th order polynomial containing the integration constants. Since $F(p)$ is finite, $P_n(p)$ comprises all divergences. They are therefore the coefficients of polynomials in the external momentum p and can be subtracted by local operators, as we claimed. This argument shows also that all non-trivial analytical structures like cuts have to be contained in $F(p)$. Moreover, choosing a different regularisation scheme or point leads to the same $f(p)$ in (11.49), and thus all the scheme dependence is contained in the polynomial $P_n(p)$. As a result, the differences caused by different schemes reside only in local terms which are absorbed in the renormalisation of the parameters.

Going to higher loop orders, non-local terms as e.g. $\ln(p^2/\mu^2)$ can be generated by subdivergences. These are divergences connected to integration regions where one or more loop momenta are finite, while the remaining ones are send to infinity. Such terms are cancelled by counter-terms determined at lower order. A sketch why this should be true goes as follows:

Green functions become singular for coinciding points, i.e. when the convergence factor e^{-kx} in the Euclidean Green function becomes one. In the simplest cases as $\langle 0|\phi(x')\phi(x)|0\rangle_{x'\to x}$, the infinities are eliminated by normal ordering, i.e. by rewriting all creation operators on the left of the annihilation operators, cf. problem 3.??. More complicated are overlapping divergences where two or more divergent loops share a propagator. Wilson suggested to expand the product of two fields as the sum of local operators O_i times coefficient functions $C_i(x-y)$ as

$$\phi(x)\phi(y)=\sum_i C_i(x-y)O_i(x)\,,$$

where the dependence on the relative distance is carried by the coefficients and the local operators O_i are of the type $O_i(x)=\phi(x)\partial_\mu\cdots\partial^\mu\phi(x)$. For a massless scalar field, dimensional analysis dictates that $C_i(x)\propto x^{-2+d_i}$, if the local operator O_i has dimension d_i. Note that only the unity operator has a singular coefficient function $1/x^2$ corresponding to the massless scalar propagator. Similarly we can expand products of operators,

$$O_n(x)O_m(y)=\sum_i C^i_{nm}(x-y)O_i(x)\,,$$

where now $C^i_{nm}(x)\propto x^{-d_n-d_m+d_i}$. Thus we can use this operator product expansion (or briefly "OPE") to rewrite the overlapping divergences in terms of (singular) coefficient functions and *local* operators. Moreover, the sub-divergence occurring at order k, when $p<k$ points coincide, are eliminated by the counter-terms found at order p.

Elaborating this argument in detail, one can conclude that non-local terms due to overlapping divergences are cancelled by the counter-terms found at lower order. We will see how this works in practice in the next section, when we calculate the vacuum energy at two-loop.

11.4.2. The $\lambda\phi^4$ theory at $\mathcal{O}(\lambda)$

There are two equivalent ways to perform perturbative renormalisation. In the one called often "conventional" perturbation theory we use the "bare" (unrenormalised) parameters in the Lagrangian,

$$\mathscr{L}=\mathscr{L}_0+\mathscr{L}_{int}=\frac{1}{2}(\partial_\mu\phi_0)^2-\frac{1}{2}m_0^2\phi_0^2-\frac{\lambda_0}{4!}\,\phi_0^4\,. \tag{11.51}$$

Then we introduce a renormalised field $\phi_R=Z_\phi^{-1/2}\phi_0$ and choose the parameters Z_ϕ, m_0 and λ_0 as function of the regularisation parameter (ε, $\Lambda,\ldots$) such that the field ϕ_R has finite Green functions. In the following, we discuss the renormalisation procedure at the one-loop level for the Green functions of the $\lambda\phi^4$ theory in this scheme. Since any 1P reducible diagram can be decomposed into 1PI diagrams which do not contain common loop integrals, we can restrict our analysis again to 1PI Green functions.

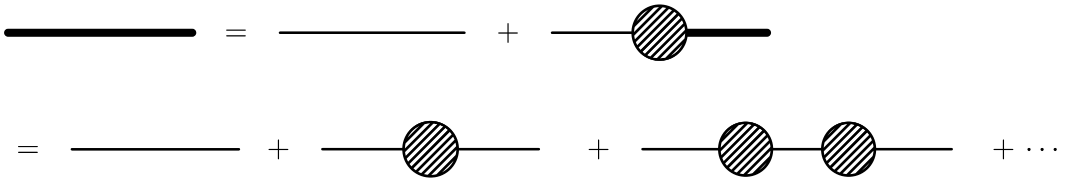
Mass and wave-function renormalisation We defined the exact or full propagator $\mathrm{i}\boldsymbol{\Delta}_F(x_1,x_2)$ in Eq. (4.2) as the path integral average of the two fields $\phi(x_1)\phi(x_2)$. Now we want to find a definition which is useful for calculations in perturbation theory: We claim that

$$[\mathrm{i}\boldsymbol{\Delta}_F(p)]^{-1}=[\mathrm{i}\Delta_F(p)]^{-1}-\boldsymbol{\Sigma}(p)=p^2-m^2-\boldsymbol{\Sigma}(p)-\mathrm{i}\varepsilon \tag{11.52}$$

is the exact propagator, where the exact self-energy $\mathbf{\Sigma}(p)$ represents the 1PI corrections to the scalar mass, $m^2_{\mathrm{p}hys} = m_0^2 + \mathbf{\Sigma}$. Multiplying this definition from the right with $\mathrm{i}\mathbf{\Delta}_F(p)$ and from the left with $\mathrm{i}\Delta_F(p)$, the so-called Dyson equation follows,

$$\mathrm{i}\mathbf{\Delta}_F(p) = \mathrm{i}\Delta_F(p) + \mathrm{i}\Delta_F(p)\mathbf{\Sigma}(p)\mathrm{i}\mathbf{\Delta}_F(p) = \mathrm{i}\Delta_F(p)\left[1 + \mathbf{\Sigma}(p)\mathrm{i}\mathbf{\Delta}_F(p)\right] . \tag{11.53}$$

Graphically, we can express this equation as



where the second line follows by iteration. Hence, $\mathrm{i}\mathbf{\Delta}_F(p)$ sums up the amplitudes to propagate at momentum p with zero, one,... self-energy Σ insertions, and corresponds therefore to the full propagator. At $\mathcal{O}(\lambda)$, we see that this relation holds comparing it to Eq. (4.33).

Next we have to show that the self-energy $\Sigma(p^2)$ is finite after renormalisation. The one-loop expression

$$-\mathrm{i}\Sigma(p^2) = \frac{-\mathrm{i}\lambda_0}{2}\int\frac{\mathrm{d}^4k}{(2\pi)^4}\,\frac{\mathrm{i}}{k^2 - m_0^2 + \mathrm{i}\varepsilon} \tag{11.54}$$

is quadratically divergent. As a particularity of the ϕ^4 theory, the p^2 dependence of the self-energy Σ shows up only at the two-loop level. We perform a Taylor expansion of $\Sigma(p^2)$ around the arbitrary point μ,

$$\Sigma(p^2) = \Sigma(\mu^2) + (p^2 - \mu^2)\Sigma'(\mu^2) + \tilde{\Sigma}(p^2)\,, \tag{11.55}$$

where $\Sigma(\mu^2) \propto \Lambda^2$, $\Sigma'(\mu^2) \propto \ln\Lambda$ and $\tilde{\Sigma}(p^2)$ is the finite remainder. A term linear in Λ is absent, since we cannot construct a Lorentz scalar out of p^μ. Note also $\tilde{\Sigma}(\mu^2) = 0$.

Now we insert (11.55) into (11.52),

$$\frac{\mathrm{i}}{p^2 - m_0^2 - \Sigma(p^2) + \mathrm{i}\varepsilon} = \frac{\mathrm{i}}{\underbrace{p^2 - m_0^2 - \Sigma(\mu^2)}_{p^2-\mu^2} - (p^2 - \mu^2)\Sigma'(\mu^2) - \tilde{\Sigma}(p^2) + \mathrm{i}\varepsilon}\,, \tag{11.56}$$

where we see that we can identify μ with the renormalised mass given by the pole of the propagator.

We aim at rewriting the remaining effect for $p^2 \to \mu^2 = m^2$ of the self-energy insertion, $\Sigma'(m^2)$, as a multiplicative rescaling. In this way, we could remove the divergence from the propagator by a rescaling of the field. At leading order in λ, we can write

$$\tilde{\Sigma}(p^2) = \left[1 - \Sigma'(m^2)\right]\tilde{\Sigma}(p^2) + \mathcal{O}(\lambda_0^2) \tag{11.57}$$

and thus

$$\mathrm{i}\mathbf{\Delta}_F(p) = \frac{1}{1 - \Sigma'(m^2)}\,\frac{\mathrm{i}}{p^2 - m^2 - \tilde{\Sigma}(p^2) + \mathrm{i}\varepsilon} = \frac{\mathrm{i}Z_\phi}{p^2 - m^2 - \tilde{\Sigma}(p^2) + \mathrm{i}\varepsilon} \tag{11.58}$$

with the wave-function renormalisation constant

$$Z_\phi = \frac{1}{1 - \Sigma'(m^2)} = 1 + \Sigma'(m^2)\,. \tag{11.59}$$

Close to the pole, the propagator is the one of a free particle with mass m,

$$\mathrm{i}\boldsymbol{\Delta}_F(p) = \frac{\mathrm{i}Z_\phi}{p^2 - m^2 + \mathrm{i}\varepsilon} + \mathcal{O}(p^2 - m^2)\,. \tag{11.60}$$

Thus the renormalisation constant Z_ϕ equals the wave-function renormalisation constant Z we had to introduce into the LSZ formalism to obtain correctly normalised states.

We define the renormalised field $\phi = Z_\phi^{-1/2}\phi_0$ such that the renormalised propagator

$$\mathrm{i}\Delta_R(p) = \int \mathrm{d}^4x\, \mathrm{e}^{\mathrm{i}px}\, \langle 0|\, T\{\phi(x)\phi(0)\}\, |0\rangle = Z_\phi^{-1}\, \mathrm{i}\Delta(p) = \frac{\mathrm{i}}{p^2 - \mu^2 - \tilde{\Sigma}(p^2) + \mathrm{i}\varepsilon} \tag{11.61}$$

is finite. Similarly, we define renormalised n-point functions by

$$G_R^{(n)}(x_1, \dots, x_n) = \langle 0|\, T\{\phi(x_1)\cdots\phi_n(x_n)\}\, |0\rangle = Z_\phi^{-n/2}\, G_0^{(n)}(x_1, \dots, x_n)\,. \tag{11.62}$$

Since the 1PI n-point Green functions miss n field renormalisation constants compared to connected n-point Green functions, the connection between renormalised and bare 1PI n-point functions is given by

$$\Gamma_R^{(n)}(x_1, \dots, x_n) = Z_\phi^{n/2}\, \Gamma_0^{(n)}(x_1, \dots, x_n)\,. \tag{11.63}$$

Coupling constant renormalisation We can choose an arbitrary point inside the kinematical region, $s + t + u = 4\mu^2$ and $s \geq 4\mu^2$, to connect the coupling to a physical measurement at this point. For our convenience and less writing work, we choose instead the symmetric point

$$s_0 = t_0 = u_0 = \frac{4\mu^2}{3}\,.$$

The bare four-point 1PI Green function is (see section 4.3.3)

$$\Gamma_0^{(4)}(s, t, u) = -\mathrm{i}\lambda_0 + \Gamma(s) + \Gamma(t) + \Gamma(u)\,, \tag{11.64}$$

the renormalised four-point function at (s_0, t_0, u_0) is

$$\Gamma_R^{(4)}(s_0, t_0, u_0) = -\mathrm{i}\lambda\,. \tag{11.65}$$

Next we expand the bare 4-point function around s_0, t_0, u_0,

$$\Gamma_0^{(4)}(s, t, u) = -\mathrm{i}\lambda_0 + 3\Gamma(s_0) + \tilde{\Gamma}(s) + \tilde{\Gamma}(t) + \tilde{\Gamma}(u) \tag{11.66}$$

where the $\tilde{\Gamma}(x)$ are finite and zero at x_0. Now we define a vertex (or coupling constant) renormalisation constant by

$$-\,\mathrm{i}Z_\lambda^{-1}\lambda_0 = -\mathrm{i}\lambda_0 + 3\Gamma(s_0) \tag{11.67}$$

Inserting this definition in (11.66) we obtain

$$\Gamma_0^{(4)}(s, t, u) = -\mathrm{i}Z_\lambda^{-1}\lambda_0 + \tilde{\Gamma}(s) + \tilde{\Gamma}(t) + \tilde{\Gamma}(u) \tag{11.68}$$

what simplifies at the renormalisation point to

$$\Gamma_0^{(4)}(s_0, t_0, u_0) = -\mathrm{i}Z_\lambda^{-1}\lambda_0\,. \tag{11.69}$$

We use now the connection between renormalised and bare Green functions,

$$\Gamma_R^{(4)}(s,t,u) = Z_\phi^2 \Gamma_0^{(4)}(s,t,u)\,, \tag{11.70}$$

and thus

$$-\,\mathrm{i}\lambda = Z_\phi^2 Z_\lambda^{-1}(-\mathrm{i}\lambda_0)\,. \tag{11.71}$$

The relation between the renormalised and bare coupling in the $\lambda\phi^4$ theory is thus

$$\lambda = Z_\phi^2 Z_\lambda^{-1}\lambda_0\,. \tag{11.72}$$

Now we have to show that $\Gamma_R^{(4)}(s,t,u)$ is finite. Inserting (11.68) into (11.70), we find

$$\begin{aligned}\Gamma_R^{(4)}(s,t,u) &= -\mathrm{i}Z_\phi^2 Z_\lambda^{-1}\lambda_0 + Z_\phi^2[\tilde\Gamma(s)+\tilde\Gamma(t)+\tilde\Gamma(u)]\\ &= -\mathrm{i}\lambda + Z_\phi^2[\tilde\Gamma(s)+\tilde\Gamma(t)+\tilde\Gamma(u)]\end{aligned} \tag{11.73}$$

Since $Z_\phi = 1+\mathcal{O}(\lambda^2)$ and $\tilde\Gamma = \mathcal{O}(\lambda^2)$, this is equivalent to

$$\Gamma_R^{(4)}(s,t,u) = -\mathrm{i}\lambda + [\tilde\Gamma(s)+\tilde\Gamma(t)+\tilde\Gamma(u)] + \mathcal{O}(\lambda^3) \tag{11.74}$$

consisting only of finite expressions. This completes the proof that at one-loop order all Green functions in the $\lambda\phi^4$ theory are finite, renormalising the field ϕ, its mass and coupling constant. as

Renormalised perturbation theory In this approach, we rescale first the bare field in the classical Lagrangian by $\phi_0 = Z_\phi^{1/2}\phi$, obtaining

$$\mathscr{L} = \frac{1}{2}Z_\phi(\partial_\mu\phi)^2 - \frac{1}{2}Z_\phi m_0^2\phi^2 - \frac{\lambda_0}{4!}\,Z_\phi^2\phi^4\,. \tag{11.75}$$

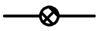
Next we introduce the renormalised mass and coupling by $m_0^2 = Z_\phi^{-1}Z_m m^2$ and $\lambda_0 = Z_\phi^{-2}Z_\lambda\lambda$, obtaining

$$\mathscr{L} = \frac{1}{2}Z_\phi(\partial_\mu\phi)^2 - \frac{1}{2}Z_m m^2\phi^2 - \frac{\lambda}{4!}\,Z_\lambda\phi^4\,. \tag{11.76}$$

The renormalisation constants Z_i vanish at tree-level and allow for a perturbative expansion. Setting $Z_i = 1+\delta_i$, we can split the Lagrangian into

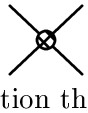
$$\begin{aligned}\mathscr{L} &= \frac{1}{2}(\partial_\mu\phi)^2 - \frac{1}{2}m^2\phi^2 - \frac{\lambda}{4!}\,\phi^4\\ &+ \frac{1}{2}\delta_\phi(\partial_\mu\phi)^2 - \frac{1}{2}\delta_m m^2\phi^2 - \frac{\lambda}{4!}\,\delta_\lambda\phi^4\,,\end{aligned} \tag{11.77}$$

where the first line contains only renormalised quantities. The terms in the second line contain the divergent renormalisation constants, and this part is called the counter-term Lagrangian $\mathscr{L}_{\mathrm{ct}}$. An advantage of renormalised perturbation theory is that now the expansion parameter is the renormalised coupling λ. Treating $\mathscr{L}_{ct}$ as a perturbation, $Z_i = 1+\sum_{n=1}^\infty \delta_i^{(n)}$, we obtain in momentum space as additional Feynman vertices



$$= \mathrm{i}[\delta_\phi p^2 - \delta_m m^2] \tag{11.78}$$

and



$$= -\mathrm{i}\delta_\lambda\lambda\,. \tag{11.79}$$

Applying renormalised perturbation theory consists of the following steps:

1. Starting from (11.76) with $n=0$, i.e. $\delta_i^{(0)}=0$, one derives propagator and vertices.
2. One calculates 1-loop 1PI diagrams and finds the divergent parts which determine the counter-terms $\delta_i^{(1)}$ at order $\mathcal{O}(\lambda)$. Then all other one-loop diagrams can be calculated.
3. Moving to two-loops, one generates two-loop 1PI diagrams using the Lagrangian with the one-loop counter-terms $\delta_i^{(1)}$. They are used to extract the counter-terms $\delta_i^{(2)}$ at order $\mathcal{O}(\lambda^2)$.
4. The procedure is iterated moving to higher orders.

We illustrate the use of renormalised perturbation theory with the calculation of the remaining loop diagram at $\mathcal{O}(\lambda)$, the vacuum energy density. This example for a two-loop diagram shows also how sub-divergences are cancelled by counter-terms found at lower loop order. Including the vacuum energy density, the Lagrangian (11.76) becomes $\mathscr{L} \to \mathscr{L} + \rho + \delta_\rho \rho$. This example shows also that the correct expansion parameter is the number of loops, not the power of the coupling constant.

Example 11.3: Vacuum energy density at two-loop:
According to step 2., we should determine first the counter-terms in $\mathscr{L}_{ct}^{(1)}$ from the already calculated one-loop 1PI diagrams. We start collecting the relevant results derived in chapter 4,

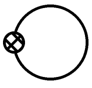
$$\rho^{(1)} = -\frac{m^4}{(4\pi)^2}\left[\frac{1}{\varepsilon} + \ln(\mu^2/m^2)\right], \qquad \delta_\rho^{(1)} = \frac{1}{(4\pi)^2}\frac{1}{\varepsilon}, \quad \textit{and} \quad \delta_m^{(1)} = \frac{\lambda}{2}\frac{1}{(4\pi)^2}\frac{1}{\varepsilon}, \tag{11.80}$$

where we use the MS scheme and re-scaled $\mu^2 \to 4\pi\mu^2\exp(-\gamma)$. Inserting the one-loop self-energy into the two-loop expression $\rho_a^{(2)} = \lambda/8\,\Delta_F^2(0)$ results in

$$\rho_a^{(2)} = \frac{\lambda}{8}\Delta_F^2(0) = \frac{\lambda}{8}\frac{m^4}{(4\pi)^4}\left[\frac{1}{\varepsilon^2} + \frac{2}{\varepsilon}\ln\left(\frac{\mu^2}{m^2}\right) + \ln^2\left(\frac{\mu^2}{m^2}\right)\right]. \tag{11.81}$$

Here a mixed term, combining a pole term $1/\varepsilon$ and a logarithm with argument μ^2/m^2, has appeared. In general, the logarithm will depend both on the masses of the loop particles and the external momenta p, $\ln[f(\mu^2/m^2, \mu^2/p^2)]$. Such terms cannot be subtracted by local polynomials in the momenta p as counter-terms. In a renormalisable theory, they have to be therefore cancelled by counter-terms determined at lower loop order.
In our concrete case, we have to add only the Feynman diagram generated by the counter-term $-\frac{1}{2}\delta_m^{(1)} m^2\phi^2$, since δ_λ contributes only from the two-loop level on. This interaction generates at $\mathcal{O}(\lambda)$ the following contribution to the vacuum energy density

$$\rho_b^{(2)} =$$



$$= \frac{1}{2}\delta_m^{(1)} m^2 \Delta_F(0) = -\frac{\lambda}{8}\frac{m^4}{(4\pi)^4}\frac{2}{\varepsilon}\left[\frac{1}{\varepsilon} + \ln(\mu^2/m^2)\right]. \tag{11.82}$$

Combining the two contributions, the mixed terms disappear as expected and the remaining $1/\varepsilon^2$ pole can be subtracted by the counter-term

$$\delta_\rho^{(2)} = \frac{\lambda}{8}\frac{1}{(4\pi)^4}\frac{1}{\varepsilon^2}. \tag{11.83}$$

Thus the two-loop contribution to the vacuum energy density is

$$\rho^{(2)} = \frac{m^4}{(4\pi)^4}\ln^2(\mu^2/m^2) = \left(\frac{m^{(1)}(\mu)}{4\pi}\right)^4. \tag{11.84}$$

Summary

Using a power counting argument for the asymptotic behaviour of the free Green functions, we singled out theories with dimensionless coupling constants: Such theories with marginal interactions are renormalisable, i.e. are theories with a finite number of primitive divergent diagrams. In this case, the multiplicative renormalisation of the finite number of parameters contained in the classical (effective) Lagrangian is sufficient to obtain finite Green functions at any order perturbation theory.

Further reading

The renormalisation of the $\lambda\phi^4$ theory at the two-loop level is performed e.g. by [Pok87]. Non-renormalisable theories are discussed by [Sch13]. [Jeg07] reviews the status of electroweak precision calculations.

12. Renormalisation II: Improving perturbation theory

We continue our discussion of renormalisation, introducing first the quantum action as the generating functional of 1PI Green functions. Then we apply the developed formalism to derive the Ward identities which imply e.g. that the exact photon propagator is transverse and that the renormalisation of the electric charge is universal. Next we introduce the renormalisation group equations which describe the evolution of n-point Green functions under a change of scale. These equations suggest to convert the parameters contained in the classical Lagrangian into "running parameter", summing up thereby the most important corrections of an infinite set of diagrams. Finally, we introduce in the last section a non-perturbative approach based on ideas developed in solid-state physics and the renormalisation group.

12.1. Quantum action

In the classical limit, the equation of motion $\delta S[\phi]/\delta\phi = -J$ allows us to determine the source $J(x)$ which produces a given field $\phi(x)$. Our aim is to find the quantum analogue of this classical equation. Let us recall first the definition for the generating functionals of a real scalar field,

$$Z[J] = \int \mathcal{D}\phi \, \mathrm{e}^{\mathrm{i}\{S[\phi]+\int \mathrm{d}^4x \, J(x)\phi(x)\}} = \mathrm{e}^{\mathrm{i}W[J]} \,. \tag{12.1}$$

Then we define the classical field $\phi_c(x)$ as $\phi_c(x) = \delta W[J]/\delta J(x)$. Performing the functional derivative in its definition, we see immediately why this definition makes sense,

$$\phi_c(x) = \frac{\delta W[J]}{\delta J(x)} = \frac{1}{\mathrm{i}Z}\frac{\delta Z[J]}{\delta J(x)} = \frac{1}{Z}\int \mathcal{D}\phi \, \phi(x) \exp \mathrm{i}\int \mathrm{d}^4y(\mathscr{L} + J\phi) \tag{12.2a}$$

$$= \frac{\langle 0|\phi(x)|0\rangle_J}{\langle 0|0\rangle_J} = \langle \phi(x)\rangle_J \,. \tag{12.2b}$$

Thus the classical field $\phi_c(x)$ is the vacuum expectation value of the quantum field $\phi(x)$ in the presence of the source $J(x)$. Now we define the quantum[1] action $\Gamma[\phi_c]$ as the Legendre transform of $W[J]$,

$$\Gamma[\phi_c] = W[J] - \int \mathrm{d}^4x \, J(x)\phi_c(x) \equiv W[J] - \langle J\phi\rangle \,, \tag{12.3}$$

where $\phi_c(x) = \delta W[J]/\delta J(x)$ should be used to replace $J(x)$ by $\phi_c(x)$ on the RHS. We compute the functional derivative w.r.t. ϕ_c of this new quantity,

$$\frac{\delta\Gamma[\phi_c]}{\delta\phi_c(y)} = \int \mathrm{d}^4x \, \frac{\delta W}{\delta J(x)}\frac{\delta J(x)}{\delta\phi_c(y)} - \int \mathrm{d}^4x \, \frac{\delta J(x)}{\delta\phi_c(y)}\phi_c(x) - J(y).$$

[1] Many authors call $\Gamma[\phi_c]$ the effective or quantum effective action.

Using the definition $\delta W/\delta J(x) = \phi_c(x)$, the first and second term cancels and we end up as desired with

$$\frac{\delta\Gamma[\phi_c]}{\delta\phi_c(y)} = -J(y)\,. \tag{12.4}$$

The analogy to the classical equation of motion suggests that $\Gamma[\phi_c]$ is the quantum version of the classical action. We will show next that its tree-level diagrams contain all loop corrections induced by the usual action $S[\phi]$. It is this feature that justifies the name quantum action for $\Gamma[\phi_c]$.

Expansion in $\hbar$ as a loop expansion In order to proceed, we perform a saddle-point expansion around the classical solution ϕ_0, given by the solution to

$$\left.\frac{\delta\{S[\phi]+\langle J\phi\rangle\}}{\delta\phi}\right|_{\phi_0} = 0 \tag{12.5}$$

or

$$\Box\phi_0 + V'(\phi_0) = J(x)\,. \tag{12.6}$$

We write the field as $\phi = \phi_0 + \tilde{\phi}$, i.e. as a classical solution with quantum fluctuations on top. Then we can approximate the path integral $Z = \exp\{\mathrm{i}W/\hbar\}$ as

$$Z \simeq \mathrm{e}^{\mathrm{i}[S[\phi_0]+\langle J\phi_0\rangle]/\hbar} \int \mathcal{D}\tilde{\phi} \exp\left\{\frac{\mathrm{i}}{\hbar}\int \mathrm{d}^4x \frac{1}{2}\left[(\partial_\mu\tilde{\phi})^2 - V''(\phi_0)\tilde{\phi}^2\right]\right\}\,. \tag{12.7}$$

We have restored Planck's constant $\hbar$ to indicate that this saddle point expansion can be viewed as an expansion in $\hbar$. Next we want to show that the expansion in $\hbar$ corresponds to a loop expansion. We introduce artificially a parameter a into our Lagrangian so that

$$\mathscr{L}(\phi, \partial_\mu\phi, a) = a^{-1}\mathscr{L}(\phi, \partial_\mu\phi)\,. \tag{12.8}$$

Let us determine the power P of a in an arbitrary 1PI Feynman graph, a^P: A propagator is the *inverse* of the quadratic form in $\mathscr{L}$ and contributes thus a positive power a, while each vertex $\propto \mathscr{L}_{int}$ adds a factor a^{-1}. The number of loops in an 1PI diagram is given by $L = I - V + 1$, cf. Eq. (11.42), where I is the number of internal lines and V is the number of vertices. Putting this together we see that

$$P = I - V = L - 1 \tag{12.9}$$

and thus[2] the power of a gives us the number of loops. We should stress that using a loop expansion does not imply a semi-classical limit, $S \gg \hbar$: Our fictitious parameter a is not small; in fact, it is one.

Quantum action as generating functional for 1PI Green functions We have now all the necessary ingredients in order to show that the tree-level graphs generated by the quantum

[2] We assume here that particle masses which carry a factor $\hbar^{-1}$ can be neglected. In few applications, as e.g. calculating quantum corrections to the Newtonian potential between two masses m_1 and m_2, this assumption is not valid.

action $\Gamma[\phi]$ correspond to the complete scattering amplitudes of the corresponding action $S[\phi]$. We compare the "true" generating functional

$$Z[J] = \int \mathcal{D}\phi \, \exp\{\mathrm{i}S + \langle J\phi\rangle\} = \mathrm{e}^{\mathrm{i}W[J]}, \tag{12.10}$$

with the functional $V_a[J]$ of a fictitious field theory whose classical action S is the quantum action $\Gamma[\phi]$ of the theory (12.10) we are interested in,

$$V_a[J] = \int \mathcal{D}\phi \, \exp\left\{\frac{\mathrm{i}}{a}\{\Gamma[\phi] + \langle J\phi\rangle\}\right\} = \mathrm{e}^{\mathrm{i}U_a[J]}. \tag{12.11}$$

Additionally, we introduced the parameter a with the same purpose as in (12.8): In the limit $a \to 0$, we can perform a saddle-point expansion and the path integral is dominated by the classical path. From (12.7), we find thus

$$\lim_{a\to 0} aU_a[J] = \Gamma[\phi] + \langle J\phi\rangle = W[J], \tag{12.12}$$

where we used the definition of the quantum action, Eq. (12.3), in the last step. The RHS is the sum of all connected Green functions of our original theory. The LHS is the classical limit of the fictitious theory $V_a[J]$, i.e. it is the sum of all connected tree graphs generated using $\Gamma[\phi]$ as action. But a connected graph which is one-particle reducible is composed of one-particle irreducible subgraphs connected by simple propagators. Hence a connected graph corresponds to a tree-level graph with 1PI subgraphs $\Gamma^{(n)}(x_1, ..., x_n)$ as (non-local) vertices. This shows that the quantum action $\Gamma[\phi]$ is the generating functional for the 1PI Green functions. Expanding $\Gamma[\phi]$ in ϕ_c gives us thus

$$\Gamma[\phi_c] = \sum_n \frac{1}{n!} \int \mathrm{d}^4x_1 \cdots \mathrm{d}^4x_n \Gamma^{(n)}(x_1, ..., x_n)\phi_c(x_1) \cdots \phi_c(x_n) \tag{12.13}$$

with $\Gamma^{(n)}$ as the one-particle-irreducible Green functions. In the following, we will need only the two- and three-point functions which we construct now explicitly.

Example 12.1: Show that $\Gamma^{(2)}$ is equal to the inverse propagator or inverse 2-point function, and derive the connection of $\Gamma^{(3)}$ to the connected 3-point function.
We write first

$$\delta(x_1 - x_2) = \frac{\delta\phi_c(x_1)}{\delta\phi_c(x_2)} = \int \mathrm{d}^4x \, \frac{\delta\phi_c(x_1)}{\delta J(x)} \frac{\delta J(x)}{\delta\phi_c(x_2)}$$

using the chain rule. Next we insert $\phi_c(x) = \delta W/\delta J(x)$ and $J(x) = -\delta\Gamma/\delta\phi_c(x)$ to obtain

$$\delta(x_1 - x_2) = -\int \mathrm{d}^4x \, \frac{\delta^2 W}{\delta J(x)\delta J(x_1)} \frac{\delta^2\Gamma}{\delta\phi_c(x)\delta\phi_c(x_2)}.$$

Setting $J = \phi_c = 0$, it follows

$$\int \mathrm{d}^4x \, \mathrm{i}G(x, x_1) \, \Gamma^{(2)}(x, x_2) = -\delta(x_1 - x_2) \tag{12.14}$$

or $\Gamma^{(2)}(x_1, x_2) = \mathrm{i}G^{-1}(x_1, x_2)$. Thus the 1PI 2-point function is the inverse propagator. Taking a further derivative $\delta/\delta J(x_3)$ of this relation, we have

$$\frac{\delta^3 W}{\delta J(x_1)\delta J(x_2)\delta J(x_3)} = -\frac{\delta}{\delta J(x_3)} \left(\frac{\delta^2\Gamma}{\delta\phi_c(x_1)\delta\phi_c(x_2)}\right)^{-1}. \tag{12.15}$$

Differentiating $MM^{-1}=1$, we find $\mathrm{d}M^{-1}=-M^{-1}\mathrm{d}MM^{-1}$ for a matrix M. Applied to (12.15), we obtain

$$\frac{\delta^3 W}{\delta J(x_1)\delta J(x_2)\delta J(x_3)}=\int \mathrm{d}^4y_1\mathrm{d}^4y_2 G^{(2)}(y_1,x_1)G^{(2)}(y_2,x_2)\frac{\delta^3\Gamma}{\delta\phi_c(x_1)\delta\phi_c(x_2)\delta J(x_3)}\,. \tag{12.16}$$

Using the chain rule, inserting $\phi_c(x_3)=\delta W/\delta J(x_3)$ and setting $J=\phi_c=0$ gives

$$G^{(3)}(x_1,x_2,x_3)=\int \mathrm{d}^4y_1\mathrm{d}^4y_2\mathrm{d}^4y_3 G^{(2)}(y_1,x_1)G^{(2)}(y_2,x_2)G^{(2)}(y_3,x_3)\Gamma^{(3)}(x_1,x_2,x_3)\,. \tag{12.17}$$

Thus the connected 3-point function $G^{(3)}(x_1,x_2,x_3)$ is obtained by appending propagators to the irreducible 3-point vertex function $\Gamma^{(3)}(x_1,x_2,x_3)$, one for each external leg. This generalises to all $n\neq 2$ and therefore one calls the $\Gamma^{(n)}$ also *amputated* Green functions.

12.2. Ward-Takahashi identities

We now turn to QED, the simplest case of a gauge theory. Discussing the quantum version of Noether's theorem, we have already shown that the vev of the electromagnetic current is conserved, $\partial_\mu\langle j^\mu\rangle=0$. Hence loop corrections respect the classical gauge symmetry and therefore we expect also that the photon remains massless. The redundancy implied by local gauge invariance leads to interrelations of Green functions. In turn, such dependencies are a necessary ingredient for the quantisation of a gauge symmetry, as we can see as follows: In QED, we define wave-function renormalisation constants for the electron and photon as

$$\psi_0\equiv Z_2^{1/2}\psi \quad and \quad A_0^\mu\equiv Z_3^{1/2}A^\mu\,. \tag{12.18}$$

Analogous to Eq. (11.72), the electric coupling is renormalised by

$$e(\mu)=Z_2Z_3^{1/2}Z_1^{-1}e_0\,, \tag{12.19}$$

where Z_1 is the charge renormalisation constant, and Z_2 and $Z_3^{1/2}$ take into account that two electron fields and one photon field enter the three-point function.

As it stands, the renormalisation condition (12.19) creates two major problems: First, the factor Z_2 will vary from fermion to fermion. For instance, the wave-function renormalisation constant of a proton includes the effects of strong interactions while the one of the electron does not. As a result, it is difficult to understand why the electric charge of an electron and an proton are renormalised such that they have the same value at $q^2=0$. Second, we see that the renormalised covariant derivative

$$D_\mu=\partial_\mu+\mathrm{i}\frac{Z_1}{Z_2}e(\mu)A_\mu \tag{12.20}$$

remains only gauge-invariant, if $Z_1=Z_2$. Clearly, this condition would also ensure the universality of the electric charge. Thus it is essential that we are able to show that $Z_1=Z_2$ holds in suitable renormalisation schemes. In a non-abelian theory as QCD, where we have to ensure that the gauge coupling in all terms of Eq. (10.76) remains after renormalisation the same, several constraints of the type $Z_1=Z_2$ arise.

We will proceed in two steps: First, we will show that the photon remains massless or, more technically, that the exact photon vacuum polarisation is transverse. Then we will derive the Ward-Takahashi identities using the quantum action which imply in particular the relation $Z_1=Z_2$.

Photon propagator Consider the generating functional of QED,

$$Z[J^\mu,\eta,\bar\eta] = \int \mathcal{D}A\mathcal{D}\bar\psi\mathcal{D}\psi \, \exp\{\mathrm{i}\int d^4x \mathscr{L}_{eff}\}, \tag{12.21}$$

where J^μ is a four-vector source, η and $\bar\eta$ are Graßmannian sources and the effective Lagrangian $\mathscr{L}_{eff}$ is composed of a classical term $\mathscr{L}_{cl}$, a gauge fixing term $\mathscr{L}_{gf}$ and a source term $\mathscr{L}_{\mathrm{s}}$,

$$\mathscr{L}_{\mathrm{cl}} = -\frac{1}{4}F_{\mu\nu}F^{\mu\nu} + \mathrm{i}\bar\psi\gamma^\mu D_\mu\psi - m\bar\psi\psi\,, \tag{12.22a}$$

$$\mathscr{L}_{\mathrm{g}f} + \mathscr{L}_{\mathrm{s}} = -\frac{1}{2\xi}(\partial_\mu A^\mu)^2 + J^\mu A_\mu + \bar\psi\eta + \bar\eta\psi\,. \tag{12.22b}$$

We consider now the renormalised version of the Lagrangian $\mathscr{L}_{eff}$, where the renormalised covariant derivative D_μ is given by Eq. (12.20) with $Z_1 \neq Z_2$ in general. This implies that an infinitesimal gauge transformation has the form

$$A_\mu \to A'_\mu = A_\mu + \partial_\mu\Lambda \tag{12.23a}$$

$$\psi \to \psi' = \psi - \mathrm{i}\bar e\Lambda\psi \qquad with \quad \bar e \equiv \frac{Z_1}{Z_2}e\,. \tag{12.23b}$$

As $\mathscr{L}_{cl}$ is gauge invariant by construction, the variation of $\mathscr{L}_{eff}$ under an infinitesimal gauge transformation consists only of

$$\delta\int \mathrm{d}^4x\;(\mathscr{L}_{\mathrm{g}f} + \mathscr{L}_{\mathrm{s}}) = \int \mathrm{d}^4x\,\left[-\frac{1}{\xi}(\partial_\mu A^\mu)\Box\Lambda + J^\mu\partial_\mu\Lambda + \mathrm{i}\bar e\Lambda(\bar\psi\eta - \bar\eta\psi)\right]. \tag{12.24}$$

Now we integrate by parts the first term twice and the second term once, to factor out the arbitrary function Λ,

$$\delta S_{eff} = \delta\int \mathrm{d}^4x\,\mathscr{L}_{eff} = \int \mathrm{d}^4x\,\left[-\frac{1}{\xi}\Box(\partial_\mu A^\mu) - \partial_\mu J^\mu + \mathrm{i}\bar e(\bar\psi\eta - \bar\eta\psi)\right]\Lambda\,. \tag{12.25}$$

Thus the variation of the generating functional $Z[J^\mu,\eta,\bar\eta]$ is

$$\delta Z[J^\mu,\eta,\bar\eta] = \int \mathcal{D}A\mathcal{D}\bar\psi\mathcal{D}\psi\,\exp\left(\mathrm{i}\int \mathrm{d}^4x\mathscr{L}_{eff}\right)\mathrm{i}\delta\int \mathrm{d}^4x\mathscr{L}_{eff}\,. \tag{12.26}$$

The fields A_μ, $\bar\psi$ and ψ are however only integration variables in the generating functional. The gauge transformation (12.23) is thus merely a change of variables which does not affect the functional $Z[J^\mu,\eta,\bar\eta]$, since the Jacobian of this transformation is one. Thus this variation has to vanish, $\delta Z[J^\mu,\eta,\bar\eta] = 0$.

If we substitute fields by functional derivatives of their sources, the change δS_{eff} can be moved outside the functional integral. Since the function $\Lambda(x)$ is arbitrary, we can drop the integration and arrive at

$$\begin{aligned} 0 &= \left[-\frac{1}{\xi}\Box\partial_\mu\frac{\delta}{\delta J_\mu} - \partial_\mu J^\mu + \mathrm{i}\bar e\left(\eta\frac{\delta}{\delta\eta} - \bar\eta\frac{\delta}{\delta\bar\eta}\right)\right]\exp\{\mathrm{i}W\} \\ &= -\frac{1}{\xi}\Box\left(\partial_\mu\frac{\delta W}{\delta J_\mu}\right) - \partial_\mu J^\mu + \mathrm{i}\bar e\left(\eta\frac{\delta W}{\delta\eta} - \bar\eta\frac{\delta W}{\delta\bar\eta}\right)\,. \end{aligned}$$

Differentiating this equation with respect to $J_\nu(y)$ and setting then the sources J^μ, η and $\bar\eta$ to zero gives us our first result,

$$-\frac{1}{\xi}\Box\Big(\partial_\mu\frac{\delta^2 W}{\delta J_\mu(x)\delta J_\nu(y)}\Big)=\partial_\mu\eta^{\mu\nu}\delta(x-y)\,. \tag{12.27}$$

The second derivative of W w.r.t. to the vector sources J_μ is the full photon propagator $\boldsymbol{D}^{\mu\nu}(x-y)$. If we go to momentum space, we have

$$\frac{\mathrm{i}}{\xi}k^2 k_\mu \boldsymbol{D}^{\mu\nu}(k)=k^\nu. \tag{12.28}$$

Splitting the propagator into a transverse part and a longitudinal part as in (10.60), the transverse part immediately drops out and we find

$$\frac{\mathrm{i}}{\xi}k^2 k_\nu \boldsymbol{D}_L(k^2)=k_\nu\,. \tag{12.29}$$

Thus the longitudinal part of the exact propagator agrees with the longitudinal part of the tree level propagator,

$$\mathrm{i}\boldsymbol{D}_L(k^2)=\mathrm{i}D_L(k^2)=\frac{\xi}{k^2}\,. \tag{12.30}$$

This implies that higher order corrections do not affect $\boldsymbol{D}_L(k^2)$. In other words, the loop correction $\Pi_{\mu\nu}$ to the photon propagator is transverse. Since we can expand all relations as power series in the coupling constant e, this holds also at any order in perturbation theory.

Ward identity $Z_1 = Z_2$ Let us go back to the constraint for the variation of the generating functional Z under gauge transformations, Eq. (12.27). We aim to derive identities between 1PI Green functions and want therefore to transform it into a constraint for the quantum action Γ. If we Legendre transform $W[J,\bar\eta,\eta]$ into $\Gamma[A,\bar\psi,\psi]$,

$$\Gamma[\psi,\bar\psi,A_\mu]=W[\eta,\bar\eta,J_\mu]-\int \mathrm{d}^4x(J_\mu A^\mu+\bar\psi\eta+\bar\eta\psi)\,, \tag{12.31}$$

we can replace the functional derivatives of W with classical fields, and the sources with functional derivatives of Γ, i.e.

$$\frac{\delta W}{\delta J_\mu}=A^\mu,\quad \frac{\delta W}{\delta\eta}=\bar\psi,\quad \frac{\delta W}{\delta\bar\eta}=\psi$$
$$\frac{\delta\Gamma}{\delta A_\mu}=-J^\mu,\quad \frac{\delta\Gamma}{\delta\bar\psi}=-\eta,\quad \frac{\delta\Gamma}{\delta\psi}=-\bar\eta\,.$$

This transforms Eq. (12.27) into

$$\frac{1}{\xi}\Box(\partial_\mu A^\mu(x))-\partial_\mu\frac{\delta\Gamma}{\delta A_\mu(x)}+\mathrm{i}\bar{e}\left(\psi\frac{\delta\Gamma}{\delta\psi(x)}-\bar\psi\frac{\delta\Gamma}{\delta\bar\psi(x)}\right)=0\,, \tag{12.32}$$

a master equation from which we can derive relations between different types of Green functions. Differentiating with respect to $\psi(x_1)$ and $\bar\psi(x_2)$ and setting then the fields to zero gives us the most important one of the *Ward-Takahashi* identities, relating the 1PI 3-point function $\Gamma^{(3)}(x,x_1,x_2)$ to the 2-point function $\Gamma^{(2)}(x_1,x_2)$ of fermions,

$$\partial^\mu \Gamma^{(3)}_\mu(x,x_1,x_2) = \mathrm{i}\bar{e}\big(\Gamma^{(2)}(x,x_1)\delta(x-x_2) - \Gamma^{(2)}(x,x_2)\delta(x-x_1)\big) \tag{12.33}$$

or, after Fourier transforming

$$k^\mu \Gamma^{(3)}_\mu(p,k,p+k) = \bar{e}\boldsymbol{S}_F^{-1}(p+k) - \bar{e}\boldsymbol{S}_F^{-1}(p)\,. \tag{12.34}$$

Taking the limit $k^\mu \to 0$, the identity $\Gamma^{(3)}_\mu(p,0,p) = \bar{e}\partial_\mu \boldsymbol{S}_F^{-1}(p)$ found originally by Ward follows. The Green functions in this equation are finite, renormalised quantities and thus Z_1/Z_2 has to be finite in any consistent renormalisation scheme too. In all schemes where we identify directly $e(0)$ with the measured electric charge, also the finite parts of Z_1 and Z_2 agree, i.e. the measured electric charge is universal.

12.3. Vacuum polarisation

The Ward identities guaranty that the gauge couplings in QED and QCD are determined solely by the loop corrections $\Pi_{\mu\nu}$ to the photon and gluon propagator, respectively. These corrections convert the classical coupling constants into running couplings. How the running coupling change as function of the scale is one of the most characteristic properties of a quantum field theory. The aim of this section is therefore to derive the one-loop corrections to the gauge boson propagator. An interpretation of these results will then be performed in the next section.

12.3.1. Vacuum polarisation in QED

We calculate next the one-loop correction to the photon propagator, the so called vacuum polarisation tensor $\Pi_{\mu\nu}$. In QED, only the first diagram of Fig. 12.1 contributes. Using the Feynman rules, we obtain for the contribution of one fermion species with mass m

$$\mathrm{i}\Pi^{\mu\nu}(q) = -(-\mathrm{i}e)^2 \int \frac{\mathrm{d}^4k}{(2\pi)^4} \frac{tr[\gamma^\mu \mathrm{i}(\not{k}+m)\gamma^\nu \mathrm{i}(\not{k}+\not{q}+m)]}{(k^2-m^2)[(k+q)^2-m^2]} = -e^2 \int \frac{\mathrm{d}^4k}{(2\pi)^4}\frac{\mathcal{N}^{\mu\nu}}{D}\,. \tag{12.35}$$

As a warm-up, we want to confirm that the vacuum polarisation tensor respects at the one-loop level gauge invariance, as we know already from Eq. (12.30). Gauge invariance implies $q_\mu \Pi^{\mu\nu}(q) = 0$ and thus the tensor structure of the vacuum polarisation tensor has to be $\Pi^{\mu\nu}(q) = (q^2\eta^{\mu\nu} - q^\mu q^\nu)\Pi(q^2)$. Hence we have to calculate only the simpler scalar function $\Pi(q^2)$ knowing that $\Pi^{\mu\nu}(q)$ is gauge invariant.

In order to show that $q_\mu \Pi^{\mu\nu}(q) = 0$, we write first

$$q = (q+k-m) - (k-m) \tag{12.36}$$

and obtain

$$q_\mu \mathcal{N}^{\mu\nu} = \mathrm{tr}\{[(\not{q}+\not{k}-m) - (\not{k}-m)](\not{k}+m)\gamma^\nu(\not{k}+\not{q}+m)\} \tag{12.37a}$$

$$= [(q+k)^2-m^2]\mathrm{tr}\{(\not{k}+m)\gamma^\nu\} - (k^2-m^2)\mathrm{tr}\{\gamma^\nu(\not{k}+\not{q}+m)\} \tag{12.37b}$$

where we used the cyclic property of the trace. Employing dimensional regularisation (DR) with $d = 4 - 2\varepsilon$ in order to obtain well-defined integrals,

$$q_\mu \mathrm{i}\Pi^{\mu\nu}(q) = -e^2\mu^{2\varepsilon}\int \frac{\mathrm{d}^d k}{(2\pi)^d}\left\{\frac{tr[(\not{k}+m)\gamma^\nu]}{(k^2-m^2)} - \frac{tr[\gamma^\nu(\not{k}+\not{q}+m)]}{(k+q)^2-m^2}\right\}, \tag{12.38}$$

we are allowed to shift the integration variable in one of the two terms. Thus $q_\mu\Pi^{\mu\nu}(q) = 0$ and hence the vacuum polarisation tensor at order $\mathcal{O}(e^2)$ is transverse as required by gauge invariance.

Let us pause a moment and summarise what we know already, before we start with the evaluation of $\Pi(q^2)$: In our power-counting analysis we found as superficial degree of divergence $D = 2$. This result was based on the assumption that the numerator $\mathcal{N}$ behaves as a constant. But the only constant available is m^2 which would lead to a mass term of the photon, $e^2 A_\mu \Pi^{\mu\nu} A_\nu \propto e^2 m^2 A_\mu \eta^{\mu\nu} A_\nu$. Thus the transversality of $\Pi^{\mu\nu}$ implies that the m^2 term in the numerator will disappear at some step of our calculation. Thereby the convergence of the polarisation tensor improves, becoming a "mild logarithmic" one.

We proceed with the explicit evaluation of $\Pi^{\mu\nu}$. Taking the trace of (12.35) and using its transversality, we find in $d = 2\omega$ dimensions

$$\Pi^\mu_\mu(q) = (q^2\delta^\mu_\mu - q^2)\Pi(q^2) = (d-1)q^2\Pi(q^2)$$

and

$$(d-1)q^2\mathrm{i}\Pi(q^2) = -e^2\mu^{2\varepsilon}\int\frac{\mathrm{d}^d k}{(2\pi)^d}\frac{tr[\gamma^\mu(\not{k}+m)\gamma_\mu(\not{k}+\not{q}+m)]}{(k^2-m^2)[(k+q)^2-m^2]} = -e^2\mu^{2\varepsilon}\int\frac{\mathrm{d}^d k}{(2\pi)^d}\frac{\mathcal{N}}{D}. \tag{12.39}$$

We combine the two propagators introducing the Feynman parameter integral $1/(AB) = \int_0^1 \mathrm{d}x\,[Ax + B(1-x)]^{-2}$ with $A = (k+q)^2 - m^2$ and $B = k^2 - m^2$. Thus the denominator becomes

$$\mathcal{D} = Ax + B(1-x) = k^2 + 2kqx + q^2x - m^2 = (k+qx)^2 + q^2x(1-x) - m^2\,. \tag{12.40}$$

Next we introduce as new integration variable $K = k + qx$,

$$\mathcal{D} = K^2 + q^2x(1-x) - m^2 = K^2 - a\,, \tag{12.41}$$

and $a = m^2 - q^2x(1-x) > 0$ as short-cut.

Evaluating the trace in the denominator using DR, we have to extend the Clifford algebra to $d = 2\omega$ dimensions. A natural choice is $tr(\gamma^\mu\gamma^\nu) = d\eta^{\mu\nu}$, giving with $\gamma^\mu\gamma_\mu = d$ and $\gamma^\mu\not{q}\gamma_\mu = (2-d)\not{q}$ as result for the trace

$$\mathcal{N} = \mathcal{N}^\mu_\mu = d[(2-d)k\cdot(k+q) + dm^2] = d\{(2-d)[K^2 - q^2x(1-x)] + dm^2\}\,. \tag{12.42}$$

In the last step, we performed the shift $k \to K = k + qx$ omitting linear terms in K that vanish after integration. Combining our results for $\mathcal{N}$ and D we arrive at

$$(d-1)q^2\mathrm{i}\Pi(q^2) = -e^2\mu^{2\varepsilon}d\int_0^1\mathrm{d}x\int\frac{\mathrm{d}^d K}{(2\pi)^d}\,\frac{(2-d)K^2 + dm^2 - (2-d)q^2x(1-x)]}{(K^2-a)^2}\,. \tag{12.43}$$

Finally, we use our results for the Feynman integrals $I_0(\omega,\alpha)$ and $I_2(\omega,\alpha)$ which were obtained performing a Wick rotation. The latter is possible as long as we do not pass a singularity. Since the prefactor $x(1-x)$ of q^2 has as maximum $1/4$, this requires $q^2 < 4m^2$. We start to look for the first two terms where we expect a cancellation of the m^2 term in the numerator,

$$\int \frac{\mathrm{d}^d K}{(2\pi)^d} \frac{(2-d)K^2 + dm^2}{(K^2-a)^2} = (2-d)I_2(\omega,2) + dm^2 I(\omega,2) \tag{12.44a}$$

$$= \frac{\mathrm{i}}{(4\pi)^\omega \Gamma(2)} \left[-2\omega(1-\omega)\Gamma(1-\omega)\, a^{\omega-1} + 2\omega m^2 \Gamma(2-\omega)\, a^{\omega-2}\right] \tag{12.44b}$$

$$= \frac{\mathrm{i}}{(4\pi)^\omega} 2\omega\Gamma(2-\omega)(-a+m^2)a^{\omega-2} \tag{12.44c}$$

$$= \frac{4\,\mathrm{i}}{(4\pi)^2} \Gamma(\varepsilon)\,(4\pi)^\varepsilon \frac{q^2 x(1-x)}{[m^2 - q^2 x(1-x)]^\varepsilon}\,. \tag{12.44d}$$

Hence the m^2 term dropped indeed out of the numerator and the whole expression is proportional to q^2, as required by the LHS of (12.43). Evaluating the third term in the same way we obtain

$$-\int \frac{\mathrm{d}^d K}{(2\pi)^d} \frac{(2-d)q^2 x(1-x)}{(K^2-a)^2} = -(2-d)q^2 x(1-x) I(\omega,2) \tag{12.45a}$$

$$= \frac{2\,\mathrm{i}}{(4\pi)^2} \Gamma(\varepsilon)\,(4\pi)^\varepsilon \frac{q^2 x(1-x)}{[m^2 - q^2 x(1-x)]^\varepsilon}\,. \tag{12.45b}$$

Adding the two contributions, we arrive at

$$\Pi(q^2) = -\frac{8e^2}{(4\pi)^2} \Gamma(\varepsilon) \int_0^1 \mathrm{d}x\, x(1-x) \left[\frac{4\pi\mu^2}{m^2 - q^2 x(1-x)}\right]^\varepsilon, \tag{12.46}$$

where the factor $\mathrm{i}q^2$ cancelled. We included also the factor $(4\pi\mu^2)^\varepsilon$ into the last term, which becomes thereby dimensionless. Now we expand the Gamma function, $\Gamma(\varepsilon) = 1/\varepsilon - \gamma + \mathcal{O}(\varepsilon)$, around $d = 4-2\varepsilon$ and because of the resulting $1/\varepsilon$ term all other ε dependent quantities,

$$\Pi(q^2) = -\frac{e^2}{12\pi^2} \left\{\frac{1}{\varepsilon} - \gamma + \ln(4\pi) - 6\int_0^1 \mathrm{d}x\, x(1-x) \ln\left[\frac{m^2 - q^2 x(1-x)}{\mu^2}\right]\right\}. \tag{12.47}$$

The prefactor $x(1-x)$ has its maximum $1/4$ for $x = 1/2$. Thus the branch cut of the logarithm starts at $q^2 = 4m^2$, i.e. when the virtuality of the photon is large enough that it can decay into a fermion pair of mass $2m$. This is a nice illustration of the optical theorem: The polarisation tensor is real below the pair creation threshold, and acquires an imaginary part above (which equals the pair creation cross section of a photon with mass $m^2 = q^2$, cf. problem 12.**??**).

The x integral can be integrated by elementary functions, but we display the result only for the two limiting cases of small and large virtualities, $q^2/m^2 \to 0$ and $|q|^2/m^2 \to \infty$. In the first case, we obtain with $\ln(1-x) \simeq -x$

$$\Pi(q^2) = -\frac{e^2}{12\pi^2}\left[\frac{1}{\varepsilon} - \gamma + \ln(4\pi) + \ln(\mu^2/m^2) + \frac{q^2}{5m^2} + \ldots\right], \tag{12.48}$$

while the opposite limit gives

$$\Pi(q^2) = -\frac{e^2}{12\pi^2}\left[\frac{1}{\varepsilon} - \gamma + \ln(4\pi) - \ln(|q^2|/\mu^2) + \ldots\right]. \tag{12.49}$$

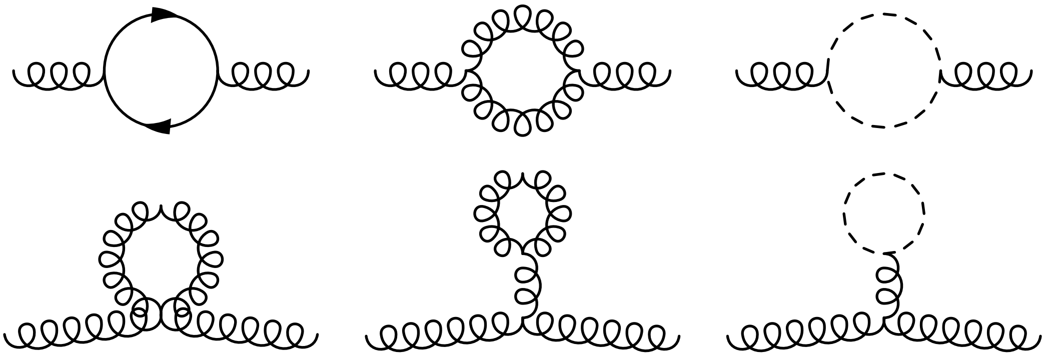


Figure 12.1.: Feynman diagrams describing vacuum polarisation in QCD at one-loop.

In the MS scheme, we obtain the renormalisation constant Z_3 for the photon field as the coefficient of the pole term,

$$Z_3 = 1 - \frac{e^2}{12\pi^2\varepsilon}\,. \tag{12.50}$$

More often the on-shell renormalisation scheme is used in QED. Here we require that quantum corrections to the electric charge vanish for $q^2 = 0$, i.e. we choose Z_3 such that $\Pi^{on}(q^2 = 0) = 0$. This is obviously achieved setting

$$\Pi^{\rm on}(q^2) = \Pi(q^2) - \Pi(0) = -\frac{e^2}{60\pi^2}\frac{q^2}{m^2} + \dots\,, \tag{12.51}$$

for $|q^2| \ll m^2$. This q^2 dependence leads to a modification of the Coulomb potential, which can be measured e.g. in the Lamb shift.

12.3.2. Vacuum polarisation in QCD

We only sketch the calculations in QCD, stressing the new or different points compared to QED. In Fig. 12.1, we show the various 1-loop diagrams contributing to the vacuum polarisation tensor in a non-abelian theory as QCD. Most importantly, the three- and four-gluon vertices allow in addition to the quark loop now also gluon loops. Since a fermion loop has an additional minus sign, we expect that the gluon loop gives a negative contribution to the beta function. This opens the possibility that non-abelian gauge theories are in contrast to QED asymptotically free, if the number of fermion species is small enough.

Quark loop The vertex changes from $-\mathrm{i}e\gamma^\mu$ in QED to $-\mathrm{i}g_sT^a\gamma^\mu$ in QCD. Since the quark propagator is diagonal in the group index, a quark loop contains for each flavor additionally the factor

$$tr\{T^aT^a\} = \frac{1}{2}\delta^{aa} = 4\,. \tag{12.52}$$

Thus we have only to replace $e \to 4n_fg_s$ in the QED result, where n_f counts the number of quark flavors. For the three light quarks, u, d and s, it is an excellent approximation to use $m = 0$. In contrast, the masses of the other three quarks (c, b and t) can not be neglected when $4m_f^2 \lesssim \mu^2$. The effect of particle masses can be approximated including in the loop only particles with mass $4m_f^2 < \mu^2$, making n_f scale dependent.

Loop with three-gluon vertex Since the three-gluon vertex connects identical particles, we have to take into account symmetry factors similar as in the case of the $\lambda\phi^4$ theory. We learnt that the imaginary part of a Feynman diagram corresponds to the propagation of real particles. Thus the imaginary part of the gluon vacuum polarisation can be connected to the total cross section of $g \to gg$ scattering. This cross section contains a symmetry factor 1/2! to account for two identical particles in the final state. Therefore the same symmetry factor should be associated to the vacuum polarisation with a gluon loop[3]. Applying the Feynman rule for the three-gluon vertex and using the Feynman-t'Hooft gauge, we find

$$\mathrm{i}\Pi^{\mu\nu}_{ab\,2}(q^2) = \frac{1}{2}(-\mathrm{i}g)^2\mathrm{i}^2 \int \frac{\mathrm{d}^4k}{(2\pi)^4}\frac{\mathcal{N}^{\mu\nu}_{ab}}{(k^2+\mathrm{i}\varepsilon)[(k+q)^2+\mathrm{i}\varepsilon]} \tag{12.53}$$

with

$$\begin{aligned}\mathcal{N}^{\mu\nu}_{ab} &= f_{bcd}[-\eta^{\nu\rho}(q+k)^\sigma + \eta^{\rho\sigma}(2k-q)^\nu + \eta^{\sigma\nu}(2q-k)^\rho] \\ &\times f_{acd}[-\eta^{\mu}{}_{\rho}(q+k)_\sigma + \eta_{\rho\sigma}(q-2k)^\mu + \eta^{\mu}{}_{\sigma}(k-2q)_\rho]\,.\end{aligned} \tag{12.54}$$

Evaluating the colour trace,

$$f^{acd}f^{cdb} = N_c\delta^{ab} \tag{12.55}$$

and extracting in the usual way the pole part using DR one obtains for $d = 4 - 2\varepsilon$

$$\Pi^{\mu\nu}_{ab\,2}(q^2) = -\frac{N_c g^2}{32\pi^2\varepsilon}\left(\frac{\mu^2}{q^2}\right)^{\varepsilon/2}\left(\frac{11}{3}q^\mu q^\nu - \frac{19}{6}\eta^{\mu\nu}q^2\right) + \mathcal{O}(\varepsilon^0)\,. \tag{12.56}$$

Thus the contribution from the three-gluon vertex alone is not transverse—demonstrating again that a covariant gauge requires the introduction of ghost particles.

Ghost loop This diagram has the same dependence on the structure constants as the previous one,

$$\mathrm{i}\Pi^{\mu\nu}_{ab,\,3}(q^2) = -(-\mathrm{i}g)^2\mathrm{i}^2 \int \frac{\mathrm{d}^4k}{(2\pi)^4}\frac{f_{bdc}(k-q)^\nu f_{acd}k_\nu}{(k^2+\mathrm{i}\varepsilon)[(k+q)^2+\mathrm{i}\varepsilon]}\,, \tag{12.57}$$

and can thus be combined with the three-gluon loop. Note the extra minus sign due to the fermionic nature of the ghost particle. Evaluating the integral results in

$$\Pi^{\mu\nu}_{ab,\,3}(q^2) = \frac{N_c g^2}{32\pi^2\varepsilon}\left(\frac{\mu^2}{q^2}\right)^{\varepsilon/2}\left(\frac{1}{3}q^\mu q^\nu + \frac{1}{6}\eta^{\mu\nu}q^2\right) + \mathcal{O}(\varepsilon^0) \tag{12.58}$$

and summing the three-gluon and ghost loops gives the expected gauge-invariant expression. Moreover, the sum has the opposite sign as the quark loop and can thus lead to the opposite behaviour of the beta function as in QED.

Four-gluon loop and tadpole diagrams The loop with the four-gluon vertex contains a massless propagator and does not depend on external momenta,

$$\mathrm{i}\Pi^{\mu\nu}_{ab,4}(q^2) \propto \int \frac{\mathrm{d}^dk}{k^2+\mathrm{i}\varepsilon}\,. \tag{12.59}$$

[3] This argument does not apply to the quark loop, since cutting leads in this case to a distinguishable $\bar{q}q$ state

Our general experience with DR tells us that this graph is zero, as gluons are massless. However, this loop integral is also in DR ambiguous: For any space-time dimension d, the integral is either UV or IR divergent. To proceed, we split therefore the integrand introducing the arbitrary mass M as

$$\frac{1}{k^2+\mathrm{i}\varepsilon} = \frac{1}{k^2-M^2+\mathrm{i}\varepsilon} - \frac{M^2}{(k^2+\mathrm{i}\varepsilon)(k^2-M^2+\mathrm{i}\varepsilon)}\,. \tag{12.60}$$

Now the IR and UV divergences are separated, and we can use $d<2$ in the first term and $2<d<4$ in the second. By dimensional analysis, both terms have to be proportional to a power of the arbitrary mass M. As the LHS is independent of M, the only option is that the two terms on the RHS cancel, as an explicit calculation confirms. The remaining tadpole diagrams (5 and 6 of Fig. 12.1) vanish by the same argument. We will come back to the combined result for the coupling $g_s(Q^2)$ in QCD, after having introduced running couplings and discussing their behaviour in general.

12.4. Renormalisation group

Renormalisation group equations Let us consider two renormalisation schemes R and R'. In the two schemes, the renormalised field will in general differ, being $\phi_R = Z_\phi^{-1/2}(R)\phi_0$ and $\phi_{R'} = Z_\phi^{-1/2}(R')\phi_0$, respectively. Hence the connection between the two renormalised fields is

$$\phi_{R'} = \frac{Z_\phi^{-1/2}(R')}{Z_\phi^{-1/2}(R)}\,\phi_R \equiv Z_\phi^{-1/2}(R',R)\,\phi_R\,. \tag{12.61}$$

As both ϕ_R and $\phi_{R'}$ are finite, also $Z_\phi(R',R)$ is finite. The transformations $Z_\phi^{-1/2}(R',R)$ form a group, called the renormalisation group. If we consider

$$G_0^{(n)}(x_1,\ldots,x_n) = Z_\phi^{n/2}\,G_R^{(n)}(x_1,\ldots,x_n)\,, \tag{12.62}$$

we know that the bare Green function is independent of the renormalisation scale μ. Taking a derivative with respect to μ, the LHS thus vanishes. To avoid clutter, we restrict ourselves to the simplest case of a massless theory with a single coupling g. Then the renormalised Green functions can depend (in a mass independent scheme) only on the renormalised coupling $g(\mu)$, and we find thus

$$0 = \frac{\mathrm{d}}{\mathrm{d}\ln\mu}\left[Z_\phi^{n/2}\,G_R^{(n)}(x_1,\ldots,x_n)\right] \tag{12.63a}$$

$$= \left[\frac{\partial}{\partial\ln\mu} + \frac{\partial g}{\partial\ln\mu}\frac{\partial}{\partial g} + \frac{n}{2}\,\frac{\partial\ln Z_\phi}{\partial\ln\mu}\right]G_R^{(n)}(x_1,\ldots,x_n) \tag{12.63b}$$

$$\equiv \left[\frac{\partial}{\partial\ln\mu} + \beta\frac{\partial}{\partial g} + \frac{n}{2}\,\gamma\right]G_R^{(n)}(x_1,\ldots,x_n)\,. \tag{12.63c}$$

Here we introduced in the last step the anomalous dimension

$$\gamma(\mu) = \mu\,\frac{\partial\ln Z_\phi(\mu)}{\partial\mu} \tag{12.64}$$

of the field ϕ and the beta function

$$\beta(g) = \mu \frac{\partial g(\mu)}{\partial \mu}, \tag{12.65}$$

which determines the logarithmic change of the coupling constant. The beta function is often re-expressed as

$$\beta(g^2) \equiv \mu^2 \frac{\partial g^2}{\partial \mu^2} = g\beta(g) \tag{12.66}$$

or as $\beta(\alpha^2)$ with $\alpha = g^2/(4\pi)$. Equations of the type (12.63c) are called generically renormalisation group equations or briefly RGE. They come in various flavors, carrying the name of their inventors: Stückelberg–Petermann, Callan–Symanzik, Gell-Mann–Low, ... We can use (11.63) to derive a similar RGE for the 1PI Green functions. Note that the sign difference between (11.63) and (11.62) in the power of the wave-function renormalisation constant induces a corresponding sign change in the RGE for 1PI Green functions.

Remark 12.1: In order to understand the term "anomalous dimensions," we look first at the canonical dimension of fields under a change of units. Note that in contrast to the scale transformations $x \to \mathrm{e}^{\alpha}x$ discussed in problem 5.??, we now change all parameters including couplings and masses. The classical action is then invariant, and a n-point Green function $G(p_1, \ldots, p_n)$ can be expressed as

$$G(p_1, \ldots, p_n) = m^a g^b p_1^{c_1} \cdots p_n^{c_n}$$

though the parameters of the classical Lagrangian. Dimensional analysis constrains the exponents as $a - c_1 - \ldots - c_n = n$, and the n-point Green function scales classically as $G \to \mathrm{e}^{n\alpha x}G$.
The renormalised Green function depends however also on the scale μ, i.e. the subtraction point of our renormalisation scheme which we keep fixed,

$$G(p_1, \ldots, p_n) = m^a g^b p_1^{c_1} \cdots p_n^{c_n} \mu^{\gamma} .$$

Hence the renormalised n-point Green function scales as $G \to \mathrm{e}^{(n-\gamma)\alpha x}G$ and satisfies therefore $\mu \mathrm{d}G/\mathrm{d}\mu = \gamma G$.

Knowing the two universal functions $\beta(\lambda)$ and $\gamma(\mu)$, we can calculate the change of any Green function under a change of the renormalisation scale μ. The general solution of (12.63c) can be found by the method of characteristics or by the analogy of $\mathrm{d}/\mathrm{d}\ln\mu$ with a convective derivative, cf. problem 12.??. Here we consider only the simplest case of a dimensionless observable R in a renormalisable theory which depends on a single physical momentum scale Q. Thus we assume that the coupling constant is dimensionless, and consider the limit that all masses can be neglected, $|Q^2| \gg m_i^2$. An important example is the e^+e^- annihilation cross section into hadrons which can be made dimensionless dividing by $\sigma(e^+e^- \to \mu^+\mu^-)$ as reference cross section. Then R can be only a function of the ratio Q^2/μ^2 and of $\alpha(\mu^2)$. A physical observable like R should be independent of the scale μ, or

$$0 = \mu^2 \frac{\mathrm{d}R}{\mathrm{d}\mu^2} = \left(\mu^2 \frac{\partial}{\partial \mu^2} + \beta(\alpha)\frac{\partial}{\partial \alpha}\right) R(Q^2/\mu^2, \alpha(\mu^2)) = \left(-\frac{\partial}{\partial \tau} + \beta(\alpha)\frac{\partial}{\partial \alpha}\right) R(\tau, \alpha(\mu^2)) , \tag{12.67}$$

where we introduced $\tau = \ln(Q^2/\mu^2)$. The two differential operators compensate each other setting

$$\tau = \int_{\alpha(\mu^2)}^{\alpha(Q^2)} \frac{\mathrm{d}x}{\beta(x)} . \tag{12.68}$$

The coupling $\alpha(Q^2) = \alpha(\tau, \alpha(\mu^2))$ defined by (12.68) as function of the physical momentum Q^2 is called the "running coupling". Setting the renormalisation scale equal to the physical scale, $\mu^2 = Q^2$, results in

$$R(Q^2/\mu^2, \alpha(\mu^2)) = R(1, \alpha(Q^2)) \,. \tag{12.69}$$

Hence all scale dependence in R can be absorbed into the running of $\alpha(Q^2)$. We have seen already an example for this behaviour, discussing $\phi\phi \to \phi\phi$ scattering in section 4.3.4, compare especially with Eq. (4.89).

Example 12.2: Any function of the running coupling, and thus in particular the running coupling itself, solves the homogeneous RGE. To show the latter claim, we differentiate first the definition Eq. (12.68) w.r.t. τ, obtaining

$$1 = \frac{\partial}{\partial\tau}\left[\int_{\alpha(\mu^2)}^{\alpha(Q^2)} \frac{\mathrm{d}x}{\beta(x)}\right] = \frac{1}{\beta(\alpha(Q^2)}\frac{\partial\alpha(Q^2)}{\partial\tau} \tag{12.70}$$

or $\partial_\tau \alpha(Q^2) = \beta(\alpha(Q^2))$. Next we differentiate (12.68) w.r.t. $\alpha(\mu^2)$, obtaining

$$0 = \frac{\partial}{\partial\alpha(\mu^2)}\left[\int_{\alpha(\mu^2)}^{\alpha(Q^2)} \frac{\mathrm{d}x}{\beta(x)}\right] = \frac{1}{\beta(\alpha(Q^2))}\frac{\partial\alpha(Q^2)}{\partial\alpha(\mu^2)} - \frac{1}{\beta(\alpha(Q^2))} \tag{12.71}$$

or

$$\frac{\partial\alpha(Q^2)}{\partial\alpha(\mu^2)} = \frac{\beta(\alpha(Q^2))}{\beta(\alpha(\mu^2))} \,. \tag{12.72}$$

Evaluating the differential operator of the homogeneous RGE (i.e. for $\gamma = 0$) for the coupling, we see that the two terms cancel as required,

$$\left(\frac{\partial}{\partial\tau} + \beta(\alpha(\mu^2))\frac{\partial}{\partial\alpha(\mu^2)}\right)\alpha(\tau, \alpha(\mu^2)) = -\beta(\alpha(Q^2)) + \beta(\alpha_s(\mu^2))\frac{\beta(\alpha(Q^2))}{\beta(\alpha(\mu^2))} = 0 \,. \tag{12.73}$$

Working through problem 11.**??**, you found that the beta-function of the $\lambda\phi^4$ theory at one loop is given by

$$\beta(\lambda) = b_1\lambda^2 + b_2\lambda^3 + \ldots = \frac{3\lambda^2}{16\pi^2} + \mathcal{O}(\lambda^3) \,, \tag{12.74}$$

so that the running coupling at leading order satisfies

$$\lambda(Q) = \frac{\lambda(\mu)}{1 - 3\lambda(\mu)/16\pi^2 \ln(Q/\mu)} \,. \tag{12.75}$$

Thus our new definitions (12.68) and (12.65) reproduce the result of our more intuitive approach in section 4.3.4. As a bonus, we know now what we should choose as the argument of the coupling, at least in the simple case considered with a single physical scale Q. Expanding the running coupling,

$$\lambda(Q) = \lambda(\mu) + b_1 \ln(Q/\mu) - [b_1 \ln(Q/\mu)]^2 + \ldots \,. \tag{12.76}$$

we see that $\lambda(Q)$ contains arbitrary powers of $[b_1\lambda(\mu \ln(Q/\mu)]^n$, although we derived the beta function only at one-loop: The RGE ensures that the running coupling sums up the largest terms in each order perturbation theory, cf. problem 12.**??**. In general, the running coupling calculated at n-loop precision contains the leading $\ln^n(Q/\mu)$ terms of loop diagrams of any order. Accordingly, one speaks of LL (leading logarithmic), NLL (next-to-leading logarithmic), NNLL, …approximations.

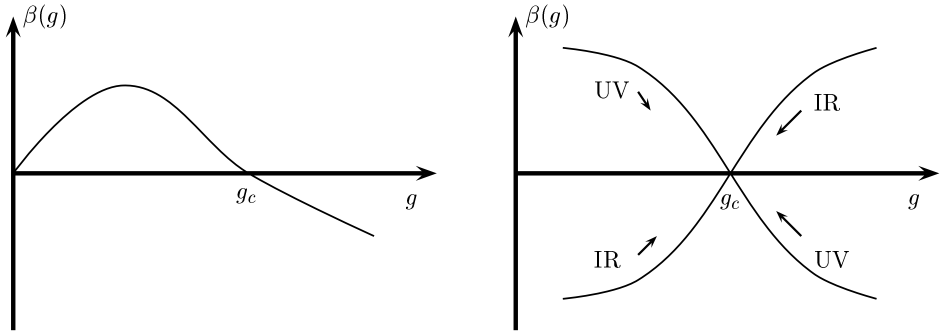


Figure 12.2.: Left: Example of a beta function with a perturbative IR and an UV fixed point. Right: General classification of UV and IR stable fixed points of the RGE flow.

Asymptotic behaviour of the beta-function The behaviour of the beta-function $\beta(\mu)$ in the limit $\mu \to 0$ and $\mu \to \infty$ provides a useful classification of quantum field theories. Consider e.g. the example shown in the left panel of Fig. 12.2: This beta function has a trivial zero at zero coupling, as we expect it in any perturbative theory, and an additional zero at g_c. How does the beta-function $\beta(\mu)$ evolve in the UV limit $\mu \to \infty$?

- Starting in the range $0 \leq g(\mu) < g_c$ implies $\beta > 0$ and thus $\mathrm{d}g/\mathrm{d}\mu > 0$. Therefore g grows with increasing μ and the coupling is driven towards g_c.
- Starting with $g(\mu) > g_c$ implies $\beta < 0$ and thus $\mathrm{d}g/\mathrm{d}\mu < 0$. Therefore g decreases for increasing μ and we are driven again towards g_c.

Fixed points g_c approached in the limit $\mu \to \infty$ are called UV fixed points, while IR fixed points are reached for decreasing μ. The range of values $[g_1 : g_2]$ which is mapped by the RGE flow on the fixed point is called its basin of attraction.

To see what happens for $\mu \to 0$, we have only to reverse the RGE flow, $\mathrm{d}\mu \to -\mathrm{d}\mu$, and are thus driven away from g_c: If we started in $0 \leq g(\mu) < g_c$, we are driven to zero, while the coupling goes to infinity for $g(\mu) > g_c$. Thus $g = 0$ is an IR fixed point. The distinction between IR and UV fixed points is sketched in the right panel of Fig. 12.2. If the beta function has several zeros, the theory consists of different phases which are not connected by the RG flow.

Looking back at our one-loop result for the beta-function of the $\lambda\phi^4$ theory, we see that the theory has $\lambda = 0$ as an IR fixed point. Thus the free states we use as asymptotic initial and final states have a direct physical meaning. On the other hand, the coupling increases for $\mu \to \infty$ formally as $\lambda \to \infty$. Clearly, we cannot trust the behaviour of $\lambda(\mu)$ based on the on-loop result in the strong-coupling limit, because perturbation theory breaks down. The solution (12.75) suggests however that the coupling explodes already for a finite value of μ: The beta function has a pole for a finite value of μ called Landau pole where the denominator of (12.75) becomes zero.

Beta function of QED We can derive the scale dependence of the renormalised electric charge from

$$e_0 = \mu^{\varepsilon} Z_3^{-1/2}\, e\,, \tag{12.77}$$

where we used $Z_1 = Z_2$. Then the beta function is given as

$$\beta(e) \equiv \mu\frac{\partial e}{\partial \mu} = \mu\frac{\partial}{\partial \mu}\left(\mu^{-\varepsilon} Z_3^{1/2} e_0\right) . \tag{12.78}$$

Since the bare charge e_0 is independent of μ, we have to differentiate only μ and Z_3,

$$\beta(e) = \mu\frac{\partial e}{\partial \mu} = -\varepsilon\mu^{-\varepsilon} Z_3^{1/2} e_0 + \mu\mu^{-\varepsilon}\,\frac{1}{2}\,Z_3^{-1/2}\frac{\partial Z_3}{\partial \mu}\,e_0 = -\varepsilon e + \frac{\mu}{2Z_3}\,\frac{\partial Z_3}{\partial \mu}\,e\,. \tag{12.79}$$

Inserting $Z_3 = 1 - e^2/(12\pi^2\varepsilon)$ and thus

$$\frac{\partial Z_3}{\partial \mu} = -\frac{1}{12\pi^2\varepsilon}\frac{2e\partial e}{\partial \mu} \tag{12.80}$$

gives

$$\beta(e) = -\varepsilon e - \frac{\mu}{12\pi^2\varepsilon Z_3}\,\frac{\partial e}{\partial \mu}\,e^2 = -\varepsilon e - \frac{1}{12\pi^2\varepsilon Z_3}\,\beta(e)\,e^2\,. \tag{12.81}$$

Note that Z_3 is scheme-dependent, while the beta-function remains scheme independent up to two loops for all mass-independent schemes (problem 12.**??**). Solving for β and neglecting higher order terms in e^2, we find in the limit $\varepsilon \to 0$

$$\beta(e) = -\varepsilon e\left(1 - \frac{e^2}{12\pi^2\varepsilon}\right) + \mathcal{O}(e^4) = \frac{e^3}{12\pi^2}\,. \tag{12.82}$$

Thus the beta function is determined by the coefficient of the pole term of Z_3. Its solution,

$$e^2(\mu) = \frac{e^2(\mu_0)}{1 - \frac{e^2(\mu_0)}{6\pi^2}\ln\left(\frac{\mu}{\mu_0}\right)} \tag{12.83}$$

shows not only explicitly the increase of e^2 with μ, but that the electric coupling has a Landau pole too. However, the scale of the Landau pole corresponds to

$$\mu = \mu_0\exp(6\pi^2/e^2(\mu_0)) = m_e\exp(3\pi/2\alpha(m_e)) \sim 10^{56}\,\mathrm{G}eV \gg M_{\mathrm{Pl}} \tag{12.84}$$

and has therefore no direct physical relevance.

Alternatively, we could derive the beta function from the renormalised vacuum polarisation, using

$$\beta(e) = \frac{e}{2}\mu\frac{\partial}{\partial \mu}\Pi^{\mathrm{M}S}(q^2)\,, \tag{12.85}$$

where in the on-shell scheme the derivative $\partial\ln\mu^2$ should be replaced with $\partial\ln q^2$.

Beta function of QCD and asymptotic freedom Deriving the beta function in QCD, we should evaluate

$$g(\mu) = \frac{Z_2 Z_3^{1/2}}{Z_1} g_0 = Z_3^{1/2} g_0\,, \tag{12.86}$$

where Z_1 is the charge renormalisation constant, and Z_2 and Z_3 are quark and gluon renormalisation constants, respectively. Without proof, we note that the generalisation of the Ward-Takahashi identities to the non-abelian case, the Slavnov-Taylor identities, ensure that

$Z_1 = Z_2$. Combining all contributions to the $1/\varepsilon$ poles as 1-loop contribution b_1 to the beta function of QCD gives (cf. problem 12.??)

$$\beta(\alpha_s) = \mu^2 \frac{\partial \alpha_s}{\partial \mu^2} = -\alpha_s^2 \left(b_1 + b_2 \alpha_s + b_3 \alpha_s^2 + \ldots \right) \tag{12.87}$$

with $\alpha_s \equiv g_s^2/(4\pi)$,

$$b_1 = \frac{11}{3} N_c - \frac{2}{3} n_f \tag{12.88}$$

and n_f as number of quark flavors. For $n_f < 17$, the beta function is negative and the running coupling decreases as $\mu \to \infty$. Asymptotic freedom of QCD explains the apparent paradox that protons are interacting strongly at small Q^2, while they can be described in deep-inelastic scattering as a collection of independently moving quarks and gluons.

Let us consider now the opposite limit, $\mu \to 0$. The solution of (12.87) at one loop,

$$\frac{1}{\alpha_s(\mu^2)} = \frac{1}{\alpha_s(\mu_0^2)} + b_1 \ln \left(\frac{\mu^2}{\mu_0^2} \right) \tag{12.89}$$

shows that the QCD coupling constant becomes formally infinite for a finite value of μ. We define Λ_{QCD} as the energy scale where the running coupling constant of QCD diverges, $\alpha_s^{-1}(\Lambda_{QCD}^2) = 0$. Experimentally, the best measurement of the strong coupling constant has been performed at the Z resonance at LEP, giving $\alpha_s(m_Z^2) \sim 0.1184$. Thus at one-loop level,

$$\Lambda_{QCD} = m_Z \exp \left(\frac{1}{2 b_1 \alpha_s(m_Z^2)} \right) . \tag{12.90}$$

Λ_{QCD} depends on the renormalisation scheme and, numerically more importantly, on the number of flavours used in b_1: For instance, $\Lambda_{QCD}^{\overline{MS}} \simeq 220\,\mathrm{MeV}$ for $n_f = 3$. The fact that the running coupling provides a characteristic energy scale is called dimensional transmutation: Quantum corrections lead to the break-down of scale-invariance of classical QCD with massless quarks and to the appearance of massive bound-states, the mesons and baryons with masses of order Λ_{QCD}. Note also that we are able to link exponentially separated scales by dimensional transmutation.

Coupling constant unification While the strong coupling $\alpha_s \equiv \alpha_3$ decreases with increasing Q^2, the electromagnetic coupling $\alpha_{\rm em} \equiv \alpha_1$ increases. Since two lines in a plane meet at one point, there is a point with $\alpha_1(Q_*) = \alpha_3(Q_*)$ and one may speculate that at this point a transition to a "Grand unified theory" (GUT) happens. Since the running is only logarithmic, unification happens at exponentially high scales, $Q_* \sim 10^{16}\,\mathrm{GeV}$, but interestingly still below the Planck scale $M_{\rm Pl}$. The problems becomes more challenging, if we add to the game the third, the weak coupling α_2. The situation in 1991 assuming the validity of the SM is shown in the left panel of Fig. 12.3. The width of the lines indicates the experimental and theoretical error, and the three couplings clearly do not meet within these errors. The right panel of the same figure assume the existence of supersymmetric partners to all SM particles with an "average mass" of around $M_{SUSY} \sim 200\,\mathrm{GeV}$. As a result, the running changes above $Q = 1\,\mathrm{TeV}$, and now the three couplings meet at $2 \times 10^{16}\,\mathrm{GeV}$.

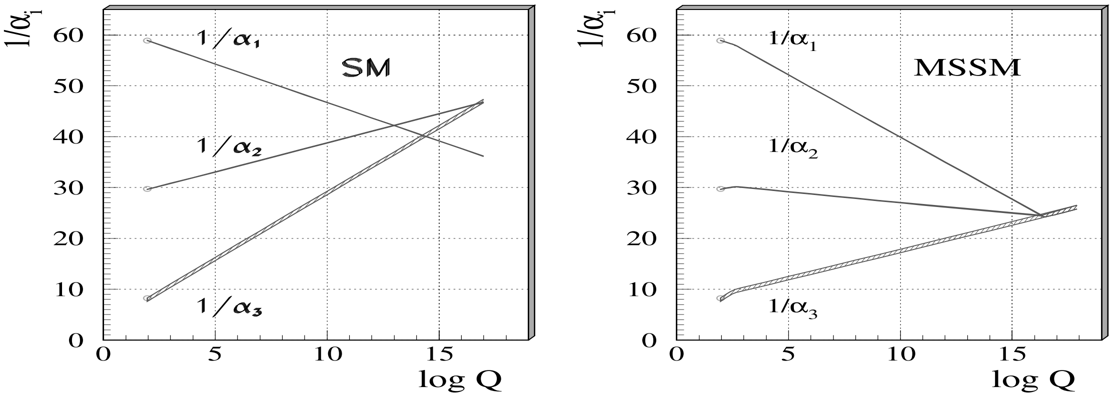


Figure 12.3.: The gauge couplings measured at low energies do not evolve towards a unified value in the SM (left), but meet assuming low-scale supersymmetry at $\simeq 2 \times 10^{16}$ GeV (right), from [Kaz00] based on [AdBF91].

12.5. Renormalisation, critical phenomena and effective theories

Overview The behaviour thermodynamical systems exhibit close to the critical points in their phase diagram are characterised as "critical phenomenon." For a fixed number of particles, we can characterise thermodynamical systems using the free energy $F = U - TS$. Ehrenfest introduced the classification of phase transitions according to the order of the first discontinuous derivative of F with respect to any thermodynamical variable ϕ. Hence a phase transition where at least one derivative $\partial^n F/\partial\phi^n$ is discontinuous while all $\partial^{n-1}F/\partial\phi^{n-1}$ are continuous is called a n.th order phase transition.

Critical phenomena are for a particle physicist interesting by at least three reasons:

- We can learn about symmetry breaking: We should look out for ideas how we can generate mass terms without violating gauge invariance. Systems like ferromagnets show that symmetries as rotation symmetry can be broken at low energies although the Hamiltonian governing the interactions is rotation symmetric.
  Another example is a plasma: Here the screening of electric charges modifies the Coulomb potential to a Yukawa potential; the photon has three massive degrees of freedom, still satisfying gauge invariance, $k_\mu \Pi^{\mu\nu}(\omega, \boldsymbol{k}) = 0$, but with $\omega^2 - \boldsymbol{k}^2 \neq 0$.

- Experimentally one finds that close to a critical point, $T \to T_c$, the correlation length ξ diverges, while otherwise correlations are exponentially suppressed. If we consider a statistical system on a lattice, then the 2-point function of a certain order parameter ϕ scales as

  $$\langle \phi_n \phi_0 \rangle \propto \exp(-n/\xi)\,,$$

  where ξ is measured in multiples of the lattice spacing a. Comparing this with $|x| = na$ to the 2-point function of an Euclidean scalar field ϕ,

  $$\langle \phi(x)\phi(0) \rangle \to \frac{4\pi}{|x|^2} \exp(-m|x|)$$

in the limit $m|x| \gg 1$, we find the correspondence $\xi = 1/(ma)$. Therefore the continuum limit $a \to 0$ is only possible for finite m, if $\xi \to \infty$.
Thus the correlation functions of a statistical system correspond for non-zero a to bare Green functions and a finite value of the regulator of the corresponding quantum field theory. The connection to renormalised Green functions ($a \to 0$ or $\Lambda \to \infty$) is only possible when the statistical system is at a critical point.

- Near a critical point, $T \to T_c$, thermodynamical systems show a universal behaviour. More precisely, they fall in different universality classes which unify systems with very different microscopic behaviour. The various universality classes can be characterised by critical exponents, i.e. by the exponents γ_i with which characteristic thermodynamical quantities X_i diverge approaching T_c, i.e. $X_i = [b(T - T_c)]^{-\gamma_i}$.
This phenomenon is similar to our realisation that e.g. two $\lambda\phi^4$ theories, one with $\lambda = 0.1$ and another one with $\lambda = 0.2$, are not fundamentally different but connected by a RGE transformation.

Landau's mean field theory Landau suggested that the free energy F of a thermodynamical system can be expanded close to a second-order phase transition as an even series in its order parameter. Considering e.g. the magnetisation M, we can write for zero external field H the free energy as

$$F = A(T) + B(T)M^2 + C(T)M^4 + \dots \tag{12.91}$$

We can find the possible value of the magnetisation M by minimising the free energy,

$$0 = \left.\frac{\partial F}{\partial M}\right|_T = 2B(T)M + 4C(T)M^3 \,. \tag{12.92}$$

The variable $C(T)$ has to be positive in order that F is bounded from below. If also $B(T)$ is positive, only the trivial solution $M = 0$ exists. If however $B(T)$ is negative, two solutions with non-zero magnetisation appear. Let us use a linear approximation, $B(T) \approx b(T - T_c)$, and $C(T) \approx c$ valid close to T_c. Then

$$M = \begin{cases} 0 & \text{for } T > T_c, \\ \pm\left[\frac{b}{2c}(T_c - T)\right]^{1/2} & \text{for } T < T_c. \end{cases} \tag{12.93}$$

Note also that the ground-state breaks the $M \to -M$ symmetry of the free energy for $T < T_c$.
Representing the thermodynamical quantity M as integral of the local spin density,

$$M = \int \mathrm{d}^3x \; s(\boldsymbol{x}) \,, \tag{12.94}$$

we can rewrite the free energy in a way resembling the Hamiltonian of a stationary scalar field,

$$F = \int \mathrm{d}^3x \;\left[(\boldsymbol{\nabla} s)^2 + b(T - T_c)s^2 + cs^4 - \boldsymbol{H}\cdot\boldsymbol{s}\right] \,. \tag{12.95}$$

Here, $(\boldsymbol{\nabla} s)^2$ is the simplest ansatz leading to an alignment of spins in the continuous language. Minimising F will give us the ground-state of the system for a prescribed external field $\boldsymbol{H}(\boldsymbol{x})$ and temperature T. For small s, we can ignore the s^4 term. The spin correlation function $\langle s(\boldsymbol{x})s(\boldsymbol{0})\rangle$ is found as response to a delta function-like disturbance $H_0\delta(x)$. Using the

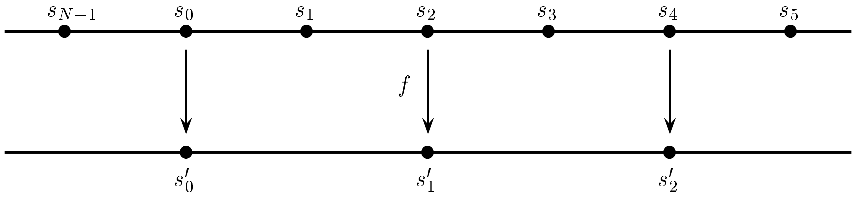


Figure 12.4.: One block spin transformation $a \to 2a$ for an one-dimensional lattice model.

analogy with the Yukawa potential after the substitution $m^2 \to b(T-T_c)$, the correlation function follows immediately as

$$\langle s(\boldsymbol{x})s(\boldsymbol{0})\rangle = \int \frac{\mathrm{d}^3k}{(2\pi)^3}\frac{H_0\mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{x}}}{k^2+b(T-T_c)} = \frac{H_0}{4\pi r}\,\mathrm{e}^{-r/\xi} \tag{12.96}$$

with

$$\xi = [b(T-T_c)]^{-1/2}\,. \tag{12.97}$$

Hence Landau's theory reproduces the experimentally observed behaviour $\xi \to \infty$ for $T \to T_c$. Moreover, the theory predicts as critical exponent 1/2 for the magnetisation. Notice that the value of the exponent depends only on the polynomial assumed in the free energy, not on the underlying micro-physics. Thus another prediction of Landau's theory is an universal behaviour of thermodynamical systems close to their critical points in the dependence on $T-T_c$. Experiments show that this prediction is too strong: Thermodynamical systems fall into different universality classes, and we should try to include some micro-physics into the description of critical phenomena.

Kadanoff's block spin transformation Close to a critical point, collective effects play a decisive role even in case of short-range interactions. In d dimension, a particle is coupled by collective effects to $(\xi/a)^d$ particles and standard perturbative methods will certainly fail for $\xi \to \infty$. Kadanoff suggested to remove the short wave-length fluctuations by the following procedure: Each step of a block spin transformation consists of i) dividing the lattice into cells of size $(2a)^d$, ii) assigning a common spin variable to the cell, and iii) of a rescaling $2a \to a$.

At each step, the number of strongly correlated spins is reduced. After n transformations, the correlation length decreases as $\xi_n = \xi/(2^n)$. When the correlation length becomes of the order of the lattice spacing, collective effects play no role: All the physics can be read off from the Hamiltonian. If the procedure is not trivial, this implies that in each step the Hamiltonian changes. In particular, the coupling constant K is changed as

$$K_2 = f(K)\,, \qquad K_3 = f(K_2) = f(f(K))\,,\dots \tag{12.98}$$

One-dimensional Ising model We illustrate the idea behind Kadanoff's block spin transformation using the example of the one-dimensional Ising model. This model consists of spins with value $s_i = \pm 1$ on a line with spacing a, interacting via nearest neighbour interactions.

We consider only the piece of six spins shown in Fig. 12.4. The corresponding partition function is

$$Z_6 = \sum_{s_{N-1},s_0,s_1,\ldots,s_4} \exp\left[K(s_{N-1}s_0 + s_0 s_1 + \ldots s_3 s_4)\right] \tag{12.99a}$$

$$= \sum_{s'_0,s'_1,s'_2} \sum_{s_{N-1},s_1,s_3} \exp\left[K(s_{N-1}s'_0 + s'_0 s_1 + \ldots s_3 s_4)\right] , \tag{12.99b}$$

where the summation is over the two spin values ±1 and $K = J/T$ is a dimensionless coupling constant. The step $a \to 2a$ requires to perform the sums over the unprimed spins. Expanding the exponentials using $(s_i s_j)^{2n} = 1$ gives terms like

$$\exp[K(s'_0 s_1)] = 1 + K s'_0 s_1 + \frac{K^2}{2!} + \frac{K^3}{3!}\, s'_0 s_1 + \ldots \tag{12.100a}$$

$$= \cosh(K) + \, s'_0 s_1 \sinh(K) = \cosh(K)[1 + \, s'_0 s_1 \tanh(K)] . \tag{12.100b}$$

The terms linear in s_1 cancel in the sum and we obtain

$$\sum_{s_1=\pm 1} \exp[K s'_0 s_1] \exp[K s_1 s'_1] = 2\cosh^2(K)[1 + \tanh^2(K)\, s'_0 s'_1] . \tag{12.101}$$

Thus the summation over the unprimed spins changes the strength of the nearest neighbourhood interaction and generates additionally a new spin-independent interaction term. We try now to rewrite the last expression in a form similar to the original one,

$$2\cosh^2(K)[1 + \tanh^2(K)\, s'_0 s'_1] = \exp[g(K) + K' s'_0 s'_1)] . \tag{12.102}$$

Using (12.100b) to replace $\exp(K' s'_0 s'_1)$ and setting $g(K) = \ln h(K)$, we find

$$2\cosh^2(K)[1 + \tanh^2(K)\, s'_0 s'_1] = h(K)\cosh(K')[1 + \tanh(K')\, s'_0 s'_1)] . \tag{12.103}$$

This determines the function $g = \ln h(K)$ as

$$h(K) = \frac{2\cosh^2(K)}{\cosh(K')} , \tag{12.104}$$

while the couplings are related by

$$\tanh(K') = \tanh^2(K) . \tag{12.105}$$

The summation over the other spins $s_3, s_5, \ldots$ can be performed in the same way. Thus the partition function on a lattice of size $2a$ has the same nearest-neighbour interactions with a new coupling $K' \equiv K_1$ determined by (12.105). Iterating this procedure generates a renormalisation flow with

$$\tanh(K_n) = \tanh^{2n}(K) . \tag{12.106}$$

Additionally, the RGE flow generates all couplings compatible with the symmetries of the fundamental Hamiltonian: Since any operator $\mathcal{O}^n$ for $n \in \mathbb{Z}$ satisfies these symmetries if $\mathcal{O}$ does, there is an infinite number of them.

Fixed point behaviour In general, we will not be able to calculate the transformation function $f(K)$. But even without the knowledge of $f(K)$, we can draw some important insight from general considerations. First, the RGE equations are of the type of a heat or diffusion equation[4]. Its flow is therefore a gradient flow which has only two possible asymptotics: a runaway solution to infinity and the approach to a fixed point defined by $K_c = f(K_c)$. With $\xi_{n+1} = \xi_n/2$ and labeling the n dependence implicitly via $\xi_n = f(K_n)$ we can write

$$\xi(f(K)) = \frac{1}{2}\xi(K)\,. \tag{12.107}$$

At a fixed point $K_c = f(K_c)$ only two solutions exist,

$$\xi(K_c) = 0 \quad \text{a}nd \quad \xi(K_c) \to \infty\,. \tag{12.108}$$

The second possibility corresponds to the approach of a critical point, allowing the limit $a \to 0$ and thus the continuum limit necessary for the transition to a QFT. This point is called a critical fixed point, while the fixed point with zero correlation length is called trivial.

We can now generalise our previous discussion of the fixed point behaviour for the beta function from one to n dimensions. The general behaviour of the RGE flow can be understood from the two-dimensional example shown in Fig. 12.5. The dashed lines show surfaces of constant correlation length, including a critical surface $\xi = \infty$. Also shown are three critical fixed points (A, B, C) and a trivial one (D). In each RGE step the correlation length decreases. Thus the trivial fixed point is an attractor, i.e. inside a small enough neighbourhood all points will flow towards it. Moreover, critical lines have at least one unstable direction, the one orthogonal to their surface: Even points infinitesimal close to the surface will flow away and eventually end in a trivial fixed point. Finally, also inside the critical surface stable and unstable directions exist: For instance, the fixed point B will attract all points in-between A and C ("its basin of attraction"). Universality classes of QFTs can be identified with stable critical fixed points and their basin of attraction.

Wilsonian action Let us now return to QFTs in the continuum limit. Wilson suggested to transfer the idea of integrating out fluctuations on small scales by including in the path integral only modes up to the cutoff scale $\Lambda \sim 1/a$. This defines an *effective field theory* which is by construction UV finite. The corresponding effective or Wilsonian action depends on the assumed value of the cutoff scale Λ, and we can generate a RGE flow by changing Λ.

Let us start from the Euclidean generating functional in momentum space restricted to wave-numbers below a cutoff, $k \leq \Lambda$. Then we split the field modes into slow modes σ and fast modes ψ,

$$Z = \int \mathcal{D}\phi\, \mathrm{e}^{-S[\phi]} = \int \mathcal{D}\sigma\mathcal{D}\phi\, \mathrm{e}^{-S[\sigma+\psi]}\,, \tag{12.109}$$

with

$$\phi = \sigma + \psi \text{ and } \begin{cases} \sigma = 0, \text{ unless } |k| \leq \Lambda/f \\ \psi = 0, \text{ unless } \Lambda/f \leq |k| \leq \Lambda \end{cases}. \tag{12.110}$$

Next we want to integrate out the fast modes ψ,

$$\mathrm{e}^{-S[\sigma]} = \int \mathcal{D}\psi\, \mathrm{e}^{-S[\sigma+\psi]}\,, \tag{12.111}$$

[4] An example is given later in remark 13.1

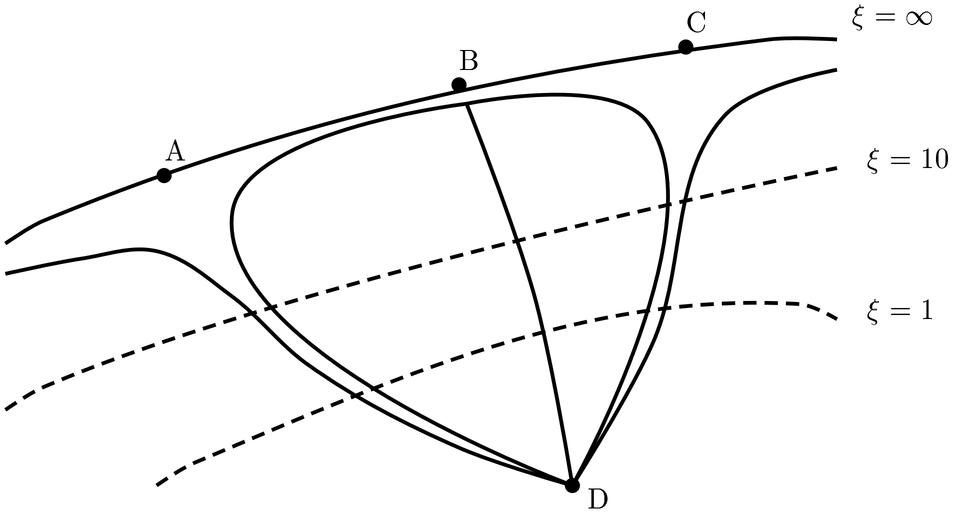


Figure 12.5.: A two-dimensional illustration of the RGE flow: A, B and C are three fixed points on the critical surface $\xi = \infty$. The fixed points A and C have a stable direction along the critical surface, while B has two unstable directions. The trivial fixed point D is a stable fixed point, attracting all points starting not on the critical surface.

lowering thereby the cutoff, $\Lambda \to \Lambda/f$. In general, we will not be able to perform this path integral. Using perturbation theory, we expand $\mathrm{e}^{-S[\sigma+\psi]}$ and exponentiate then again the result. As illustrated with a toy model in problem 12.??, this procedure will renormalise the values of the parameters contained in the original Lagrangian and introduce an infinite set of irrelevant operator $\mathcal{O}_{d,i}$ with dimension $d > 5$.

Having integrated out ψ, we relabel σ as ϕ. Next we have to recover the canonical normalisation for its kinetic energy. If we rescale distances by $x \to x' = x/f$, the functional integral is again over modes $\phi(x')$ with $x > a$, complying with step iii) of the Kadanoff-Wilson prescription. Keeping then the kinetic term invariant,

$$\int \mathrm{d}^4x \, (\partial_\mu \phi)^2 = \int \mathrm{d}^4x' \, (\partial'_\mu \phi')^2 = \frac{1}{f^2} \int \mathrm{d}^4x \, (\partial_\mu \phi')^2 \,, \tag{12.112}$$

requires a rescaling of the field as $\phi' = f\phi$. Let us consider now an irrelevant interaction, e.g. $g_6\phi^6$. Then

$$g_6 \int \mathrm{d}^4x \, \phi^6 = \frac{g_6}{f^2} \int \mathrm{d}^4x' \, \phi'^6 \tag{12.113}$$

shows that the new coupling g'_6 is rescaled as $g'_6 = g_6/f^2$: As f grows and the cutoff scale Λ decreases, the value of an irrelevant coupling is driven to zero. Clearly, a relevant operator as the cosmological constant ρ or a mass term $m^2\phi^2$ shows the opposite behaviour and grows. As result, irrelevant couplings are in our low-energy world suppressed and as first approximation a renormalisable theory emerges at low energies. Note that both directions of the RGE flow—towards the UV or the IR—are useful discussing QFTs. The point of view of a RGE flow towards the IR is useful, if we want to connect a theory at high energy scales to a theory valid at lower scales. An example for this approach is chiral perturbation theory where one connects QCD to an effective theory of mesons and baryons at low energies. In the opposite view, we may look e.g. at the SM as an effective theory known to be valid up to scales around TeV and ask what happens if we increase the cutoff.

We can generalise now our earlier discussion of the two-dimensional RGE flow sketched in Fig. 12.5. The RGE flow stops at fixed points on critical surfaces of low dimension. All initial values on the critical surface inside the basin of attraction flow to the critical fixed point. Directions perpendicular to the critical surface are controlled by the irrelevant interactions; flows beginning off the surface are driven to the trivial fixed point. On the way towards $\xi \to \infty$ only the relevant and marginal interactions survive. Insisting not on $\xi \to \infty$ and keeping a finite cutoff (somewhere between TeV and M_{Pl}, depending on the limit of validity of the theory we assume) we can keep irrelevant interactions which are however suppressed.

Effective theories While integrating out momentum shells in the path integral, as required in Eq. (12.111), is useful to demonstrate the concept of effective theories, it is not the method applied by practitioners. Instead, either functional RGE methods are used for which we will give later an example in remark 13.1. Or one evaluates effective theories using the same apparatus as we have developed for renormalisable ones. In the following, we want first to illustrate how this is done and, second, to show that effective theories are predictive, despite of being non-renormalisable.

We note first that the fact that divergences are polynomial in the external momenta guarantees also for non-renormalisable theories that all divergences can be subtracted by local counter-terms. In contrast to renormalisable theories, the number of counter-terms and thus of a priori required measurements is infinite. Thus it is important to show that a truncation scheme for effective theories including loop correction exists which makes them predictive. As usually, we use as an example a real scalar field in $d = 4$ space-time dimensions. Then an operator $\mathcal{O}_{n,m}$ consisting of n fields and m derivatives contributes to the action

$$S \approx \frac{g_{n,m}}{\Lambda^{p+q-4}} \int \mathrm{d}^4x \, \mathcal{O}_{n,m} \approx c_{n,m} \left(\frac{E}{\Lambda}\right)^{n+m-4} , \tag{12.114}$$

where the $g_{n,m}$ are dimensionless couplings to be determined by experiment. We see again that irrelevant operators, $n+m > 4$, are suppressed at energies $E \ll \Lambda$. Clearly, an increase of the mass dimension $n+m$ results in a stronger suppression of the operator $\mathcal{O}_{n,m}$. This allows us to truncate the effective theory at some chosen dimension $D = n + m$. The truncated theory contains only a finite number of operators with unknown couplings $g_{n,m}$ which have to be determined from experiment. The predictions of the truncated theory become exact in the limit $E \to 0$. For finite $E \ll \Lambda$, we can increase the precision including higher dimensional operators, at the expense of more calculational work and additional experimental input with sufficient precision. In contrast, for $E \gtrsim \Lambda$ an infinite number of operators contribute a priori equally and the effective theory approach breaks down. Finally, note that in contrast to the Wilsonian action the value of the cutoff scale Λ is determined experimentally: For instance, calculating weak processes using a four-fermion interaction fixes the scale Λ^2 through the Fermi constant G_F.

Our estimate (12.114) holds at tree-level, and we should consider next the effect of loop corrections. To be specific, we estimate the order of the one-loop corrections induced by the operators $g_6\phi^6/\Lambda^2$, $g_8\phi^4(\partial\phi)^2/\Lambda^4$, ..., to the basic $\lambda\phi^4$ vertex. Cutting the momentum integrals at Λ results in

$$\delta\lambda \approx \frac{g_6}{\Lambda^2}\int^\Lambda \frac{\mathrm{d}^4k}{(2\pi)^4}\frac{1}{k^2-m^2} \approx \frac{c_6}{\Lambda^2}\Lambda^2\,, \tag{12.115}$$

$$\delta\lambda \approx \frac{g_8}{\Lambda^4}\int^\Lambda \frac{\mathrm{d}^4k}{(2\pi)^4}\frac{k^2}{k^2-m^2} \approx \frac{c_8}{\Lambda^4}\Lambda^4\,,\ldots \tag{12.116}$$

Thus all one-loop corrections are of $\mathcal{O}(1)$, since the factor $1/\Lambda^{D-4}$ contained in $\mathcal{O}_{n,m}$ is compensated by a factor Λ^{D-4} generated in the momentum integrations. Going to higher loops makes things even worse. The solution to this problem is to use a mass independent regulator as the MS or $\overline{\mathrm{MS}}$ schemes in DR. Then the factors Λ^n originating from the momentum integration are replaces by m^n. With $E \gg m$, the loop corrections preserve thus the expansion scheme $(E/\Lambda)^{n+m-4}$ found at tree-level. As a result, the truncation of an effective theories at a chosen order D leads to a predictive theory including loop corrections at energies below the cutoff scale.

Summary

Interactions can be characterised by the asymptotic behaviour of their coupling constants. Gauge theories with a sufficiently small number of fermions are the only renormalisable interactions which are asymptotically free, i.e. their running coupling constant goes to zero for $\mu \to \infty$. The scale dependence of renormalised Green functions can be interpreted as a running of coupling constants and masses. The use of a running coupling sums up the leading logarithms of type $\ln^n(\mu^2/\mu_0^2)$, and a suitable choice of the renormalisation scale in a specific problems reduces the remaining scale dependence of perturbative results. The non-perturbative approach of Wilson provides an argument why the SM as description of our low-energy world is renormalisable: Integrating out high-energy degrees of freedom, irrelevant couplings are driven to zero and thus it is natural that a renormalisable theory emerges at low energies.

Further reading

Our discussion of the renormalisation of gauge theories left out most details. I recommend those interested to fill the gaps to start with [Ram94] and [Pok87]. A useful text-book to learn about critical phenomena is [LB92].

13. Symmetries and symmetry breaking

The analogy of Landau's mean field model for a ferromagnet with a scalar $\lambda\phi^4$ theory suggests that we can hide its $\phi \to -\phi$ symmetry at low temperatures, if we choose a negative mass term in the Lagrangian. Although such a choice seems at first sight unnatural, we will investigate this case in the following in detail. Our main motivation is the expectation that hiding a symmetry by choosing a non-invariant ground-state retains the "good" properties of the symmetric Lagrangian. Coupling then such a scalar theory to a gauge theory, we hope to break gauge invariance in a "gentle" way which allows e.g. gauge boson masses without spoiling the renormalisability of the unbroken theory. As an additional motivation we recall that couplings and masses are not constants but depend on the scale considered. Thus it might be that the parameters determining the Lagrangian of the Standard Model at low energies originate from a more complete theory at high scales, where the squared mass parameter μ^2 is originally positive. In such a scenario, $\mu^2(Q^2)$ becomes negative only after running it down to the electroweak scale $Q = m_Z$.

13.1. Symmetry breaking and Goldstone's theorem

We consider in the following systems where the Lagrangian contains an exact symmetry which is not shared by its ground-state. Since particle masses in a field theory are determined by the ground-state, the symmetry of the Lagrangian is thus not visible in the mass spectrum of physical particles. This opens the possibility of having a gauge invariant Lagrangian despite of massive gauge bosons. If the ground-state breaks the original symmetry because one or several scalar fields acquire a non-zero vacuum expectation value, one calls this spontaneous symmetry breaking (SSB). As the symmetry is not really broken on the Lagrangian level, a perhaps more appropriate name would be "hidden symmetry".

In this and the following chapter, we discuss the case of SSB, first in general and then applied to the electroweak sector of the SM. Since the breaking of an internal symmetry should leave Poincaré symmetry intact, we can give only scalar quantities ϕ a non-zero vacuum expectation value. This excludes non-zero expectation values for fields with spin, which would single out a specific direction. On the other hand, we can construct scalar expectation values as $\langle 0|\phi|0\rangle = \langle 0|\bar{\psi}\psi|0\rangle \neq 0$ out of the product of multiple fields. In the following, we will always treat ϕ as an elementary field, but we should keep in mind the possibility that ϕ is a composite object, e.g. a condensate of fermion fields, $\langle\phi\rangle = \langle\bar{\psi}\psi\rangle$, similar to the case of superconductivity.

Spontaneous breaking of discrete symmetries We start with the simplest example of a theory with a broken symmetry: A single scalar field with a discrete reflection symmetry. Consider the familiar $\lambda\phi^4$ Lagrangian, but with a negative mass term which we include into the potential $V(\phi)$,

$$\mathscr{L} = \frac{1}{2}(\partial_\mu\phi)^2 + \frac{1}{2}\mu^2\phi^2 - \frac{\lambda}{4}\phi^4 = \frac{1}{2}(\partial_\mu\phi)^2 - V(\phi)\,. \tag{13.1}$$

The Lagrangian is for both signs of μ^2 invariant under the discrete Z_2 symmetry, $\phi \to -\phi$. The field configuration with the smallest energy is a constant field ϕ_0, chosen to minimise the potential

$$V(\phi) = -\frac{1}{2}\mu^2\phi^2 + \frac{\lambda}{4}\phi^4, \tag{13.2}$$

which has the two minima

$$\phi_0 = \pm v \equiv \pm\sqrt{\mu^2/\lambda}\,. \tag{13.3}$$

In quantum mechanics, we learnt that the wave-function of the ground state for the potential $V(x) = -\frac{1}{2}\mu^2 x^2 + \lambda x^4$ will be a symmetric state, $\psi(x) = \psi(-x)$, since the particle can tunnel through the potential barrier. In field theory, such a tunnelling can happen in principle too. However, the tunnelling probability is inversely proportional to the volume L^3 occupied by the system, and vanishes in the limit $L \to \infty$: In order to transform $\phi(x) = -v$ into $\phi(x) = +v$ we have to switch an infinite number of oscillators, which clearly costs an infinite amount of energy. Thus in quantum field theory, the system has to choose between the two vacua $\pm v$ and the symmetry of the Lagrangian is broken in the ground state. Had we used the ϕ^4 Lagrangian with a positive mass term, the vacuum expectation value of the field would have been zero, and the ground state would respect the symmetry.

Quantising the theory (13.1) with the negative mass around the usual vacuum, $|0\rangle$ with $\langle 0|\phi|0\rangle = \phi_c = 0$, we find modes behaving as

$$\phi_k \propto \exp(-\mathrm{i}\omega t) = \exp(-\mathrm{i}\sqrt{-\mu^2 + |\boldsymbol{k}|^2}\,t)\,, \tag{13.4}$$

which can grow exponentially for $|\boldsymbol{k}|^2 < \mu^2$. More generally, exponentially growing modes exist, if the potential is concave at the position of ϕ_c, i.e. for

$$m^2_{eff}(\phi_c) = V''(\phi_c) = -\mu^2 + 3\lambda\phi_c^2 < 0 \tag{13.5}$$

or $|\phi_c| < \sqrt{\mu^2/(3\lambda)}$.

Clearly, the problem arises because we should expand the field around the ground-state v. This requires that we shift the field as

$$\phi(x) = v + \xi(x)\,, \tag{13.6}$$

splitting it into a classical part $\langle\phi\rangle = v$ and quantum fluctuations $\xi(x)$ on top of it. Then we express the Lagrangian as function of the field ξ,

$$\mathscr{L} = \frac{\mu^4}{4\lambda} + \frac{1}{2}(\partial_\mu\xi)^2 - \frac{1}{2}(2\mu^2)\xi^2 - \mu\sqrt{\lambda}\xi^3 - \frac{\lambda}{4}\xi^4\,. \tag{13.7}$$

In the new variable ξ, the Lagrangian describes a scalar field with *positive* mass $m_\xi = \sqrt{2}\mu > 0$. The original symmetry is no longer apparent: Since we had to select one out of the two possible ground states, a ξ^3 term appeared and the $\phi \to -\phi$ symmetry is broken. The new cubic interaction term rises now the question, if our scalar $\lambda\phi^4$ theory becomes non-renormalisable after SSB: As we have no corresponding counter-term at our disposal, the renormalisation of μ and λ has to cure also the divergences of the ξ^3 interaction.

Finally, we note that the contribution $\mu^4/(4\lambda)$ to the energy density of the vacuum is, in contrast to the vacuum loop diagrams generated by $Z[0]$, classical and finite. We see later that symmetries will be restored at high temperatures or at early times in the evolution of the Universe. Even if we take the freedom to shift the vacuum energy density, we have either before or after SSB an unacceptable large contribution to the vacuum energy (problem 13.1).

Spontaneous breaking of continuous symmetries Our main aim is to understand the SSB of the electroweak gauge symmetry. As next step we look therefore at a system with a global continuous symmetry. We discussed already in section 5.1 the case of N real scalar fields described by the Lagrangian

$$\mathscr{L} = \frac{1}{2}\left[(\partial_\mu \boldsymbol{\phi})^2 + \mu^2 \boldsymbol{\phi}^2\right] - \frac{\lambda}{4}(\boldsymbol{\phi}^2)^2 . \tag{13.8}$$

Since $\boldsymbol{\phi} = \{\phi_1, \ldots, \phi_N\}$ transforms as a vector under rotations in field space, $\phi_i \to R_{ij}\phi_j$ with $R_{ij} \in \mathrm{O}(n)$, the Lagrangian is invariant under orthogonal transformations. Before we consider the general case of arbitrary N, we look at the case $N = 2$ for which the potential is shown in Fig. 13.1. Without loss of generality, we choose the vacuum pointing in the direction of ϕ_1: Thus $v = \langle\phi_1\rangle = \sqrt{\mu^2/\lambda}$ and $\langle\phi_2\rangle = 0$. Shifting the field as in the discrete case gives

$$\mathscr{L} = \frac{\mu^4}{4\lambda} + \frac{1}{2}(\partial_\mu \boldsymbol{\xi})^2 - \frac{1}{2}(2\mu^2)\xi_1^2 + \mathscr{L}_{int} , \tag{13.9}$$

i.e. the two degrees of freedom of the field $\boldsymbol{\phi}$ split after SSB into one massive and one massless mode.

Since the mass matrix consists of the coefficients of the terms quadratic in the fields, the general procedure for the determination of physical masses is the following: Find first the minimum of the potential $V(\boldsymbol{\phi})$. Expand then the potential up to quadratic terms,

$$V(\boldsymbol{\phi}) = V(\boldsymbol{\phi}_0) + \frac{1}{2}(\phi - \phi_0)_i(\phi - \phi_0)_j \underbrace{\frac{\partial^2 V}{\partial\phi_i \partial\phi_j}}_{M_{ij}} + \ldots \tag{13.10}$$

The second derivatives form a symmetric matrix with elements $M_{ij} \geq 0$, because we evaluate the mass matrix by assumption at the minimum of V. Diagonalising M_{ij} gives as eigenvalues the squared masses of the fields. The eigenvectors of M_{ij} are called the mass eigenstates or physical states. Propagators and Green functions describe the evolution of fields with definite masses and should be therefore build up on these states. If the potential has $n > 0$ flat directions, the vacuum is degenerated and n massless modes appear.

Looking at Fig. 13.1 suggests to use polar instead of Cartesian coordinates in field space. In this way, the rotation symmetry of the potential and the periodicity of the flat direction is reflected in the variables describing the scalar fields. Introducing first the complex field $\phi = (\phi_1 + \mathrm{i}\phi_2)/\sqrt{2}$, the Lagrangian becomes

$$\mathscr{L} = \partial_\mu \phi^\dagger \partial^\mu \phi + \mu^2 \phi^\dagger \phi - \lambda(\phi^\dagger \phi)^2 . \tag{13.11}$$

Next we set

$$\phi(x) = \rho(x) \mathrm{e}^{\mathrm{i}\vartheta(x)} \tag{13.12}$$

and use $\partial_\mu \phi = [\partial_\mu \rho + \mathrm{i}\rho\partial_\mu\vartheta]\mathrm{e}^{\mathrm{i}\vartheta}$ to express the Lagrangian in the new variables,

$$\mathscr{L} = (\partial_\mu \rho)^2 + \rho^2(\partial_\mu \vartheta)^2 + \mu^2\rho^2 - \lambda\rho^4 . \tag{13.13}$$

Shifting finally again the fields as $\rho = v + \xi$ with $v = \sqrt{\mu^2/2\lambda}$, we find

$$\begin{aligned}\mathscr{L} = {} & \frac{\mu^4}{4\lambda} + \frac{\mu^2}{2\lambda}(\partial_\mu\vartheta)^2 + (\partial_\mu\xi)^2 - 2\mu^2\xi^2 - 2\mu\sqrt{2\lambda}\xi^3 - \lambda\xi^4 \\ & + \left[\sqrt{2\mu^2/\lambda}\xi + \xi^2\right](\partial_\mu\vartheta)^2 .\end{aligned} \tag{13.14}$$

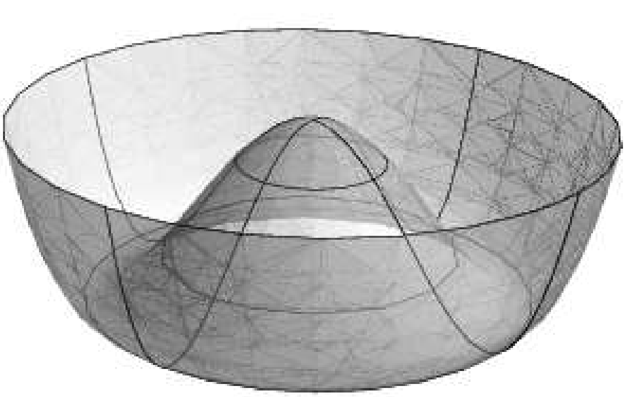


Figure 13.1.: Scalar potential with “Mexican hat” shape symmetric under O(2).

The phase ϑ which parametrises the flat direction of the potential $V(\vartheta,\xi)$ remained massless. This mode is called Goldstone (or Nambu-Goldstone) boson and has derivative couplings to the massive field ξ, given by the last term in Eq. (13.14). This is a general result, implying that static Goldstone bosons do not interact. Another general property of Goldstone bosons is that they carry the quantum number of the corresponding symmetry generator. They are therefore (pseudo-) scalar particles, if an internal symmetry is broken,

Let us now discuss briefly the case of general N for the Lagrangian (13.8). The lowest energy configuration is again a constant field. The potential is minimised for any set of fields $\boldsymbol{\phi}_0$ that satisfies $\boldsymbol{\phi}_0^2 = \mu^2/\lambda$. This equation only determines the length of the vector, but not its direction. It is convenient to choose a vacuum such that $\boldsymbol{\phi}_0$ points along one of the components of the field vector. Aligning $\boldsymbol{\phi}_0$ with its Nth component,

$$\boldsymbol{\phi}_0 = \left(0,\ldots,0,\sqrt{\mu^2/\lambda}\right) , \tag{13.15}$$

we now follow the same procedure as in the previous example. First we define a new set of fields, with the Nth field expanded around the vacuum

$$\boldsymbol{\phi}(x) = (\phi^k(x), v + \xi(x)) , \tag{13.16}$$

where k now runs from 1 to $N-1$. Then we insert this, and the value $v = \sqrt{\mu^2/\lambda}$ for the vacuum expectation value into the Lagrangian, and obtain

$$\begin{aligned}\mathscr{L} = &\frac{1}{2}(\partial_\mu \phi^k)^2 + \frac{1}{2}(\partial_\mu \xi)^2 - \frac{1}{2}(2\mu^2)\xi^2 + \frac{1}{4}\frac{\mu^4}{\lambda}\\ &- \sqrt{\lambda}\mu\xi^3 - \sqrt{\lambda}\mu\xi(\phi^k)^2 - \frac{\lambda}{2}\xi^2(\phi^k)^2 - \frac{\lambda}{4}\left[(\phi^k)^2\right]^2 - \frac{\lambda}{4}\xi^4 .\end{aligned} \tag{13.17}$$

This Lagrangian describes $N-1$ massless fields and a single massive field ξ, with cubic and quartic interactions. The O(N) symmetry is no longer apparent, leaving as symmetry group the subgroup O(N-1), which rotates the ϕ^k fields among themselves. This rotation describes movements along directions where the potential has a vanishing second derivative, while the massive field corresponds to oscillations in the radial direction of V.

Goldstone's theorem The observation that massless particles appear in theories with spontaneously broken *continuous* symmetries is a general result, known as Goldstone's theorem. The first example for such particles was suggested by Nambu in 1960: He showed that a massless quasi-particle appears in a magnetised solid, because the magnetic field breaks rotation invariance. Goldstone applied soon after that this idea to relativistic QFTs and showed that massless scalar elementary particles appear in theories with SSB. Since no massless scalar particles are known to exist, this theorem appeared to be a dead end for the application of SSB to particle physics. So our task is two-fold: First we should derive Goldstone's theorem and then we should find out how we can bypass the theorem applying it to our case of interest, gauge theories.

The theorem is obvious at the classical level: Consider a Lagrangian with a symmetry G and a vacuum state invariant under a subgroup H of G. For instance, choosing a Lagrangian invariant under $G = \mathrm{O}(3)$ and picking out a vacuum along ϕ_3, the subgroup $H = \mathrm{O}(2)$ of rotation around ϕ_3 keeps the vacuum invariant. Let us denote with $U(g)$ a representation of G acting on the fields $\boldsymbol{\phi}$ and with $U(h)$ a representation of H, respectively. Since we consider constant fields, derivative terms in the fields vanish and the potential V alone has to be symmetric under G, i.e.

$$V(U(g)\boldsymbol{\phi}) = V(\boldsymbol{\phi})\,. \tag{13.18}$$

Moreover, we know that the vacuum is kept invariant for all h, $\boldsymbol{\phi}_0' = U(h)\boldsymbol{\phi}_0$, but changes for some g, $\boldsymbol{\phi}_0' \neq U(g)\boldsymbol{\phi}_0$. Using the invariance of the potential and expanding $V(U(g)\boldsymbol{\phi}_0)$ for an infinitesimal group transformation gives

$$V(\boldsymbol{\phi}_0) = V(U(g)\boldsymbol{\phi}_0) = V(\boldsymbol{\phi}_0) + \frac{1}{2}\left.\frac{\partial^2 V}{\partial\phi_i\partial\phi_j}\right|_0 \delta\phi_i\delta\phi_j + \dots\,, \tag{13.19}$$

where $\delta\phi_i$ denotes the resulting variation of the field. Equation (13.19) implies that

$$M_{ij}\delta\phi_i\delta\phi_j = 0\,. \tag{13.20}$$

The variation $\delta\phi_i$ depends on whether the transformation belong to $U(h)$ or not: In the former case, the vacuum $\boldsymbol{\phi}_0$ is unchanged, $\delta\phi_i = 0$ and (13.20) is automatically satisfied. If on the other hand g does not belong to H, i.e. is a member of the left coset G/H, then $\delta\phi_i \neq 0$, implying that the mass matrix M_{ij} has a zero eigenvalue. It is now clear that the number of massless particles is simply determined by the dimensions of the two groups G and H: The number of Goldstone bosons equals $\dim(G) - \dim(H)$, or the dimension $\dim(G/H)$ of the left coset.

Quantum case The previous discussion was based on the classical potential. Thus we should address the question if this picture survives quantum corrections. Noether's theorem tells us that every continuous symmetry has associated to its generators g_i conserved charges Q_i. On the quantum level this means the operators Q_i commute with the Hamiltonian, $[H, Q_i] = 0$. Subtracting the vacuum energy, we have $H\,|0\rangle = 0$. If the vacuum is invariant under the symmetry Q, then $\exp(\mathrm{i}\vartheta Q)\,|0\rangle = |0\rangle$. For the infinitesimal form of the symmetry transformation, $\exp(\mathrm{i}\vartheta Q) \approx 1 + \mathrm{i}\vartheta Q$, we conclude that the charge annihilates the vacuum,

$$Q\,|0\rangle = 0\,. \tag{13.21}$$

Or, in simpler words, the vacuum has the charge 0.

Now we came to the case we are interested in, namely that the symmetry is spontaneously broken and thus $Q\,|0\rangle \neq 0$. We first determine the energy of the state $Q\,|0\rangle$. From

$$HQ\,|0\rangle = (HQ - \underbrace{QH)\,|0\rangle}_{H|0\rangle=0} = [H,Q]\,|0\rangle \;\; = 0\,, \tag{13.22}$$

we see that at least another state $Q\,|0\rangle$ exists which has as the vacuum $|0\rangle$ zero energy. We represent the charge operator as the volume integral of the time-like component of the corresponding current operator,

$$Q = \int \mathrm{d}^3x J^0(t,\boldsymbol{x})\,. \tag{13.23}$$

The state

$$|s\rangle = \int \mathrm{d}^3x\, \mathrm{e}^{\mathrm{i}\boldsymbol{p}\cdot\boldsymbol{x}} J^0(t,\boldsymbol{x})\,|0\rangle \to Q\,|0\rangle \qquad \mathrm{f}or \quad \boldsymbol{p}\to 0 \tag{13.24}$$

becomes in the zero-momentum limit equal to the state $Q\,|0\rangle$ we are searching for. Moreover, applying the momentum operator $\boldsymbol{P}$ on $|s\rangle$ gives (problem 13.5)

$$\boldsymbol{P}\,|s\rangle = \boldsymbol{p}\,|s\rangle\;. \tag{13.25}$$

Thus the SSB of the vacuum, $Q\,|0\rangle \neq 0$, implies excitations of the system with a frequency that vanishes in the limit of long wavelengths. In the relativistic case, Goldstone's theorem predicts massless states, while in the non-relativistic case relevant for solid states the theorem predicts collective excitations with zero energy gap.

13.2. Renormalisation of theories with SSB

When we went through the SSB of the scalar field, we saw that new ϕ^3 interactions were introduced. The question then arises, are new renormalisation constants needed when a symmetry is spontaneously broken? This would make these theories non-renormalisable. We can address this questions in two ways. One possibility is to repeat our analysis of the renormalisability of the scalar theory in section 11.4.2, but now for the broken case with a negative mass term. Then we would find that the ϕ^3 term becomes finite, renormalising fields, mass and coupling as in the unbroken case. This is not unexpected, because shifting the field, which is an integration variable in the generating functional, by $\phi \to \tilde{\phi} = \phi - v$ should not affect physics. On the other hand, such a shift reshuffles the splitting $\mathscr{L} = \mathscr{L}_0 + \mathscr{L}_{int}$ in our standard perturbative expansion in the coupling constant. To avoid this problem, we will use the quantum action employing a loop expansion. Additionally of being not affected by a shift of the fields, this formalism allows us to calculate the potential including all quantum corrections in the limit of constant fields.

Effective potential Let us start recalling the definition of the quantum action[1]

$$\Gamma[\phi_c] = W[J] - \int \mathrm{d}^4x' J(x')\phi_c(x') \equiv W[J] - \langle J\phi\rangle\,. \tag{13.26}$$

[1] We will suppress the subscript c on the classical field from now on and use brackets $\langle J\phi\rangle$ to indicate integration.

In general we will not be able to solve the quantum action. Studying SSB, we can however make use of a considerable simplification: The fields we are interested in are constant. Performing then a gradient expansion of the quantum action $\Gamma[\phi]$,

$$\Gamma[\phi] = \int \mathrm{d}^4x \left[-V_{eff}(\phi) + \frac{1}{2}F_2(\phi)(\partial_\mu\phi)^2 + \ldots \right], \tag{13.27}$$

only the zeroth order term $V_{eff}(\phi)$ of the expansion in $(\partial_\mu\phi)^2$ survives. If we now choose the source $J(x)$ to be constant, the field $\phi(x)$ has to be uniform too, $\phi(x) = \phi$, by translation invariance. Together this implies that

$$\Gamma = -\Omega V_{eff}\,, \quad \mathrm{a}nd \quad -J = \frac{\delta\Gamma[\phi]}{\delta\phi} = -\Omega\,\frac{\partial V_{eff}(\phi)}{\partial\phi}\,, \tag{13.28}$$

where Ω denotes the space-time volume. In the absence of external sources, $J = 0$, Eq. (13.28) simplifies to $V'_{eff}(\phi) = 0$. This is the quantum version of our old approach where we minimised the classical potential $V(\phi)$ in order to find the vacuum expectation value of ϕ. Therefore $V_{eff}(\phi)$ is called the effective potential, which includes all quantum corrections to the classical potential in the limit of zero gradients.

In order to proceed, we use that we know the classical potential and we assume that quantum fluctuations are small. Then we can perform a saddle-point expansion around the classical solution ϕ_0, given by the solution to $\Box\phi_0 + V'(\phi_0) = J(x)$. We split the field as $\phi = \phi_0 + \tilde{\phi}$ and approximate the path integral as

$$Z = \mathrm{e}^{\mathrm{i}W} \simeq \exp\{\mathrm{i}\,[S[\phi_0] + \langle J\phi_0\rangle]\} \int \mathcal{D}\tilde{\phi}\, \exp\left\{ \mathrm{i} \int \mathrm{d}^4x \frac{1}{2}\left[(\partial_\mu\tilde{\phi})^2 - V''(\phi_0)\tilde{\phi}^2 \right] \right\}. \tag{13.29}$$

The neglected terms are of order $\mathcal{O}(\hbar^2)$ and correspond thus to two- and higher loop contributions. The functional integral over $\tilde{\phi}$ is quadratic and is formally given by $\mathrm{D}et(\Box + V'')^{-1/2}$. Using then the identity $\ln \mathrm{D}et A = \mathrm{T}r \ln A$, we find

$$W = S[\phi_0] + \langle J\phi_0\rangle + \frac{\mathrm{i}}{2}\mathrm{T}r\ \ln[\Box + V''(\phi_0)] + \mathcal{O}(\hbar^2)\,. \tag{13.30}$$

The trace implies a summation over all discrete and an integration over the continuous quantum numbers. In the case of a scalar particle, no discrete quantum numbers exist and we have to integrate the matrix element $\langle x|\ln[\Box + V'']|x\rangle$ only over space-time. To find the eigenvalues of the operator, we insert a complete sets of plane waves,

$$\begin{aligned} \mathrm{T}r\ln[\Box + V''] &= \int \mathrm{d}^4x \langle x|\ln[\Box + V'']|x\rangle = \int \mathrm{d}^4x \frac{\mathrm{d}^4k}{(2\pi)^4} \langle x|\ln[\Box + V'']|k\rangle\langle k|x\rangle = \\ &= \int \mathrm{d}^4x \frac{\mathrm{d}^4k}{(2\pi)^4} \ln[-k^2 + V'']\langle x|k\rangle\langle k|x\rangle = \Omega \int \frac{\mathrm{d}^4k}{(2\pi)^4} \ln[-k^2 + V'']\,. \end{aligned} \tag{13.31}$$

Performing the Legendre transformation and using $S[\phi_0] = -\Omega V(\phi_0)$, we obtain for the effective potential $V_{eff}(\phi_0)$ including the first quantum corrections

$$V_{eff}(\phi_0) = V(\phi_0) - \frac{\mathrm{i}}{2} \int \frac{\mathrm{d}^4k}{(2\pi)^4} \ln\left[-k^2 + V''(\phi_0)\right] + \mathcal{O}(\hbar^2)\,. \tag{13.32}$$

As an example we can use the $\lambda\phi^4$ theory: From

$$V''(\phi_0) = \mu^2 + \frac{1}{2}\lambda\phi_0^2\,, \tag{13.33}$$

we see that $V''(\phi_0)$ can be interpreted as an effective mass, consisting of μ^2 and the contribution $\lambda\phi_0^2/2$ due to the constant background field ϕ_0. The total effective potential at order $\mathcal{O}(\hbar^2)$ consists of the classical potential $V(\phi)$, i.e. the classical energy density of a scalar field with vacuum expectation value ϕ, while the first quantum correction is given by the zero-points energies[2] of a scalar particle with effective mass $V''(\phi_0)$.

Not surprisingly, the effective potential is divergent and we have to introduce counter-terms that eliminate the divergent parts. Our effective potential is then

$$V_{eff}(\phi_0) = V(\phi) + \frac{1}{2}\int \frac{\mathrm{d}^4k_E}{(2\pi)^4} \ln\left(\frac{k_E^2 + V''(\phi_0)}{k_E^2}\right) + B\phi_0^2 + C\phi_0^4 + \mathcal{O}(\hbar^2)\,. \tag{13.34}$$

Here, we Wick rotated the integral to Euclidean space and subtracted an infinite constant in order to make the logarithm dimensionless. (Equivalently we could have added an additional constant counter-term A renormalising the vacuum energy density.) The integral can be evaluated in different regularisation schemes. Here we will expand the logarithm,

$$\ln\left(1 + \frac{V''}{k_E^2}\right) = \sum_{n=1}^{\infty} \frac{1}{n}\left(\frac{V''}{k_E^2}\right)^n\,, \tag{13.35}$$

and cutoff the integral at some large momentum Λ. The first two terms of the sum will depend on the cutoff, being proportional to Λ^2 and $\ln(\Lambda^2/V'')$, respectively. Performing the integral and neglecting terms that vanish for large Λ, we obtain

$$V_{eff}(\phi_0) = V(\phi_0) + \frac{\Lambda^2}{32\pi^2}V''(\phi_0) + \frac{V''(\phi_0)^2}{64\pi^2}\ln\left(\frac{V''(\phi_0)}{\Lambda^2}\right)\,. \tag{13.36}$$

Now we see that if we start out with a massless $\lambda\phi^4$ theory, our cutoff-dependent terms are

$$V'' = \frac{1}{2}\lambda\phi_0^2, \quad and \quad (V'')^2 = \frac{\lambda^2}{4}\phi_0^4, \tag{13.37}$$

which both can be absorbed into the counter-terms B and C by imposing appropriate renormalisation conditions.

Let us stress the important point in this result: The renormalisation of the $\lambda\phi^4$ theory using the effective potential approach is not affected by a shift of the field: We are free to use both signs of μ^2 and any value of the classical field ϕ_0 in Eq. (13.36). Independently of the sign of μ^2, we need only symmetric counter-terms, as a cubic term does not appear. We can rephrase this point as follows: If we renormalise before we shift the fields, we know that we obtain finite renormalised Green functions. But shifting the fields does not change the total Lagrangian. Thus the quantum action and the effective potential are unchanged too. Consequently the theory has to stay renormalisable after SSB.

Let us now discuss what happens with a non-renormalisable theory in the effective potential approach. Including e.g. a ϕ^6 term leads to $(V'')^2 \propto \phi^8$ which requires an additional counter-term $D\phi^8$, generating in turn even higher order terms and so forth. Thus in this case an

[2] Integrating $\mathrm{i}\Delta_F(0)$ w.r.t. m^2 reproduces the one-loop term.

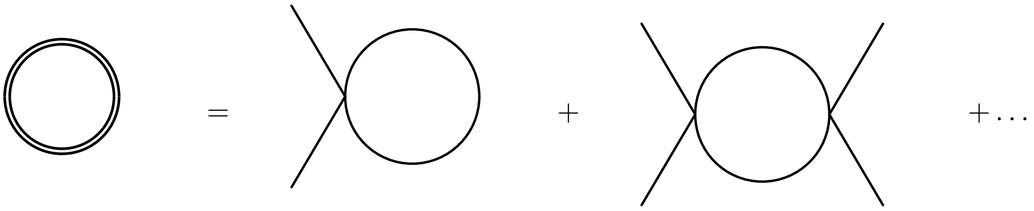


Figure 13.2.: Perturbative expansion of the one-loop effective potential $V_{eff}^{(1)}$ for the $\lambda\phi^4$ theory; all external legs have zero momentum.

infinite number of counter-terms is needed for the calculation of $V_{eff}^{(1)}$. The reason for this behaviour becomes clear, if we look again at the series expansion of the logarithm in the one loop contribution $V_{eff}^{(1)}$,

$$V_{eff}^{(1)} = \mathrm{i}\sum_{n=1}^{\infty}\int\frac{\mathrm{d}^4k}{(2\pi)^4}\frac{1}{2n}\left[\frac{V''(\phi_0)}{k^2}\right]^n . \tag{13.38}$$

This contribution is an infinite sum of single loops with progressively more external legs with zero-momentum attached, see Fig. 13.2 for the case of $V''(\phi_0) = \frac{1}{2}\lambda\phi_0^2$. (We added the factor i, because we returned to Minkowski space; the symmetry factor $2n$ appearing automatically in this approach accounts for the symmetry of a graph with n vertices under rotations and reflection.) As we saw, the superficial degree of divergence increases with the number of external particles for a $\lambda\phi^n$ theory and $n > 4$. Hence every single diagram in the infinite sum contained in $V_{eff}^{(1)}$ diverges for $n > 4$ and requires a counter-term of higher order. As discussed in section 12.5, we should treat non-renormalisable theories as effective theories only valid below a physical cutoff Λ. Calculating then loop corrections as in $V_{eff}^{(1)}$, we have to us a mass independent renormalisation scheme instead of a dimensionfull cutoff. An alternative is the functional RGE approach, for which we give next a brief example:

Remark 13.1: A RGE flow equation for the effective potential:
Let us define V_k as the effective potential cutting off the loop integrals at the scale k. We can repeat the saddle point expansion, splitting ϕ as $\phi = \phi_0 + \psi$, where now ϕ_0 contains the modes $k < p$ and ψ the modes $p \leq k \leq \Lambda$. For simplicity, we assume that the minimum of the potential is at zero and treat the slow field ϕ_0 as uniform. Integrating out the field ψ, we obtain then

$$V_k(\phi_0) = V_\Lambda(\phi_0) + \frac{\hbar}{2}\int_k^\Lambda \frac{\mathrm{d}^4k}{(2\pi)^4}\ln\left(\frac{k_E^2 + V''(\phi_0)}{k_E^2 + V''(0)}\right) . \tag{13.39}$$

Next we differentiate w.r.t. k and obtain

$$k\frac{\partial}{\partial k}V_k(\phi_0) = -\frac{\hbar k^4}{16\pi^2}\ln\left(\frac{k_E^2 + V''(\phi_0)}{k_E^2 + V''(0)}\right) , \tag{13.40}$$

i.e. a differential equation describing how the effective potential changes integrating out UV modes with momentum k. Similar equations, called RGE flow equations, can be derived also for the quantum action. If we consider now an asymptotically free theory, then for sufficiently high scales, $\Gamma[\phi] \simeq S[\phi]$. Then also V_k can be approximated by the classical potential, fixing our initial condition.

Finally, we note that the result for a single, real scalar generalises as

$$V_{\mathrm{e}ff}^{(1)} = (-1)^s g_i V_{\mathrm{e}ff,scalar}^{(1)} \tag{13.41}$$

to a particle with spin s and g_i internal degrees of freedom. This should come as now surprise, since the one-loop expression of the effective potential sums up the logarithm of the zero-point energies.

Another proof of the Goldstone theorem With the help of the effective potential we can give another simple proof of the Goldstone theorem. We know that the zero of the inverse propagator determines the mass of a particle. From Eq. (12.14), the exact inverse propagator in momentum space for a set of scalar fields is given by

$$\boldsymbol{\Delta}_{ij}^{-1}(p^2) = \int \mathrm{d}^4x\, \mathrm{e}^{-\mathrm{i}p(x-x')} \frac{\delta^2\Gamma}{\delta\phi_i(x)\delta\phi_j(x')} \,. \tag{13.42}$$

Massless particles correspond to zero eigenvalues of this matrix equation for $p^2 = m^2$. If we set $\boldsymbol{p} = 0$, the fields are constant. But differentiating the quantum action w.r.t. to constant fields is equivalent to differentiating simply the effective potential,

$$\frac{\partial^2 V_{\mathrm{e}ff}}{\partial\phi_i(x)\partial\phi_j(x')} = 0\,. \tag{13.43}$$

Thus our previous analysis of Goldstone's theorem using the classical potential holds also in the quantum case, if we replace the classical by the effective potential.

Coleman-Weinberg Problem We can use the effective potential to investigate if quantum fluctuations can trigger SSB in an initially massless theory. Rewriting the effective potential (and going back to our normalisation $\lambda\phi/4!$) we have

$$V_{\mathrm{e}ff}(\phi) = \left[\frac{\Lambda^2}{64\pi^2}\lambda + B\right]\phi^2 + \left[\frac{\lambda}{4!} + \frac{\lambda^2}{(16\pi)^2}\ln\frac{\phi^2}{\Lambda^2} + C\right]\phi^4\,. \tag{13.44}$$

Now we impose the renormalising conditions, first

$$\left.\frac{\mathrm{d}^2 V_{\mathrm{e}ff}}{\mathrm{d}\phi^2}\right|_{\phi=0} = 0\,, \tag{13.45}$$

which implies that

$$B = -\frac{\lambda\Lambda^2}{64\pi^2}\,. \tag{13.46}$$

Renormalising the coupling constant, we have to pick a different point than $\phi = 0$, because the logarithm is ill-defined there. This means that we have to introduce a scale μ. Taking the fourth derivative and ignoring terms that are independent of ϕ, we find

$$\left.\frac{\mathrm{d}^4 V_{\mathrm{e}ff}}{\mathrm{d}\phi^4}\right|_{\phi=\mu} = \lambda = 24\frac{\lambda^2}{(16\pi)^2}\ln\frac{\mu^2}{\Lambda^2}\,. \tag{13.47}$$

We can convince ourselves that this expression gives the correct beta function,

$$\beta(\mu) = \mu\frac{\partial\lambda}{\partial\mu} = \frac{3}{16\pi^2}\lambda^2 + \mathcal{O}(\lambda^3)\,. \tag{13.48}$$

Using the complete expression for Eq. (13.47), we can determine C and obtain for the renormalised effective potential (problem 13.6)

$$V_{eff}(\phi) = \frac{\lambda(\mu)}{4!}\phi^4 + \frac{\lambda^2(\mu)}{(16\pi)^2}\phi^4\left[\ln\frac{\phi^2}{\mu^2} - \frac{25}{6}\right] + \mathcal{O}(\lambda^3)\,. \tag{13.49}$$

This potential has two minima outside of the origin, so it seems that SSB does indeed happen. These minima lie however outside the expected range of validity of the one loop approximation: Rewriting the potential as $V_{eff}(\phi) = \lambda\phi^4/4!(1 + a\lambda\ln(\phi^2/\mu^2) + \ldots$ suggest that we can trust the one-loop approximation only as long as $(3/32\pi^2)\lambda\ln(\phi^2/\mu^2) \ll 1$.

13.3. Abelian Higgs model

After we have shown that the renormalisability is not affected by SSB, we now try to apply this idea to a the case of a gauge symmetry. First of all, because we aim to explain the masses of the W and Z bosons as consequence of SSB. Secondly, we saw that SSB of global symmetries leads to massless scalars which are however not observed. As SSB cannot change the number of physical degrees of freedom, we hope that each of the two diseases is the cure of the other: The Goldstone bosons which would remain massless in a global symmetry hopefully disappear becoming the required additional longitudinal degrees of freedom of massive gauge bosons in case of a spontaneously broken gauge symmetry.

The Abelian Higgs model, which is the simplest example for this mechanism, is obtained by gauging a complex scalar field theory. Introducing in the Lagrangian (13.11) the covariant derivative

$$\partial_\mu \to D_\mu = \partial_\mu + \mathrm{i}eA_\mu \tag{13.50}$$

and adding the free Lagrangian of an U(1) gauge field gives

$$\mathscr{L} = -\frac{1}{4}F_{\mu\nu}F^{\mu\nu} + (D_\mu\phi)^\dagger(D^\mu\phi) + \mu^2\phi^\dagger\phi - \lambda(\phi^\dagger\phi)^2\,. \tag{13.51}$$

The symmetry breaking and Higgs mechanism is best discussed changing to polar coordinates in field-space, $\phi = \rho\exp\{\mathrm{i}\vartheta\}$. Then we insert

$$D_\mu\phi = \left[\partial_\mu\rho + \mathrm{i}\rho(\partial_\mu\vartheta + eA_\mu)\right]e^{\mathrm{i}\vartheta} \tag{13.52}$$

into the Lagrangian, obtaining

$$\mathscr{L} = -\frac{1}{4}F_{\mu\nu}F^{\mu\nu} + \rho^2(\partial_\mu\vartheta + eA_\mu)^2 + (\partial_\mu\rho)^2 + \mu^2\rho^2 - \lambda\rho^4\,. \tag{13.53}$$

The only difference to the ungauged model is the appearance of the gauge field in the prospective mass term $\rho^2(\partial_\mu\vartheta + eA_\mu)^2$. This allows us to eliminate the angular mode ϑ which shows up nowhere else by performing a gauge transformation on the field A_μ: The action of a $U(1)$ gauge transformation $A_\mu \to A'_\mu = A_\mu - \partial_\mu\Lambda$ on the original field ϕ is just a phase shift, hence ρ is unchanged and ϑ is shifted by a constant, $\vartheta \to \vartheta' = \vartheta + e\Lambda$. This means that if we consider the gauge invariant combination

$$B_\mu = A_\mu + \frac{1}{e}\partial_\mu\vartheta \tag{13.54}$$

as new variable, we eliminate ϑ completely, as $F_{\mu\nu}(A_\mu) = F_{\mu\nu}(B_\mu)$ is gauge invariant,

$$\mathscr{L} = -\frac{1}{4}F_{\mu\nu}F^{\mu\nu} + e^2\rho^2(B_\mu)^2 + (\partial_\mu\rho)^2 + \mu^2\rho^2 - \lambda\rho^4 \,. \tag{13.55}$$

It is now evident that the Goldstone mode ϑ has disappeared. Eliminating the field ρ in favour of fluctuations χ around the vacuum $v = \sqrt{\mu^2/\lambda}$, i.e. shifting as usually the field as

$$\rho = \frac{1}{\sqrt{2}}(v + \chi) \,, \tag{13.56}$$

we find as new Lagrangian

$$\begin{aligned} \mathscr{L} = &-\frac{1}{4}F_{\mu\nu}F^{\mu\nu} + \frac{1}{2}M^2(B_\mu)^2 + e^2v\chi(B_\mu)^2 + \frac{1}{2}e^2\chi^2(B_\mu)^2 \\ &+ \frac{\mu^4}{4\lambda} + \frac{1}{2}(\partial_\mu\chi)^2 - \frac{1}{2}(\sqrt{2}\mu)^2\chi^2 - \sqrt{\lambda}\mu\chi^3 - \frac{\lambda}{4}\chi^4 \,. \end{aligned} \tag{13.57}$$

As in the ungauged model we obtain a χ^3 self-interaction and a contribution to the vacuum energy density. The gauge field B_μ acquired the mass $M = ev$, therefore having now three spin degrees of freedom. The additional longitudinal one has been delivered by the Goldstone boson which in turn disappeared: The gauge field has "eaten" the Goldstone boson. We also see that the number of degrees of freedom before SSB $(2 + 2)$ matches the number afterwards $(3 + 1)$. The phenomenon that breaking spontaneously a gauge symmetry does not lead to massless Goldstone bosons because they become the longitudinal degree of freedom of massive gauge bosons is called the Higgs effect.

The gauge transformation we used to eliminate the ϑ field corresponds to the Higgs model in the unitary gauge, where only physical particles appear in the Lagrangian. The massive gauge boson is described by the Procca Lagrangian and we know that the resulting propagator becomes constant for large momenta. Hence, this gauge is convenient for illustrating the concept of the Higgs mechanism, but not suited for loop calculations. For several reasons we expect that the renormalisability of this model is only hidden in the unitary gauge: The model before shifting the fields is renormalisable, and our discussion of SSB using the effective potential has taught us that such a shift has no impact. Moreover, we should be able to neglect v in a scattering process like $B^\mu + B^\mu \to B^\mu + B^\mu$ for $s \gg v^2$. Thus the broken theory should have the same UV behaviour as the unbroken one.

We should therefore explore alternative gauges of the same model. Avoiding the unitary gauge, we re-start using Cartesian fields $\phi = (\phi_1 + \mathrm{i}\phi_2)/\sqrt{2}$ for the complex scalar. Then the Lagrangian is

$$\begin{aligned} \mathscr{L} = &-\frac{1}{4}F_{\mu\nu}F^{\mu\nu} + \frac{1}{2}\left[(\partial_\mu\phi_1 - eA_\mu\phi_2)^2 + (\partial_\mu\phi_2 + eA_\mu\phi_1)^2\right] \\ &+ \frac{\mu^2}{2}(\phi_1^2 + \phi_2^2) - \frac{\lambda}{4}(\phi_1^2 + \phi_2^2)^2 \,. \end{aligned} \tag{13.58}$$

Performing the shift due to the SSB, $\phi_1 = v + \tilde\phi_1$ and $\phi_2 = \tilde\phi_2$, the Lagrangian becomes

$$\begin{aligned} \mathscr{L} = &-\frac{1}{4}F_{\mu\nu}F^{\mu\nu} + \frac{1}{2}M^2A_\mu^2 + evA^\mu\partial_\mu\tilde\phi_2 \\ &+ \frac{1}{2}\left[(\partial_\mu\tilde\phi_1)^2 - 2\mu^2\tilde\phi_1^2\right] + \frac{1}{2}(\partial_\mu\tilde\phi_2)^2 + \dots \,, \end{aligned} \tag{13.59}$$

where we have omitted interaction and vacuum terms not relevant to the discussion. As we see, the Goldstone boson $\tilde{\phi}_2$ does not disappear and it couples to the gauge field A_μ. On the other hand, the mass spectrum of the physical particles is the same as in the unitary gauge. The degrees of freedom before and after breaking the symmetry do not match, hence there is an unphysical degree of freedom in the theory, namely that corresponding to $\tilde{\phi}_2$.

Gauge fixing and gauge boson propagator In order to make the generating functional $Z[J^\mu, J, J^*]$ of the abelian Higgs model well-defined, we have to remove the gauge freedom of the classical Lagrangian. Using the Faddeev-Popov trick to achieve this implies to add a gauge-fixing and a Faddeev-Popov ghost term to the classical Lagrangian,

$$\mathscr{L}_{eff} = \mathscr{L}_{\mathrm{cl}} + \mathscr{L}_{\mathrm{g}f} + \mathscr{L}_{\mathrm{FP}} = \mathscr{L}_{\mathrm{cl}} - \frac{1}{2\xi}G^2 + \bar{c}\frac{\partial G}{\partial \vartheta}c\,. \tag{13.60}$$

Here $G(A^\mu, \phi) = 0$ is a suitable gauge condition, ϑ is the generator of the gauge symmetry and $c, \bar{c}$ are Graßmannian ghost fields.

In the unbroken abelian case we used as gauge condition $G = \partial_\mu A^\mu$. With the gauge transformation $A^\mu \to A^\mu - \partial^\mu \vartheta$ the ghost term becomes simply $\mathscr{L}_{\mathrm{FP}} = \bar{c}(-\Box)c$. Thus the ghost fields completely decouple from any physical particles, and the ghost term can be absorbed in the normalisation. In the present case of a theory with SSB, we want to use the Faddeev-Popov term to cancel the mixed $A^\mu \partial_\mu \phi_2$ term. Therefore we include the Goldstone boson ϕ_2 in the gauge condition,

$$G = \partial_\mu A^\mu - \xi e v \phi_2 = 0\,, \tag{13.61}$$

what defines the R_ξ gauge. From $\phi_2 = \partial_\mu A^\mu/(\xi e v)$ we see that the unitary gauge corresponds to $\xi \to \infty$, while we come back to $G = \partial_\mu A^\mu$ for $v \to 0$. We calculate first G^2,

$$G^2 = (\partial_\mu A^\mu)(\partial_\nu A^\nu) - 2\xi e v \phi_2 \partial_\mu A^\mu + \xi^2 e^2 v^2 \phi_2^2\,, \tag{13.62}$$

integrate by parts the cross term and insert the result into $\mathscr{L}_{\mathrm{g}f}$,

$$\mathscr{L}_{\mathrm{g}f} = -\frac{1}{2\xi}G^2 = -\frac{1}{2\xi}(\partial_\mu A^\mu)^2 - evA_\mu \partial^\mu \phi_2 - \frac{1}{2}\xi(ev)^2\phi_2^2\,. \tag{13.63}$$

Now we see that the second term cancels the unwanted mixed term in $\mathscr{L}_{\mathrm{cl}}$, while a ξ dependent mass term ξM^2 for ϕ_2 appeared.

If we write out the terms in $\mathscr{L}_{eff}$ quadratic in A_μ and ϕ_2,

$$\mathscr{L}_{eff,2} = -\frac{1}{4}F_{\mu\nu}F^{\mu\nu} + \frac{1}{2}M^2 A_\mu^2 - \frac{1}{2\xi}(\partial_\mu A^\mu)^2 + \frac{1}{2}(\partial_\mu \phi_2)^2 - \frac{1}{2}\xi M^2 \phi_2^2\,, \tag{13.64}$$

we can find the boson propagator. Using the antisymmetry of $F_{\mu\nu}$ and a partial integration, we transform $F^2/4$ into standard form, $A_\mu(\eta^{\mu\nu}\Box - \partial^\mu\partial^\nu)A_\nu/2$. The part of the Lagrangian quadratic in A_μ then reads

$$\mathscr{L}_A = \frac{1}{2}A_\mu\left[\eta^{\mu\nu}\Box - \partial^\mu\partial^\nu\right]A_\nu + \frac{1}{2}A_\mu\eta^{\mu\nu}M^2 A_\nu + \frac{1}{2\xi}A_\mu\partial^\mu\partial^\nu A_\nu \tag{13.65a}$$

$$= \frac{1}{2}A_\mu\left[\eta^{\mu\nu}(\Box + M^2) - (1-\xi^{-1})\partial^\mu\partial^\nu\right]A_\nu\,. \tag{13.65b}$$

To find the propagator we want to invert the term in the bracket, which we denote by $P^{\mu\nu}$. If we go to momentum space, then

$$P^{\mu\nu}(k^2) = -(k^2 - M^2)\eta^{\mu\nu} + (1 - \xi^{-1})k^\mu k^\nu \,. \tag{13.66}$$

Next we split $P^{\mu\nu}(k^2)$ into a transverse and a longitudinal part by factoring out terms proportional to $P_T^{\mu\nu} = \eta^{\mu\nu} - k^\mu k^\nu/k^2$,

$$\begin{aligned} P^{\mu\nu} &= -(k^2 - M^2)\left(P_T^{\mu\nu} + \frac{k^\mu k^\nu}{k^2}\right) + (1 - \xi^{-1})k^\mu k^\nu && (13.67a) \\ &= -(k^2 - M^2)P_T^{\mu\nu} - \xi^{-1}(k^2 - \xi M^2)P_L^{\mu\nu} \,, && (13.67b) \end{aligned}$$

with the longitudinal part given by $P_L^{\mu\nu} = k^\mu k^\nu/k^2$. We can invert the two parts separately and obtain

$$\begin{aligned} \mathrm{i}D_F^{\mu\nu}(k^2) &= \frac{-\mathrm{i}P_T^{\mu\nu}}{k^2 - M^2 + \mathrm{i}\varepsilon} + \frac{-\mathrm{i}\xi P_L^{\mu\nu}}{k^2 - \xi M^2 + \mathrm{i}\varepsilon} && (13.68a) \\ &= \frac{-\mathrm{i}}{k^2 - M^2 + \mathrm{i}\varepsilon}\left[\eta^{\mu\nu} - (1 - \xi)\frac{k^\mu k^\nu}{k^2 - \xi M^2 + \mathrm{i}\varepsilon}\right] . && (13.68b) \end{aligned}$$

The transverse part propagates with mass M^2, while the longitudinal part propagates with mass ξM^2. The limit $\xi \to \infty$ corresponds again to the unitary gauge and $\xi = 1$ corresponds to the easier Feynman-'t Hooft gauge. For finite ξ, the propagator is proportional to $1/k^2$ as in the massless case. The Goldstone boson ϕ_2 has the usual propagator of a scalar particle, however with gauge-dependent mass ξM^2. As usual in a covariant gauge, the unphysical gauge-dependent modes have to be cancelled by ghosts which we discuss next.

Ghosts Using the Faddeev-Popov ansatz introduces ghosts field through the term $\mathscr{L}_{\mathrm{F}P} = \bar{c}\,(\delta G/\delta\vartheta)\,c$ into the Lagrangian. To calculate $\delta G/\delta\vartheta$, we have to find out how the gauge fixing condition G changes under an infinitesimal gauge transformation. Looking first at the change of the complex field,

$$\phi \to \tilde{\phi} = \phi + \mathrm{i}e\vartheta\phi = \phi + \mathrm{i}e\vartheta\frac{1}{\sqrt{2}}(v + \phi_1 + \mathrm{i}\phi_2) \,, \tag{13.69}$$

we see that the fields ϕ_1 and ϕ_2 are mixed under the gauge transformation.

$$\begin{aligned} A_\mu &\to \tilde{A}_\mu = A_\mu - \partial_\mu\vartheta && (13.70a) \\ \phi_1 &\to \tilde{\phi}_1 = \phi_1 - e\vartheta\phi_2 && (13.70b) \\ \phi_2 &\to \tilde{\phi}_2 = \phi_2 + e\vartheta(v + \phi_1) \,. && (13.70c) \end{aligned}$$

Inserting this into the gauge fixing condition (13.61) and differentiating with respect to the generator, we obtain

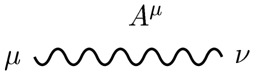
$$\frac{\delta G}{\delta\vartheta} = \frac{\delta}{\delta\vartheta}\left(\partial_\mu\tilde{A}^\mu - \xi ev\tilde{\phi}_2\right) = -\Box - \xi e^2 v(v + \phi_1) \,. \tag{13.71}$$

Thus after spontaneous symmetry breaking the ghost particles receive a ξ-dependent mass and interact with the Higgs field ϕ_1. To see this explicitly we insert $\delta G/\delta\vartheta$ into the ghost Lagrangian,

$$\mathscr{L}_{\mathrm{F}P} = -\bar{c}\left[\Box + \xi e^2 v(v + \phi_1)\right]c = (\partial^\mu\bar{c})(\partial_\mu c) - \xi M^2\bar{c}c - \xi e^2 v\phi_1\bar{c}c \,. \tag{13.72}$$

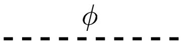
The second term on the RHS corresponds to the mass $\xi e^2 v^2 = \xi M^2$ for the ghost field, while the third one describes the ghost-ghost-Higgs interaction.

In summary, we have the following propagators in the R_ξ gauge, where we denote with h the physical Higgs with mass $M = ev$, with ϕ the Goldstone boson and with c the ghost:

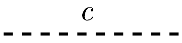


$$\frac{\mathrm{i}}{k^2 - M^2 + \mathrm{i}\varepsilon}\left[-\eta_{\mu\nu} + (1-\xi)\frac{k_\mu k_\nu}{k^2 - \xi M^2}\right] \tag{13.73}$$

$$\frac{\mathrm{i}}{p^2 - M^2 + \mathrm{i}\varepsilon} \tag{13.74}$$



$$\frac{\mathrm{i}}{p^2 - \xi M^2 + \mathrm{i}\varepsilon} \tag{13.75}$$



$$\frac{\mathrm{i}}{p^2 - \xi M^2 + \mathrm{i}\varepsilon} \tag{13.76}$$

Before closing this chapter, we should answer why the Goldstone theorem does not apply to the case of the Higgs model. The characteristic property of gauge theories that no manifestly covariant gauge exists which eliminates all gauge freedom is also responsible for the failure of the Goldstone theorem: In the first version of our proof, we may either choose a gauge as the Coulomb gauge. Then only physical degrees of freedom of the photon propagate, but the potential $A^0(x)$ drops only as $1/|\boldsymbol{x}|$ and the charge Q defined in (13.23) becomes ill-defined. Alternatively, we can use a covariant gauge as the Lorentz gauge. Then the charge is well-defined, but unphysical scalar and longitudinal photons exist. The Goldstone theorem does apply, but the massless Goldstone bosons do not couple to physical modes. In the second version of our proof, the effective potential for the scalar and for the gauge sector do not decouple and mix by the same reason after SSB. This invalidates our analysis including only scalar fields.

Summary

Examining spontaneous symmetry breaking of internal symmetries, we found three qualitatively different types of behaviours: For a broken global continuous symmetry, Goldstone's theorem predicts the existence of massless scalars. In the case of broken approximate symmetries, this can explain the existence of light scalar particles—an example are pions. The case of broken global continuous symmetry which are exact seems to be not realised in nature, since no massless scalar particles are observed. If we gauge the broken symmetry, the would-be massless Goldstone bosons become the longitudinal degrees of freedom required for massive spin-1 bosons. Finally, neither Noether's nor Goldstone's theorems apply to the case of discrete symmetries; therefore the breaking of discrete symmetries does not change the mass spectrum of the theory.

The effective potential is a convenient tool to study the renormalisability of spontaneously broken theories: This approach allows the calculation of all quantum corrections to the classical potential in the limit of constant fields and is invariant under a shift of fields. Thereby we could establish that renormalisability is not affected by SSB.

Further reading

Our discussion of the effective potential is based on the 1966 Erice lecture “Secret Symmetry” of Coleman198802.

Problems

13.1 Contribution to the vacuum energy density from SSB.

Calculate the difference in the vacuum energy density before and after SSB in the SM using $v = 256\,\text{GeV}$ and $m_h^2 = 2\mu^2 = (125)^2\,\text{GeV}^2$. Compare this to the observed value of the cosmological constant.

13.2 Scalar Lagrangian after SSB.

Derive Eq. (13.9) and write down the explicit form of $\mathscr{L}_{int}$.

13.3 Quantum corrections to $\langle\phi\rangle$.

We implicitly assumed that quantum corrections are small enough that the field stays at the chosen classical minimum. Calculate $\langle\phi(0)^2\rangle$ for d space-time dimensions and show that this assumption is violated for $d \leq 2$.

13.4 Instability of $\langle\phi\rangle$.
Calculate the imaginary part of the self-energy for a scalar field with the Lagrangian (13.1), i.e. with a negative squared mass $\mu^2 < 0$. Discuss the physical interpretation.

13.5 Goldstone mode as zero mode.
Show that the state $|s\rangle$ defined in Eq. (13.24) has zero energy for $\boldsymbol{k} \to 0$.

13.6 Coleman-Weinberg problem.
Derive Eq. (13.49), find the minima of the potential and discuss the validity of the one-loop approximation.

13.7 Effective potential in DR.
Repeat the calculation of the effective potential using DR.

14. Thermal field theory

The Green functions we have considered so far were defined as the expectation value of field operators in a pure state, the vacuum $|0\rangle$ in the absence of real particles. Out of these Green functions, we could built up our quantities of prime interest, decay or scattering amplitudes for $1 \to n$ and $2 \to n$ particles via the LSZ formalism. In this chapter, we discuss how Green functions should be calculated for a system which is not in its ground state but described by a density matrix ρ. Examples for such systems are the early Universe or the dense, hot interior of a star. The simplest and at the same time most important cases of thermal systems are those in equilibrium. The appendix of this chapter collects a few basic formulas from statistical physics we will need later.

14.1. Overview

In equilibrium statistical physics, the partition function Z is of central importance as all thermodynamic quantities can be derived from it. In the grand canonical ensemble (where both particles and energy may be exchanged between the system and the reservoir) the partition function takes the form

$$Z(V,T,\mu_1,\mu_2,\dots) = \sum_n \langle n| \, \mathrm{e}^{-\beta H-\mu_i N_i} \, |n\rangle = \mathrm{e}^{-\beta\Omega} \,, \tag{14.1}$$

where $\beta \equiv 1/T$ is the inverse temperature of the system, n denotes a complete set of quantum numbers, μ_i the chemical potential and N_i the number of particles of type i. The Landau or grand canonical free energy

$$\Omega(V,T,\mu_i) = -T\ln Z = U - TS - \mu_i N_i = -PV \tag{14.2}$$

connects the microscopic partition function to thermodynamics: While the partition function is closely related to the generating functional of the corresponding field theory in Euclidean space, we can derive from Ω all relevant thermodynamical quantities. For instance, we obtain the pressure P from Ω as $P = \partial\Omega/\partial V|_{T,\mu_i}$. In addition, the expectation value of any observable O is given as

$$\langle O\rangle = Z^{-1}\mathrm{T}r\left[\mathrm{e}^{-\beta H-\mu_i N_i}\, O\right] . \tag{14.3}$$

In the following, we will always set $\mu_i = 0$. Then the partition function $Z(N,V,T)$ determines the free (Helmholtz) energy F as $F = -T\ln Z$.

Calculational approaches Two main approaches to calculations are used in thermal field theory:

- In the real-time formalism, one applies the formula (14.3) valid for any observable directly to Green functions, that is one evaluates

$$\begin{aligned} G(x_1,\ldots,x_n) &= \langle T(\phi(x_1)\cdots\phi(x_n))\rangle \\ &= Z^{-1}\mathrm{Tr}\left[\mathrm{e}^{-\beta H}\,T(\phi(x_1)\cdots\phi(x_n))\right]. \end{aligned} \tag{14.4}$$

  The main advantage of this method is that it can be extended to the non-equilibrium case. In particular, one can investigate the time evolution of a system towards thermal equilibrium. The proper definition of the propagators becomes, however, more involved than in the vacuum.

- In the imaginary time formalism, we perform a Wick rotation from Minkowski to Euclidean space, $t \to t_E = \mathrm{i}t$, so that the transition amplitude from an initial state, $|q(t_i)\rangle$, to a final state, $|q(t_f)\rangle$, is given by

$$\langle q(t_f)|\,\mathrm{e}^{-(t_f-t_i)H}\,|q(t_i)\rangle = \int_{q(t_i)}^{q(t_f)} \mathcal{D}q\,\mathrm{e}^{-S}\,, \tag{14.5}$$

  where S is now the Euclidean action. If we set the evolution time, $t_f - t_i$, equal to the inverse temperature β and integrate over all periodic paths $q(t_f) = q(t_i+\beta)$, we obtain

$$Z = \sum_q \langle q|\,\mathrm{e}^{-\beta H}\,|q\rangle = \int_{q(t)}^{q(t+\beta)} \mathcal{D}q\,\mathrm{e}^{-S}\,. \tag{14.6}$$

  We now see that we have formally connected the path integral formulation of quantum mechanics (in Euclidean space) to the partition function of statistical mechanics. In contrast to field theories at zero temperature, the partition function in the Euclidean and the resulting Euclidean Green functions are not merely a mathematical tool but our main objects of interest. Since the Euclidean Green functions depend on temperature instead of time, we are not able to describe time-dependent phenomena in this approach.

Thermal Green functions The trace in the partition function of statistical physics implies that we have to sum over configurations connecting the same physical state at t and $t+\beta$. In the path integral corresponding to statistical mechanics, the periodicity condition $q(t) = q(t+\beta)$ for the real coordinate q is clearly the only possible choice. In contrast, fields may only be observable through bilinear quantities, as e.g. $\bar\psi\Gamma\psi$ for a fermion field. This raises the question, if we should require periodic or anti-periodic boundary conditions.

We start by considering thermal Green functions G_β for a free scalar field. We split the Feynman propagator into two pieces, setting $G^+(x,x') = \langle\phi(x)\phi(x')\rangle$ for $t > t'$ and $G^-(x,x') = \langle\phi(x')\phi(x)\rangle$ for $t < t'$. From the Heisenberg equation for the field operator,

$$\phi(t,\boldsymbol{x}) = \mathrm{e}^{\mathrm{i}Ht}\phi(0,\boldsymbol{x})\mathrm{e}^{-\mathrm{i}Ht}\,, \tag{14.7}$$

we find inserting $1 = \mathrm{e}^{\beta H}\mathrm{e}^{-\beta H}$ into the definition (14.3) and using then (14.7),

$$G_\beta^+(t',\boldsymbol{x}';t,\boldsymbol{x}) = \mathrm{Tr}\,[\mathrm{e}^{-\beta H}\phi(t',\boldsymbol{x}')\phi(t,\boldsymbol{x})]/Z \tag{14.8a}$$

$$= \mathrm{Tr}\,[\mathrm{e}^{-\beta H}\phi(t',\boldsymbol{x}')\mathrm{e}^{\beta H}\mathrm{e}^{-\beta H}\phi(t,\boldsymbol{x})]/Z \tag{14.8b}$$

$$= \mathrm{Tr}\,[\phi(t'+\mathrm{i}\beta,\boldsymbol{x}')\mathrm{e}^{-\beta H}\phi(t,\boldsymbol{x})]/Z \tag{14.8c}$$

$$= \mathrm{Tr}\,[\mathrm{e}^{-\beta H}\phi(t,\boldsymbol{x})\phi(t'+\mathrm{i}\beta,\boldsymbol{x}')]/Z = G_\beta^-(t'+\mathrm{i}\beta,\boldsymbol{x}';t,\boldsymbol{x})\,. \tag{14.8d}$$

Hence Green functions satisfy $G^{\pm}_{\beta}(t',\boldsymbol{x}';t,\boldsymbol{x}) = G^{\mp}_{\beta}(t'+\mathrm{i}\beta,\boldsymbol{x}';t,\boldsymbol{x})$. Repeating the procedure, we obtain $G^{\pm}_{\beta}(t',\boldsymbol{x}';t,\boldsymbol{x}) = G^{\mp}_{\beta}(t'+\mathrm{i}\beta,\boldsymbol{x}';t,\boldsymbol{x}) = G^{\pm}_{\beta}(t'+2\mathrm{i}\beta,\boldsymbol{x}';t,\boldsymbol{x})$, implying periodic boundary condition for the thermal propagators of bosonic fields,

$$G_{\beta}(t',\boldsymbol{x}';t,\boldsymbol{x}) = G_{\beta}(t'+2\mathrm{i}\beta,\boldsymbol{x}';t,\boldsymbol{x})\,. \tag{14.9}$$

We define now the Matsubara propagator $G_{\beta}(\tau,\boldsymbol{x})$, i.e. the analogue to the Feynman propagator in the imaginary-time formalism, setting the real part of the temporal argument to zero and applying time-ordering to its imaginary part $\tau = \Im(t)$,

$$G_{\beta}(\tau',\boldsymbol{x}';\tau,\boldsymbol{x}) = G_{\beta}(\tau'+2\beta,\boldsymbol{x}';\tau,\boldsymbol{x}) = \mathrm{Tr}\,\{\mathrm{e}^{-\beta H}T_{\tau}[\phi(\tau',\boldsymbol{x}')\phi(\tau,\boldsymbol{x})]\}/Z\,. \tag{14.10}$$

The derivation (14.8d) goes through unchanged for fermionic fields,

$$S^{\pm}_{\beta}(t',\boldsymbol{x}';t,\boldsymbol{x}) = S^{\mp}_{\beta}(t'+\mathrm{i}\beta,\boldsymbol{x}';t,\boldsymbol{x})\,. \tag{14.11}$$

But now the anti-commuting nature of fermionic fields, i.e. the minus sign in the definition of time-ordered product (8.85), leads to anti-periodic boundary condition for their thermal propagators,

$$S_{\beta}(t',\boldsymbol{x}';t,\boldsymbol{x}) = -S_{\beta}(t'+2\mathrm{i}\beta,\boldsymbol{x}';t,\boldsymbol{x})\,. \tag{14.12}$$

Both periodicity conditions discretise the frequency spectrum of the thermal wave-functions and propagator. Moreover, they constrain the set of allowed frequencies, such that bosonic fields contain only even frequencies, while fermionic fields contain only odd frequencies, cf. problem 15.**??**. Thus the Fourier transform of thermal fields contained in a box of size $\beta\times V$ is given by

$$\phi(t,\boldsymbol{x}) = \frac{1}{\sqrt{\beta V}}\sum_{n=-\infty}^{\infty}\sum_{\boldsymbol{p}}\phi_{n,\boldsymbol{p}}\,\mathrm{e}^{-\mathrm{i}(\omega_n t+\boldsymbol{p}\cdot\boldsymbol{x})} \tag{14.13}$$

with $\omega_n = 2n\pi T$ for bosonic and $\omega_n = (2n+1)\pi T$ for fermionic fields, respectively, and $n\in\mathbb{Z}$. The frequencies ω_n are called Matsubara frequencies. Similarly, the Green functions for a free scalar field is given in the limit $V\to\infty$ by

$$G_{\beta}(t,\boldsymbol{x}) = \frac{1}{\beta}\sum_{n=-\infty}^{\infty}\int\frac{\mathrm{d}^3p}{(2\pi)^3}\,G_n(\omega_n,\boldsymbol{p})\,\mathrm{e}^{-\mathrm{i}(\omega_n t+\boldsymbol{p}\cdot\boldsymbol{x})} \tag{14.14}$$

with

$$G_n(\omega_n,\boldsymbol{p}) = \frac{1}{\omega_n^2+\boldsymbol{p}^2+m^2}\,. \tag{14.15}$$

Thermal Green functions and vertices come without imaginary units, because we have transformed the path integral to Euclidean time.

14.2. Scalar gas

We will illustrate the basics of thermal field theory considering the simplest example, a gas of scalar particles, evaluating its free energy density $\mathcal{F} = F/V$ as a power series in λ,

$$\mathcal{F} = \mathcal{F}_0 + \lambda\mathcal{F}_1 + \lambda^2\mathcal{F}_2 + \ldots \tag{14.16}$$

The free energy is determined by connected vacuum diagrams. The lowest order contribution in perturbation theory is given by a one-loop vacuum diagram and corresponds to the Stefan-Boltzmann law valid for a free, non-interacting gas. Going on to the two-loop vacuum diagrams, we will be able to derive the first quantum correction to the Stefan-Boltzmann law.

14.2.1. Free scalar gas

In the case of a non-interacting scalar field, we can perform the path integral in the free partition function Z, obtaining as formal solution in Euclidean space

$$\int \mathcal{D}\phi\, \mathrm{e}^{-S_0} = \mathcal{N}\, \mathrm{D}et(-\partial^2 + m^2)^{-1/2} . \tag{14.17}$$

Neglecting the temperature independent normalisation constant $\mathcal{N}$ and using the identity $\ln \mathrm{D}et A = \mathrm{T}r \ln A$, we arrive at

$$\beta F = -\ln Z = \frac{1}{2} \mathrm{T}r \ln(-\partial^2 + m^2) . \tag{14.18}$$

We evaluate the operator trace as in the case of the effective potential in section 13.2, but take into account the changes

$$\int_{-\infty}^{\infty} \mathrm{d}t \to \frac{1}{\beta} \int_0^\beta \mathrm{d}\tau \quad and \quad \int_{-\infty}^{\infty} \mathrm{d}k_0 \to \frac{1}{\beta} \sum_{n=-\infty}^{\infty}$$

in the completeness relation for thermal states. Inserting a complete set of plane waves, we find

$$\begin{aligned} \mathrm{T}r \ln[-\partial^2 + m^2] &= \int_0^\beta \mathrm{d}\tau \int \mathrm{d}^3x \langle \tau, \boldsymbol{x}| \ln[-\partial^2 + m^2] |\tau, \boldsymbol{x}\rangle \\ &= \int_0^\beta \mathrm{d}\tau \int \mathrm{d}^3x \frac{1}{\beta} \sum_n \frac{\mathrm{d}^3k}{(2\pi)^3} \langle \tau, \boldsymbol{x}| \ln[-\partial^2 + m^2] |\omega_n, \boldsymbol{k}\rangle \langle \omega_n, \boldsymbol{k}|\tau, \boldsymbol{x}\rangle \\ &= V \sum_n \frac{\mathrm{d}^3k}{(2\pi)^3} \ln[\omega_n^2 + \boldsymbol{k}^2 + m^2] . \end{aligned} \tag{14.19}$$

The free energy density $\mathcal{F} = F/V$ follows as

$$\beta \mathcal{F} = \frac{1}{2} \sum_{n=-\infty}^{\infty} \int \frac{\mathrm{d}^3k}{(2\pi)^3} \ln\left[\omega_n^2 + \boldsymbol{k}^2 + m^2\right] . \tag{14.20}$$

In order to evaluate the sum, it is sufficient to consider

$$A = T \sum_n \ln\left[-(\mathrm{i}\omega_n)^2 + E_{\boldsymbol{k}}^2\right] \tag{14.21}$$

with $E_{\boldsymbol{k}}^2 = \boldsymbol{k}^2 + m^2$. First we eliminate the logarithm by differentiating A w.r.t. $E_{\boldsymbol{k}}$,

$$\frac{\mathrm{d}A}{\mathrm{d}E_{\boldsymbol{k}}} = -2TE_{\boldsymbol{k}} \sum_n \frac{1}{(\mathrm{i}\omega_n)^2 - E_{\boldsymbol{k}}^2} . \tag{14.22}$$

The function $\mathrm{d}A/\mathrm{d}E_{\boldsymbol{k}}$ has poles in the complex ω plane along the imaginary axis at $\omega = \mathrm{i}\omega_n = 2\pi nT\mathrm{i}$ plus two poles on the real axis at $\omega = \pm E_{\boldsymbol{k}} \mp \mathrm{i}\varepsilon$, cf. Fig. 14.1. We convert the sum into a contour integral using Cauchy's theorem in "reverse order". Because of $\coth(z) = \sinh'(z)/\sinh(z)$, we see that $\coth(z)$ has simple poles at $k\pi\mathrm{i}$ with residue 1 and $k \in \mathbb{Z}$. Thus $\coth(\beta\omega/2)$ has poles at $\mathrm{i}\omega_n$ with residua $2/\beta$, and we obtain

$$\frac{\mathrm{d}A}{\mathrm{d}E_{\boldsymbol{k}}} = -2TE_{\boldsymbol{k}} \frac{\beta}{2} \sum_n \mathrm{r}es_{\omega_n} \left\{ \frac{\coth(\beta\omega/2)}{\omega_n^2 - E_{\boldsymbol{k}}^2} \right\} = \frac{E_{\boldsymbol{k}}}{2\pi\mathrm{i}} \oint_{\mathcal{C}} \mathrm{d}\omega\, \frac{\coth(\beta\omega/2)}{\omega^2 - E_{\boldsymbol{k}}^2} . \tag{14.23}$$

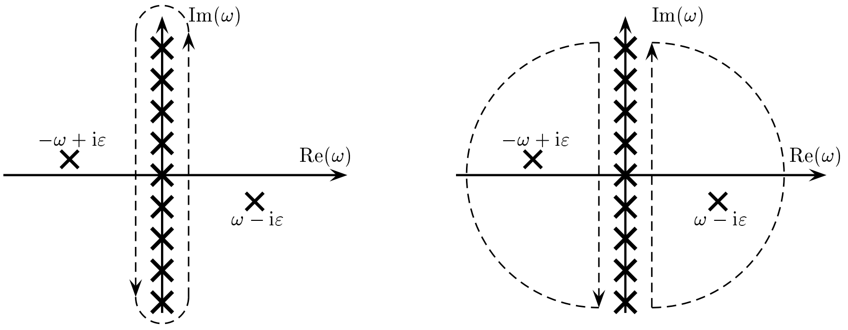


Figure 14.1.: Poles and contours in the complex ω plane used for the evaluation of the free energy $\mathcal{F}$.

The integrand vanishes as $1/|\omega|^2$ for $|\omega| \to \infty$, and thus we can break up the contour $\mathcal{C}$ into two pieces which we close at $\pm i\infty$. Note that we pick up thereby a minus sign, since we change the orientation of the integration path. Now we can use Cauchy's theorem in the "normal order" to evaluate the residua of the two enclosed poles at $\pm E_{\boldsymbol{k}} \mp i\varepsilon$,

$$\frac{\mathrm{d}A}{\mathrm{d}E_{\boldsymbol{k}}} = E_{\boldsymbol{k}} \sum_{\pm} res_{\pm E_{\boldsymbol{k}}} \left\{ \frac{\coth(\beta\omega/2)}{\omega^2 - E_{\boldsymbol{k}}^2} \right\} = \coth(\beta E_{\boldsymbol{k}}/2) = 1 + \frac{2}{\mathrm{e}^{\beta E_{\boldsymbol{k}}} - 1} . \tag{14.24}$$

Integration gives

$$\mathcal{F} = \frac{1}{2} \int \frac{\mathrm{d}^3k}{(2\pi)^3} \left[E_{\boldsymbol{k}} + 2T \ln(1 - \mathrm{e}^{-\beta E_{\boldsymbol{k}}}) + c \right] . \tag{14.25}$$

The integration constant c cancels against the normalisation constant of the path integral; dropping also the $T = 0$ vacuum part and taking the high-temperature limit $T \gg m$ gives

$$\begin{aligned} \mathcal{F} &= T \int \frac{\mathrm{d}^3k}{(2\pi)^3} \ln(1 - \mathrm{e}^{-\beta E_{\boldsymbol{k}}}) \\ &= \frac{T^4}{2\pi^2} \int_0^\infty \mathrm{d}x\, x^2 \ln\left(1 - \exp\{-\sqrt{x^2 + (\beta m)^2}\}\right) \simeq -\frac{\pi^2 T^4}{90} + \frac{m^2 T^2}{24} . \end{aligned} \tag{14.26}$$

With $\mathcal{F} = -P$, the Stefan-Boltzmann law $P = \pi^2 T^4/90$ for a massless real scalar gas follows.

Example 14.1: Derive the high-temperature expansion (14.26) of the free energy $\mathcal{F}$.
Performing a Taylor expansion of $\ln(1 - \exp\{-[x^2 + (\beta m)^2]^{1/2}\})$ around $\beta m = 0$, we find

$$\int_0^\infty \mathrm{d}x\, x^2 \ln(1 - \mathrm{e}^{-[x^2+(\beta m)^2]^{1/2}}) = \int_0^\infty \mathrm{d}x\, x^2 \ln(1 - \mathrm{e}^{-x}) + \frac{(\beta m)^2}{2} \int_0^\infty \mathrm{d}x\, \frac{x}{\mathrm{e}^x - 1} + \mathcal{O}(\beta m)^4 .$$

In the first integral we expand the logarithm,

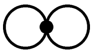
$$\int_0^\infty \mathrm{d}x\, x^2 \ln(1 - \mathrm{e}^{-x}) = -\sum_{n=1}^\infty \int_0^\infty \mathrm{d}x\, x^2 \mathrm{e}^{-nx} = -2 \sum_{n=1}^\infty \frac{1}{n^4} = -2\zeta(4) = -2\,\frac{\pi^4}{90} ,$$

while we use in the second integral $1/(\mathrm{e}^x - 1) = \mathrm{e}^{-x}/(1 - \mathrm{e}^x) = \sum_{n=1}^\infty \mathrm{e}^{-nx}$. With the definition (A.24) for the Gamma function, the second integral results in $\Gamma(2)\zeta(2) = \pi^2/6$.

14.2.2. Interacting scalar gas

As main application of the formalism of thermal field theory we next calculate the first quantum correction to the equation of state of a gas of scalar particles interacting with a $\lambda\phi^4$ interaction. This $\mathcal{O}(\lambda)$ correction is given by the two-loop vacuum diagram we considered already at $T = 0$; the overlapping divergences of the type $1/\varepsilon \ln(p^2/\mu^2)$ correspond now to UV divergent terms multiplied by temperature-dependent factors. Again, such terms have to be cancelled by counter-terms found at lower order.

Equation of state of a scalar gas The first-order contribution $\lambda\mathcal{F}_1$ to the free energy density corresponds to the two-loop vacuum diagram,



$$= \frac{\lambda}{4!}\langle\phi^4\rangle = \frac{3\lambda}{4!}\langle\phi^2\rangle\langle\phi^2\rangle = \frac{\lambda}{8}\Big(\Delta(0)\Big)^2 ,$$

Recall from section 4.2 that the factor three accounts for the three possible ways of joining the lines to form the two loops. Alternatively, we can determine this factor using that $\langle\phi^n\rangle$ is the expectation value of ϕ^n times a Gaussian. We could now insert into this formula the Matsubara Green function derived in the imaginary time formalism. Instead, we use a more intuitive argument to obtain the propagator (for coinciding points) in the real-time formalism. At $T = 0$, we can express the propagator at coincident points as the sum over the zero-point energies,

$$\Delta(0) = \langle 0|\phi(x')\phi(x)|0\rangle_{x'\searrow x} = \int \frac{\mathrm{d}^3k}{(2\pi)^3 2\omega_k}\mathrm{e}^{-\mathrm{i}k(x'-x)}\Big|_{x'\searrow x} = \int \frac{\mathrm{d}^3k}{(2\pi)^3}\frac{1}{2\omega_{\boldsymbol{k}}} . \tag{14.27}$$

The vacuum contains no real particles, and the $a^\dagger_{\boldsymbol{k}} a_{\boldsymbol{k}}$ term in $H_0 = \sum_{\boldsymbol{k}} \omega_{\boldsymbol{k}}(1/2 + a^\dagger_{\boldsymbol{k}} a_{\boldsymbol{k}})$ gives zero contribution. For $T > 0$, the expectation value of the number operator $N_{\boldsymbol{k}} = a^\dagger_{\boldsymbol{k}} a_{\boldsymbol{k}}$ is just the number distribution $n_{\boldsymbol{k}}$ of ϕ particles,

$$\Delta(t = 0, x = 0) = \int \frac{\mathrm{d}^3k}{(2\pi)^3}\frac{1}{\omega_{\boldsymbol{k}}}\left(\frac{1}{2} + \left\langle a^\dagger_{\boldsymbol{k}} a_{\boldsymbol{k}}\right\rangle\right) = \int \frac{\mathrm{d}^3k}{(2\pi)^3}\frac{1 + 2n_{\boldsymbol{k}}}{2\omega_{\boldsymbol{k}}} . \tag{14.28}$$

In thermal equilibrium the number density $n_{\boldsymbol{k}}$ of a scalar field is a Bose-Einstein distribution,

$$n_{\boldsymbol{k}} = \frac{1}{e^{\beta\omega_{\boldsymbol{k}}} - 1} . \tag{14.29}$$

This result can be derived directly from the periodicity condition of the Green functions (problem 15.**??**). Note that we can view the propagator as the sum of a vacuum part ("1/2") and a thermal part ("$n_{\boldsymbol{k}}$"). In the latter, the high energy modes are exponentially suppressed and thus no UV divergences should appear in the temperature dependent parts of physical observables. Thus our standard renormalisation program at $T = 0$ should apply in the same way at $T > 0$.

Continuing the derivation of the first-order correction to $\mathcal{F}$ we have

$$\frac{\lambda}{4!}\langle\phi^4\rangle = \frac{\lambda}{8}\Big[\sum_{\boldsymbol{k}} \frac{1 + 2n_{\boldsymbol{k}}}{2\omega_{\boldsymbol{k}}}\Big]^2 \tag{14.30}$$

$$= \frac{\lambda}{8}\Bigg[\underbrace{\Big(\sum_{\boldsymbol{k}} \frac{1}{2\omega_{\boldsymbol{k}}}\Big)^2}_{\text{vacuum}} + \underbrace{\Big(\sum_{\boldsymbol{k}} \frac{n_{\boldsymbol{k}}}{\omega_{\boldsymbol{k}}}\Big)^2}_{T \text{ dependent}} + \underbrace{2\Big(\sum_{\boldsymbol{k}} \frac{1}{2\omega_{\boldsymbol{k}}}\Big)\Big(\sum_{\boldsymbol{k}} \frac{n_{\boldsymbol{k}}}{\omega_{\boldsymbol{k}}}\Big)}_{\text{vacuum} \times T \text{ dependent}}\Bigg] . \tag{14.31}$$

The mixed term gives rise to a temperature dependent UV divergence, which from the previous argument should not appear in our calculation. Similarly as at $T = 0$, this term should be cancelled by a one-loop counter-term. We consider therefore the one-loop self-energy Σ, which is given by

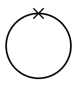
$$\Sigma = \frac{\lambda}{2}\Delta(t = 0, \boldsymbol{x} = 0) = \frac{\lambda}{2}\sum_{\boldsymbol{k}} \frac{1 + 2n_{\boldsymbol{k}}}{2\omega_{\boldsymbol{k}}}\,, \tag{14.32}$$

and contains the usual $T = 0$ divergence coming from the unsuppressed sum over frequencies. We use renormalised perturbation theory, adding the counter-term $\delta_m m^2\phi^2$ to the Lagrangian, where δ_m is chosen such that m corresponds to the physical mass. Thus $\delta_m m^2$ is determined by

$$\delta_m m^2 + \frac{\lambda}{2}\sum_{\boldsymbol{k}} \frac{1}{2\omega_{\boldsymbol{k}}} = 0\,. \tag{14.33}$$

Because the product $\delta_m m^2\Delta(0)$ is of $\mathcal{O}(\lambda)$, we have to add the following vacuum diagram to Ω_1,

$$\frac{1}{2}\delta_m m^2\Delta(0,0) =$$



$$= \frac{1}{2}\left(-\frac{\lambda}{2}\sum_{\boldsymbol{k}} \frac{1}{2\omega_{\boldsymbol{k}}}\right)\left(\sum_{\boldsymbol{k}} \frac{1 + 2n_{\boldsymbol{k}}}{2\omega_{\boldsymbol{k}}}\right). \tag{14.34}$$

Comparing this expression to the troublesome mixed term, we see that they agree but have the opposite sign. Thus the one-loop subdivergence cancels the temperature dependent UV divergence at two-loop in Ω_1. As result, we obtain the consistent expression

$$\frac{\lambda}{4!}\langle\phi^4\rangle = \frac{\lambda}{8}\left(\sum_{\boldsymbol{k}} \frac{n_{\boldsymbol{k}}}{\omega_{\boldsymbol{k}}}\right)^2 + \text{vac}\,, \tag{14.35}$$

which we can now calculate explicitly. For simplicity, we restrict ourselves to the high-temperature limit setting $m = 0$,

$$\sum_{\boldsymbol{k}} \frac{n_{\boldsymbol{k}}}{\omega_{\boldsymbol{k}}} = \int \frac{\mathrm{d}^3k}{(2\pi)^3}\frac{1}{\omega}\frac{1}{\mathrm{e}^{\beta\omega} - 1} = \frac{T^2}{2\pi^2}\underbrace{\int_0^\infty \mathrm{d}x \frac{x}{\mathrm{e}^x - 1}}_{\pi^2/6} = \frac{T^2}{12}\,, \tag{14.36}$$

Hence our final result for $\mathcal{F}_1$ is

$$\mathcal{F}_1 = \frac{\lambda}{8}\left(\frac{T^2}{12}\right)^2 = \frac{\lambda}{1152}T^4\,, \tag{14.37}$$

which we may compare to the non-interacting result, $\Omega_0 = \pi^2T^4/90$: The ratio $\Omega_1/\Omega_0 \approx 10^{-2}\lambda$ seems to indicate a fast convergence of the perturbative expansion of the pressure for any reasonable value of the coupling.

We simply quote the result to three loops or second order in λ from the literature,

$$P = \frac{\pi^2T^4}{9}\left[\frac{1}{10} - \frac{1}{8}\frac{\lambda}{16\pi^2} + \frac{1}{8}\left(3\ln\frac{\mu}{4\pi T} + \frac{31}{35} + C\right)\left(\frac{\lambda}{16\pi^2}\right)^2\right]. \tag{14.38}$$

The parameter μ in Eq. (14.38) is not the chemical potential but as usually the renormalisation scale. As the pressure is a physical quantity, it should not depend upon such a parameter.

However, the truncation of the perturbative series leads to a residual μ dependence of $\mathcal{O}(\lambda^3)$. Examining this μ dependence of P we obtain

$$\mu\frac{\mathrm{d}P}{\mathrm{d}\mu} = \frac{\pi^2 T^4}{9}\left[-\frac{1}{8}\frac{1}{16\pi^2}\,\mu\frac{\mathrm{d}\lambda}{\mathrm{d}\mu} + \frac{3}{8}\left(\frac{\lambda}{16\pi^2}\right)^2 + \mathcal{O}(\lambda^3)\right] = \mathcal{O}(\lambda^3)\,, \tag{14.39}$$

where we have used $\mu\mathrm{d}\lambda/\mathrm{d}\mu = \beta = 3\lambda^2/16\pi^2$. Thus the dependence on the renormalisation scale μ is indeed of higher order in λ than the order of perturbation theory we are working with. Still we can try to minimise the remaining dependence by a suitable choice of μ: Clearly, a sensible choice in this case is $\mu = 4\pi T$, for which the logarithm vanishes. Still, any other choice is mathematically as correct as this one. Instead of being worried about this dependence on the renormalisation scale, we may take advantage of it as follows: Varying the renormalisation scale μ in a "reasonable range", say between $\mu = 2\pi T$ and $\mu = 8\pi T$, we obtain an error estimate for the missing higher-order corrections. Finally, note that $\mu \propto T$ implies that a QCD plasma becomes in the large temperature limit an asymptotically free gas of quarks and gluons. Since we know that at $T = 0$ quarks and gluons are confined in hadrons, we expect that at $T = \mathcal{O}(\Lambda_{QCD})$ a phase transition from a quark-gluon gas to a gas of colourless mesons and baryons takes place. Both analytical calculations and lattice QCD simulations confirm this picture.

IR behaviour We have found a fast convergence of the perturbative expansion from the results for the pressure of a scalar gas. For applications particularly interesting is the case of a massless particle and we consider now as a toy-model for QCD a $\lambda\phi^4$ theory with $m = 0$.

Looking back at our previous result for the self-energy, Eq. (14.32), setting $m \to 0$ and dropping the vacuum term, we have

$$\Sigma = \frac{\lambda}{2}\Delta(t = 0, \boldsymbol{x} = 0) \to \frac{\lambda}{2}\int \frac{\mathrm{d}^3 k}{(2\pi)^3}\frac{n_{\boldsymbol{k}}}{\omega_{\boldsymbol{k}}} = \frac{\lambda}{2}\frac{T^2}{12}\,. \tag{14.40}$$

Thus the thermal part of the self-energy induces at first order in λ a thermal or Debye mass,

$$m_{\mathrm{D}}^2 = \frac{\lambda}{2}\frac{T^2}{12}\,. \tag{14.41}$$

Switching to the covariant form of the thermal propagator, Eq. (14.14), the contribution shown in the left panel of Fig. 14.2 at second order is

$$\Sigma_2 = -\frac{\lambda}{2}T\sum_n\int\frac{\mathrm{d}^3 k}{(2\pi)^3}\frac{m_{\mathrm{D}}^2}{(\omega_n^2 + \boldsymbol{k}^2)^2}\,. \tag{14.42}$$

While in the terms with $n \neq 0$ the ω_n act as a IR cutoff, we see that for $n = 0$ and thus $\omega_0 = 0$ the integral is proportional to $\int \mathrm{d}k/\boldsymbol{k}^2$ and thus IR divergent. If we go to higher terms in the expansion and add additional loops to the "primary loop" as shown in the right panel of Fig. 14.2, then the degree of the IR divergence increases.

The solution to this problem is to account for the thermal mass of the scalar particle properly. If we use an effective propagator which includes the Debye mass of the particle,

$$\frac{1}{\omega_n^2 + \boldsymbol{k}^2} \to \frac{1}{\omega_n^2 + \boldsymbol{k}^2 + m_{\mathrm{D}}^2}\,, \tag{14.43}$$

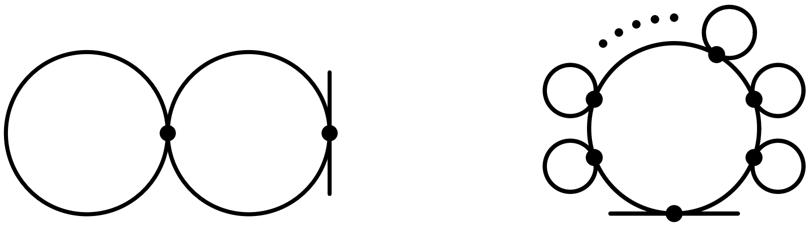


Figure 14.2.: Left: A second order correction to the mass; Right: Ring diagram, each external bubble corresponds to an insertion of $m_{\rm D}^2$.

then the $n = 0$ term of Eq. (14.42) with $\omega_0 = 0$ is given by

$$\Sigma_{2,n=0} = -\frac{\lambda}{2} T \int \frac{\mathrm{d}^3 k}{(2\pi)^3} \frac{m_{\rm D}^2}{(\boldsymbol{k}^2 + m_{\rm D}^2)^2} = -\frac{\lambda^2}{4} \left(\frac{T^2}{12}\right) \left(\frac{T}{8\pi m_{\rm D}}\right) \tag{14.44}$$

and thus finite. Since the Debye mass scales as $m_{\rm D} \propto \lambda^{1/2}$, the contribution of $\Sigma_{2,n=0}$ in the perturbative expansion is not of order λ^2 but of order $\lambda^{3/2}$. Thus we obtained a term which is non-analytic in the coupling—what can not happen, if we sum a finite number of terms. As explanation, we have to look at the expansion of the effective propagator for small $m_{\rm D}$ (restricting ourselves again to the $n = 0$ term),

$$\frac{1}{\boldsymbol{k}^2 + m_{\rm D}^2} = \frac{1}{\boldsymbol{k}^2} - \frac{m_{\rm D}^2}{\boldsymbol{k}^4} + \frac{m_{\rm D}^4}{\boldsymbol{k}^6} + \ldots \tag{14.45}$$

Here we can view e.g. the $m_{\rm D}^4/\boldsymbol{k}^6$ term as a ring diagram with three massless propagators $\boldsymbol{k}^{-2}$ and two factors $m_{\rm D}^2$ produced by self-energy insertions. Thus including the thermal mass corresponds to summing up the infinite sum of diagrams shown in the right panel of Fig. 14.2.

We can formalise the inclusion of the Debye mass in the originally massless scalar theory as follows: We reorganise perturbation theory by adding a mass term to the free Lagrangian and subtracting it from the the interaction term,

$$\mathscr{L}_0 = \frac{1}{2} (\partial_\mu \phi)^2 - \frac{1}{2} m^2 \phi^2 \quad \textit{and} \quad \mathscr{L}_{int} = -\frac{\lambda}{4!} \phi^4 + \frac{1}{2} m^2 \phi^2 \,. \tag{14.46}$$

Here we may set $m^2 = m_D^2$ or keep it as a free parameter to be determined by, for example, that the free energy is independent of this parameter, $\mathrm{d}\mathcal{F}/\mathrm{d}m^2 = 0$. This reformulation of the perturbative expansion in thermal field theories is called screened or optimised perturbation theory.

Symmetry restoration at high temperature We have chosen the sign of the mass term such that the $\lambda\phi^4$ theory is in the unbroken phase, and the minimal energy is obtained for $\langle\phi\rangle = 0$. In this case thermal effects simply increase the effective mass of the ϕ particle,

$$V(\phi, T) = \frac{1}{2} \left(m^2 + \frac{\lambda}{24} T^2\right) \phi^2 + \frac{\lambda}{4!} \phi^4 \,, \tag{14.47}$$

Something more interesting happens, if we consider the broken phase choosing $m^2 < 0$. Then for $T = 0$ the minimal energy is obtained for $\phi_0 = \pm\sqrt{-6m^2/\lambda}$, and the $\phi \to -\phi$ symmetry

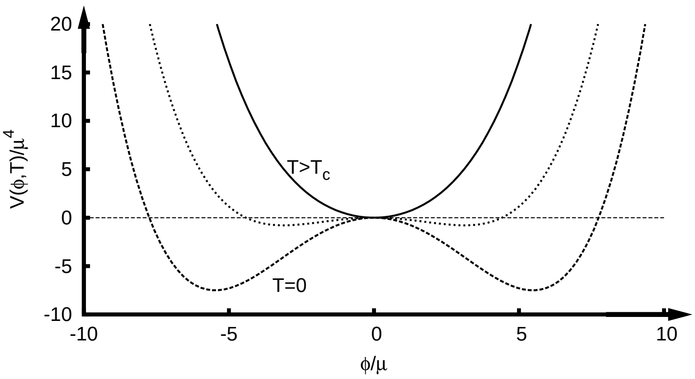


Figure 14.3.: Potential with negative mass term showing restoration of symmetry for $T > T_c$.

of the Lagrangian is broken. Increasing the temperature, the positive thermal mass grows. Above some critical temperature, the effective mass becomes therefore positive and the minimal energy configuration becomes $\phi_0 = 0$, cf. with Fig. 14.3. This behaviour resembles the one we know from a ferromagnet: Below some critical temperature, spontaneous magnetisation breaks rotation invariance, which is above T_c restored. In the context of cosmology, this observation suggests that symmetries that are broken today might have been unbroken in the early, hot universe.

14.A. Appendix: Equilibrium statistical physics in a nut-shell

The one-particle distribution function $f(p)$ of a free gas in *kinetic equilibrium* is given by

$$f(p) = \frac{1}{\exp[\beta(E-\mu)] \pm 1} \tag{14.48}$$

where $\beta = 1/T$ is the inverse temperature, $E = \sqrt{m^2 + \boldsymbol{p}^2}$, and $+1$ refers to fermions and -1 to bosons, respectively. A species X stays in kinetic equilibrium, if in the reaction $X + Y \rightleftharpoons X + Y$ the energy exchange with at least one other species Y, which is in thermal equilibrium is fast enough. The chemical potential μ is the average energy needed, if an additional X particle is added, $\mathrm{d}U = \mu_X \mathrm{d}N_X$. If the species X is also in *chemical equilibrium* with other species, e.g. via the reaction $X + \bar{X} \rightleftharpoons \gamma + \gamma$, then their chemical potentials are related by $\mu_X + \mu_{\bar{X}} = 2\mu_\gamma = 0$.

The number density n, energy density ρ and pressure P of a species X (which may be not in equilibrium) are connected to its one-particle distribution function $f(p)$ as

$$n = g \int \frac{\mathrm{d}^3p}{(2\pi)^3}\, f(p)\,, \qquad \rho = g \int \frac{\mathrm{d}^3p}{(2\pi)^3}\, Ef(p)\,, \tag{14.49}$$

$$P = g \int \frac{\mathrm{d}^3p}{(2\pi)^3}\, \frac{p^2}{3E}\, f(p)\,. \tag{14.50}$$

The factor g takes into account the internal degrees of freedom like spin or colour. Thus for a photon, a massless spin-1 particle $g = 2$, for an electron $g = 2$, a quark $g = 6$, etc.

Table 14.1.: The number of relativistic degrees of freedom g_* present in the Universe as function of its temperature T.

Temperature	new particles	$4\Delta g_*$	$4g_*$
$T < m_e$	$\gamma + \nu_i$	$4 \times (2 + 3 \times 2 \times 7/8)$	29
$m_e < T < m_\mu$	$e^\pm$	14	43
$m_\mu < T < m_\pi$	$\mu\pm$	14	57
$m_\pi < T < T_{\rm QCD}$	$\pi^\pm, \pi^0$	12	69
$T_{QCD} < T < m_s$	$u, \bar{u}, d, \bar{d}, g$	$6 \times 14 + 4 \times 8 \times 2 - 12$	205
$m_s < T < m_c$	$s, \bar{s}$	$3 \times 14 = 42$	247
$m_c < T < m_\tau$	$c, \bar{c}$	42	289
$m_\tau < T < m_b$	$\tau^\pm$	14	303
$m_b < T < m_{W,Z}$	$b, \bar{b}$	42	345
$m_{W,Z} < T < m_h$	$W^\pm, Z$	$4 \times 3 \times 3 = 36$	381
$m_h < T < m_t$	h	4	385
$m_t < T < ?$	$t, \bar{t}$	42	427

In the non-relativistic limit $T \ll m$ or $\mathrm{e}^{\beta(m-\mu)} \gg 1$, and thus differences between bosons and fermions disappear,

$$n = \frac{g}{2\pi^2}\mathrm{e}^{-\beta(m-\mu)} \int_0^\infty \mathrm{d}p\, p^2 \mathrm{e}^{-\beta\frac{p^2}{2m}} = g\left(\frac{mT}{2\pi}\right)^{3/2} \exp[-\beta(m-\mu)]\,, \tag{14.51}$$

$\rho = mn$ and $P = nT \ll \rho$. These expressions correspond to the classical Maxwell-Boltzmann statistics. The number of non-relativistic particles is exponentially suppressed, if their chemical potential is small. In the relativistic limit $T \gg m$ with $T \gg \mu$, all properties of a gas are determined by its temperature T,

$$n = \frac{g\,T^3}{2\pi^2} \int_0^\infty \mathrm{d}x \frac{x^2}{\mathrm{e}^x \pm 1} = \varepsilon_1 \frac{\zeta(3)}{\pi^2}\, gT^3\,, \tag{14.52}$$

$$\rho = 3P = \frac{g\,T^4}{2\pi^2} \int_0^\infty \mathrm{d}x \frac{x^3}{\mathrm{e}^x \pm 1} = \varepsilon_2 \frac{\pi^2}{30}\, gT^4\,, \tag{14.53}$$

where for bosons $\varepsilon_1 = \varepsilon_2 = 1$ and for fermions $\varepsilon_1 = 3/4$ and $\varepsilon_2 = 7/8$, respectively.

Since the energy density and the pressure of non-relativistic species is exponentially suppressed, the total energy density and the pressure of all species present in the universe can be approximated including only the relativistic ones,

$$\rho_{\rm rad} = 3P_{\rm rad} = \frac{\pi^2}{30}\, g_* T^4\,, \tag{14.54}$$

where

$$g_* = \sum_{bosons} g_i \left(\frac{T_i}{T}\right)^4 + \frac{7}{8} \sum_{fermions} g_i \left(\frac{T_i}{T}\right)^4 . \tag{14.55}$$

We denote the relativistic species also collectively as radiation. Table 14.1 shows the number of relativistic degrees of freedom g_* in the SM as function of the temperature. Here T_{QCD} denotes the temperature of the QCD phase transition, above which quarks and gluons as free particles exist.

The total entropy density $s \equiv S/V$ of the universe can again approximated by the relativistic species,

$$s = \frac{2\pi^2}{45}\, g_{*S} T^3\,, \tag{14.56}$$

where now

$$g_{*,S} = \sum_{bosons} g_i \left(\frac{T_i}{T}\right)^3 + \frac{7}{8} \sum_{fermions} g_i \left(\frac{T_i}{T}\right)^3 . \tag{14.57}$$

The entropy S is a useful quantity, because it is conserved if the universe evolves adiabatically. Then the conservation of S implies that $S \propto g_{*,S} a^3 T^3 = const.$, where $a(t)$ is the scale factor describing the expansion of the universe. Thus the temperature of the universe evolves as $T \propto g_{*,S}^{-1/3} a^{-1}$, and when $g_* = const.$, the temperature scales as $T \propto 1/a$. Consider now the case that a particle species, e.g. electrons, becomes non-relativistic at $T \sim m_e$. Then the particles annihilate, $e^+e^- \to \gamma\gamma$, and their entropy is transferred to photons. Formally, $g_{*,S}$ decreases and therefore the temperature decreases for a short period less slowly than $T \propto 1/a$.

Since the net number density n of particles with a conserved charge and the entropy density scale both as $\propto a^{-3}$, the ratio n/s is constant. Therefore it is often convenient to consider the time evolution of the dimensionless variable $Y = n/s$.

Summary

Thermal Green functions are (anti-) periodic functions in imaginary time, leading to discrete energies with $\omega_n = 2n\pi T$ for bosonic and $\omega_n = (2n+1)\pi T$ for fermionic fields, respectively. No new UV divergences appear for $T > 0$, since the thermal distribution function vanish exponentially for $E/T \to \infty$. In a plasma, even massless particles can acquire a temperature dependent (Debye) mass and symmetries of the Lagrangian may be hidden at low temperatures.

Further reading

This chapter could give only a flavour of what thermal field theory is. The lecture notes of [Bla11] are a useful starting point to learn more, before turning to text books dedicated to thermal field theory as [LV16] or [KG11].

15. Phase transitions and topological defects

We introduced at the end of the last chapter the idea that spontaneously broken symmetries of field theories could be restored at high temperatures. In this case, the hot early Universe would be in a symmetric phase, followed by transitions to phases with more and more broken symmetries. Phase transitions in the early Universe can lead to observable consequences for mainly two reasons: First, phase transitions lead often to the formation of topological defects. These defects are zero-, one- or two-dimensional extended solutions of the classical equations of motion for the Higgs-gauge sector which contain in their core the unbroken $\langle\phi\rangle = 0$ vacuum. Depending on the symmetry breaking scale and their dimensionality, they lead to (un-) desirable cosmological consequences and can thus be used to constrain particle physics models beyond the SM. Second, the state of the Universe deviates from thermodynamical equilibrium during a first-order phase transition. Thus processes like the generation of a baryon asymmetry which require out-of-equilibrium conditions may take place during first-order phase transitions.

15.1. Phase transitions

Effective potential at $T > 0$ We have seen that the ground-state of a quantum field theory including quantum fluctuations is determined by the effective potential V_{eff}. We can include additionally thermal fluctuations by studying the temperature-dependent effective potential. The latter is obtained in a rather straight-forward way in the imaginary-time formalism, where we only have to replace vacuum expectation values with thermal averages in the definition of classical fields, and to use periodic boundary conditions in the Euclidean effective action and effective potential. In particular, we can transform the $T = 0$ effective potential (13.32) defined in Euclidean space,

$$V_{eff}(\phi) = V(\phi) + \frac{1}{2}\int \frac{\mathrm{d}^4k}{(2\pi)^4} \ln\left[k^2 + V''(\phi)\right] + \mathcal{O}(\hbar^2)\,, \tag{15.1}$$

into the temperature-dependent effective potential replacing the integration over the continuous energy k_0 by a summation over discrete Matsubara frequencies ω_n,

$$\beta V_{eff}(\phi) = \beta V(\phi) + \frac{1}{2}\sum_n \int \frac{\mathrm{d}^3k}{(2\pi)^3} \ln\left[\omega_n^2 + \boldsymbol{k}^2 + V''(\phi)\right] + \mathcal{O}(\hbar^2)\,. \tag{15.2}$$

The sum over n is performed in the same way as in the calculation of the free energy $\mathcal{F}$ of a non-interacting scalar gas in section 14.2.1,

$$\beta V_{eff}^{(1)}(\phi) = \frac{1}{2}\int \frac{\mathrm{d}^3k}{(2\pi)^3} \left[\beta E_{\boldsymbol{k}} + 2\ln(1 - \mathrm{e}^{-\beta E_{\boldsymbol{k}}})\right] \tag{15.3}$$

with $E_{\boldsymbol{k}}^2 = \boldsymbol{k}^2 + V''(\phi)$. Next we split the one-loop contribution $\beta V_{eff}^{(1)}(\phi)$ into the $T = 0$ vacuum part $V_{eff}^{(1)}(\phi, 0)$ and a temperature dependent thermal part $V_{eff}^{(1)}(\phi, T)$,

$$V_{eff}^{(1)}(\phi) = \frac{1}{2}\int \frac{\mathrm{d}^4k}{(2\pi)^4} \ln\left[k^2 + V''(\phi)\right] + T\int \frac{\mathrm{d}^3k}{(2\pi)^3} \ln\left(1 - \mathrm{e}^{-\beta E_{\boldsymbol{k}}}\right), \tag{15.4}$$

where we dropped a ϕ-independent constant.

In order to simplify the discussion we consider separately the two limiting cases that quantum or thermal fluctuations dominate. In the latter case, $V_{eff}^{(1)}(\phi, 0) \ll V_{eff}^{(1)}(\phi, T)$, we evaluate $V_{eff}^{(1)}(\phi, T)$ in the high-temperature limit[1]

$$V_{eff}^{(1)}(\phi, T) = T\int \frac{\mathrm{d}^3k}{(2\pi)^3} \ln(1 - \mathrm{e}^{-\beta E_{\boldsymbol{k}}}) = \frac{T^4}{2\pi^2}\int_0^\infty \mathrm{d}x\, x^2 \ln(1 - \mathrm{e}^{-\sqrt{x^2+a}}) \tag{15.5}$$

$$= -\frac{\pi^2 T^4}{90} + \frac{aT^4}{24} + \dots . \tag{15.6}$$

with $a = \beta^2 V''(\phi)$. For the choice $V''(\phi) = -\mu^2 + \frac{1}{2}\lambda\phi^2$, the effective potential becomes

$$V(\phi) + V_{eff}^{(1)}(\phi, T) = -\frac{1}{2}\mu^2\phi^2 + \frac{\lambda}{24}\phi^4 - \frac{\pi^2 T^4}{90} - \frac{1}{24}\mu^2 T^2 + \frac{1}{48}\lambda\phi^2 T^2 + \frac{\lambda}{12}\phi^4 . \tag{15.7}$$

The terms quadratic in the field ϕ reproduce the thermal or Debye mass $m_D^2 = \lambda T^2/24$ which we found in (14.41). The minimum of $V_{eff}(\phi)$ moves smoothly away from $\phi = 0$ below the critical temperature and no barrier is formed, as shown in Fig. 14.3. Thus the phase transition is in this case of second order. More interesting is a first order phase transition which is shown schematically in Fig. 15.1. As the system cools down, a local minimum $\langle\phi\rangle \neq 0$ develops at the temperature T_1. The characteristic feature of a first order transition is that in the temperature range $T_2 < T < T_1$ the two local minima at $\langle\phi\rangle = 0$ and $\langle\phi\rangle \neq 0$ coexist. Thus a potential barrier has to separate them. At the critical temperature T_c, the minimum at $\langle\phi\rangle \neq 0$ becomes the global minimum. Because of the barrier, the transition from $\langle\phi\rangle = 0$ to $\langle\phi\rangle \neq 0$ cannot proceed classically, but has to proceed via a quantum or thermal fluctuation. This tunnelling process may lead to deviations from thermal equilibrium which in turn may lead to traces observable today.

An example for a first order phase transition is given by the SM for a small Higgs mass, $m_h \lesssim 70\,\text{GeV}$. In this (unphysical) limit, $V_{eff}(\phi, T)$ is given for large T by

$$V_{eff}(\phi, T) \simeq \frac{1}{2}a(T^2 - T_1^2)\phi^2 - \frac{1}{3}bT\phi^3 + \frac{1}{4}\lambda\phi^4 \tag{15.8}$$

with

$$a = \frac{3}{16}g^2 + \left(\frac{1}{2} + \frac{m_t^2}{m_h^2}\right)\lambda, \qquad b = \frac{9g^3}{32\pi}, \qquad T_1 = \frac{m_h}{2\sqrt{a}} . \tag{15.9}$$

The critical temperature of the first-order phase transition follows as $T_c = T_1/\sqrt{1 - 2b^2/(9a\lambda)} > T_1$. Between T_c and T_1, the effective potential $V_{eff}(\phi, T)$ has two degenerate minima at $\phi = 0$ and $\phi_c = 2bT_c/(3\lambda)$ separated by a barrier. This corresponds to the effective potential sketched in Fig. 15.1.

[1] The derivation follows the one of Eq. (14.26); recall also that a resummation of ring diagrams may be necessary.

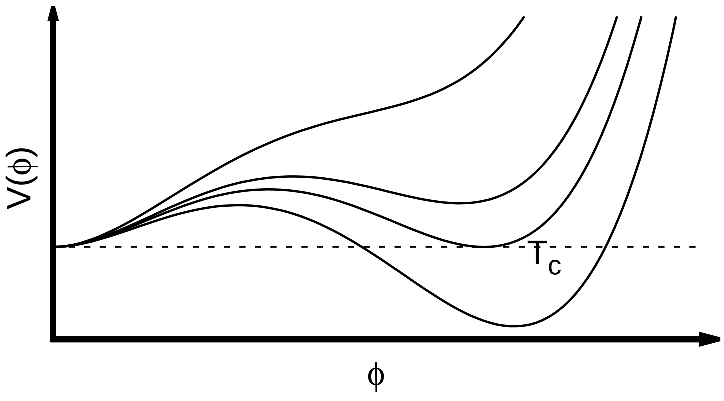


Figure 15.1.: The effective potential $V_{eff}(\phi,T)$ in case of a first-order phase transition.

15.2. Decay of the false vacuum

When the Universe goes through a first order phase transition as it cools down, the fields sitting in the false vacuum have to tunnel to the true minimum. The true vacuum nucleates at a localised position in space-time, when a quantum (or thermal) fluctuation tunnels through (or above) the barrier. It forms an extending bubble which contains the energetically favoured ground-state. In the case of thermal fluctuations, this phenomenon is known to everybody from boiling water. The equivalent problem in quantum mechanics is the tunnelling through the barrier in a double-well potential $V(x)$ depicted in the left panel of Fig. 15.2. The tunnelling probability P can be calculated using the WKB method as

$$P \sim \exp\left(-\mathrm{i}\int_a^b \mathrm{d}x\sqrt{2m[V(x)-E]}\right), \tag{15.10}$$

where a and b are the turning points of the tunnelling trajectory. In order to translate this prescription to a field theory, we rewrite the tunnelling probability P first as an Euclidean path integral. The exponent is the integral over the (imaginary) momentum of the particle, which we can express as

$$\mathrm{i}\int \mathrm{d}x\, p = \mathrm{i}\int \mathrm{d}t\, p\dot{x} = \mathrm{i}\int \mathrm{d}t\,(E+L) = \int \mathrm{d}t_E\,(-L_E)\,. \tag{15.11}$$

In the last step we assumed that the energy of the particle is normalised to zero and changed to Euclidean time $t_E = \mathrm{i}t$. Thus the tunnelling probability is given by the Euclidean path integral

$$P = \int \mathcal{D}x \exp(-S_E[x]) \tag{15.12}$$

with

$$S_E[x] = \int \mathrm{d}x\left(\frac{1}{2}\frac{\mathrm{d}^2x}{\mathrm{d}t_E^2} + V(x)\right)\,. \tag{15.13}$$

From (15.10), we know that the (Minkowski) path integral is dominated by the path depicted on the left of Fig. 15.2. Since the potential V changes sign performing a Wick rotation, a classically forbidden path in Minkowski space corresponds to an allowed Euclidean path shown in the right panel of Fig. 15.2. Thus finding the tunnelling probability between two vacua amounts to finding solutions in the Euclidean theory which connect the two vacua and minimise the action.

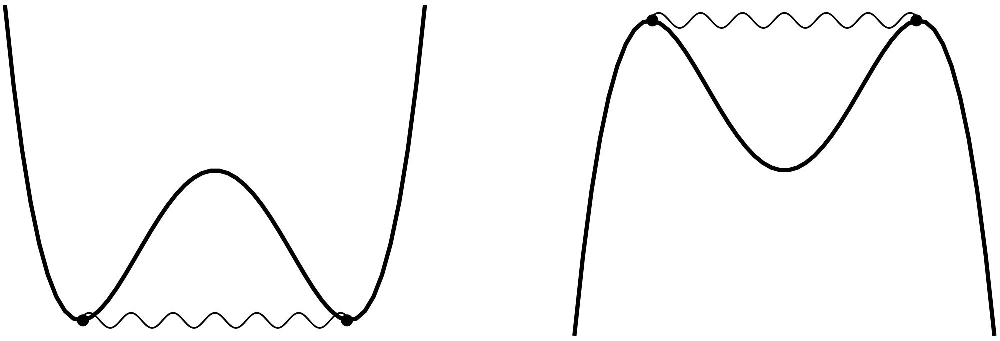


Figure 15.2.: Tunnelling through a double-hill potential in Minkowski space corresponds to the classically allowed path from one maxima to another one in Euclidean space.

Bounce solution We now turn to the case we are interested in, namely the tunnelling of a scalar field ϕ sitting in the false, metastable vacuum a into the true vacuum b, cf. Fig. 15.3. In Euclidean space, we should find solutions of the classical action starting from b which bounce at the classical turning point a back to b. These solutions are called "bounce" or instanton, since the tunnelling happens instantaneously in Minkowski space. Moreover, these solutions should have a finite action such that they give a non-zero contribution to the semi-classical limit of the path integral. In principle, we should find all solution with a finite action and then integrate their contribution. Since the tunnelling probability is however exponentially suppressed, the integral is dominated by the solution which minimises the action.

Solutions which are symmetric under rotations minimise the gradient energy. We use therefore an O(4) symmetric ansatz, where the real scalar field ϕ is only a function of the radial coordinate $r^2 = \boldsymbol{x}^2 + t_E^2$. Then the Euclidean action for a real scalar field becomes

$$S = 2\pi^2 \int_0^\infty \mathrm{d}r\, r^3 \left[\frac{1}{2}\left(\frac{\mathrm{d}\phi}{\mathrm{d}r}\right)^2 + V(\phi)\right] \tag{15.14}$$

and the field equation simplifies to (problem 16.2)

$$\frac{\mathrm{d}^2\phi}{\mathrm{d}r^2} + \frac{3}{r}\frac{\mathrm{d}\phi}{\mathrm{d}r} - \frac{\mathrm{d}V}{\mathrm{d}\phi} = 0\,. \tag{15.15}$$

The boundary conditions $\phi \to 0$ for $r \to \infty$ and $\mathrm{d}\phi/\mathrm{d}r|_{r=0} = 0$ ensure a regular solution at $r = 0$ and a finite action. Viewing (15.15) as a mechanical problem, it describes the motion of a particle in a potential with the friction term $\frac{3}{r}\frac{\mathrm{d}\phi}{\mathrm{d}r}$. Thus the two parts $a \to b$ and $b \to a$

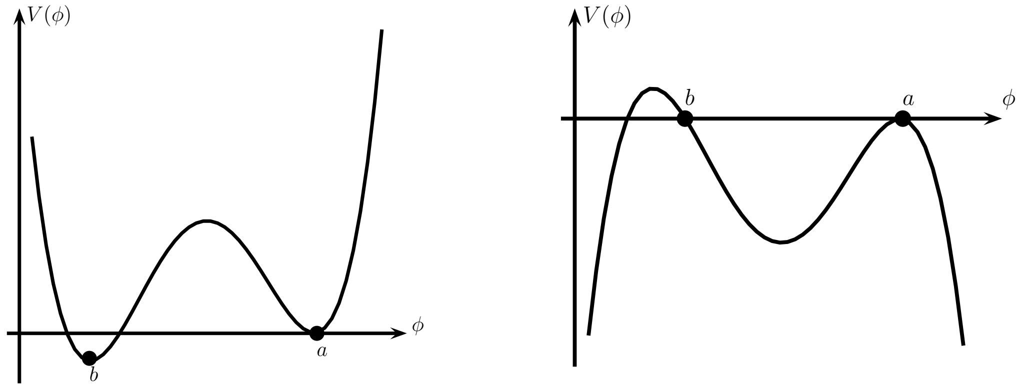


Figure 15.3.: Left: Transition from the false vacuum (a) to the true one (b) requires in Minkowski space tunnelling through a barrier. Right: In Euclidean space, a classically allowed path starts from b and bounces at a back to b; the zero point of the potential is chosen to agree with a.

of the bounce are not equivalent: We should choose the starting position $\phi(t=-\infty)$ uphill of b such that at $t=+\infty$ the field comes to rest at a.

We can proceed analytically, if we consider the limiting case that either the potential difference between the true and the false vacuum is much smaller or much larger than the potential barrier which separates them. We will consider here the first case, which corresponds to the so called thin-wall approximation. This allows us to approximate the potential as

$$V(\phi) = V_0(\phi) + \mathcal{O}(\varepsilon)\,, \tag{15.16}$$

where $V_0(\phi)$ is the symmetric potential and the difference $\varepsilon = V(\phi_+) - V(\phi_-)$ between the false and the true minima is a small parameter. For small ε, the volume energy $\propto \varepsilon r^4$ of a bubble dominates the surface energy $\propto r^3$ only, if the bubble is sufficiently large. In this case, the thickness of the bubble wall, i.e. the region where $\mathrm{d}\phi/\mathrm{d}r$ deviates significantly from zero, is also much smaller than the bubble size. Thus we can neglect the $\frac{3}{r}\frac{\mathrm{d}\phi}{\mathrm{d}r}$ term in (15.15), and the field equation simplifies to the one of a one-dimensional problem,

$$\phi'' \equiv \frac{\mathrm{d}^2\phi}{\mathrm{d}r^2} = \frac{\mathrm{d}V_0}{\mathrm{d}\phi}\,. \tag{15.17}$$

Using the chain rule, $\mathrm{d}V_0/\mathrm{d}\phi = V_0'/\phi'$ and $(\phi'^2)' = 2\phi''\phi'$, we find after one integration

$$\frac{1}{2}\phi'^2 - V_0 = c\,. \tag{15.18}$$

We determine the integration constant by asking that the contribution to the action S goes to zero for $r\to\infty$. With $\phi'(\infty)=0$, this gives $c=-V(\phi_+)$. Separating variables in (15.18) leads then to

$$r = \int \frac{\mathrm{d}\phi}{\sqrt{2[V_0(\phi) - V(\phi_+)]}}\,. \tag{15.19}$$

To gain more insight, we consider now a specific potential. We choose our favourite $\lambda\phi^4$ potential,

$$V_0(\phi) = \frac{\lambda}{4}\left(\phi^2 - \eta^2\right)^2\,. \tag{15.20}$$

Compared to chapter 13, we shifted the potential such that the two minima at $\phi = \pm\eta = \pm\sqrt{\mu^2/\lambda}$ have the value zero, $V_0(\phi_\pm) = 0$, as required. Inserting this potential into Eq. (15.19), we find

$$r = \mp\frac{\sqrt{2}}{\eta\sqrt{\lambda}} arctanh\,(\phi/\eta) + r_0\,. \tag{15.21}$$

The integration constant r_0 determines the position of the bubble wall. Inverting (15.21) gives as field profile

$$\phi_\mp(r) = \mp\eta\tanh\left(\frac{\eta\sqrt{\lambda}}{\sqrt{2}}(r - r_0)\right)\,. \tag{15.22}$$

The solution $\phi_\mp$ interpolate between the two vacua $\pm\eta$ of the symmetric potential $V_0(\phi)$. The thickness of the bubble wall is determined by the argument of the tanh and is given by $\delta \sim \sqrt{2}/(\eta\sqrt{\lambda})$.

Knowing the solution $\phi(r)$, we can calculate the action. For $r \gg r_0$, the field is constant, $\phi \approx \eta$, and $V(\phi_+) = 0$. Therefore, the contribution $S_>$ from this region to the action is zero. For $r \ll r_0$, the field $\phi \approx -\eta$ is again constant, but the potential contributes $V(\phi_-) \simeq -\varepsilon$ and thus

$$S_< = 2\pi^2 \int_0^{r_0} \mathrm{d}r\, r^3 \left[\frac{1}{2}\left(\phi'\right)^2 + V(\phi)\right] \simeq -\frac{1}{2}\pi^2\varepsilon r_0^4\,. \tag{15.23}$$

Finally, the contribution from the bubble wall is

$$\begin{aligned} S_\mathrm{w} &= 2\pi^2 r_0^3 \int \mathrm{d}r \left[\frac{1}{2}\left(\phi'\right)^2 + V(\phi)\right] = 2\pi^2 r_0^3 \int \mathrm{d}r\, 2V(\phi) = \\ &= 2\pi^2 r_0^3 \int_{\phi_1}^{\phi_2} \mathrm{d}\phi\, \sqrt{2V(\phi)} = 2\pi^2 r_0^3 I\,, \end{aligned} \tag{15.24}$$

with $\phi_{1/2} \approx \phi(r_0 \mp \delta)$. The quantity

$$I = \int_{\phi_1}^{\phi_2} \mathrm{d}\phi\, \sqrt{2V(\phi)} \tag{15.25}$$

has the interpretation of the surface tension of the bubble. Adding the two contributions, the total action follows as

$$S = S_< + S_\mathrm{w} = -\frac{1}{2}\pi^2\varepsilon r_0^4 + 2\pi^2 r_0^3 I\,. \tag{15.26}$$

We find the solution which gives the largest tunnelling probability minimising the action S w.r.t. r_0, resulting in the condition

$$r_0 = \frac{3I}{\varepsilon}\,. \tag{15.27}$$

For our choice (15.19) for the potential, the surface tension is $I = 2\mu^3/(3\lambda)$ and the action

$$S = \frac{27\pi^2 I^4}{2\varepsilon^3} = \frac{8\pi^2\mu^{12}}{3\lambda^4\varepsilon^3}\,. \tag{15.28}$$

For $\varepsilon \to 0$, the action becomes infinite. Equivalently, the tunnelling from one to another ground-state $\pm\eta$ of the symmetric potential $V_0(\phi)$ costs an infinite amount of energy, as we argued in chapter 13.

The tunnelling probability per time and volume follows then as

$$p = A \int \mathcal{D}\phi \exp(-S[\phi]) \,, \tag{15.29}$$

where the prefactor A is determined by $\mathrm{D}et(\Box - V''(\phi))^{-1/2}$ and its zero-modes [CC77]. In practice, the result is rather insensitive to the exact value of A, and setting $A \sim V''(\phi_0)$ is sufficient for simple estimates.

We can now justify the assumptions made: Our starting assumption that ε is small implies that r_0 is large, $r_0 \propto 1/\varepsilon$. Moreover, a small ε implies that the bubble of true vacuum is separated by a thin wall from the metastable vacuum, $\delta \sim (\varepsilon r_0)^{-1/3}$. This is the reason for the name thin-wall approximation.

In order to describe the time evolution of a bubble, we have to continue analytically the O(4) symmetric solution $r_0^2 = \boldsymbol{x}^2 + t_E^2$ back to Minkowski space. From $r_0^2 = \boldsymbol{x}^2 - t^2$, we see that the bubble extends close to the speed of light for $t \gg r_0$. The resulting O(1,3) symmetry means that all inertial observers[2] will measure the same expansion law.

Scalar instantons For the case of a massless $\lambda\phi^4$ theory we can find a class of exact solutions for the bounce (15.15), which are often called scalar instantons. For a massless particle, the classical solution should fall off as a power-law. Then dimensional analysis tell us that ϕ has the dimension of an inverse length. Since the massless theory is scale-invariant, the solutions should be parametrised by an arbitrary parameter ρ characterising their size. This suggests to insert as ansatz $\phi(r) \propto \rho/(r^2+\rho^2)$ into Eq. (15.15) what results in (problem 16.3)

$$\phi(r) = \left(\frac{8}{\lambda}\right)^{1/2} \frac{\rho}{r^2+\rho^2} \,. \tag{15.30}$$

Thus the action of the bounce becomes

$$S = \frac{27\pi^2 I^4}{2\varepsilon^3} = \frac{8\pi^2}{3\lambda} \,. \tag{15.31}$$

Example 15.1: Decay of the metastable SM vacuum:
We use $V = \lambda_{eff}\phi^4/4$ as crude approximation for the scalar potential of the SM, where λ_{eff} is the running Higgs self coupling at the scale $\mu = \phi$. Using the thin-wall approximation, we can check that the result (15.31) makes sense: The surface tension is $I \sim (|\lambda_{eff}|/2)^{1/2}\phi^3/3$ and $\varepsilon \sim |\lambda_{eff}|\phi^4/4$. Thus the action of the bounce becomes

$$S = \frac{27\pi^2 I^4}{2\varepsilon^3} = \frac{8\pi^2}{3|\lambda_{eff}|}$$

agreeing with (15.31). An analysis of the RGE of the SM gives for the effective Higgs boson coupling at the Planck scale $\lambda_{eff}(M_{\mathrm{Pl}}) \approx -0.01$, using the central experimental values for m_h and m_t. For the estimate of the probability P that a bubble of the true vacuum has nucleated in the past-light cone of an observer, we can set $VT \sim t_0^4$ with t_0 as the present age of the universe. Setting the prefactor A by dimensional reasons equal to $A \sim \phi^4 \sim M_{\mathrm{Pl}}^4$, the tunnelling probability is

$$P \sim (t_0 M_{\mathrm{Pl}})^4 \exp\{-8\pi^2/(3|\lambda_{eff}|)\} \sim 10^{-900} \,. \tag{15.32}$$

[2] The facts that the bubble extends with $v \sim 1$ and that the bubble is thin imply (un?)-fortunately that any potential observers will be dissolved without having the time to notice the arrival of the wall.

Despite of the enormous volume factor, $(t_0 M_{\mathrm{P}l})^4 \sim (8 \times 10^{60})^4$ the tunnelling probability is exceedingly small.

Although the SM vacuum is unstable when extrapolated to the Planck scale, the probability that inside the past-light cone of the observed universe a tunnelling has happened is practically zero. This corresponds to the "metastable region" of the SM shown in Fig. **??**.

15.3. Topological defects

A rather generic consequence of phase transitions is the formation of extended solutions to the field equations which are stable by virtue of a topological quantum number. The latter arises if the space of possible vacuum field configurations consists of two or more subspaces which cannot be connected by solutions with a finite classical action. We call two field configurations ϕ_1 and ϕ_2 topologically equivalent, if they can be transformed continuously into each other keeping the the potential energy and the Euclidean action finite during the transformation. Thus topologically equivalent field configurations form equivalence classes which can be characterised by a topological quantum number and are separated by an infinite energy barrier. Within each equivalence class, the minimal energy solution will be stable. The spatially uniform ground-state we have assumed up to now as our vacuum in perturbative calculations may thus be disconnected from other stable solutions which cannot spread out to become spatially uniform due to their non-trivial topology.

As a system crosses its critical temperature cooling down, it performs a transition from its symmetric state to a state with broken symmetry. While the correlation length $\xi = \langle \phi(0)\phi(x) \rangle$ becomes formally infinite at the phase transition, the order parameter in the broken phase can take the same value only in a finite volume, restricted by the finite propagation speed of the relevant waves. In the case of the expanding Universe, we expect the formation of order one topological defect per causally connected region. This process was first suggested by Kibble, and is therefore called Kibble mechanism. In a big bang model, it follows thus $\xi(t) < t$ with t as the age of the universe. The number density n of topological defects created in a phase-transition at time t is therefore bounded from below as $n > \xi^{-3} \sim t^{-3}$.

Domain walls are two-dimensional topological defects which separate three-dimensional volumes containing a different vacua. Examples are ferromagnets where domains of uniform magnetisation exist. Depending on the gauge group and the pattern of symmetry breaking, topological defects with one and zero dimensions are also possible: In the first case, an one-dimensional line or string contains a vacua different from the surrounding, while it is the second case a point-like object called monopole. We will proceed as in chapters 13 and **??**, starting with a Higgs model with a single scalar field and increasing then the number of scalar fields. At the same, the dimensionality of the topological defects formed will decrease. As start, we will consider however the Sine-Gordon model which is defined in 1+1 space-time dimensions.

Sine-Gordon solitons We have chosen in general as potential $V(\phi)$ a polynomial in the field ϕ. In the case of SSB, the periodicity of the angular variable in $\phi = \eta e^{i\vartheta}$ leads to a periodicity of the potential. Another example for a Lagrangian with a periodic potential is the Sine-Gordon model defined by

$$\mathscr{L} = \frac{1}{2}(\partial_\mu \phi)^2 - V(\phi) = \frac{1}{2}(\partial_\mu \phi)^2 - \frac{a}{b^2}[1 - \cos(b\phi)] \tag{15.33}$$

with $\mu = \{0, 1\}$ and two real parameters a and b. The potential $V(\phi)$ has zeros for $\phi = 2\pi n/b$. Expanding $V(\phi)$ for small ϕ,

$$\mathscr{L} = \frac{1}{2}(\partial_\mu \phi)^2 - \frac{1}{2}a\phi^2 - \frac{ab^2}{4!}\phi^4 + \dots \tag{15.34}$$

and comparing to our usual $\lambda\phi^4$ potential, we see that they agree for small ϕ if we identify[3] $a = m^2$ and $b^2 = \lambda$.

From the Lagrangian (15.33), the Sine-Gordon equation

$$\frac{\partial^2\phi}{\partial t^2} - \frac{\partial^2\phi}{\partial x^2} + \frac{a}{b}\sin(b\phi) = 0 \tag{15.35}$$

follows. It admits static and travelling wave solutions, $\phi(x,t) = f(\pm(x - vt)) \equiv f(\pm\xi)$. You can check that

$$f(\xi) = \frac{4}{b}\arctan[\exp(\pm\gamma\sqrt{a/b}\xi)] \tag{15.36}$$

is a solution with $\gamma = (1 - v^2)^{-1/2}$ as usual Lorentz factor (problem 16.4). The solution with the plus sign is called a kink, the one with the minus an anti-kink. The kink interpolates between the $n = 0$ ground-state at $x \to -\infty$ and $n = 1$ ($\phi = 2\pi/\sqrt{\lambda}$) at $x \to \infty$. The extension ℓ of the kink is determined by the argument of the arctan and is given by $\ell = \sqrt{b}/(\gamma\sqrt{a}) = \lambda/(\gamma m)$, cf. Fig. 15.4.

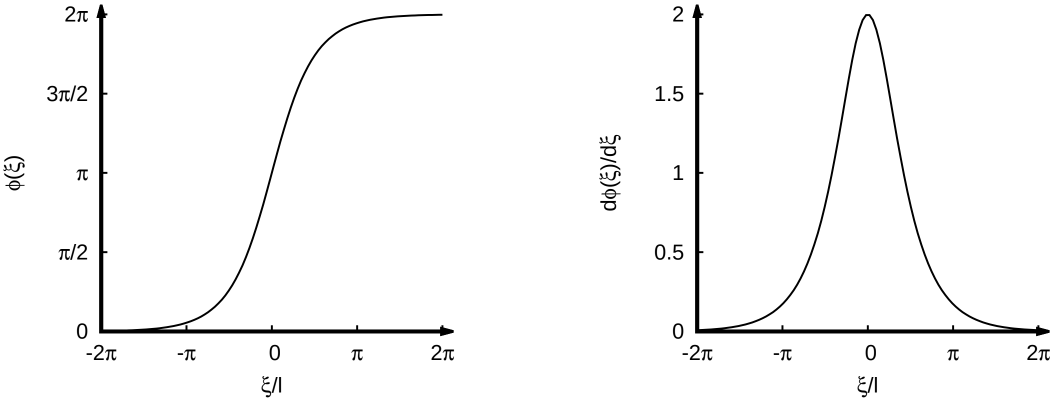


Figure 15.4.: Left: Kink solution $\phi(\xi)$ for $a = b = 1$; Right: its derivative $\mathrm{d}\phi(\xi)/\mathrm{d}\xi$.

The energy (or mass) of a static solution interpolating between $n = 0$ and $n = 1$ is

$$E = \int_{-\infty}^{\infty} \mathrm{d}x \left[\frac{1}{2}\left(\frac{\partial\phi}{\partial x}\right)^2 + V(\phi)\right] = \int_{-\infty}^{\infty} \mathrm{d}x \; 2V(\phi) = \int_0^{2\pi/b} \mathrm{d}\phi \, \frac{\partial x}{\partial\phi} V(\phi) =$$
$$= \int_0^{2\pi/b} \mathrm{d}\phi \, \sqrt{2V(\phi)} = \frac{\sqrt{2a}}{b^2}\int_0^{2\pi} \mathrm{d}\vartheta \, [1 - \cos(\vartheta)]^{1/2} = \frac{8m}{\lambda}\,. \tag{15.37}$$

[3] Recall that a scalar field has dimension $[\phi] = (d-2)/2$ in d space-time dimensions. Thus the argument of $\cos(b\phi)$ is dimensionless while the ϕ^4 interaction requires a m^2 prefactor.

Here we used in the second step the field equation together with the requirement that the energy E is finite, i.e. Eq. (15.18) with $c = 0$. More generally, a static solution connecting n_1 and n_2 has the energy $E = 8|n_1 - n_2|m/\lambda$. Thus the energy spectrum of these solutions is discrete and inverse proportionally to the coupling constant. The latter property implies that these solutions are not accessible in a perturbative calculation which is a power series in the coupling constant λ. In the weak coupling regime, the mass $8m/\lambda$ of the kink is much larger than the mass m of the elementary field ϕ, while in the strong coupling limit the kink is the lightest excitation.

We can imagine the Sine-Gordon model as a chain of arrows, which the force $F = -\partial_x V$ tries to orient e.g. downwards. Then a kink is a solution which points from $x = -\infty \ldots - \ell$ downwards, turning at $\xi \sim 0$ by 2π, and pointing at $x \gtrsim \ell$ again downwards. To untwist the chain, we would have to turn the arrows on one of the two sides of $\xi \approx 0$ which would cost an infinite amount of energy. Therefore the winding number $\nu \equiv n_1 - n_2$ is a conserved quantum number and the solutions are stable. We can understand why the winding number is called a *topological* quantum numbers as follows: Identifying the points $x = -\infty$ and $x = \infty$, we map $\mathbb{R}$ on a compact interval. Then the kink becomes a Möbius band, which cannot be smoothly transformed into a circle S^1.

A conserved quantum number implies the existence of a conserved current: The condition $E < \infty$ requires that the field approaches for $x \pm \infty$ one of the vacuum states, and thus

$$\phi(\infty) - \phi(-\infty) = \frac{2\pi}{\sqrt{\lambda}}\,\nu\,. \tag{15.38}$$

We can rewrite this is an integral,

$$\int_{-\infty}^{\infty} \mathrm{d}x\, \partial_x \phi = \frac{2\pi}{\sqrt{\lambda}}\,\nu\,. \tag{15.39}$$

Since the solution is two-dimensional, we can set as current

$$j^\mu = \frac{\sqrt{\lambda}}{2\pi}\varepsilon^{\mu\nu}\partial_\nu \phi\,. \tag{15.40}$$

The antisymmetry of $\varepsilon^{\mu\nu}$ ensures then that $\partial_\mu j^\mu = 0$. The conserved charge follows with $\varepsilon^{01} = 1$ as

$$Q = \int \mathrm{d}x\, j^0 = \frac{\sqrt{\lambda}}{2\pi}\int \mathrm{d}x\, \partial_x\phi = \frac{\sqrt{\lambda}}{2\pi}[\phi(\infty) - \phi(-\infty)] = n_1 - n_2\,, \tag{15.41}$$

i.e. it equals the winding number ν. In contrast to the conserved charges we have encountered up to now, the current is not the Noether current of a global symmetry and we did not have to use the equations of motion in its derivation. Instead, the charge has a topological origin.

The solutions of the Sine-Gordon equation maintain their shape, although the equation is non-linear. Classical solutions with this property are called *solitons*. The distinctive property of the Sine-Gordon solitons in 1+1 space-time dimensions is that this holds also for the asymptotic regions of a scattering event: In particular, two solitons emerge from a collision unchanged except possibly for a phase shift, although the superposition principle is not valid for non-linear equations of motion.

Domain walls We now move to topological defects in $3+1$ dimensions considering two-dimensional topological defects called domain walls. Assuming that the domain wall is uniform in the $y-z$ plane, its energy (per area) is

$$E[\phi] = \int_{-\infty}^{\infty} \mathrm{d}x \left[\frac{1}{2}(\phi')^2 + V(\phi)\right] . \tag{15.42}$$

We assume that the potential energy $V(\phi)$ is shifted such that $\min(V(\phi)) = 0$. Then the energy is the sum of two (semi-) positive terms, $E[\phi] \geq 0$. Instead of integrating the equation of motions, as in Eqs. (15.17) to (15.19), we will use now an argument due to Bogomolnyi: He suggested to rewrite the energy "completing the square" as

$$E[\phi] = \frac{1}{2}\int_{-\infty}^{\infty} \mathrm{d}x \left[\phi' \mp \sqrt{2V(\phi)}\right]^2 \pm \int_{\phi(-\infty)}^{\phi(\infty)} \mathrm{d}\tilde{\phi}\sqrt{2V(\tilde{\phi})} . \tag{15.43}$$

The second integral depends only on the boundary values of the field at infinity. For field configurations which connect the same ground-state at $x=-\infty$ and $+\infty$, this boundary term vanishes, and the minimum $E[\phi]=0$ of the energy is attained for constant fields. If the field configurations connect however different ground-states, then we can associate the second integral to a non-zero topological charge. The winding number in the Sine-Gordon case is a specific example for such a topological charge.

A lower bound for the energy of fields with non-zero topological charge is

$$E[\phi] \geq \left| \int_{\phi(-\infty)}^{\phi(\infty)} \mathrm{d}\tilde{\phi}\sqrt{2V(\tilde{\phi})} \right| , \tag{15.44}$$

which is attained when the first integral vanishes. In this case, the field satisfies a first-order equation $\phi' = \pm\sqrt{2V(\phi)}$ which is generally much easier to solve than the original second-order equation of motion. Separating again variables, we arrive at

$$x = \pm \int \frac{\mathrm{d}\phi}{\sqrt{2V(\phi)}} . \tag{15.45}$$

To proceed, we have to choose a definite potential. For the choice (15.20), we come back to the solution

$$\phi_{\mp}(x) = \mp\eta \tanh\left(\frac{\eta\sqrt{\lambda}}{\sqrt{2}}(x-x_0)\right) \tag{15.46}$$

which we have used in the calculation of the bounce. The solution $\phi_{\mp}$ interpolates between the two vacua $\pm\eta$ of the symmetric potential $V_0(\phi)$ and contains at $x=x_0$ a two-dimensional plane with the unbroken vacuum $\phi=0$. More generally, the two-dimensional surfaces which separate domains with opposite values of η can have a finite size.

Global cosmic strings We continue our way through possible topological defects looking at the consequences of a global continuous symmetry. Thus we replace the single real field by a set of two real or one complex field $\phi = \phi_1 + \mathrm{i}\phi_2$ with potential

$$V(\phi) = \frac{\lambda}{4}\left(\phi^\dagger\phi - \eta^2\right)^2 . \tag{15.47}$$

We use cylinder coordinates ρ, ξ, z, and search for static solutions using as ansatz

$$\phi = \eta \mathrm{e}^{\mathrm{i}n\vartheta} f(\rho) . \tag{15.48}$$

The phase $\mathrm{e}^{in\vartheta}$ can be chosen arbitrarily at $\rho = 0$, unless the field is zero at the origin. In order to ensure a single-valued field, we have therefore to impose the boundary condition $f(\rho) \to 0$ for $\rho \to 0$. In the opposite limit, $\rho \to \infty$, the field should approach one of its minima,

$$\phi \to \eta \mathrm{e}^{in\vartheta} \quad and \quad f(\rho) \to 1 , \tag{15.49}$$

in order to minimise the energy.

For the ansatz (15.48), the field equations are separable and we find from $\Delta\phi = \lambda(\phi^2 - \eta^2)\phi$ as differential equation for f

$$\frac{\mathrm{d}^2 f}{\mathrm{d}\xi^2} + \frac{1}{\xi}\frac{\mathrm{d}f}{\mathrm{d}\xi} - \frac{n^2}{\xi^2} f = \xi(f^2 - 1) f . \tag{15.50}$$

Here we have introduced also the new dimensionless variable $\xi = \lambda^{1/2}\eta\rho$. This differential equation together with the two boundary conditions for $\rho \to 0$ and $\rho \to \infty$ has to be solved numerically.

We can however estimate the extension of a cosmic string by dimensional arguments. The scale of the problem is set by $\rho/\xi = \lambda^{-1/2}\eta^{-1}$. Thus the core containing the unbroken vacuum $\phi = 0$ has the radius $\mathcal{O}(\lambda^{-1/2}\eta^{-1})$, while for larger distances $\rho \gg \lambda^{-1/2}\eta^{-1}$ the field approaches the broken phase $|\phi| = \eta$. The extended solution should have a finite energy E per length L,

$$\frac{E}{L} = \int_0^\infty \mathrm{d}\rho\rho \int_0^{2\pi} \mathrm{d}\vartheta \left[|\boldsymbol{\nabla}\phi|^2 + V(\phi) \right] \tag{15.51a}$$

$$= \int_0^\infty \mathrm{d}\rho\rho \int_0^{2\pi} \mathrm{d}\vartheta \left[\partial_\rho\phi^\dagger \partial_\rho\phi + \frac{1}{\rho^2}\frac{\partial\phi^\dagger}{\partial\vartheta}\frac{\partial\phi}{\partial\vartheta} + V(\phi) \right] . \tag{15.51b}$$

The first and the last term in (15.51b) give a finite contribution to E/L, since $\partial_\rho f(\rho) \to 0$ and $V(\phi) \to 0$ for $\rho \to \infty$. In contrast, the middle term contributes $\propto \int_0^R \mathrm{d}\rho\rho^{-1} f(\rho)$, i.e. a logarithmically diverging term $\ln(R/\delta)$ to the linear energy density. The scale δ has to be determined by the typical extension of the string, and thus the energy density inside the radius R around the string is $E/L \sim \ln[R/(\lambda^{1/2}\eta)]$. If we consider instead of the idealisation of an isolated global string the realistic case of a string network, then R should be given by the typical distance of strings.

Local cosmic strings We can avoid the (formal) problem of the infinite energy associated with a string if we gauge the model. In this case we obtain the abelian Higgs model,

$$\mathscr{L} = -\frac{1}{4}F_{\mu\nu}F^{\mu\nu} + (D_\mu\phi)^\dagger(D^\mu\phi) + \mu^2\phi^\dagger\phi - V(\phi) . \tag{15.52}$$

The kinetic term $D_\mu\phi = \partial_\mu\phi + \mathrm{i}eA_\mu\phi$ contains now two contributions which can cancel for $\rho \to \infty$. We require therefore that A_μ is a pure gauge field for $\rho \to \infty$ with

$$A_\mu \sim -\frac{\mathrm{i}}{e}\,\partial_\mu \ln(\phi/\eta) . \tag{15.53}$$

Then $D_\mu\phi$ and $F_{\mu\nu}$ approach zero for $\rho \to \infty$, and the energy density per length of a local string is finite.

A local string carries magnetic flux Φ of the corresponding gauge field. Integrating (15.53) around a circle in a plane with $z = const.$ and large ρ gives

$$\Phi = \int \boldsymbol{B} \cdot \mathrm{d}\boldsymbol{S} = \oint \boldsymbol{A} \cdot \mathrm{d}\boldsymbol{l} = -\frac{\mathrm{i}}{e}\int_0^{2\pi} \mathrm{d}\vartheta\, \partial_\vartheta(\mathrm{i}n\vartheta) = \frac{2\pi n}{e}\,. \tag{15.54}$$

Thus the magnetic flux carried by local strings is quantised in units of $2\pi/e$. If we go back to cgs units, then the unit of magnetic flux becomes hc/e, determined completely by the three fundamental constants of a relativistic quantum theory. A local cosmic string with winding number n carries n flux quanta, analogous to quantised tubes of magnetic flux in a superconductor.

Global monopoles Let us try to condense our results before we proceed: If we split complex scalar fields intro real ones, $\phi = \phi_1 + \mathrm{i}\phi_2$, then we considered in $d = 4$ potentials of the type

$$V(\phi) = \frac{\lambda}{4}\left(\sum_{i=1}^{n} \phi_i^2 - \eta^2\right)^2 . \tag{15.55}$$

For $n = 1$, the possible ground-states $\phi^2 = \eta^2$ correspond to the zero-dimensional sphere S^0. The single constraint $\phi(x_1, x_2, x_3) = 0$ defines a two-dimensional surface in $\mathbb{R}^3$, which corresponds to a domain wall. The abelian Higgs model with one complex scalar doublet has $n = 2$. Now $\phi_1^2 + \phi_2^2 = \eta^2$ defines an one-dimensional sphere S^1 as manifold of the possible ground-states, while a topological defect defined by $\phi_1 = \phi_2 = 0$ is as the section of two surfaces a line. These results are summarised in the first two entries of Table 15.1.

The logical next step is to consider three scalar fields which transform under SO(3). Then the vacuum manifold is the sphere S^2 and we expect zero-dimensional topological defects which are called monopoles. We start again examining the global case. Using spherical coordinates, we search for static solutions using as ansatz

$$\phi_i = \eta h(r)\frac{x_i}{r}\,, \tag{15.56}$$

which satisfies the requirement

$$\phi_i\phi_i = \eta^2 \text{ for } |\boldsymbol{x}| \to \infty\,, \tag{15.57}$$

if $h(r) \to \pm 1$ for $r \to \infty$. Additionally, the function $h(r)$ should satisfy the boundary condition $h(r) \to 0$ for $r \to 0$ to ensure a non-singular $\phi(0)$. Estimating again the energy density at

defect	n	d	homotopy group
domain wall	1	2	$\pi_0(M_0)$
cosmic string	2	1	$\pi_1(M_0)$
monopole	3	0	$\pi_2(M_0)$
texture	4	–	$\pi_3(M_0)$

Table 15.1.: The number n of scalars determines the dimension d of the vacuum manifold M_0 and the dimension $3-n$ of the hypersurface containing the unbroken vacuum.

large r by simple dimensional analysis gives

$$\rho \sim \frac{1}{2}(\partial_i \phi)^2 \sim \frac{3}{2}\frac{\eta^2}{r^2} \tag{15.58}$$

and thus the energy contained inside a sphere of radius R diverges linearly, $E \approx 3\eta^2 R/2$. In contrast to the mild logarithmic divergence in case of global strings, the behaviour $E \approx 3\eta R/2$ is clearly disastrous.

The natural way-out is to go on to local monopoles and to check if the gauge fields can compensate the scalar gradient energy. Before we do so, we ask however if there is a way to ensure that the total energy of global monopoles is finite. Imagine a monopole ($h(r) = 1$ for $r \to \infty$) and an anti-monopole ($h(r) = -1$ for $r \to \infty$) pair separated by the distance d: Their fields $\phi_i(r)$ will cancel for $r \gg d$, leaving over only higher multipole moments $\phi_i \sim r^{-2}$. Thus the energy density of a monopole-antimonopole pair scales as $\rho \sim \eta r^{-4}$ and therefore their total energy E is finite. Separating a monopole-antimonopole pair would cost an infinite amount of energy—instead a new monopole-antimonopole pair will be created as soon as the potential energy between the pair exceeds a certain threshold. We can view this behaviour therefore as a model for the confinement of coloured particles as quarks and gluons in QCD, where the potential energy $V(r)$ for distances $r \gtrsim \Lambda_{QCD}^{-1}$ scales also linearly.

't Hooft-Polyakov monopoles The simplest model exhibiting local monopoles is the Georgi-Glashow model. It contains a SO(3) gauge field A_μ^a supplemented by a triplet of real Higgs scalars ϕ^a. Choosing a uniform vev as $\phi^a = (0, 0, v)$ leaves a residual U(1) symmetry unbroken, corresponding to rotations around the 3-axis in isospin space. Thus one can view the Georgi-Glashow model as a toy model for the electroweak sector of the SM, where the gauge field A_μ^3 plays the role of the photon and the Z boson is missing.

't Hooft and Polyakov showed first that this model contains extended classical solutions which have finite energy and correspond to local magnetic monopoles. Using the adjoint representation $(T^a_{\rm adj})_{bc} = -{\rm i} f^{abc}$ for the scalar triplet of real Higgs scalars ϕ^a, the covariant derivative becomes

$$D_i\phi^a = \partial_i \phi^a - e\varepsilon^{abc}A_i^b\phi^c \tag{15.59}$$

or in vector notation $D_i\boldsymbol{\phi} = \partial_i\boldsymbol{\phi} - e\boldsymbol{A}_i \times \boldsymbol{\phi}$. Here, we assumed for simplicity that the monopole carries no electric charge, setting A_0^a. We fix again the asymptotic behaviour of the gauge fields by requiring that the kinetic term $D_i\phi^a$ vanishes for $r \to \infty$. From

$$\partial_i\phi^a \sim \eta \frac{\delta^{ai} - x^i x^a}{r^2},$$

we conclude that A_i^b should be constant, while (15.59) implies that $\boldsymbol{A}_i$ is perpendicular to $\boldsymbol{\phi}$. Evaluating $D_i\phi^a$ using $\varepsilon^{aij}\varepsilon^{akl} = \delta^{ik}\delta^{jl} - \delta^{il}\delta^{jk}$ shows that

$$A_i^a = \frac{\eta}{e}[1 - f(r)]\varepsilon^{aij}\frac{x^j}{r} \tag{15.60}$$

with $f(r) \to 0$ for $r \to \infty$ leads to the desired asymptotic behaviour of the covariant derivative. Now the regularity of the solution A_i^a at $r = 0$ requires $f(0) = 1$.

Inserting the two Ansätze (15.56) and (15.60) into the energy functional of the Georgi-Glashow model and minimising the energy results in two coupled differential equations for the functions $h(r)$ and $f(r)$. In order to be be able to proceed analytically, we consider only the limit $\lambda/g^2 = m_h^2/(2m_W^2) \to 0$. Then the potential energy $V(\phi)$ is negligible,

$$V(\phi) = \frac{\lambda}{4}\left(\sum_{a=1}^{3}\phi_a^2 - \eta^2\right)^2 \to 0 \qquad \text{for} \qquad \lambda \to 0\,, \tag{15.61}$$

if the fields satisfy the constraint (15.57). Thus the boundary conditions (15.57) remain valid in the limit considered. Neglecting $V(\phi)$, we can use Bogomolnyi's trick to derive an exact solution. We express the static energy as

$$E = \frac{1}{2}\int \mathrm{d}^3x\,\left[(B_i^a)^2 + (D_\mu\phi^a)^2\right] = \frac{1}{2}\int \mathrm{d}^3x\,\left[(B_i^a \mp D_i\phi^a)^2 \pm 2B_i^a D_i\phi^a\right] \geq 0\,. \tag{15.62}$$

Then we obtain the bound

$$E \geq \int \mathrm{d}^3x\, B_i^a D_i\phi^a \tag{15.63}$$

which is attained, if the fields satisfy $D_i\phi^a = \pm B_i^a$. Now we have to solve only the simpler first-order equation $D_i\phi^a = \frac{1}{2}\varepsilon_{ijk}F_{jk}^a$. Its solution is called a Bogomolnyi-Parad-Sommerfield (BPS) monopole and, generalising, all solitons which minimise the classical action are denoted as BPS states or BPS solitons.

Example 15.2: Find the solution of the first-order equation $D_i\phi^a = \pm B_i^a$.
We evaluate with (15.56) the LHS,

$$D_i\phi^a = \eta\frac{\delta^{ai} - x^i x^a}{r}fh + \eta\frac{x^i x^a}{r^2}h' \tag{15.64}$$

and with (15.60) the RHS,

$$B_i^a = \frac{1}{2}\varepsilon_{ijk}F_{jk}^a = \frac{1}{2}\varepsilon_{ijk}\left[\partial_j A_k^a - \partial_k A_j^a - e\varepsilon_{abc}A_j^b A_k^c\right] \tag{15.65a}$$

$$= \varepsilon_{ijk}\partial_j A_k^a - \frac{1}{2}e\varepsilon_{ijk}\varepsilon_{abc}A_j^b A_k^c = \frac{1}{g}\frac{r^2\delta^{ai} - x^i x^a}{r^2}f' + \frac{1}{g}\frac{x^i x^a}{r^3}(1-f^2)\,. \tag{15.65b}$$

Since the two tensors $r^2\delta^{ai} - x^i x^a$ and $x^i x^a$ are orthogonal, the two coupled differential equations for the profile functions $h(r)$ and $f(r)$ simplify to

$$f'(r) = \pm g\eta f(r)h(r) \quad \text{and} \quad h'(r) = \frac{1}{g\eta r^2}[1 - f(r)^2]\,. \tag{15.66}$$

Taking into account the boundary conditions, the solution expressed through the dimensionless variable $\xi = g\eta r$ is given by

$$h(r) = \coth(\xi) - \frac{1}{\xi} \quad \text{and} \quad f(r) = \frac{\xi}{\sinh(\xi)}\,. \tag{15.67}$$

Finally, we want to determine the charge of a BPS monopole. While the massive fields fall off exponentially for $r \to \infty$, the component of F_{ij}^a connected to the massless photon decrease as a power-law. Moreover, the fields (15.56) and (15.60) are practically uniform at large distances r from the center of the monopole. Thus in a laboratory at $\boldsymbol{x}$, the field $\boldsymbol{F}_{ij}$ corresponds to the massless photon, while the two states orthogonal to $\boldsymbol{F}_{ij}$ are the massive weak gauge bosons. This implies that the part of F^a_{ij} which corresponds to the electromagnetic field is parallel to ϕ^a,

$$\mathcal{F}_{ij} = F^a_{ij}\frac{\phi^a}{v}\,. \tag{15.68}$$

The magnetic part of the SU(2) field-strength tensor is

$$B^a_i = \frac{1}{2}\varepsilon_{ijk}F^a_{jk} = \varepsilon_{ijk}\partial_j A^a_k - \frac{1}{2}e\varepsilon_{ijk}\varepsilon_{abc}A^b_j A^c_k\,, \tag{15.69}$$

where we inserted the definition of F^a_{jk}. If we are far away from the center of the soliton, we can use the asymptotic expression for the fields, setting $f(r) = 0$ and $h(r) = 1$. Performing the differentiations (problem 16.8) we arrive at

$$B^a_i = \frac{1}{g}\frac{n_i n_a}{r^2}\,. \tag{15.70}$$

Thus the U(1) magnetic field is given by

$$\mathcal{B}_i = \frac{1}{g}\frac{n_i}{r^2}\,. \tag{15.71}$$

Comparing this to the expression $qn_i/(4\pi r^2)$ for the field of a monopole, we conclude that the magnetic charge of the BPS monopole is $q_m = 4\pi/g$.

Textures Finally we should comment on the case $n = 4$ which corresponds to the electroweak model. The four equations $\phi_i(x_1, x_2, x_3) = 0$ have in general no solution in $\mathbb{R}^3$. Thus regions which contain the unbroken vacuum $\phi_i = 0$ will be not formed during the electroweak phase transition. Still, correlations can not exist beyond the horizon scale and thus non-zero gradients $\partial_\mu\phi_i$ can be produced. In the case of global textures, static solutions with positive energy density can exist. In the case of a broken gauge symmetry, gauge fields will compensate the $\partial_\mu\phi_i$ term in $D_\mu\phi_i$. Thus local textures as predicted by the SM have no observable consequences.

Homotopy groups and winding numbers The topological quantum numbers we have met discussing extended classical solutions can be associated to winding numbers of maps between the ground states of a field theory and its configuration space. Let us define the ground-state or vacuum manifold M_0 of a theory as the set of all global minima $V(\phi) = 0$ of its potential,

$$M_0 = \{\phi : V(\phi) = 0\}\,. \tag{15.72}$$

The condition that the potential energy is finite requires that

$$\lim_{|\boldsymbol{x}|\to\infty} \phi(x) = \phi \in M_0\,. \tag{15.73}$$

We can compactify $\mathbb{R}^n$ (where n is the number of spatial dimensions) to the sphere S^n using e.g. a stereographic projection as shown in Fig. 15.5. Then we can view the condition that the potential energy is finite as a mapping $S^n \to \mathcal{M}_0$. We ask now the question when two such mappings are topologically distinguished.

We consider only the mappings $S^1 \to S^1 \approx \mathbb{R}^2\backslash\{0\}$ which are easiest to visualise. Two closed loops with base point $\boldsymbol{x}_0$, i.e. curves $\boldsymbol{x}(t)$ with $t \in [0:1]$ and $\boldsymbol{x}(0) = \boldsymbol{x}(1) = \boldsymbol{x}_0$,

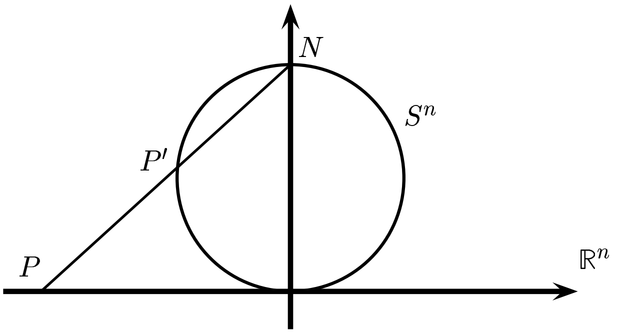


Figure 15.5.: A stereographic projection maps points $P \in \mathbb{R}^n$ onto points $P' \in S^n$. The north pole N of the sphere S^n corresponds to the sphere S_R^{n-1} at spatial infinity, $x_i^2 + \cdots + x_n^2 = R^2 \to \infty$, of $\mathbb{R}^n$.

are shown in Fig. 15.6. While the dashed loop is contractable, the solid one is wrapped once around $\{0\}$ and therefore not contractable to its base-point $\boldsymbol{x}_0$ by continuous transformations. We define the maps

$$U^{(\nu)} = \mathrm{e}^{\mathrm{i}\vartheta\nu} = \left[U^{(1)}\right]^{\nu}, \tag{15.74}$$

which count the number of times a loop is wrapped around the origin. The integer ν is called the winding number of the map and can be rewritten as an integral,

$$\nu = \frac{\mathrm{i}}{2\pi}\int_0^{2\pi} \mathrm{d}\vartheta\, U\partial_\vartheta U^\dagger . \tag{15.75}$$

Clearly, this formula reproduces the correct values for the mappings defined above. Moreover, it is useful for the proof that ν is invariant under continuous transformations. It is sufficient to investigate an infinitesimal change δU. Unitarity $UU^\dagger = 1$ implies that $\delta UU^\dagger + U\delta U^\dagger = 0$ and $\delta U^\dagger = -U^\dagger\delta UU^\dagger$. Since we will interested later in the non-abelian version of the winding number, we will not use that the U are commuting complex numbers.

We calculate the variation of the integrand in the integral formula (15.75) for ν,

$$\delta(U\partial_\vartheta U^\dagger) = \delta U\partial_\vartheta U^\dagger + U\partial_\vartheta\delta U^\dagger \tag{15.76a}$$

$$= \delta U\partial_\vartheta U^\dagger - U\partial_\vartheta U^\dagger\delta UU^\dagger - UU^\dagger\partial_\vartheta\delta UU^\dagger - UU^\dagger\delta U\partial_\vartheta U^\dagger \tag{15.76b}$$

$$= -U\left[\partial_\vartheta U^\dagger\delta U + U^\dagger\partial_\vartheta\delta U\right]U^\dagger = -U\partial_\vartheta\left[U^\dagger\delta U\right]U^\dagger . \tag{15.76c}$$

Here, we inserted first $\delta U^\dagger = -U^\dagger\delta UU^\dagger$ and performed the differentiations. Then the first and fourth terms cancel, and finally we combined the remaining two terms using the product rule. In case of the abelian winding number ν of (15.75), we obtain then

$$\delta\nu = \frac{\mathrm{i}}{2\pi}\int_0^{2\pi}\mathrm{d}\vartheta\,\delta(U\partial_\vartheta U^\dagger) = -\frac{\mathrm{i}}{2\pi}\int_0^{2\pi}\mathrm{d}\vartheta\,\partial_\vartheta\left[U^\dagger\delta U\right] = 0 . \tag{15.77}$$

Thus the winding number ν is an integer which is invariant under infinitesimal deformations of the loop. Since any continuous transformation can be built up out of infinitesimal ones,

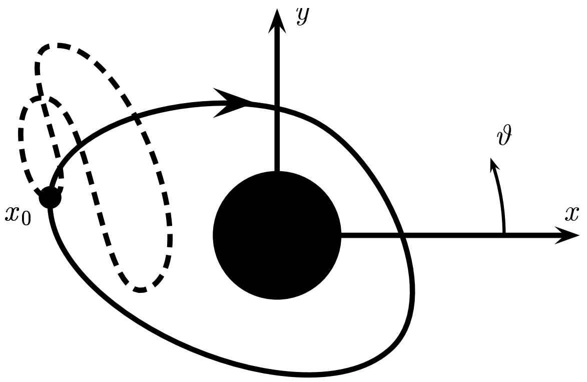


Figure 15.6.: Two loops in $\mathbb{R}^2\backslash\{0\}$ with common base point $\boldsymbol{x}_0$: the dashed line is contractable and has winding number $\nu = 0$, the solid loop is wrapped once around $\{0\}$ and has winding number $\nu = -1$.

the maps $S^1 \to S^1$ can be divided into different equivalence classes, each characterised by the value $\nu \in \mathbb{Z}$. Only maps within each class can be continuously transformed into each other. Mathematicians say that such maps are homotopic. The theory of homotopy groups addresses the question into how many homotopy classes $\pi_n(X)$ the set of maps $S^n \to X$ can be divided. If $\pi_n(\mathcal{M}_0) \neq 1$, then the theory with vacuum manifold $\mathcal{M}_0$ contains stable topological defects, cf. table 15.1. Expressed in the language of homotopy groups, we have shown that $\pi_1(S^1) = \mathbb{Z}$. Since $S^1 \simeq U(1)$, this shows also that $\pi_1(U(1)) = \mathbb{Z}$.

A general results from the theory of homotopy groups says that the second homotopy group $\pi_2(G/H)$ of the quotient group G/H equals the first homotopy group $\pi_1(H)$ of H. If we identify the SM with H, then $H = SU(3) \otimes U(1)$ and $\pi_1(U(1)) = \mathbb{Z}$ is non-trivial. Thus the sequence

$$\pi_2(G/H) = \pi_1(H) = \pi_1(SU(3) \otimes U(1)) = \mathbb{Z} \neq 1 \tag{15.78}$$

shows that, if $U_{\rm em}(1)$ is unified at a higher scale within a larger semi-simple group G, then magnetic monopole solutions exist. Thus a grand unified theory implies that magnetic monopoles are produced at the GUT phase transition. We will see later that such monopoles would overclose the universe, leading to a very short life-time of the universe. The wish to dilute the density of monopoles was a prime motivation for the invention of inflation.

Summary

Classical static solutions of theories with SSB can fall into different equivalence classes which are separated by an infinite potential energy barrier. This can lead to two-, one- or zero-dimensional topological defects which contain in their core the unbroken vacuum of the symmetric phase. Instantons or bounces are solutions of the classical field equation in Euclidean space which evolve in Euclidean time between two different vacua. If their action is finite, they describe in Minkowski space instantaneous tunnelling from the false to the true vacuum.

Further reading

A clear account of topologically non-trivial solutions of the classical field equations is given by [Rub02]. For a calculation of the effective potential of the SM and the resulting tunnelling probability see [Esp14] and the references therein.

Problems

15.1 Order of the phase transition.
Determine the critical temperature T_c of the $\lambda\phi^4$ theory from Eq. (15.7). Calculate the pressure and the heat capacity above and below T_c and confirm thereby that the phase transition is of second order.

15.2 Field equation for the bounce solution.
Show that Eq. (15.15) agrees with the Klein-Gordon equation $\Delta\phi = V'(\phi)$ for a radial symmetric ϕ. (Use Eq. (6.61) to evaluate $\Delta\phi = \nabla^a\nabla_a\phi$.)

15.3 Scalar instantons.
Show that the ansatz $\phi(r) = A\rho/(r^2 + \rho^2)$ solves (15.15). Determine the action S.

15.4 Soliton solution of the Sine-Gordon equation.
Show that Eq. (15.36) solves the Sine-Gordon equation and discuss the behaviour under Lorentz transformations. .

15.5 Domain wall
Estimate the extension of a domain wall using dimensional analysis.

15.6 Derrick's theorem.
Consider the behaviour of the energy functional $E[\phi] = T[\phi] + V[\phi]$ for a scalar field ϕ under scale transformation $x \to \lambda x$ in D space dimensions. Show that for $D \geq 2$ no stable solutions with finite energy exist.

15.7 D_μ for the adjoint representation.
Show that $D_\mu = \partial_\mu + \mathrm{e}A_\mu$ results for SO(3) in Eq. (15.59) for the adjoint representation.

15.8 Duality of Maxwell
Show that the source-free Maxwell equations are invariant under the duality transformation $E' = E\cos\alpha + B\sin\alpha$ and $B' = -E\sin\alpha + B\cos\alpha$. Show that the invariance of the Maxwell equations with sources requires the existence of magnetic monopoles.

15.9 Magnetic field of a monopole
Derive Eq. (15.70).

16. Anomalies, instantons and axions

At the classical level, any global continuous symmetry of a system described by a Lagrangian leads to a locally conserved current. In order to decide if these symmetries do survive quantisation, we have to study if the generating functional $Z[J]$ retains the symmetries of the classical Lagrangian $\mathscr{L}$. Several examples where a quantised system does not share the symmetries of its classical counter-part have been found. As this behaviour came as a surprise, it was called "anomalous" and the non-zero terms violating on the quantum level the classical conservation laws were called "anomalies". A case we encountered already is the breaking of scale invariance (or conformal symmetry) in the process of renormalising massless theories; we will calculate this anomaly in the next chapter. The only other example of anomalous theories in $d = 4$ space-time dimensions are models containing chiral fermions, i.e. theories where left- and right-chiral fermions interact differently. This anomaly is called axial or chiral anomaly and shows up in loop graphs containing chiral fermions coupled to gauge fields or gravitons. In the SM, the axial anomaly leads to three important phenomenological consequences: First, the anomaly in the electroweak currents vanishes, if the electric charges of all left-handed fermions sum up to zero. Thus the fact that electroweak Ward identities hold restricts the particle content of the theory. Second, the axial anomaly in conjunction with topologically non-trivial solutions of the Yang-Mills equations leads to the violation of CP invariance in strong interactions, in contradiction to observations. This is the so-called "strong CP problem". Finally, the violation of CP is a necessary condition for the generation of a baryon asymmetry, and thus these topologically non-trivial solutions play an important role in models of baryogenesis.

The fact that a classical symmetries is broken by quantum effects is often described as "the anomaly breaks the classical symmetry". One should keep in mind that on the quantum level there is no symmetry to start with. Thus there is no Goldstone boson associated to a symmetry broken by an anomaly.

16.1. Axial anomalies

Anomaly from non-invariance of the path integral measure The simplest model exhibiting an axial anomaly is axial electrodynamics, i.e. a fermion coupled via its vector and axial current to two different gauge fields. The Lagrangian for this system reads

$$\mathscr{L} = \bar{\psi}\mathrm{i}\gamma^\mu(\partial_\mu + \mathrm{i}qV_\mu + \mathrm{i}gA_\mu\gamma^5)\psi - \frac{1}{4}F^2 - \frac{1}{4}G^2 \tag{16.1}$$

with

$$F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu \quad and \quad G_{\mu\nu} = \partial_\mu V_\nu - \partial_\nu V_\nu \,. \tag{16.2}$$

Performing a combined $\mathrm{U}_V(1) \otimes \mathrm{U}_A(1)$ gauge transformation,

$$V_\mu \to V'_\mu = V_\mu - \partial_\mu\Lambda(x) \quad and \quad A_\mu \to A'_\mu = A_\mu - \partial_\mu\alpha(x)\,, \tag{16.3}$$

induces the following change of the fermion fields,

$$\psi(x) \to \psi'(x) = \exp\{\mathrm{i}q\Lambda(x) + \mathrm{i}g\alpha(x)\gamma^5\}\psi(x)\,, \tag{16.4a}$$

$$\bar\psi(x) \to \bar\psi'(x) = \bar\psi(x)\exp\{-\mathrm{i}q\Lambda(x) + \mathrm{i}g\alpha(x)\gamma^5\}\,. \tag{16.4b}$$

In order to determine the transformation properties of the fermionic path integral measure, we introduce first eigenfunctions of the gauge-invariant Dirac operator $\mathrm{i}\slashed{D}$,

$$\mathrm{i}\slashed{D}\phi_n = \lambda_n\phi_n \quad \mathit{and} \quad \sum_n \phi_x(x)\phi_n^\dagger(y) = \delta(x-y)\,\mathbf{1}\,. \tag{16.5}$$

Then we can expand the fermion fields as

$$\psi(x) = \sum_n a_n\phi_n(x) \quad \mathit{and} \quad \bar\psi(x) = \sum_n \phi_n^\dagger(x)\bar b_n\,, \tag{16.6}$$

where the coefficients a_n and $\bar b_n$ are Graßmann variables. Thus we can rewrite the path integral measure as $\mathcal{D}\psi\mathcal{D}\bar\psi = \prod_n \mathrm{d}a_n\,\mathrm{d}\bar b_n$. Next we relate the change in the fermion fields to a change in the expansion coefficients,

$$\psi'(x) = \sum_n a_n'\phi_n(x) = \sum_n a_n \exp\{\mathrm{i}q\Lambda(x) + \mathrm{i}g\alpha(x)\gamma^5\}\phi_n(x) \tag{16.7a}$$

$$\bar\psi'(x) = \sum_n \phi_n^\dagger(x)\bar b_n' = \sum_n \phi_n^\dagger(x)\exp\{-\mathrm{i}q\Lambda(x) + \mathrm{i}g\alpha(x)\gamma^5\}\bar b_n\,. \tag{16.7b}$$

Thus the variation introduced by a vector $\mathrm{U}_V(1)$ gauge transformation cancels in the fermionic measure $\mathcal{D}\psi\mathcal{D}\bar\psi$, while the corresponding change under the axial $\mathrm{U}_A(1)$ gauge transformation adds up. We can determine the latter taking into account that the ϕ_n are an orthonormal basis as

$$a_m' = \sum_n \int \mathrm{d}^4x\,\phi_m^\dagger(x)\exp\{\mathrm{i}g\alpha(x)\gamma^5\}\phi_n(x)a_n \equiv C_{mn}a_n\,. \tag{16.8}$$

Recalling the transformation rules for Graßmann integrals, $\mathcal{D}\psi \to \mathcal{D}\psi\mathcal{J}^{-1} = \mathcal{D}\psi[\mathrm{D}et(C)]^{-1}$, the product $\mathcal{D}\psi\mathcal{D}\bar\psi$ of the fermionic measure changes as

$$\mathcal{D}\psi\mathcal{D}\bar\psi \to \mathcal{D}\psi\mathcal{D}\bar\psi[\mathrm{D}et(C)]^{-2} = \mathcal{D}\psi\mathcal{D}\bar\psi\exp\left(-2\mathrm{i}g\sum_n\int \mathrm{d}^4x\;\alpha(x)tr\;\phi_n^\dagger(x)\gamma^5\phi_n(x)\right)\,. \tag{16.9}$$

Here we used also the identity $\mathrm{D}et\exp(A) = \exp\mathrm{T}r(A)$. The sum contains a trace over the spinor indices which is zero, but also a divergent sum over the eigenfunctions of $\mathrm{i}\slashed{D}$. Therefore we have to regularise the expression. We add as gauge-invariant regulator the function

$$f = \exp\{-\lambda_n^2/M^2\} = \exp\left\{\slashed{D}^2/M^2\right\}, \tag{16.10}$$

which approaches zero fast for $\lambda_n^2 \to \infty$. The limit $f \to 1$ of the regulator corresponds to $M \to \infty$, which will allow us later an expansion of our result for

$$\mathcal{A}(x) \equiv \sum_n \mathrm{tr}\left\{\phi_n^\dagger(x)\gamma^5\exp\left(\slashed{D}^2/M^2\right)\phi_n(x)\right\} \tag{16.11}$$

in powers of $1/M^2$. Note that we have expressed the change of the measure as an additional term $\delta\mathscr{L} = 2g\alpha(x)\mathcal{A}(x)$ in the Lagrangian.

As the vector $\mathrm{U}_V(1)$ gauge transformation keeps the integration measure invariant, we ignore it in the following calculations. Thus we need to calculate $\not{D}^2$ including only the axial gauge field A,

$$\not{D}\not{D} = \gamma_\mu\gamma_\nu D^\mu D^\nu = D^2 + \frac{g}{2}\sigma_{\mu\nu}F^{\mu\nu} \tag{16.12}$$

with

$$D^2 = (\partial_\mu + \mathrm{i}gA_\mu)(\partial^\mu + \mathrm{i}gA^\mu) = \partial^2 + 2\mathrm{i}gA^\mu\partial_\mu + \mathrm{i}g\big(\partial_\mu A^\mu\big) - g^2 A_\mu A^\mu\,. \tag{16.13}$$

We evaluate D^2 using for ϕ_n plane-waves, $\phi_n = \mathrm{e}^{-ikx}\,\mathbf{1}$. Then we can replace the differentiations by momenta,

$$\exp\left(\frac{D^2}{M^2}\right)\phi_n(x) = \exp\left(-\frac{(k_\mu + gA_\mu)^2}{M^2} + \frac{\mathrm{i}g(\partial_\mu A^\mu)}{M^2}\right)\phi_n(x)\,. \tag{16.14}$$

Taking the continuum limit of the sum and writing out the full regulator, we obtain

$$\mathcal{A}(x) = \int\frac{\mathrm{d}^4k}{(2\pi)^4}\,tr\left\{\mathrm{e}^{\mathrm{i}kx}\gamma^5\exp\left[\not{D}^2/M^2\right]\mathrm{e}^{-\mathrm{i}kx}\right\} \tag{16.15a}$$

$$= \int\frac{\mathrm{d}^4k}{(2\pi)^4}\,tr\left\{\gamma^5\exp\left[-\frac{(k_\mu + gA_\mu)^2}{M^2} + \frac{\mathrm{i}g\partial_\mu A^\mu}{M^2} + \frac{g}{2}\frac{\sigma_{\mu\nu}F^{\mu\nu}}{M^2}\right]\right\}\,. \tag{16.15b}$$

The second term in $\{\cdots\}$ is zero because it does not depend on k^μ and $tr[\gamma^5] = 0$. Shifting variables to the dimensionless $MK_\mu = k_\mu + gA_\mu$ we obtain

$$\mathcal{A}(x) = M^4\int\frac{\mathrm{d}^4K}{(2\pi)^4}\,\mathrm{e}^{-K^2}tr\left[\gamma^5\exp\left\{\frac{g}{2}\frac{\sigma_{\mu\nu}F^{\mu\nu}}{M^2}\right\}\right]\,. \tag{16.16}$$

If we expand the exponential in powers of $1/M^2$, only the terms up to $\mathcal{O}(M^{-4})$ will survive in the limit $M \to \infty$. Using the antisymmetry of $F^{\mu\nu}$, we can also replace $\sigma_{\mu\nu}$ by $\mathrm{i}\gamma_\mu\gamma_\nu$ and find then

$$\exp\{\dots\} = 1 + \frac{\mathrm{i}g}{2}\gamma_\mu\gamma_\nu F^{\mu\nu}\frac{1}{M^2} + \frac{1}{2}\left(\frac{\mathrm{i}g}{2}\right)^2\gamma_\mu\gamma_\nu\gamma_\alpha\gamma_\beta F^{\mu\nu}F^{\alpha\beta}\frac{1}{M^4} + \mathcal{O}\left(\frac{1}{M^6}\right)\,. \tag{16.17}$$

The trace properties of the gamma matrices inform us that the first two terms vanish, while the third results in a term proportional to the totally antisymmetric tensor $\varepsilon_{\mu\nu\alpha\beta}$. Introducing the dual field-strength tensor $\tilde{F}_{\alpha\beta} = \frac{1}{2}\varepsilon_{\alpha\beta\mu\nu}F^{\mu\nu}$ and then performing the Gaussian integral over K we are left with

$$\delta S = \frac{g^2}{8\pi^2}\int\mathrm{d}^4x\,\alpha(x)\tilde{F}_{\mu\nu}F^{\mu\nu}\,. \tag{16.18}$$

It is sometimes stated that this result is exact, because our derivation seems not to rely on perturbation theory. However, we replaced in the evaluation of $\not{D}^2\phi_n$ the correct solutions ϕ_n accounting for the external gauge fields by plane-waves. Therefore our derivation corresponds to an one-loop result, similar to our calculation of the effective potential in section 13.2.

In a classical theory without gauge fields, a local axial gauge transformation leads to the change

$$\mathscr{L} \to \mathscr{L} + g(\partial_\mu\alpha)\bar{\psi}\gamma^\mu\gamma^5\psi\,. \tag{16.19}$$

Thus the non-invariance of the measure in the path integral violates the classical conservation of the axial current,

$$\partial_\mu j_5^\mu = \partial_\mu(\bar{\psi}\gamma^\mu\gamma^5\psi) = \partial_\mu(j_R^\mu - j_L^\mu) = \frac{g^2}{8\pi^2}\tilde{F}_{\mu\nu}F^{\mu\nu} \,. \tag{16.20}$$

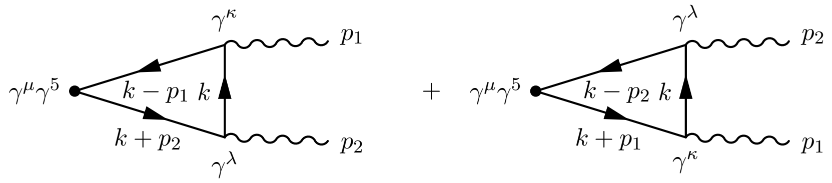
This equation is also known as the Adler-Bell-Jackiw anomaly equation.

The extra term introduced in the Lagrangian by an axial gauge transformation is proportional to $\tilde{F}_{\mu\nu}F^{\mu\nu}$ and transforms thus odd under CP. If this interaction were physically relevant, electrodynamics coupled to chiral fermions would violate CP. The $\tilde{F}_{\mu\nu}F^{\mu\nu}$ term is gauge invariant, has dimension four and corresponds therefore to a renormalisable interaction. The reason why we have not considered it earlier is that it is a total derivative,

$$\begin{aligned} F_{\mu\nu}\tilde{F}^{\mu\nu} &= \frac{1}{2}\varepsilon_{\mu\nu\rho\sigma}(\partial^\mu A^\nu - \partial^\nu A^\mu)(\partial^\rho A^\sigma - \partial^\sigma A^\rho) = 2\varepsilon_{\mu\nu\rho\sigma}\partial^\mu A^\nu \partial^\rho A^\sigma = \\ &= 2\varepsilon_{\mu\nu\rho\sigma}\partial^\mu(A^\nu\partial^\rho A^\sigma) = \partial^\mu(2\varepsilon_{\mu\nu\rho\sigma}A^\nu\partial^\rho A^\sigma) \equiv \partial^\mu K_\mu \,. \end{aligned} \tag{16.21}$$

Since the four-divergence K_μ is not gauge invariant, it cannot be an observable. Therefore it is not excluded that K_μ is singular, leading after integration to non-zero effects in the action. Before we discuss in more detail when it is justified to neglect a total derivative term, we will re-derive the anomaly using a diagrammatic approach.

Perturbative appearance of anomalies Historically, the chiral anomaly was first encountered calculating the process $\pi^0 \to \gamma\gamma$ using perturbation theory. Describing the neutral pion within a non-relativistic quark picture as a $\pi^0 = (\bar{u}u + \bar{d}d)/\sqrt{2}$ state, we can view the process as



Here, the γ^5 matrix accounts for the fact that the pion is a pseudoscalar particle. The two diagrams are connected by the crossing symmetry $\kappa \leftrightarrow \lambda$, $p_1 \leftrightarrow p_2$. The total matrix element of this process is thus given by the sum

$$A_{\kappa\lambda\mu}(p_1,p_2) = S_{\kappa\lambda\mu}(p_1,p_2) + S_{\lambda\kappa\mu}(p_2,p_1) \,, \tag{16.22}$$

where the matrix element $S_{\kappa\lambda\mu}$ describing the first diagram (neglecting coupling constants) is given by

$$S^{\kappa\lambda\mu} = -(-\mathrm{i})^3 \int \frac{\mathrm{d}^4 k}{(2\pi)^4} tr\left[\gamma^\kappa \frac{\mathrm{i}}{\not{k} - \not{p}_1}\gamma^\mu\gamma^5 \frac{\mathrm{i}}{\not{k} + \not{p}_2}\gamma^\lambda \frac{1}{\not{k}}\right] \,. \tag{16.23}$$

It is sufficient to consider only massless fermions: The anomaly is connected to the UV divergences of these diagrams and in this limit masses play no rôle.

We check now, if the classical conservation law for the vector and the axial current hold also at the one-loop level. Current conservation implies the following three relations in momentum space,

$$p_1^\kappa A_{\kappa\lambda\mu} = p_2^\lambda A_{\kappa\lambda\mu} = 0 \tag{16.24a}$$

$$(p_1 + p_2)^\mu A_{\kappa\lambda\mu} = 0\,, \tag{16.24b}$$

where (16.24a) are the equations for the conservation of the vector current, and (16.24b) for the conservation of the axial current. Crossing symmetry implies that these relations must also hold for the two individual amplitudes S.

We check first, if the axial current is conserved. In evaluating

$$I_{\kappa\lambda} \equiv (p_1 + p_2)^\mu S_{\kappa\lambda\mu} = -\int \frac{\mathrm{d}^4k}{(2\pi)^4} \frac{tr\left[\gamma_\kappa(\not{k} - \not{p}_1)(\not{p}_1 + \not{p}_2)\gamma^5(\not{k} + \not{p}_2)\gamma_\lambda \not{k}\right]}{(k - p_1)^2(k + p_2)^2 k^2} \tag{16.25}$$

we use

$$(\not{p}_1 + \not{p}_2)\gamma^5 = \not{p}_1\gamma^5 - \gamma^5\not{p}_2 = -(\not{k} - \not{p}_1)\gamma^5 - \gamma^5(\not{k} + \not{p}_2)$$

and $(\not{k} \pm \not{p}_i)^2 = (k \pm p_i)^2$, so that we can split the integral into two parts,

$$I_{\kappa\lambda} = \underbrace{\int \frac{\mathrm{d}^4k}{(2\pi)^4} \frac{tr\left[\gamma_\kappa\gamma^5(\not{k} + \not{p}_2)\gamma_\lambda\not{k}\right]}{(k + p_2)^2 k^2}}_{A_{\kappa\lambda}(p_2)} + \underbrace{\int \frac{\mathrm{d}^4k}{(2\pi)^4} \frac{tr\left[\gamma_\kappa(\not{k} - \not{p}_1)\gamma^5\gamma_\lambda\not{k}\right]}{(k - p_1)^2 k^2}}_{B_{\kappa\lambda}(p_1)}\,. \tag{16.26}$$

Observe now that the LHS is a pseudo-tensor of rank 2, while the RHS is the sum of the two quantities $A_{\kappa\lambda}$ and $B_{\kappa\lambda}$ which each only depend on one momentum. As it is not possible to create a pseudo-tensor of rank 2 from that, the RHS must be zero. We have thus shown (ostensibly!) that the axial current is conserved.

Next we verify the conservation of the vector current, following the same line of argument as in the case of the axial current. Starting from

$$J_{\lambda\mu} \equiv p_1^\kappa S_{\kappa\lambda\mu} = -\int \frac{\mathrm{d}^4k}{(2\pi)^4} \frac{tr\left[\not{p}_1(\not{k} - \not{p}_1)\gamma_\mu\gamma^5(\not{k} + \not{p}_2)\gamma_\lambda\not{k}\right]}{(k - p_1)^2(k + p_2)^2 k^2}\,, \tag{16.27}$$

we shift the integration variable $k' = k + p_2$ and reorder the terms in the trace, obtaining

$$J_{\lambda\mu} = -\underbrace{\int \frac{\mathrm{d}^4k'}{(2\pi)^4} \frac{tr\left[(\not{k}' - \not{p}_1 - \not{p}_2)\gamma_\mu\gamma^5\not{k}'\gamma_\lambda\right]}{(k' - p_1 - p_2)^2 k'^2}}_{A_{\lambda\mu}(q=p_1+p_2)} + \underbrace{\int \frac{\mathrm{d}^4k'}{(2\pi)^4} \frac{tr\left[(\not{k}' - \not{p}_2\gamma_\mu\gamma^5\not{k}'\gamma_\lambda\right]}{k'^2(k' - p_2)^2}}_{B_{\lambda\mu}(p_2)}\,. \tag{16.28}$$

Thus we conclude that $J_{\lambda\mu}$ also vanishes. For the $p_2^\lambda S_{\kappa\lambda\mu}$ part, we follow the same argument, but we shift the integration variable by $k' = k - p_1$ instead.

We have now shown that both the axial current and the vector current are conserved. Using these results in the calculation for the decay width of a π^0 leads however to a width which vanishes in the limit of a massless pion. Including the small pion mass leads still to life-time much longer than observed. Looking back at our calculation, we should check therefore if all our manipulations were legitimate, although we have not regularised the divergent loop integrals. In particular, the superficial divergence of the diagrams $S_{\kappa\lambda\mu}$ is worse than the logarithmic divergences we have become accustomed to,

$$S_{\kappa\lambda\mu} = -\int \frac{\mathrm{d}^4k}{(2\pi)^4} \frac{tr\left[\gamma_\kappa\not{k}\gamma_\mu\gamma^5\not{k}\gamma_\lambda\not{k}\right]}{k^6} + \text{ subleading terms} \tag{16.29}$$

producing a linearly divergent term.

An essential ingredient for the derivation of the current conservation was the shift of our integration variable. Such a shift can only be done if the integral is properly convergent (after regularisation, if required) or only logarithmic divergent. By contrast, in the case of a linearly divergent integral, the shift $k' = k - a$

$$\int \mathrm{d}^4 k' f(k') \overset{?}{=} \int \mathrm{d}^4 k f(k-a) = \int \mathrm{d}^4 k \left\{ f(k) - a_\mu \frac{\partial f}{\partial k_\mu} + \ldots \right\} \tag{16.30}$$

changes the value of the integral: Using Gauss' theorem, we can convert the gradient term into a surface integral which will lead to a finite change of the integral because of $\mathrm{d}S_\mu \propto k^3$ and $f \propto k^{-3}$. You should show in problem 17.?? that the shift $k^\mu - a^\mu$ in (16.29) changes $S_{\kappa\lambda\mu}$ as

$$S_{\kappa\lambda\mu} \to S'_{\kappa\lambda\mu} = S_{\kappa\lambda\mu} + \frac{1}{8\pi^2} \epsilon_{\kappa\lambda\mu\nu} a^\nu \tag{16.31}$$

where $\epsilon_{\kappa\lambda\mu\nu}$ is again the totally anti-symmetric tensor.

We can look back at the corresponding proof of gauge invariance for the vacuum polarisation, Eqs. (12.36–12.38). There we used dimensional regularisation which respects gauge symmetry. Since γ^5 is not well-defined for $d \neq 4$, we cannot apply this regularisation method here. Using Pauli-Villars regularisation as an alternative would break axial symmetry, since it consists of adding massive particles. Thus, from a technical point of view, anomalies arise, if no regularisation procedure exists which respects the classical symmetry.

Since the shifts of the integration variables required to obtain current conservation differ for the vector and axial current, we can absorb only one of the resulting boundary terms (16.31) by a suitably chosen counter-term. Clearly, we will choose the vector current to be conserved: Otherwise electric charge conservation would be violated, while the axial current is anyway broken by mass effects. If we consider as amplitude including a counter-term

$$A_{\kappa\lambda\mu} = S_{\kappa\lambda\mu}(p_1, p_2) + S_{\lambda\kappa\mu}(p_2, p_1) + \frac{1}{4\pi^2} \epsilon_{\kappa\lambda\mu\nu}(p_1^\nu + p_2^\nu) , \tag{16.32}$$

the vector current will still be conserved, but the axial current will not

$$(p_1 + p_2)^\mu A_{\kappa\lambda\mu} = \frac{1}{2\pi^2} \epsilon_{\kappa\lambda\mu\nu} p_2^\mu p_1^\nu . \tag{16.33}$$

The added contribution gives a non-zero contribution to the divergence of the axial current which is identical to our previous result (16.18),

$$\partial_\mu j_5^\mu = \frac{e^2}{8\pi^2} F_{\mu\nu} \tilde{F}^{\mu\nu} . \tag{16.34}$$

Let us come back to the question, if these results are exact. Studying higher-order corrections, one can show that these corrections vanish and thus the perturbative one-loop result for the anomaly is exact. As an heuristic argument we can use that adding additional propagators will reduce the superficial degree of divergence, while the anomaly is connected to linearly divergent diagrams. As a consequence, processes which are dominated by the anomaly like $\pi^0 \to 2\gamma$ can be calculated reliably in lowest-order perturbation theory, although for $Q^2 = m_\pi^2 \sim \Lambda_{QCD}^2$ the strong coupling constant $\alpha_s(Q^2)$ is certainly not small and higher-order corrections are naively expected to be large.

Cancellations of anomalies Anomalies may be useful when they can break "accidental" global symmetries like baryon and lepton number. As we will see later, this opens the possibility to explain why the universe consists mainly of matter. Similarly, the explicit breaking of global symmetries in the quark Lagrangian can explain why in certain cases no Goldstone bosons are observed. In contrast, symmetries and the resulting Ward identities between Green functions are crucial for the renormalisability of gauge theories. If these identities are not satisfied, physical observables in renormalisable gauges depend on the gauge-fixing parameter ξ, while Lorentz invariance is violated in unitary gauges. The excellent agreement of electroweak precision data with experiment is a strong argument that the underlying theory is renormalisable. As the V–A structure of the electroweak interactions is however similar to our toy model of axial electrodynamics, we can only expect that the anomalies of individual loop diagrams cancel after summing over all contributions.

Compared to axial electrodynamics, we have to consider in the GSW model both $\mathrm{U}_Y(1)$ and SU(2) currents. Then each vertex i comes with the corresponding generator, $T_i = Y\,\mathbf{1}$ and $T_i = \sigma^a/2$, respectively. The anomaly is thus proportional to

$$A_{abc} = \mathrm{tr}\big(T_a T_b T_c\big) + \mathrm{tr}\big(T_a T_c T_b\big) = \mathrm{tr}\big(T_a\{T_b T_c\}\big)\,, \tag{16.35}$$

where the trace is over SU(2) doublets and the second term corresponds to the exchange diagram $b \leftrightarrow c$. Concentrating on diagrams containing SU(2) currents, we have to consider $A^{(1)}_{abc} = \mathrm{tr}\big(\tau_a\{\tau_b,\tau_c\}\big)$, $A^{(2)}_{abc} = \mathrm{tr}\big(\tau_a\{Y_b,Y_c\}\big)$ and $A^{(3)}_{abc} = \mathrm{tr}\big(\tau_a\{Y_b,\tau_c\}\big)$. In the first two cases, the properties of the Pauli matrices imply that the anomaly vanishes automatically,

$$A^{(1)}_{abc} \propto \mathrm{tr}\left(\tau_a\left\{\tau_b,\tau_c\right\}\right) = 2\delta_{bc}\mathrm{tr}(\tau_a) = 0 \qquad and \qquad A^{(2)}_{abc} \propto \mathrm{tr}(\tau_a) = 0\,. \tag{16.36}$$

In the third case, we use the Gell-Mann–Nishijima relation to replace the hypercharge, $Y = 2(T_3 - Q)$, obtaining

$$A^{(3)}_{abc} \propto \mathrm{tr}\left(\tau_a\left\{Q_b,\tau_c\right\}\right) = \mathrm{tr}\left(Q_b\left\{\tau_a,\tau_c\right\}\right) = 2\delta_{ab}\sum_i Q_i\,. \tag{16.37}$$

Recalling that only the left-chiral fermions couple to the W boson, the condition that this anomaly vanishes is thus $\sum_i Q_i^L = 0$. If we now only look at the leptons and quarks separately, then

$$Q_e + Q_\nu = -1 \neq 0 \quad and \quad Q_u + Q_d = \frac{2}{3} - \frac{1}{3} = \frac{1}{3} \neq 0\,. \tag{16.38}$$

Including both leptons and quarks where we account by the factor three for their colour quantum number we find as required for the anomaly cancellation

$$Q_e + Q_\nu + 3(Q_u + Q_d) = -1 + 3\,\frac{1}{3} = 0\,. \tag{16.39}$$

The last remaining triangle contribution to the anomaly contains three $\mathrm{U}_Y(1)$ currents, $A^{(4)}_{abc} = \mathrm{tr}\big(Y_a\{Y_b,Y_c\}\big)$. Again, the anomaly vanishes, if the condition $\sum_i Q_i = 0$ is met, cf. problem 17.**??**. Thus chiral anomalies are cancelled within each full fermion generation. In the SM, there is no explanation for this conspiracy between the quark and lepton sector, and this has been one of the major motivations to consider GUTs. Note also that if a single member of a hypothetical fourth generation of fermions would be found, anomaly cancellation would require the existence of a complete set of quarks and leptons.

We have restricted our analysis of the anomaly to an abelian model. In the non-abelian case, the field-strength contains term linear and quadratic in the gauge fields. As a result, additional anomalies in square and pentagon diagrams appear. However, the absence of anomalies in the triangle diagrams guaranties also their absence in square and pentagon diagrams.

Remark 16.1: Grand Unified Theories: The gauge group of the SM contains four commuting generators, i.e. it has rank four. Because SU(n) has rank $n-1$, SU(5) is thus the smallest SU(n) group encompassing the SM. In the 5-dimensional fundamental representation, we can set $T^a = \left(\begin{smallmatrix} \lambda^a/2 & 0 \\ 0 & 0 \end{smallmatrix}\right)$ and $T^i = \left(\begin{smallmatrix} 0 & 0 \\ 0 & \sigma^i/2 \end{smallmatrix}\right)$, with λ^a and σ^i as the Gell-Mann and Pauli matrices. Thus SU(5) contains SU(3) and SU(2) as subgroups. The remaining diagonal matrix can be chosen proportional to hypercharge, $T^0 \propto \mathrm{d}iag(-2,-2,-2,3,3)$. Additionally, there are 12 off-diagonal matrices.
We combine three colour states of down-like quarks together with a lepton doublet in $\psi = (\bar{d}^r, \bar{d}^b, \bar{d}^g, \nu_e, e)^T$, while the remaining ten states can be fitted into an antisymmetric 5×5 matrix. From this, we can draw immediately three conclusions: First, the 12 off-diagonal matrices correspond to new gauge boson (called X^μ and XY^μ) which interchange quarks and leptons. Thus baryon and lepton number is generically broken in GUTs, and the X^μ and XY^μ bosons have to be sufficiently heavy such that proton decay is suppressed. Second, the electric charge is diagonal, and thus any multiplet has zero electric charge. Applied to $\psi = (\bar{d}^r, \bar{d}^b, \bar{d}^g, \nu_e, e)^T$, this leads to charge quantisation, $Q_{\bar{d}} = -Q_e/3$. Finally, above the energy scale M_{GUT} where the SU(5) symmetry is restored, a single gauge coupling controls all interactions. In such a picture, one expects that the SM couplings meet in one point at the scale M_{GUT}.

16.2. Instantons and the strong CP problem

The effective Lagrangian induced by the chiral anomaly is a surface term. This suggests that a term of the same structure can be obtained as the topological charge of a Yang-Mills theory, following the argument of Bogomol'nyi in Eq. (15.43). If such a topological charge exists, then an unbroken Yang-Mills theory has a non-trivial vacuum. Such a non-trivial vacuum structure has however only physical consequences if the corresponding classical tunnelling solutions have a finite action. We should therefore search for classical solutions of a pure, Euclidean Yang-Mills theory with $S_{\mathrm{YM}} < \infty$. These solutions are the non-abelian analogue of the scalar bounce solution we have considered earlier.

Instantons We define an Euclidean Yang-Mills theory by the action

$$S = \frac{1}{2} \int \mathrm{d}^4x \, \mathrm{tr}\{F_{\mu\nu} F_{\mu\nu}\} \,, \tag{16.40}$$

where $F_{\mu\nu} = F^a_{\mu\nu} T^a = \partial_\mu A_\nu - \partial_\nu A_\mu + \mathrm{i}g[A_\mu, A_\nu]$ and derivatives and integrations are with respect to Euclidean coordinates $(x_1, x_2, x_3, x_4 = \mathrm{i}x_0)$ with $x_i \in \mathbb{R}$. Since the metric tensor is $\eta_{\mu\nu} = \delta_{\mu\nu}$, we do not need to distinguish between lower and upper indices. Note that in Euclidean space the dual of the dual field-strength tensor is again the field-strength tensor, $\tilde{\tilde{F}}_{\mu\nu} = F_{\mu\nu}$, while in Minkowski space it is its negative, $\tilde{\tilde{F}}_{\mu\nu} = -F_{\mu\nu}$. Next we define instantons as self-dual and anti-self-dual solutions,

$$\tilde{F}_{\mu\nu} = F_{\mu\nu} \quad \mathit{and} \quad \tilde{F}_{\mu\nu} = -F_{\mu\nu} \,, \tag{16.41}$$

of the classical Yang-Mills equations. Using Bogomol'nyi's trick, we can show that they correspond to the topologically non-trivial solutions with the lowest energy—if such non-trivial solutions exist. We write first

$$\mathrm{t}r\{(F_{\mu\nu}-\tilde{F}_{\mu\nu})^2\}=\mathrm{tr}\{F_{\mu\nu}^2+\tilde{F}_{\mu\nu}^2-2F_{\mu\nu}\tilde{F}_{\mu\nu}\}\geq 0\,. \tag{16.42}$$

For the calculation of $\tilde{F}^2$ we use $\epsilon_{\mu\nu\rho\sigma}\epsilon_{\mu\nu\kappa\lambda}=2(\delta_{\rho\kappa}\delta_{\sigma\lambda}-\delta_{\rho\lambda}\delta_{\sigma\kappa})$ and end up with $\tilde{F}^2=F^2$. Thus

$$\mathrm{t}r\{F_{\mu\nu}F_{\mu\nu}\}\geq \mathrm{t}r\{F_{\mu\nu}\tilde{F}_{\mu\nu}\}\,, \tag{16.43}$$

which is minimised if F is self-dual. We obtain the same bound for an antiself-dual solution, if we choose a plus sign in Eq. (16.42).

We first examine if QED, i.e. an abelian Yang-Mills theory, contains instantons. A finite action,

$$\left|\int_{\Omega(\tau)}\mathrm{d}^4x\,\tilde{F}F\right|\leq\left|\int_{\Omega(\tau)}\mathrm{d}^4x\,FF\right|<\infty \tag{16.44}$$

requires that F decreases faster than τ^{-2} in the limit $\tau^2=\boldsymbol{x}^2+x_4^2\to\infty$. But as a classical Yang-Mills theory contains no scale, $\tilde{F}F$ must be a polynomial in τ: Thus $F\sim\mathcal{O}(\tau^{-3})$, $A\sim\mathcal{O}(\tau^{-2})$, and then the total derivative (16.21) behaves as

$$K_\mu=2\epsilon_{\mu\nu\rho\sigma}A_\nu\partial_\rho A_\sigma\sim\mathcal{O}(\tau^{-5})\,. \tag{16.45}$$

As a result, the surface term in

$$\int_\Omega\mathrm{d}^4x\,\tilde{F}F=\int_{\partial\Omega}\mathrm{d}S_\mu K_\mu\to 0 \tag{16.46}$$

vanishes and we see that in an abelian theory $\tilde{F}F$ does not influence physical quantities. This argument justifies our usual practise to neglect surface terms in QED.

We now turn to the non-abelian case. Then we can express $\mathrm{t}r\,\{\tilde{F}_{\mu\nu}F_{\mu\nu}\}$ again as a four-divergence,

$$\partial_\mu K_\mu=2\,\mathrm{tr}\,\tilde{F}_{\mu\nu}F_{\mu\nu} \tag{16.47}$$

where now

$$K_\mu=2\varepsilon_{\mu\nu\sigma\rho}\mathrm{t}r\left(A_\nu F_{\sigma\rho}+\frac{2}{3}\mathrm{i}gA_\mu A_\sigma A_\rho\right)\,. \tag{16.48}$$

Choosing at infinity a pure gauge field,

$$A_\mu=\frac{\mathrm{i}}{g}(\partial_\mu U)U^\dagger\,, \tag{16.49}$$

results in $F_{\mu\nu}=0$ for $\tau\to\infty$ and ensures that the action is finite. On the other hand, a gauge transformation U which becomes constant for $\tau\to\infty$, i.e. depends in this limit only on the angles, gives $A\sim\mathcal{O}(\tau^{-1})$ and thus $K\sim\mathcal{O}(\tau^{-3})$. As a result, the surface integral may become non-zero. One may wonder if we can gauge away $A_\mu\propto(\partial_\mu U)U^\dagger$ on $\tau=\infty$ by performing a suitable gauge transformation $\tilde{U}$: Since $\tilde{U}$ has to be regular in all $\mathbb{R}^4$, it must be constant a $\tau=0$ and independent of the angles. Thus $\tilde{U}$ is continuously connected to the identity and can be used only to gauge away fields A_μ in the same homotopy class. Thus the surface term $F\tilde{F}$ has only physical significance, if the gauge fields at $\tau\to\infty$ are split into non-trivial topological classes.

We next move to the specific case of SU(2). Any SU(2) matrix can be written as $U = a + \mathrm{i}\boldsymbol{b}\cdot\boldsymbol{\sigma}$ with $a^2 + |\boldsymbol{b}|^2 = 1$. Thus SU(2) is isomorphic to S^3. Compactifying the boundary $\partial\mathbb{R}^4 = \mathbb{R}^3$ of Euclidean space at large τ to S^3, the function (16.49) defines a map $S^3 \to S^3$. The question if non-trivial instanton solutions exist is thus equivalent to the existence of topologically non-trivial mappings $S^3 \to S^3$, what in turn requires that $\pi_3(S^3)$ is not the identity. We can generalise the winding number (15.75) from $S^1 \to S^1$ to the case $S^3 \to S^3$ as follows: If we use as coordinates the three Euler angles specifying a point on S^3, then each angle i contributes a factor $U\partial_i U^\dagger$ as in the S^1 case. Since the winding number is a pseudo-scalar, we have to contract ijk with the Levi-Civita tensor and to take the trace over the SU(2) indices, arriving at

$$\nu = -\frac{1}{24\pi^2}\int \mathrm{d}x_1\mathrm{d}x_2\mathrm{d}x_3\;\varepsilon_{ijk}\mathrm{tr}\left[\left(U\partial_i U^\dagger\right)\left(U\partial_j U^\dagger\right)\left(U\partial_k U^\dagger\right)\right]. \tag{16.50}$$

Because of $\mathrm{d}x_i\partial_i = \tilde{\mathrm{d}x}_i\tilde{\partial}_i$, the expression is equally valid using Cartesian coordinates. In order to show that the winding number is invariant under continuous transformations, $\delta\nu = 0$, it is sufficient to consider the variation of a single factor in the trace. Now we can profit from (15.75) where we derived already the variation of this factor in the non-abelian case as

$$\delta(U\partial_i U^\dagger) = -U\partial_i\left(U^\dagger\delta U\right)U^\dagger.$$

Inserting this relation into the integrand results in

$$E \equiv \varepsilon_{ijk}\mathrm{tr}\left[\left(U\partial_i U^\dagger\right)\left(U\partial_j U^\dagger\right)\delta\left(U\partial_k U^\dagger\right)\right] = -\varepsilon_{ijk}\mathrm{tr}\left[\partial_i U^\dagger U\partial_j U^\dagger U\partial_k\left(U^\dagger\delta U\right)\right]. \tag{16.51}$$

Then we perform a partial integration and use that terms like $\partial_k\partial_i U^\dagger$ vanish contracted with ε_{ijk}, obtaining

$$E = \varepsilon_{ijk}\mathrm{tr}\left[\partial_i U^\dagger\partial_k U\partial_j U^\dagger\delta U + \partial_i U^\dagger U\partial_j U^\dagger\partial_k UU^\dagger\delta U)\right]. \tag{16.52}$$

In the second term, we use $U\partial_j U^\dagger = -\partial_j UU^\dagger$ and $\partial_k UU^\dagger = -U\partial_k U^\dagger$ to symmetrise the expression in j and k,

$$E = \varepsilon_{ijk}\mathrm{tr}\left[\partial_i U^\dagger\partial_k U\partial_j U^\dagger\delta U + \partial_i U^\dagger\partial_j U\partial_k U^\dagger\delta U\right] = 0. \tag{16.53}$$

Thus the winding number ν is invariant under infinitesimal and thus under continuous transformations.

Next we try to express the winding number as a volume integral over $\mathrm{tr}\,\{\tilde{F}_{\mu\nu}F_{\mu\nu}\}$. We write (16.50) as a surface integral,

$$\begin{aligned}\nu &= \frac{1}{24\pi^2}\int \mathrm{d}S_\mu\;\varepsilon_{\mu\nu\sigma\tau}\mathrm{tr}\left[\left(U\partial_\nu U^\dagger\right)\left(U\partial_\sigma U^\dagger\right)\left(U\partial_\tau U^\dagger\right)\right] &\text{(16.54a)}\\ &= -\frac{\mathrm{i}g^3}{24\pi^2}\int \mathrm{d}S_\mu\;\varepsilon_{\mu\nu\sigma\tau}\mathrm{tr}\left[A_\nu A_\sigma A_\tau\right], &\text{(16.54b)}\end{aligned}$$

where we could use $U\partial_\mu U^\dagger = \mathrm{i}gA_\mu$ because A_μ is a pure gauge field for $\tau \to \infty$. Since then also $F^a_{\sigma\rho} = 0$, only the second term in the expression (16.48) for the four-divergence K^μ survives and we obtain

$$\nu = \frac{g^2}{32\pi^2}\int \mathrm{d}S_\mu\,K^\mu = \frac{g^2}{16\pi^2}\int \mathrm{d}^4x\;\mathrm{tr}\,\{\tilde{F}_{\mu\nu}F_{\mu\nu}\}. \tag{16.55}$$

Thus we have shown that gauge fields A^a_μ exist which are solutions of the classical Yang-Mills equations, have a finite action and fall into distinct equivalence classes characterised by the winding number[1] ν. Our remaining task is to write down an explicit form of the mappings $S^3 \to S^3$ and to show that the definition (16.50) results in a integer winding number. We choose to write $U(x)$ as

$$U^{(n)}(x) = \left(\frac{x_4 + \mathrm{i}\boldsymbol{x}\cdot\boldsymbol{\sigma}}{r}\right)^n \tag{16.56}$$

where x_μ is the unit vector $x_\mu = (\sin\chi\boldsymbol{e}, \cos\chi)$. Evaluation of (16.50) for $U^{(1)}(\boldsymbol{x})$ confirms then that $\nu = 1$. Integrating now $\mathrm{tr}\{F_{\mu\nu}F_{\mu\nu}\} \geq \mathrm{tr}\{F_{\mu\nu}\tilde{F}_{\mu\nu}\}$ over space, we find on the LHS twice the Euclidean action, while the RHS equals $16\pi^2\nu/g^2$. Thus the Euclidean action is bounded by

$$S \geq \frac{8\pi^2|\nu|}{g^2}\,, \tag{16.57}$$

and instantons as the non-trivial solutions with the lowest energy have the action $S = 8\pi^2/g^2$.

We used for our discussion of instantons in non-abelian Yang-Mills theories the specific example of SU(2). A theorem of Brott states that for any simple Lie group G containing SU(2) the maps $S^3 \to G$ can be deformed continuously to the ones of $S^3 \to \mathrm{SU}(2)$. Thus all our results apply identically to the case of strong interactions, SU(3). In the remaining part of this chapter, we will discuss the impact of instantons on the QCD vacuum, while we postpone the electroweak case to chapter 21.

Tunnelling interpretation We consider the four-dimensional cylinder defined by $|x_4| \leq T$ and $|\boldsymbol{x}| \leq R$ in the limit $T, R \to \infty$. At $x_4 = -T$, we choose A_μ as a pure gauge field with winding number ν_1, and at $x_4 = T$ with ν_2. On the boundary $|\boldsymbol{x}| = R$, we choose U constant and thus A_μ is zero. Calculating the total winding number, we find

$$\begin{aligned}\frac{g^2}{16\pi^2}\int_\Omega \mathrm{d}^4x\, \mathrm{tr}(F\tilde{F}) &= \frac{g^2}{32\pi^2}\int_{\partial\Omega} \mathrm{d}^3S_\mu K^\mu = \\ = \frac{g^2}{32\pi^2}\int \mathrm{d}^3x\, &\left[K^0(t=-\infty) - K^0(t=+\infty)\right] = \nu_1 - \nu_2 = \nu\,.\end{aligned} \tag{16.58}$$

The minus sign appears, because of the opposite orientations of the two caps. Thus the classical solutions we have determined interpolate between a vacuum with winding number ν_1 at time $t = -\infty$ and a vacuum with winding number ν_2 at time $t = +\infty$. The two vacua are separated by a finite energy barrier, and thus the solutions describe the quantum tunnelling between different vacua. This agrees with our finding in chapter 15.2 that tunnelling solutions correspond to solutions of the Euclidean field equations. These solutions were dubbed instantons by 't Hooft, since the tunnelling they describe happens instantaneously in Minkowski space.

The ϑ vacuum The tunnelling interpretation indicates that the true vacuum of a pure Yang-Mills theory is the superposition of all vacua with fixed winding number ν. Let us call

[1] Mathematicians call the winding number of the mapping $S^3 \to S^3$ the Pontryagin index, while mathematical physicists use often the term Chern-Simon number.

these vacua $|\nu\rangle$ and the true one $|\vartheta\rangle$. Applying a gauge transformation $U \equiv U^{(1)}$ with unit winding number results in

$$U\,|\nu\rangle = |\nu + 1\rangle \ . \tag{16.59}$$

On the other hand, the Yang-Mills Hamiltonian is invariant under gauge transformations,

$$UHU^{\dagger} = H\,, \tag{16.60}$$

or $[H, U] = 0$. Thus the true vacuum $|\vartheta\rangle$ is a common eigenstate of H and U. Since the vacuum is normalised, the eigenvalue of U has to be a phase,

$$U\,|\vartheta\rangle = \mathrm{e}^{\mathrm{i}\vartheta}\,|\vartheta\rangle \ . \tag{16.61}$$

The angle ϑ is a conserved quantum number. Thus a classical Yang-Mills theory is characterised by two numbers, the coupling g and the angle ϑ.

The true vacuum $|\vartheta\rangle$ is a linear superposition of the vacua with fixed winding number n given by

$$|\vartheta\rangle = \sum_{n=-\infty}^{\infty} \mathrm{e}^{-\mathrm{i}n\vartheta}\,|n\rangle \ , \tag{16.62}$$

since

$$U\,|\vartheta\rangle = \sum_{n=-\infty}^{\infty} \mathrm{e}^{-\mathrm{i}n\vartheta}\,|n+1\rangle = \mathrm{e}^{\mathrm{i}\vartheta} \sum_{n=-\infty}^{\infty} \mathrm{e}^{-\mathrm{i}n\vartheta}\,|n\rangle \ . \tag{16.63}$$

The matrix elements $\langle n'|\,H\,|n\rangle$ depend only on the difference $\nu = n' - n$, because (16.60) gives

$$\langle n' + 1|\,H\,|n+1\rangle = \langle n'|\,H\,|n\rangle \ . \tag{16.64}$$

Using also the parity properties, $P\,|n\rangle = -\,|n\rangle$ and $PHP^{-1} = H$, we obtain

$$\langle n'|\,H\,|n\rangle = \langle -n'|\,H\,|-n\rangle \ . \tag{16.65}$$

Hence the matrix elements $\langle n'|\,H\,|n\rangle$ depend only on the absolute value $|\nu| = |n' - n|$. We conclude that instantons lead to an effective potential $V_{eff}(\vartheta)$ which is periodic and even in ϑ

$$H\,|\vartheta\rangle = L^3 V_{eff}(\vartheta)\,|\vartheta\rangle \ , \tag{16.66}$$

with $V_{eff}(\vartheta) = V_{eff}(\vartheta + 2\pi)$ and $V_{eff}(\vartheta) = V_{eff}(-\vartheta)$, where L^3 is the considered volume. A general argument due to Weinberg shows that points of enhanced symmetry are stationary points of the action. Thus we expect the minimum of $V_{eff}(\vartheta)$ to coincide with the CP conserving point $\vartheta = 0$.

We included into our definition of the path integral using the Faddeev-Popov trick only gauge fields which are continuously connected with the identity. Thus our next task is to add the effect of the ϑ vacuum to the path integral. The path integral in the presence of external sources is identical to the vacuum persistence amplitude,

$$\langle \vartheta' | \vartheta \rangle_J = \sum_{n,n'} \mathrm{e}^{\mathrm{i}(n'\vartheta' - n\vartheta)} \langle n' | n \rangle_J \,. \tag{16.67}$$

Now we introduce the difference $\nu = n' - n$ so that we can rewrite the phase as $n'\vartheta' - n\vartheta = n(\vartheta' - \vartheta) + \nu\vartheta'$. But $\langle n'|n\rangle_J$ depends only on ν and thus we can perform the sum over n. This leads to a factor $\delta_{\vartheta',\vartheta}$, which expresses the fact that ϑ is conserved. Thus

$$\langle\vartheta|\vartheta\rangle_J = \sum_\nu \mathrm{e}^{\mathrm{i}\nu\vartheta} \int \mathcal{D}A^{(\nu)}\, \mathrm{e}^{-S+\langle JA\rangle}\,, \tag{16.68}$$

where $\mathcal{D}A^{(\nu)}$ denotes the integration over all gauge field configurations with the fixed winding number ν. Replacing ν with the help of Eq. (16.55) and introducing $\mathcal{D}A \equiv \sum_\nu \mathcal{D}A^{(\nu)}$ gives

$$\langle\vartheta|\vartheta\rangle_J = \int \mathcal{D}A\, \mathrm{e}^{-S+\langle JA\rangle} \exp\left\{\mathrm{i}\,\frac{\vartheta g^2}{16\pi^2}\,\mathrm{tr}(F\tilde{F})\right\}. \tag{16.69}$$

Thus instantons induce an additional term $\mathscr{L}_\vartheta$ to the classical Yang-Mills Lagrangian,

$$\mathscr{L}_{eff} = \mathscr{L} + \frac{\vartheta g^2}{16\pi^2}\,\mathrm{tr}(F\tilde{F})\,, \tag{16.70}$$

which depends on the arbitrary parameter $\vartheta \in [0, 2\pi[$. In order to discuss observable effects of this additional term, we have to add fermions to the pure Yang-Mills theory we discussed up to now.

Fermionic contribution to ϑ We have derived in Eq. (16.20) the axial anomaly for an abelian gauge theory. The corresponding result for a Yang-Mills theory coupled to a single massless fermion is

$$\partial_\mu j_5^\mu = \partial_\mu(j_R^\mu - j_L^\mu) = \frac{g^2}{8\pi^2}\,\mathrm{tr}(F\tilde{F})\,. \tag{16.71}$$

Thus the axial anomaly leads to the additional term

$$\mathscr{L}_{eff} = \mathscr{L} + \frac{\alpha n_f g^2}{8\pi^2}\,\mathrm{tr}(F\tilde{F})\,, \tag{16.72}$$

in the effective QCD Lagrangian, if we perform a chiral $\mathrm{U}_A(1)$ transformation $q_{L,R} \to \mathrm{e}^{\mathrm{i}\alpha\gamma^5} q_{L,R}$ on the n_f quark fields. This term has the same structure as the instanton contribution, and if we choose

$$\alpha = -\frac{\vartheta}{2n_f} \tag{16.73}$$

the two terms cancel. Thus for massless quarks the ϑ parameter is unphysical and we can choose the $\vartheta = 0$ vacuum.

We can understand this, if we consider the effect of an instanton transition. Integrating Eq. (16.71) gives as change of the axial charge $Q_5 = N_R - N_L$ per massless quark flavor

$$\Delta Q_5 = Q_5(t = -\infty) - Q_5(t = +\infty) = 2\nu\,. \tag{16.74}$$

Thus an instanton process $\nu = \pm 1$ changes the axial quark number by two units, creating a left-chiral and destroying a right-chiral quark and vice versa. As chirality is a conserved quantum number for massless particles, at least one of the two states connected by the instanton process can therefore not correspond to the vacuum. By contrast, for $m > 0$ the mass term mixes left- and right-chiral fields: the quark-antiquark pair can annihilate via the mass term and the states can be identified with the vacuum.

We now consider massive quarks. Then the conservation of the axial current is additionally to the axial anomaly explicitly broken by the quark masses, since a Dirac mass term transform as

$$m\bar{q}q \to m\bar{q}\mathrm{e}^{2\mathrm{i}\alpha\gamma^5}q = \cos(2\alpha)m\bar{q}q + \mathrm{i}\sin(2\alpha)m\bar{q}\gamma^5 q \tag{16.75}$$

under a chiral transformation $q_{L,R} \to \mathrm{e}^{\mathrm{i}\alpha\gamma^5}q_{L,R}$. The second term violates T (CP) invariance, as the original $\mathscr{L}_\vartheta$ we wanted to eliminate. Therefore, CP violation implied by $\mathscr{L}_\vartheta$ is a physical effect for massive quarks. If we consider n_f flavor of quarks, then the mass matrix M_{ij} will change as $M_{ij} \to e^{2\mathrm{i}\alpha}M_{ij}$ and thus

$$\arg\det M \to \arg\det M + 2n_f\alpha\,. \tag{16.76}$$

Therefore only the combination

$$\bar{\vartheta} \equiv \vartheta + \arg\det M \tag{16.77}$$

is an observable quantity: It is a question of convenience, if we choose real mass matrices and rotate all CP violation via the axial anomaly into the ϑ term. Or if we eliminate $\mathscr{L}_\vartheta$ and transfer its effect into CP violating complex mass matrices.

Example 16.1: Two Higgs doublet model: Let assume that the Higgs doublet $\Phi_1 = (\phi^+, \phi^0)$ generates the masses for down-like fermions, while a second Higgs doublet $\Phi_2 = (\phi^0, \phi^-)$ generates the masses for up-like fermions,

$$\mathscr{L}_Y = -X_{\alpha\beta}\bar{Q}_\alpha\Phi_1 d_{R,\beta} - Y_{\alpha\beta}\bar{Q}_\alpha\Phi_2 u_{R,\beta} + \text{h.c.} \tag{16.78}$$

with

$$\langle 0|\Phi_1|0\rangle = \frac{1}{\sqrt{2}}\begin{pmatrix} 0 \\ v_1\mathrm{e}^{\mathrm{i}\delta_1}\end{pmatrix} \quad and \quad \langle 0|\Phi_2|0\rangle = \frac{1}{\sqrt{2}}\begin{pmatrix} v_2\mathrm{e}^{\mathrm{i}\delta_2} \\ 0\end{pmatrix}. \tag{16.79}$$

The resulting quark mass matrix M_{ij}, $i,j = 1,\ldots,6$ has the determinant

$$\det(M) = \frac{(v_1v_2)^2}{8}\mathrm{e}^{3\mathrm{i}(\delta_1+\delta_2)}\det(X)\det(Y) \tag{16.80}$$

and thus

$$\arg\det M = \arg\det(XY) + 3\mathrm{i}(\delta_1+\delta_2)\,. \tag{16.81}$$

By an SU(2) gauge transformation $U = \exp(\mathrm{i}\alpha\tau_3/2)$ we can eliminate one of the two phases. Thus a general two Higgs doublet model has one CP violating phase. In contrast, the SM uses $\Phi_2 = \mathrm{i}\tau_2\phi^*$. Then $\delta_1 = -\delta_2$, and no CP violation arises in the Higgs sector.

We look now for observable consequences of the ϑ vacuum in the low-energy interactions of hadrons. Since the ϑ term and the axial anomaly are flavour blind, the change c in the quark masses,

$$c \equiv \sin(2\alpha_q)m_q\,, \tag{16.82}$$

is the same for all quarks. In order to shift the effects of the ϑ term completely into the mass matrix of the quarks, we need also $\sum_{q=1}^{n_f} 2\alpha_q = -\vartheta$. Eliminating α_q in the limit of small ϑ gives

$$c = -\frac{\vartheta}{\sum_q m_q^{-1}}\,, \tag{16.83}$$

and thus the ϑ dependent, CP violating part of the mass term becomes

$$\mathscr{L}_m^{(\vartheta)} = -\mathrm{i}\vartheta\Big(\sum_q m_q^{-1}\Big)^{-1}\sum_q \bar{q}\gamma^5 q\,. \tag{16.84}$$

For light nucleons and mesons which consist only of u and d quarks this simplifies to

$$\mathscr{L}_m^{(\vartheta)} = -\mathrm{i}\vartheta\frac{m_u m_d}{m_u+m_d}\left(\bar{u}\gamma^5 u + \bar{d}\gamma^5 d\right)\,. \tag{16.85}$$

This mass term generates a CP violating effective pion-nucleon interaction which in turn leads to an electric dipole moment d_n of the neutron. The corresponding limit on d_n bounds the value of $\bar{\vartheta}$ as $|\bar{\vartheta}| \leq 2\times 10^{-10}$. Although the value of $\bar{\vartheta}$ is a free parameter within the SM, it seems natural to ask for an explanation why $\bar{\vartheta}$ is so small.

The most straight-forward explanation would be that one current quark mass is zero, i.e. that one quark has no Yukawa coupling to the SM Higgs. Then (16.85) shows for $n_f = 2$ clearly that the CP violating effect disappears. This holds also for $n_f > 2$, because $\arg\det M = 0$ if one mass eigenvalue is zero. While it has been debatable, if $m_u \sim 5\,\mathrm{MeV}$ deduced from chiral perturbation theory for the current u quark mass might be lowered to $m_u = 0$, this possibility was closed by lattice data around 2005. The question why $\bar{\vartheta}$ is so small is called the strong CP problem of the SM.

16.3. Axions

Peceei and Quinn proposed to promote the parameter ϑ to a dynamical variable which settles automatically at its minimum zero. The basic ingredient of this proposal is a new massless pseudo-scalar field a which couples to gluons with an interaction of the same structure as the ϑ term,

$$\mathscr{L}_a = \frac{1}{2}(\partial_\mu a)^2 - \frac{g^2}{16\pi^2}\frac{a}{f_a}F\tilde{F}\,. \tag{16.86}$$

Since a is a pseudo-scalar, the interaction term $aF\tilde{F}$ conserves CP. Adding $\mathscr{L}_a$ to the effective Lagrangian (16.70) of QCD means that observables depend only on the combination $\vartheta - a/f_a$. If $\mathscr{L}_a$ is invariant under the shift $a \to a + const.$, then we can use this arbitrariness to absorb the ϑ parameter into a redefinition of the field a. Such a shift symmetry is typical for a Goldstone boson which has only derivative couplings. Thus the pseudo-scalar particle a should be the Goldstone boson of a spontaneously broken global symmetry, and the $aF\tilde{F}$ interaction suggests to choose this symmetry as a chiral U(1) symmetry. This symmetry is called Peceei-Quinn symmetry $U_{\mathrm{PQ}}(1)$ and its Goldstone boson a is the axion.

Weinberg and Wilzcek realised that there is an additional twist in this proposal. The effective potential (16.66) generated via the $F\tilde{F}$ term has two effects: First, instanton effects break the shift symmetry, generating a mass term $m_a^2 = \partial^2 V_{\mathrm{e}ff}/\partial a^2$ for the axion. Thus the axion becomes a massive pseudo Goldstone boson. Second, the vacuum expectation value of the axion field will relax to the minimum of the potential, $\partial V_{\mathrm{e}ff}/\partial a = 0$. But we have argued above that the minimum of $V_{\mathrm{e}ff}(\vartheta)$ is situated at $\vartheta = 0$. Thus also in the case of a massive axion the strong CP problem is solved.

Let us illustrate the key points of this idea with a simple model. We add a complex scalar field ϕ and a set of heavy fermions ψ to the SM,

$$\mathscr{L} = \mathrm{i}\bar{\psi}\partial\!\!\!/\psi + \frac{1}{2}\partial_\mu\phi^\dagger\partial^\mu\phi - y_i(\bar{\psi}_L\psi_R\phi + \mathrm{h.c.}) - V(\phi)\,. \tag{16.87}$$

While ϕ is a SM singlet, the fermions are charged under SU(3) and U(1). The Lagrangian is invariant under the global chiral $\mathrm{U}_{PQ}(1)$ gauge transformation

$$\phi \to \mathrm{e}^{\mathrm{i}\alpha}\phi \qquad and \qquad \psi_{L/R} \to \mathrm{e}^{\pm\mathrm{i}\alpha/2}\psi_{L/R} \,. \tag{16.88}$$

We now break spontaneously the Peccei-Quinn $\mathrm{U}_{PQ}(1)$ symmetry, choosing the usual Mexican hat potential for $V(\phi)$. Splitting ϕ into its vacuum expectation value and the fluctuating fields,

$$\phi = \frac{f_a + \rho}{\sqrt{2}}\mathrm{e}^{\mathrm{i}a/f_a}\begin{pmatrix}0\\1\end{pmatrix}\,, \tag{16.89}$$

generates a mass term for ρ while a remains massless. We assume that f_a is much larger than the energy scale E we are interested in, and thus we can neglect the field ρ,

$$\mathscr{L} = \frac{1}{2}(\partial_\mu a)^2 - m_i\bar{\psi}\mathrm{e}^{\mathrm{i}\gamma^5 a/f_a}\psi - V(\phi) + \mathcal{O}(E^2/f_a^2)\,. \tag{16.90}$$

The combined expression is clearly invariant under $\mathrm{U}_{PQ}(1)$ transformations $a \to a + \alpha f_a$. Expanding the exponential, we generate mass terms $m_i = y_i f_a/\sqrt{2}$ for the heavy fermions plus fermion-axion interactions,

$$\mathscr{L}_{int} = -m_i\bar{\psi}\psi - \mathrm{i}\frac{m_i}{f_a}a\bar{\psi}\gamma^5\psi - \dots \tag{16.91}$$

The latter lead to a AVV triangle graph for the process $a \to 2g$ which in turn induces via the axial anomaly the desired $F\tilde{F}$ term. In the same way, an effective $a \to 2\gamma$ coupling is generated. Thus the characteristic features of an axion are its two-gluon and two-photon couplings. Moreover, the parameters of the pion and axion sector are connected by

$$m_a f_a \approx m_\pi f_\pi\,, \tag{16.92}$$

since the two Goldstone bosons mix via their two-gluon coupling, $a \leftrightarrow 2g \leftrightarrow \pi^0$.

Summary

The CP-odd term $\tilde{F}_{\mu\nu}F^{\mu\nu}$ is a gauge invariant renormalisable interaction. Terms of this type are produced by instanton transitions between Yang-Mills vacua with different winding numbers and by the chiral anomaly. While we can rotate classically all CP violating phases contained in the quark mass matrices into the single CP violating phase of the CKM matrix, the chiral anomaly leads additionally to the change $\vartheta \to \bar{\vartheta} = \vartheta + \arg\det M$ in the coefficient of the $\tilde{F}_{\mu\nu}F^{\mu\nu}$ term. Since the physics origin of both contributions seem to be disconnected, it is puzzling that they sum up to $|\bar{\vartheta}| \lesssim 10^{-10}$. A possible solution is the Peceei-Quinn symmetry which promotes the parameter ϑ to the field $a = f_a\vartheta$ which settles automatically at the minimum at $\bar{\vartheta} = 0$ of the instanton potential V_{eff}.

Further reading

Instantons are discussed in more detail by [Rub02]. For an introduction into the physics of axions see [KN13], while [CL88] contains a concise introduction to grand unification.

17. Hadrons, partons and QCD

We introduced QCD as a non-abelian gauge theory that describes strong interactions in terms of quarks and gluons. The characteristic feature of a Yang-Mills theory, asymptotic freedom, implies however that perturbation theory breaks down for small momentum transfer: The strong coupling $\alpha_s(Q^2)$ diverges in perturbation theory at a finite value, introducing Λ_{QCD} as a new mass scale. Thereby the scale invariance of QCD, classically valid in the limit of zero quark masses, is broken. Another consequence of asymptotic freedom is confinement[1]: Quarks and gluons are not observed as isolated particles, but exist only in bound-states which are colour-singlets. Aim of this chapter is to discuss how calculations in perturbation theory using quarks and gluons can be connected to experiments which observe hadrons. For instance, we aim to describe processes like $\bar{X}X \to \bar{q}q \to$ hadrons, where the initial state could be e.g. a pair of leptons or dark matter particles. We start the chapter however elucidating the surprising fact that the masses of hadrons are much larger than expected from the Higgs effect.

17.1. Trace anomaly and hadron masses

The classical QCD Lagrangian is scale and conformally invariant in the limit of zero quark masses. We introduce first these symmetries before we calculate the anomalous term introduced by quantum corrections which breaks these symmetries. It is this "correction" which is responsible for 95% of the mass of ordinary matter in the Universe.

Scale and conformal invariance We defined in chapter 7.4 the dynamical energy-momentum stress tensor $T_{\mu\nu}$ as the response of the matter action $S_{\rm m}$ under an infinitesimal change of the metric tensor, $T_{\mu\nu} = (2/\sqrt{|g|})\delta S_{\rm m}/\delta g^{\mu\nu}$. This procedure implies that we leave temporarily Minkowski space, even if we are only interested in the behaviour of a Lorentz invariant field theory. Having performed the variation $\delta g^{\mu\nu}$, we move back to Minkowski space setting $g^{\mu\nu} \to \eta^{\mu\nu}$.

We want to study when we can extend the Poincaré group as the symmetry group of Minkowski space acting on local fields to the larger group of conformal transformations. We start considering *scale transformations* of the coordinates, $x \to x' = {\rm e}^{\omega}x$ with fixed ω. As a result of this coordinate transformation, the metric tensor changes as

$$g_{\mu\nu}(x) \to g_{\mu\nu}(x') = {\rm e}^{2\omega} g_{\mu\nu}(x)\,. \tag{17.1}$$

Thus distances are rescaled by a constant factor, while angles and hence the light-cone structure given by ${\rm d}s^2 = 0$ are conserved. Since only the latter is important for theories without mass parameters, we expect them to be invariant under such transformations. For an infinitesimal scale change, $x \to x' = (1 + \delta\omega)x$, the metric varies as $\delta g_{\mu\nu} = 2g_{\mu\nu}\delta\omega$. The matter

[1] While confinement is an observational fact, it has not been derived from first principles and is part of one of the six open "millennium problems".

action $S_{\rm m}$ changes as a result of the variation of the metric as

$$\begin{aligned}\delta S_{\rm m} &= \int \mathrm{d}^4x \, \frac{\delta S_{\rm m}}{\delta g^{\mu\nu}} \delta g^{\mu\nu} = \frac{1}{2} \int \mathrm{d}^4x \sqrt{|g|}\, T_{\mu\nu} \delta g^{\mu\nu} = \\ &= \int \mathrm{d}^4x \sqrt{|g|}\, T_{\mu\nu} g^{\mu\nu} \delta\omega = \delta\omega \int \mathrm{d}^4x \sqrt{|g|}\, T_{\mu}{}^{\mu} \, .\end{aligned} \tag{17.2}$$

The variation $\delta S_{\rm m}$ of the action remains unchanged, if we add a total derivative $\nabla^\mu K_\mu$ to $T_\mu{}^\mu$. Thus we can conclude for $\omega = const.$ only that the trace of the stress tensor has to equal a total derivative,

$$T_\mu{}^\mu = g^{\mu\nu} \frac{2}{\sqrt{|g|}} \frac{\delta S_{\rm m}}{\delta g^{\mu\nu}} = \nabla^\mu K_\mu \, , \tag{17.3}$$

which may be zero. If we promote ω to a local function $\omega(x)$, the resulting space-time dependent transformations are called *special conformal transformations*. Now we cannot pull out $\delta\omega(x)$ from the integral and consequently the requirement $\delta S_{\rm m} = 0$ implies that the trace of the stress tensor has to vanish, $T_\mu{}^\mu = 0$. Clearly, conformal invariance implies scale invariance. The opposite direction holds in practically all interesting cases, but exceptions exist.

These additional symmetries should lead according to Noether's theorem to conserved currents. We recall that the stress tensor has the properties $T^{\mu\nu} = T^{\nu\mu}$ and $\partial_\mu T^{\mu\nu} = 0$. For a scale invariant action, we can define the additional conserved current $D^\mu = T^{\mu\nu}x_\nu - K^\mu$, because then

$$\partial_\mu D^\mu = \partial_\mu (T^{\mu\nu} x_\nu - K^\mu) = T^{\mu\nu}\eta_{\mu\nu} - \partial_\mu K^\mu = T_\mu^\mu - \partial_\mu K^\mu = 0 \, . \tag{17.4}$$

The conserved quantity D^μ is called the dilatation current, the four-divergence K^μ the virial current. For a conformally invariant action, we obtain additionally the four conserved currents

$$S^\rho = \mathrm{d}_\nu [\eta^{\mu\nu} x^2 - 2x^\mu x^\nu)] T_\mu^\rho \, , \tag{17.5}$$

which are parametrised by the vector d_μ. The condition that the stress tensor is traceless (and symmetric) implies the conservation of these currents,

$$\partial_\rho S^\rho = \mathrm{d}_\nu \partial_\rho [\eta^{\mu\nu} x^2 - 2x^\mu x^\nu] T_\mu^\rho = 2\mathrm{d}_\nu [x_\rho \eta^{\mu\nu} - x^\mu \delta^\nu_\rho - x^\nu \delta^\mu_\rho] T_\mu^\rho = 0 \, . \tag{17.6}$$

In problem 6.??, we encountered already the conformal Killing equation. Its solutions, the conformal Killing vector fields of Minkowski space, agree with the infinitesimal generators of the group of conformal transformations, cf. appendix B.3. This generalises our results for the Poincaré group from example 6.2.

Theories with traceless stress tensors we met already are electrodynamics (or more generally pure Yang-Mills theories) and the massless Dirac field. Thus the invariance group of these theories in Minkowski space is the 15-dimensional conformal group. Now we look at the case of a scalar field. In problem 5.??, we found that a massless scalar field in invariant under scale transformations. The trace of its stress tensor is given by

$$T_\mu{}^\mu = (\partial_\mu \phi)^2 - \delta^\mu_\mu \mathscr{L}_0 = \left(1 - \frac{d}{2}\right) (\partial_\mu \phi)^2 \, . \tag{17.7}$$

The virial current follows using $\Box\phi = 0$ as $K_\mu = (1 - d/2)\phi\partial_\mu\phi$ and thus the action for a free massless scalar field is scale invariant for all d. By contrast, the action is only in $d = 2$ dimensions conformally invariant, in which case the field is dimensionless and does not scale[2].

Connection to hadron masses The matrix element of the stress tensor $T^{\mu\nu}$ for a hadron $h(\boldsymbol{k})$ can depend for zero momentum transfer only on the combination $k^\mu k^\nu$. Assuming hadron states normalised as $\langle h(\boldsymbol{k})|h(\boldsymbol{k})\rangle = 1$, the matrix element of the stress tensor is therefore

$$\langle h(\boldsymbol{k})|T^{\mu\nu}|h(\boldsymbol{k})\rangle = 2k^\mu k^\nu \,. \tag{17.8}$$

Thus the vanishing of the trace $T_\mu{}^\mu$ would imply that all hadrons are massless,

$$\langle h(\boldsymbol{k})|T_\mu{}^\mu|h(\boldsymbol{k})\rangle = 2m_h^2 = 0 \,. \tag{17.9}$$

In reality, the masses of the light quarks are non-zero and thus the scale invariance of strong interactions is broken already on the classical level. Nevertheless, the relation (17.9) poses a problem, since the masses of the light quarks due to the Higgs effect, $m_{u,d} \approx (4 - 10)$ MeV, are much smaller than the masses of the lightest hadrons, the pions. Thus the dominant contribution to hadron masses should come from strong interactions, but Eq. (17.9) informs us that conformal invariance forbids such a contribution. The resolution to this apparent problem lies in the fact that quantum corrections spoil classical conformal invariance, because we have to introduce a dimensionful parameter calculating quantum corrections.

Trace anomaly The phenomenon that the trace of the stress tensor receives non-zero corrections due to loop effects is called trace anomaly. This effect can be analysed using the same approach as for the axial anomaly, and we summarise therefore only the main steps: If the Jacobian $\mathcal{J}$ for the conformal transformation

$$\psi(x) = \exp\{-3\omega(x)/2\}\psi'(x) \tag{17.10}$$

is not the identity, an anomalous contribution to T_μ^μ in addition to the usual mass term appears,

$$\mathrm{i}\int \mathrm{d}^4x\, \omega\, T_\mu^\mu = \ln\mathcal{J}/3 + \mathrm{i}\int \mathrm{d}^4x\, \omega\, m\bar{\psi}\psi \,. \tag{17.11}$$

Using the same regulator in the evaluation of the Jacobian as in the case of the chiral anomaly,

$$\mathcal{J} = \exp\omega \sum_n \int \mathrm{d}^4x\, \phi_n^*(x) \exp\left\{-\not{D}^2/M^2\right\} \phi_n \,, \tag{17.12}$$

the only change compared to Eq. (16.11) is the absence of the γ_5 matrix. Following the same steps but keeping track of the non-abelian part of the QCD field-strength $F^a_{\mu\nu}$, one finds

$$T_\mu{}^\mu = \frac{g_s^2}{48\pi^2}\, F^a_{\mu\nu}F^{a\mu\nu} + m\bar{\psi}\psi \,. \tag{17.13}$$

[2] The fact that we can express the trace in (17.7) as $T_\mu{}^\mu = \partial_\mu\partial_\nu L^{\mu\nu}$ (where $L^{\mu\nu} = (2 - d)/4\eta^{\mu\nu}\phi^2$ using the equation of motions) signals that we can "improve" the stress tensor adding appropriate terms such that the stress tensor becomes traceless and the action conformally invariant.

This the $\mathcal{O}(g_s^2)$ contribution to the trace anomaly. We can obtain the full non-perturbative expression for the trace anomaly by the following, more intuitive argument: We rescale first the Lagrange density by $A^a_\mu \to \bar{A}^a_\mu = A^a_\mu / g_s$ as

$$\mathscr{L} = -\frac{1}{4} F^a_{\mu\nu} F^{a\mu\nu} + \mathrm{i}\bar{\psi}(\not{\partial} + \mathrm{i}g_s \not{A})\psi \to -\frac{1}{4g_s^2} \bar{F}^a_{\mu\nu} \bar{F}^{a\mu\nu} + \mathrm{i}\bar{\psi}(\not{\partial} + \mathrm{i}\bar{\not{A}})\psi\,. \tag{17.14}$$

If we view this Lagrange density as the renormalised Lagrange density, then g_s depends on the renormalisation scale μ and the action is not longer scale invariant: The change δS under an infinitesimal scale transformations results solely from the dependence via the renormalised coupling,

$$\delta S = -\frac{1}{4} \int \mathrm{d}^4x \, \frac{\partial}{\partial\omega} \left(\frac{1}{4g_s^2(\mu)} \right) \bar{F}^a_{\mu\nu} \bar{F}^{a\mu\nu} \to \int \mathrm{d}^4x \, \frac{\beta(g_s)}{2g_s} F^a_{\mu\nu} F^{a\mu\nu}\,. \tag{17.15}$$

Thus the full result for the trace of the stress tensor in QCD is, including the anomaly and the mass terms of the quarks that break scale invariance explicitly, given by

$$T^{\mu}_{\mu} = \frac{\beta^{(6)}_{\mathrm{QCD}}}{2g} F^a_{\mu\nu} F^{a\mu\nu} + m_u \bar{u}u + m_d \bar{d}d + m_s \bar{s}s + \sum_h m_h \bar{q}q\,. \tag{17.16}$$

Since the anomaly is due to an UV divergence, it is independent of the quark masses and we should use the beta function $\beta_{\mathrm{Q}CD}$ for six flavours, $b_{\mathrm{Q}CD} = 11 - 2n_f/3 = 7$. We have singled out in (17.16) the contribution of the heavy quarks, $\sum_{h=c,b,t} m_h \bar{q}q$, since they are present in a hadron only as virtual fluctuations. Because of $m_h \gg \Lambda_{QCD}$, a systematic expansion in Q^2/m_h^2 of the term $\sum_h m_h \bar{q}q$ is possible. At leading order, these fluctuations connect m_h via a quark loop to two gluons. Therefore one can rewrite the mass term for heavy quarks using the effective Lagrangian derived in problem 14.?? for the Higgs-gluon-gluon coupling,

$$m_h \bar{q}q \to \frac{2}{3} \frac{\alpha_s}{8\pi} F^a_{\mu\nu} F^{a\mu\nu}\,. \tag{17.17}$$

Now we see that the mass terms cancel the contribution of the heavy quarks in the beta function. As a result, we obtain for the matrix element between on-shell nucleons

$$m_N^2 = \langle N | \, T^{\mu}_{\mu} \, | N \rangle = \langle N | \, \frac{\beta^{(3)}}{2g_s} F^a_{\mu\nu} F^{a\mu\nu} \, | N \rangle + \sum_{l=u,d,s} \langle N | \, m_l \bar{q}q \, | N \rangle \tag{17.18}$$

with $b^{(3)}_{\mathrm{Q}CD} = 9$. The matrix elements $\langle N | \, m_l \bar{q}q \, | N \rangle \equiv f_q^{(N)} m_N$ can be estimated using results from lattice QCD simulations: Numerical values for the so-called mass fractions $f_q^{(N)}$ are shown in Table 17.1. Thus only $\sim 5\%$ of the mass of proton is given by the current quarks masses, or in other words by the Higgs effect, while 95% of the proton mass is a consequence of the gluon condensate via the trace anomaly.

17.2. DGLAP equations

Experiments studying electron-nucleon scattering at large momentum transfer $Q^2 \gg m_N^2$ revealed in the late 1960s that a nucleon can be described as a collection of nearly massless, freely interacting scattering centers which share the total nucleon momentum. These point-like scattering centers were called partons and are formed by the valence quarks (which

Proton		Neutron	
$f_u^{(p)}$	0.019(5)	$f_u^{(n)}$	0.013(3)
$f_d^{(p)}$	0.027(6)	$f_d^{(n)}$	0.040(9)
$f_s^{(p)}$	0.009(22)	$f_s^{(n)}$	0.009(22)

Table 17.1.: The mass fractions $f_q^{(N)}$ (with errors) contributed by the u, d and s quarks to the nucleon mass deduced by [HNN15] from lattice QCD simulations.

quantum numbers sum up to those of the nucleon), a sea of virtual $\bar{q}q$ pairs and gluons. Moreover, cross sections were found in this limit to follow approximately scaling invariant: For instance, the energy spectrum of hadrons produced with energy E_h in a reaction at cm energy $\sqrt{s} \gg m_N$ can be approximated by

$$\frac{E_h}{\sigma}\frac{\mathrm{d}\sigma(s,E_h)}{\mathrm{d}E_h} \approx \frac{x}{\sigma}\frac{\mathrm{d}\sigma(x)}{\mathrm{d}x}\,, \tag{17.19}$$

introducing the scaling variable $x = 2E_h/\sqrt{s}$. With hindsight, we can connect these features easily to the general properties of QCD: Asymptotic freedom explains why the constituents of the nucleon behave at large momentum transfer as unbound particles. Moreover, we can treat the light quarks for $Q^2 \gg m_N^2 \gg m_q^2$ as massless, and scattering on quark and gluons contains therefore no mass scale. Hence we expect only logarithmic scaling violations, caused by the running of the coupling $\alpha_s(Q^2)$, in quantities which are not sensitive to IR singularities. As we will discuss at the end of this section, these are so called “inclusive observables” like the e^+e^- annihilation cross section where at fixed order perturbation theory IR divergences from real and virtual gluon emission cancel.

In contrast, “exclusive quantities” like a differential cross section to produce a fixed number of partons contain large IR logarithms. They are caused by the emission of soft and collinear partons, similar to the case of QED discussed in section 9.4, and spoil naive perturbation theory. What comes to our rescue, however, is that these logarithms are connected to the emission of additional partons in the semi-classical regime: Therefore, we should be able to develop an approximate probabilistic picture where quantum mechanical interference terms can be neglected. Additionally, we have to add to this semi-classical picture a connection between the perturbative description using partons, valid at $Q^2 \gg \Lambda_{QCD}^2$, and a description using hadrons, valid at $Q^2 \lesssim \Lambda_{QCD}^2$: This connection cannot (yet) be derived from first principles and therefore one has to employ phenomenological models.

Our aim in this section is to describe processes like $\bar{X}X \to \bar{q}q \to$ hadrons, where the initial state could be e.g. a pair of leptons or DM particles. We can break this process into several steps: The calculation of the “hard” process $\bar{X}X \to \bar{q}q$ uses standard perturbation theory, with which we are familiar by now. Subsequently, a parton cascade develops, $q \to q + g \to q + g + g \to \dots$ In each splitting process, the virtuality t of the partons decreases, until we have to stop at $t \sim few \times \Lambda_{QCD}^2$ the perturbative evolution before $\alpha_s(t)$ becomes too large. Hadrons are then formed out of partons using a phenomenological model. For the calculation of this perturbative parton cascade, we need two main ingredients: An evolution equation which determines the probability that a parton evolves from t_n to t_{n+1} without splitting, and splitting functions $P_{ij}(z)$ which describe how the energy is shared between the daughter partons in a splitting process.

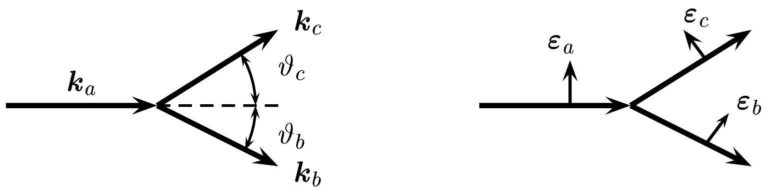


Figure 17.1.: Left: The momenta and angles used to describe the parton splitting $a \to bc$; Right: Choice of the polarisation vectors.

Splitting functions We start determining the splitting functions describing the time-like evolution of a parton cascade. We consider the branching of a parton a into the parton pair bc with $t \equiv k_a^2 \gg k_b^2, k_c^2 > 0$ in the small angle approximation, $\vartheta = \vartheta_b + \vartheta_c \ll 1$, cf. Fig. 17.1. Defining then the energy fraction taken by the parton b as $z = E_b/E_a = 1 - E_c/E_a$, we obtain

$$t = 2E_bE_c(1 - \cos\vartheta) \approx z(1-z)E_a^2\vartheta^2 . \tag{17.20}$$

Conservation of transverse momentum implies $\vartheta_b E_b = \vartheta_c E_c$ or

$$\frac{\vartheta_b}{1-z} = \frac{\vartheta_c}{z} . \tag{17.21}$$

Solving (17.20) for ϑ and using (17.21) together with $\vartheta = \vartheta_b + \vartheta_c$, we find

$$\vartheta = \sqrt{\frac{t}{z(1-z)}}\frac{1}{E_a} = \frac{\vartheta_b}{1-z} = \frac{\vartheta_c}{z} . \tag{17.22}$$

To proceed, we have to fix the parton type and we consider here as an example the process of gluon splitting, $g \to gg$. The triple gluon vertex is given by

$$V_{ggg} = \mathrm{i}gf^{ABC}\varepsilon_a^\alpha\varepsilon_b^\beta\varepsilon_c^\gamma[\eta_{\alpha\beta}(k_a - k_b)_\gamma + \eta_{\beta\gamma}(k_b - k_c)_\alpha + \eta_{\gamma\alpha}(k_c - k_a)_\beta] , \tag{17.23}$$

where all momenta are defined as outgoing, $k_a = -k_b - k_c$, and ε_a^α is the polarisation vector of gluon a. We assume that all three gluons are close to mass shell, $k^2 \approx 0$, what allows us to use transverse polarisation vectors. In order to use these conditions to simplify V_{ggg}, we replace $k_a = -k_b - k_c$ in the first and third, and $k_c = -k_a - k_b$ in the second factor, obtaining

$$V_{ggg} = -2\mathrm{i}gf^{ABC}[(\varepsilon_a \cdot \varepsilon_b)(\varepsilon_c \cdot k_b) - (\varepsilon_b \cdot \varepsilon_c)(\varepsilon_a \cdot k_b) - (\varepsilon_c \cdot \varepsilon_a)(\varepsilon_b \cdot k_c)] . \tag{17.24}$$

Then we evaluate the scalar products $\varepsilon_i \cdot k_j$ in the limit of small ϑ, choosing the two transverse polarisation states as ε_i^{in} in the plane spanned by $\boldsymbol{k}_b$ and $\boldsymbol{k}_c$ as shown in the right panel of Fig. 17.1 and ε_i^{out} perpendicular to this plane,

$$\varepsilon_i^{\mathrm{in}} \cdot \varepsilon_j^{\mathrm{in}} = \varepsilon_i^{\mathrm{out}} \cdot \varepsilon_j^{\mathrm{out}} = -1 \quad and \quad \varepsilon_i^{\mathrm{in}} \cdot \varepsilon_j^{out} = \varepsilon_i^{out} \cdot k_j = 0 .$$

Now we express the three remaining scalar products as functions of E_a and ϑ,

$$\varepsilon_a^{\mathrm{in}} \cdot k_b = -E_b\vartheta_b = -z(1-z)E_a\vartheta , \tag{17.25}$$

$$\varepsilon_b^{\mathrm{in}} \cdot k_c = E_c\vartheta = (1-z)E_a\vartheta , \tag{17.26}$$

$$\varepsilon_c^{\mathrm{in}} \cdot k_b = -E_b\vartheta = -zE_a\vartheta . \tag{17.27}$$

a	b	c	$F(a,b,c;z)$
in	in	in	$\frac{1-z}{z}+\frac{z}{1-z}+z(1-z)$
in	out	out	$z(1-z)$
in	in	out	$\frac{1-z}{z}$
out	out	in	$\frac{z}{1-z}$

Table 17.2.: The function $F^{abc}(z)$ for different gluon polarisations.

Each of the three terms in V_{ggg} contains one factor ϑ. Combined with the propagator $1/t \propto 1/\vartheta^2$, this results in a collinear $1/t$ singularity of the squared amplitude. We express $|\mathcal{A}_{n+1}|^2$ as

$$|\mathcal{A}_{n+1}|^2 \approx \frac{4g^2}{t} C_A \, F(a,b,c;z) |\mathcal{A}_n|^2 \,, \tag{17.28}$$

where the function $F(a,b,c;z)$ for the various non-zero combinations of polarisation vectors is given in table 17.2. The colour factor C_A equals for the gauge group SU(N) $C_A = \sum_{abc} f^{abc} f^{abc} = N$, see Eq. (B.14).

Example 17.1: We evaluate the vertex as an example for the case of $\{abc\} = \{\text{in, out, out}\}$ polarisations,

$$\begin{aligned} V_{ggg} &= -2\mathrm{i}gf^{abc}[(\varepsilon_a^{\mathrm{in}} \cdot \varepsilon_b^{out})(\varepsilon_c^{out} \cdot k_b) - (\varepsilon_b^{out} \cdot \varepsilon_c^{out})(\varepsilon_a^{\mathrm{in}} \cdot k_b) - (\varepsilon_c^{out} \cdot \varepsilon_a^{\mathrm{in}})(\varepsilon_b^{out} \cdot k_c)] \\ &= -2\mathrm{i}gf^{abc}[0 - (-1)(-z(1-z)E_a\vartheta) - 0] = 2\mathrm{i}gf^{abc}[z(1-z)E_a\vartheta] \,. \end{aligned}$$

Thus the squared amplitude for this combination of polarisation states is

$$|\mathcal{A}_{n+1}|^2 = \left|\frac{V_{ggg}}{t}\mathcal{A}_n\right|^2 \approx \frac{4g^2}{t} C_A \, z(1-z) \, |\mathcal{A}_n|^2 \,,$$

and thus $F(\text{in, out, out}; z) = z(1-z)$.

The enhancement of the amplitude for $z \to 0$ (gluon b is soft) and $z \to 1$ (c is soft) comes from the emission of soft gluons polarised in the plane of branching. Therefore $\mathcal{A}_{n+1}$ carries some ϕ dependence. As this dependence is small and washed out considering several successive splittings, one uses unpolarised splitting functions $P_{i\to j}(z)$ defined for $g \to gg$ by

$$P_{g\to g}(z) = \frac{1}{2} C_A \sum_{a,b,c} F(a,b,c;z) = C_A \left[\frac{1-z}{z} + \frac{z}{1-z} + z(1-z)\right] \,. \tag{17.29}$$

The gluon splitting function $P_{g\to g}(z)$ is obviously symmetric under an exchange of the two gluons produced, $P_{g\to g}(z) = P_{g\to g}(1-z)$. It is called an unregularised splitting function, because it becomes infinite for $z \to 1$ (and $z \to 0$). These IR divergences can be cancelled adding the effect of virtual gluon emission, and one obtains thereby regularised splitting functions. Proceeding in the same way for the processes $g \to qq$ and $q \to qg$, you should be able to derive the other three splitting functions of QCD, $P_{g\to q}(z)$, $P_{q\to q}(z)$ and $P_{q\to g}(z)$ (problem 18.**??**).

We can now compute the probability for the emission of an additional parton. We compare the cross section $\mathrm{d}\sigma_{n+1}$ and $\mathrm{d}\sigma_n$ for the process with and without emission of an additional parton. Their phase space is given by

$$\mathrm{d}\Phi_n \propto \frac{\mathrm{d}^3 k_a}{(2\pi)^3 2E_a} \tag{17.30}$$

$g \to gg$	$P_{g\to g}(z) = C_A\left[\frac{1-z}{z}+\frac{z}{1-z}+z(1-z)\right]$
$g \to qq$	$P_{g\to q}(z) = n_f T_F\left[z^2+(1-z)^2\right]$
$q \to qg$	$P_{q\to q}(z) = C_F\left[\frac{1+z^2}{1-z}\right]$
$q \to gq$	$P_{q\to g}(z) = C_F\left[\frac{1+(1-z)^2}{z}\right]$

Table 17.3.: The four unregularised splitting functions of QCD, with $C_F = 4/3$, $T_F = 1/2$, $N_C = 3$ and n_f as the number of active quark flavors.

and

$$\mathrm{d}\Phi_{n+1} \propto \frac{\mathrm{d}^3k_b}{(2\pi)^3 2E_b}\frac{\mathrm{d}^3k_c}{(2\pi)^3 2E_c}\,. \tag{17.31}$$

With $k_c = k_a - k_b$ and $\mathrm{d}^3k_c = \mathrm{d}^3k_a$ for fixed k_b, we obtain trading the variables $\{E_b, \vartheta_b\}$ versus $\{z, t\}$ in the small-angle approximation

$$\begin{aligned}\mathrm{d}\Phi_{n+1} &= \mathrm{d}\Phi_n\frac{1}{2(2\pi)^3}\int E_b\mathrm{d}E_b\vartheta_b\mathrm{d}\vartheta_b\mathrm{d}\phi\mathrm{d}t\frac{\mathrm{d}z}{1-z}\delta(t-E_bE_c\vartheta^2)\delta(z-E_b/E_a)\\ &= \mathrm{d}\Phi_n\frac{1}{4(2\pi)^3}\mathrm{d}t\mathrm{d}z\mathrm{d}\phi\,.\end{aligned} \tag{17.32}$$

Performing the ϕ integral, we find for their ratio,

$$\frac{\mathrm{d}\sigma_{n+1}}{\mathrm{d}\sigma_n} = \frac{\mathrm{d}t}{t}\mathrm{d}z\frac{\alpha_s}{2\pi}P_{i\to j}(z)\,. \tag{17.33}$$

This ratio is the relative probability density for the emission of an additional collinear parton with virtuality t and energy fraction z by the state n. Integrating this probability using $P(z) \sim 1/z$ for the emission of a soft gluon results in $\mathrm{d}\sigma_{n+1}/\mathrm{d}\sigma_n \propto \ln^2(t/t_{\min})$. Thus the probability for the emission of an additional parton increases with t, spoiling naive perturbation theory. Moreover, the preference for the emission of collinear partons is the basis for the explanation of jets observed in hadronic final states.

Evolution equation We want to derive an equation which determines how the number density $f_i(x,t)$ of partons of type i with a certain energy fraction x evolves by parton branching from t to $t+\delta t$. In the case of a space-like evolution, $t_{n+1} > t_n$ and $x_{n+1} < x_n$, such an equation can be used to determine the number of partons inside a hadron, if it is probed e.g. by a photon with virtuality t. The functions $f_i(x,t)$ are then called parton distribution functions. For a time-like evolution, $t_{n+1} < t_n$ and $x_{n+1} < x_n$, the same equation determines the number of partons or hadrons which are produced in a process like $e^+e^- \to$ hadrons as function of the squared cm energy s. The functions $f_i(x,t)$ are then called fragmentation or hadronisation functions.

To be concrete, we consider the space-like evolution of a single parton type. Then the change δf consists of all paths arriving from $x' > x$ and all paths leaving f, as shown in Fig. 17.2. We obtain the number of incoming paths integrating from $x' = x$ to $x' = 1$ under

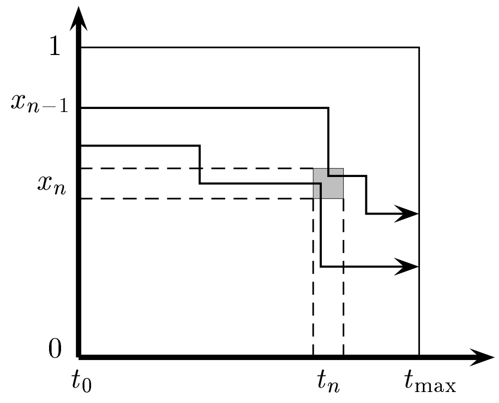


Figure 17.2.: Possible paths with $t_{n+1} > t_n$ and $x_{n+1} < x_n$ describing the space-like evolution of partons.

the constraint $x' = x/z$,

$$\begin{aligned}\delta f_{\text{in}}(x,t) &= \frac{\delta t}{t}\int_x^1 \mathrm{d}x'\mathrm{d}z\,\frac{\alpha_s}{2\pi}\,P(z)f(x',t)\delta(x-zx')\\ &= \frac{\delta t}{t}\int_0^1 \frac{\mathrm{d}z}{z}\,\frac{\alpha_s}{2\pi}\,P(z)f(x/z,t)\,.\end{aligned} \tag{17.34}$$

Here, we extended the integration range from $[x,1]$ to $[0,1]$ in the second step, which is possible because of $f(x/z,t)=0$ for $z<x$. In the same way, we find the number of leaving paths as

$$\begin{aligned}\delta f_{out}(x,t) &= \frac{\delta t}{t}f(x,t)\int_x^0 \mathrm{d}x'\mathrm{d}z\,\frac{\alpha_s}{2\pi}\,P(z)\delta(x'-zx)\\ &= \frac{\delta t}{t}f(x,t)\int_0^1 \frac{\mathrm{d}z}{z}\,\frac{\alpha_s}{2\pi}\,P(z)\,.\end{aligned} \tag{17.35}$$

Combing both expressions, we obtain for the total change

$$\delta f(x,t) = \delta f_{\text{in}}(x,t) - \delta f_{out}(x,t) = \frac{\delta t}{t}\int_0^1 \mathrm{d}z\,\frac{\alpha_s}{2\pi}\,P(z)\left[\frac{f(x/z,t)}{z} - f(x,t)\right]\,. \tag{17.36}$$

Now we introduce the “plus prescription”, defining

$$\int_0^1 \mathrm{d}x\,\left.\frac{g(x)}{x}\right._+ \equiv \int_0^1 \mathrm{d}x\,\left[\frac{g(x)-g(1)}{x}\right]\,, \tag{17.37}$$

for any sufficiently regular function $g(x)$, and regularised splitting functions $\hat{P}(z)$ by

$$\hat{P}(z) = P(z)_+\,. \tag{17.38}$$

Thus the regularised splitting function $\hat{P}(z)$ equals the unregularised splitting function $P(z)$ everywhere except at $z=1$ where a delta function is added so that Eq. (17.37) is satisfied. For $z<1$, both splitting function describe the emission of real partons and allow a probabilistic interpretation. The emission of virtual partons does not change the energy and corresponds therefore to the contribution at $z=1$. The coefficient of the delta function can be explicitly determined requiring momentum conservation, see problem 18.**??**, without the need to evaluate virtual processes. Rewriting the evolution equation as a differential equation, we obtain the DGLAP equation[3],

$$t\,\frac{\partial f(x,t)}{\partial t}=\frac{\alpha_s(t)}{2\pi}\int_x^1\frac{\mathrm{d}z}{z}\,\hat{P}(z)\,f(x/z,t)\,. \tag{17.39}$$

Given the initial values on $[x_{\min}:1]$ of the function f at a fixed scale t_0, we can calculate within perturbative QCD its evolution as function of the virtuality t. The initial values $f(x,t_0)$, however, have to be determined from measurements. These functions depend on the hadron probed, but are independent of a specific process. Thus one can determine e.g. $f_i^p(x,t)$ for a proton from a mixture of pp and e^-p reactions, and apply them then to neutrino-nucleon scattering.

Monte Carlo methods The evolution equation (17.39) admits a probabilistic interpretation, since interference terms disappeared in the semi-classical limit. This allows a straightforward application of Monte Carlo methods to solve it. Compared to a numerical integration of (17.39), the Monte Carlo approach has the advantage of being more flexible, allowing e.g. in a straightforward way to add hadronisation models and experimental cuts. Moreover, this method reproduces event-by-event fluctuation which is an important ingredient to estimate the statistical significance of an observed signal.

We can either start from Eq. (17.36) or use as short-cut the analogy to the radioactive decay problem with a time-varying decay constant. Using the latter picture, we denote the differential probability for the decay of an atom (or the branching of a parton) at t by $\mathcal{P}(t)$. It equals minus the time-derivative of the survival probability which we call $\Delta(t)$. Moreover, the differential probability $\mathcal{P}(t)$ that something happens at t is proportional to the survival probability $\Delta(t)$, with a proportionality constant $F(t)$ which equals the decay rate in the radioactive decay problem,

$$\mathcal{P}(t)=-\frac{\mathrm{d}\Delta(t)}{\mathrm{d}t}=F(t)\Delta(t)\,. \tag{17.40}$$

Integration gives

$$\Delta(t)=\Delta(0)\exp\left\{-\int_0^t\mathrm{d}tF(t)\right\}\,, \tag{17.41}$$

if the evolution starts at $t=0$, and thus

$$\mathcal{P}(t)=F(t)\Delta(t)=F(t)\exp\left\{-\int_0^t\mathrm{d}t'F(t')\right\}\,. \tag{17.42}$$

In the case of a parton cascade, $\mathcal{P}(t)$ denotes the differential probability that a splitting happens at the virtuality t, while the probability that a parton survives, i.e. does not split between the virtuality t and t_{min}, is called the Sudakov form factor $\Delta(t)$. In turn, the probability for a parton to survive between t' and t is given by the ratio $\Delta(t')/\Delta(t)$. The function $F(t)$ playing the role of the decay constant is given by the ratio $\mathrm{d}\sigma_{n+1}/\mathrm{d}\sigma_n$ from Eq. (17.33). Because of its $1/t$ singularity, we have to introduce a cutoff t_0 which should be

[3] Until the 1990s, equations of this type were called Altarelli-Parisi equations. The name evolved then from Gribov-Lipatov-AP to its present form, Dokshitzer-Gribov-Lipatov-Altarelli-Parisi or DGLAP equations.

moreover large enough that we can trust perturbative calculations. Combining all this, the differential probability distribution for the splitting of the parton i into the channels j is

$$\mathrm{d}\mathcal{P}_i(t,z) = \sum_j \frac{\mathrm{d}t}{t}\frac{\mathrm{d}z}{2\pi}\,\alpha_s P_{ij}(z)\,\frac{\Delta_i(t')}{\Delta_i(t)}\,. \tag{17.43}$$

Here, the sum includes all possible branching channels j, and Δ_i is given by

$$\Delta_i(t) = \exp\left[-\int_{t_{\min}}^{t}\frac{\mathrm{d}t'}{t'}\sum_j\int_{z_{\min}}^{z_{max}}\frac{\mathrm{d}z}{2\pi}\,\alpha_s P_{ij}(z)\right]\,. \tag{17.44}$$

As the use of distributions like $\delta(1-x)$ is problematic doing numerical calculations, we have used the unregularised splitting functions $P_{ij}(z)$ given in Table 17.3 together with an IR cutoff $t_{\min}$. The latter serves both as a boundary to the non-perturbative regime, $t_{\min} \gg \Lambda^2_{QCD}$, and as a regulator of the IR singularities in the splitting functions,

$$z_{\min} = \sqrt{t_{\min}/t} \quad and \quad z_{max} = 1 - \sqrt{t_{\min}/t}\,. \tag{17.45}$$

We can now set up a simple Monte Carlo scheme for the simulation of parton cascades. In each step $(t_n, x_n) \to (t_{n+1}, x_{n+1})$, we determine first the new t_{n+1} from the Sudakov factor as follows: We choose a random number r from an uniform distribution between $[0,1]$. For a space-like parton cascade, we compare then this random number with the probability for the evolution from t_n to $t_{n+1} > t_n$,

$$r = \frac{\Delta(t_{n+1})}{\Delta(t_n)}\,, \tag{17.46}$$

and solve for t_{n+1}. If t_{n+1} is larger than the hard scale in the process considered, the cascade stops. Otherwise we determine the splitting channel j and the energy fraction $z = x_{n+1}/x_n$ according to the probability distribution $\alpha_s P_{i\to j}(z)$ and continue.

For a time-like parton cascades, where the virtuality decreases, we compare a random number r with

$$r = \frac{\Delta(t_n)}{\Delta(t_{n+1})}\,. \tag{17.47}$$

If $r < \Delta(t_n, t_0)$, the equation has no solution and no further splitting happens. Otherwise we solve again for the parton type and the energy fraction z and continue. As final outcome of the time-like cascade, we obtain a set of quarks and gluons whose virtualities are close to our cutoff scale $t_{\min}$.

This description of a parton cascade has been schematic in few respects: First, we have not yet specified the argument of α_s. A more careful analysis shows that choosing $p_\perp^2 = z(1-z)t$ sums up partially NLL effects. Second, in a time-like cascade coherence effects lead to an angular-ordered cascade. As a result, the evolution parameter used differs for the two cases.

Hadronisation models The quarks and gluons produced as the final state of a parton shower have to be converted into hadrons using a phenomenological model. An ansatz which requires no additional theoretical input is to determine fragmentation functions $f_i^h(z, t_0)$ from experimental data. These functions give the probability that a parton i with energy E produces a hadron h with energy zE. Convoluting then the fragmentation functions with the parton

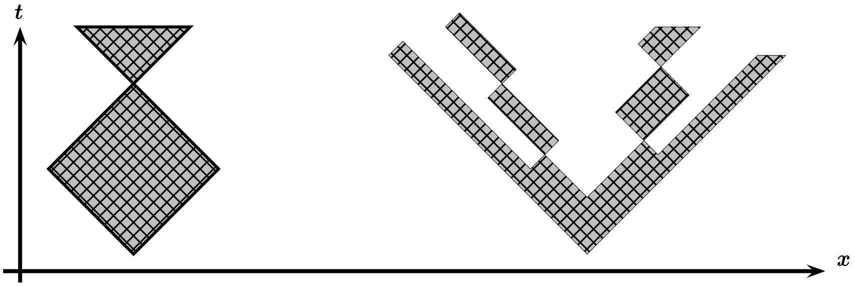


Figure 17.3.: Left: Yo-yo movement of a $\bar{q}q$ pair in the massless limit. Right: Break-up of the string into fragments.

spectra $D_i(x/z,\sqrt{s},t_0)$ obtained from the perturbative evolution from $\sqrt{s}$ down to t_0, one obtains the hadron spectra as

$$D_h(x,\sqrt{s}) = \sum_{i=q,g} \int_x^1 \frac{\mathrm{d}z}{z}\, D_i(x/z,\sqrt{s},t_0) f_i^h(z,t_0)\,. \tag{17.48}$$

An alternative to this purely phenomenological method is to develop models that try to capture basic properties of QCD like confinement. The cluster hadronisation model used first in HERWIG is based on the idea of pre-confinement, i.e. it assumes that the parton spectra at the end of the cascade resemble closely the hadron spectra. Since the P_{gg} splitting function is the most singular one, the majority of partons in the cascade are gluons. In order to create the valence quarks necessary to form mesons and baryons, one adds therefore an artificial splitting $g \to \bar{q}q$ after the parton cascade stopped at $t_{\min}$.

The Lund string model which is another scheme often employed in Monte Carlo simulations is based on a rather different ansatz: Lattice QCD calculations show that the potential between two static quarks can be approximated at large distance by $V(r) \approx \kappa r$ with $\kappa \approx 0.2\,\mathrm{GeV}^2$. Thus in a confining theory like QCD, the force lines between a $\bar{q}q$ pair are concentrated in a narrow tube connecting the pair. This tube of colour field can be viewed as a string with tension κ, trying to pull the quark pair together. The string model uses this picture to describe the hadronisation of partons into hadrons. Let us discuss the simplest case that a $\bar{q}q$ pair is created in an e^+e^- annihilation. As we will show in section 22.3, a constant acceleration $\ddot{x} = -\kappa$ leads to a hyperbola as trajectory. In the limit of massless quarks, we can approximate the hyperbola by straight line segments as shown in Fig. 17.3: Neglecting further interactions, the two quarks would oscillate back and forth in a yo-yo mode. Instead, a new $\bar{q}q$ pair will be created when the energy $\kappa\Delta x$ in the string tube is sufficiently large such that two colour singlet states can be formed, $(\bar{q}q)(\bar{q}q)$. If one adds the assumption that all breakings happen during the initial expansion phase of the yo-yo modes, a formulation of the break-up process as a probabilistic process is possible. Moreover, this process is Lorentz invariant, since it is based on the dynamics of a relativistic string.

While the basic picture underlying the cluster hadronisation or the string model are theoretically well motivated, a rather large number of additional assumptions and parameters is required to model the momentum distributions of the various types of mesons and baryons produced and their branching ratios. These parameters are partly obtained from fits to experimental data and interfere with the properties of the basic hadronisation model. It is therefore difficult to differentiate which one of these two models better describes nature.

17.3. Corrections to e^+e^- annihilation

In this section, we examine the total annihilation cross section of an e^+e^- pair into hadrons. While the scheme (17.43–17.44) allows us to calculate the relative fraction of final-states in a QCD process, it accounts for the effect of virtual processes only indirectly, imposing flavour conservation, $\sum_i \int \mathrm{d}x P_{q\to q}(x) = 1$, or momentum conservation, $\sum_i \int \mathrm{d}x x P_i = 1$, on the regularised splitting functions. As a consequence, we cannot calculate the impact of higher-order corrections to the *total* rate within this formalism. Instead, a perturbative QCD calculation for the total annihilation cross section is sufficient and, as we will see, free IR divergences. This requires that the large logarithms which are present in individual processes cancel in the total e^+e^- annihilation cross section.

Let us give first a heuristic argument why the total annihilation cross section into hadrons can be calculated using partons: The partonic process contains the time scale $1/\sqrt{s}$, while hadronisation occurs on the time scale $1/\Lambda_{QCD}$. This means that the production of partons proceeds independently of their latter hadronisation, if we consider processes with $s \gg \Lambda^2_{QCD}$. More precisely, we expect from hadronisation corrections of order Λ^2_{QCD}/s to the result of a perturbative QCD calculations using partons. We can make this argument a bit more quantitative: Let us denote the amplitude to create a quark state by a hadronic QCD current $J(t)$ which evolves then into a final-state of hadrons $|h\rangle$ schematically by $\mathcal{A}_h \propto \int \mathrm{d}t \, \langle 0| \, J(t)U(t,\infty) \, |h\rangle$. The probability determining the total annihilation cross section is then $|\mathcal{A}_h|^2$ summed over all possible hadronic states,

$$\begin{aligned}\sigma_{ann} &\propto \sum_h \mathcal{A}_h \mathcal{A}_h^* \propto \sum_h \langle 0| \, J(t')U(t',\infty) \, |h\rangle \, \langle h| \, U(t,\infty)J(t) \, |0\rangle = \\ &= \langle 0| \, J(t')U(t',\infty)U(\infty,t)J(t) \, |0\rangle = \langle 0| \, J(t')U(t',t)J(t) \, |0\rangle \, .\end{aligned} \tag{17.49}$$

Thus unitarity, $\sum_h |h\rangle \langle h| = 1$, leads to a cancellation of the complicated long-distance physics, and only the short-distance scale $t - t' \sim 1/\sqrt{s}$ enters the total e^+e^- annihilation cross section. As a result, such quantities can be calculated within perturbative QCD.

After these preliminaries, let us move to the specific process $e^+e^- \to$ h*adrons*. We split the annihilation cross section into

$$\sigma_{\mathrm{h}adrons} = \sigma_{\mathrm{q}\bar{q}} + \sigma_{\mathrm{q}\bar{q}g} + \ldots \tag{17.50}$$

according to the partonic final state. Factoring out explicitly the strong coupling, the perturbative series up to α_s is

$$\sigma_{\mathrm{q}\bar{q}} = \frac{1}{4I}\int \mathrm{d}\Phi_2 \, |\mathcal{A}_{\mathrm{q}\bar{q}}|^2 = \frac{1}{4I}\int \mathrm{d}\Phi_2 \, \left[|\mathcal{A}^{(0)}_{\mathrm{q}\bar{q}}|^2 + 2\alpha_s \Re\left(\mathcal{A}^{(0)}_{\mathrm{q}\bar{q}}\mathcal{A}^{(2)*}_{\mathrm{q}\bar{q}}\right) + \ldots\right] , \tag{17.51a}$$

$$\sigma_{\mathrm{q}\bar{q}g} = \frac{1}{4I}\int \mathrm{d}\Phi_3 \, |\mathcal{A}_{\mathrm{q}\bar{q}g}|^2 = \frac{1}{4I}\int \mathrm{d}\Phi_3 \, \left[\alpha_s|\mathcal{A}^{(1)}_{\mathrm{q}\bar{q}g}|^2 + \ldots\right] . \tag{17.51b}$$

We expect that the IR divergences of the real gluon emission process are cancelled by those of the interference term between the tree-level process and the virtual correction. At order $\mathcal{O}(\alpha_s)$, a finite total annihilation cross section requires

$$\int \mathrm{d}\Phi_2 \, 2\Re\left(\mathcal{A}^{(0)}_{\mathrm{q}\bar{q}}\mathcal{A}^{(2)*}_{\mathrm{q}\bar{q}}\right) + \int \mathrm{d}\Phi_3 \, |\mathcal{A}^{(1)}_{\mathrm{q}\bar{q}g}|^2 = \text{finite}. \tag{17.52}$$

As we are used from the case of UV divergences, we have to apply also to IR divergences a regularisation scheme in order to make our mathematical manipulations well-defined. The simplest scheme, introducing a finite gluon mass, breaks gauge invariance. We can maintain gauge invariance using again DR, now however in the limit $d = 2\omega = 4 + 2\varepsilon$, i.e. extrapolating to more than four dimensions.

Before embarking onto calculations, let us discuss two features of this process that can be used to simplify the calculations. First, we can express $|\bar{\mathcal{A}}|^2 = L_{\mu\nu}H^{\mu\nu}/Q^4$ always as the product of a leptonic and hadronic tensor, $L_{\mu\nu}$ and $H_{\mu\nu}$, if we are interested only in QCD corrections. Moreover, QCD corrections affect only the hadronic tensor $H_{\mu\nu}$. Then we can write the cross sections as

$$\sigma = \frac{1}{2Q^2} L_{\mu\nu} \frac{1}{Q^4} \int \mathrm{d}\Phi \, H^{\mu\nu} \,, \tag{17.53}$$

where we used the flux factor $4I = 4p_{cms}\sqrt{s} = 2s$ and $s = Q^2$. Second, gauge invariance implies $Q_\mu L^{\mu\nu} = \cdots = Q_\nu H^{\mu\nu} = 0$, where Q_μ denotes the four-momentum of the photon or Z boson in the s channel. Hence the tensor structure of the leptonic and hadronic tensor has to be of the form $-\eta^{\mu\nu} + Q^\mu Q^\nu/Q^2$. This allows us to simplify the phase space integration, introducing

$$\int \mathrm{d}\Phi \, H^{\mu\nu}(Q) = \frac{1}{d-1}\left(-\eta^{\mu\nu} + \frac{Q^\mu Q^\nu}{Q^2}\right) H(Q^2) \tag{17.54}$$

with

$$H(Q^2) = -\eta_{\mu\nu} \int \mathrm{d}\Phi \, H^{\mu\nu}(Q) = -H_\mu^{\ \mu}(Q^2)\Phi \,. \tag{17.55}$$

The last step is valid, when $H^{\mu\nu}$ depends only on Q^2, as it is the case after averaging over spins. Thus our task is reduced to the calculation of the trace of the hadronic tensor. As additional simplification, we include only the photon diagrams which give the dominant contribution to the total annihilation cross section of an e^+e^- pair into hadrons, except close to the Z pole when $s \simeq m_Z^2$.

Tree-level process We consider first the tree-level process $e^+e^- \to \bar{q}q$, denoting the four-momenta of the particles by $l^+, l^-, \bar{q}, q$. The leptonic and hadronic tensor, $L_{\mu\nu}$ and $H_{\mu\nu}$, are given by

$$L_{\mu\nu} = e^2[l_\mu^+ l_\nu^- + l_\mu^- l_\nu^+ - Q^2\eta_{\mu\nu}/2] \tag{17.56}$$

and

$$H_{\mu\nu} = (e_q e)^2 4N_c[\bar{q}_\mu q_\nu + q_\mu \bar{q}_\nu - Q^2\eta_{\mu\nu}/2] \,, \tag{17.57}$$

where e_q denotes the quark charge in units of the positron charge e. Evaluating $L_\mu^\mu = e^2[2(l^+ \cdot l^-) - Q^2 d/2] = e^2 Q^2(1 - d/2)$, we find as general expression for the cross section

$$\sigma = \frac{e^2}{4Q^4}\frac{d-2}{d-1} H(Q^2) \,. \tag{17.58}$$

The specific ingredient of the tree-level process is the trace of the hadronic tensor. In the limit of massless quarks, we obtain for its trace

$$H_\mu^{\ \mu} = -2(e_q e)^2 N_c (2 - d) Q^2 \,. \tag{17.59}$$

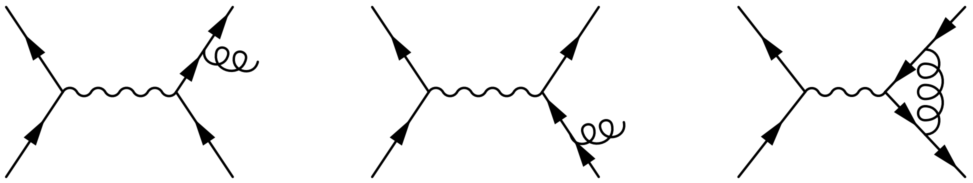


Figure 17.4.: Feynman diagrams for real gluon emission and vertex correction.

Adding also the two-particle phase space in DR, cf. Eq. (17.82), with $p^2 = s/4 = Q^2/4$ gives for the tree-level cross section

$$\begin{aligned}\sigma^{(0)} &= \frac{e^2}{2Q^4}\frac{1-\varepsilon}{3-2\varepsilon} \times 4(e_q e)^2 N_c Q^2(1-\varepsilon) \times \frac{1}{4\pi}\frac{1}{2}\left(\frac{4\pi}{Q^2}\right)^{\varepsilon}\frac{\Gamma(1-\varepsilon)}{\Gamma(2-2\varepsilon)} \\ &= \frac{4\pi\alpha^2}{2Q^2} e_q^2 N_c \left(\frac{4\pi\mu^2}{Q^2}\right)^{\varepsilon}\frac{(1-\varepsilon)^2}{3-2\varepsilon}\,\frac{\Gamma(1-\varepsilon)}{\Gamma(2-2\varepsilon)}\,.\end{aligned} \tag{17.60}$$

Simple inspection shows that this formula reproduces for $\varepsilon \to 0$ the standard result (problem 18.??).

Real emission The hadronic current including the emission of an additional gluon with momentum g and polarisation vector $\varepsilon_\sigma^{(r)}$ as shown by the first two graphs in Fig. 17.4 is given by

$$J^\mu = \varepsilon_\sigma^{(r)*} J^{\mu\sigma} = -\mathrm{i}\mu^{2\varepsilon} e e_q g_s T^a_{ij} \varepsilon_\sigma^{(r)*} \bar{u}(q)\left[\gamma^\sigma \frac{\not{q}+\not{g}}{(q+g)^2}\gamma^\mu + \gamma^\mu \frac{-\not{\bar{q}}-\not{g}}{(\bar{q}+g)^2}\gamma^\sigma\right] v(\bar{q})\,.$$

The hadronic tensor is obtained contracting the squared currents with the polarisation vector of the gluon,

$$H^{\mu\nu} = \sum_r \varepsilon_\sigma^{(r)*}\varepsilon_\rho^{(r)} \sum J^{\mu\sigma} J^{\rho\nu *} = -J^{\mu\sigma} J_\sigma^{\ \nu *}\,. \tag{17.61}$$

Here we could evaluate the polarisation sum using $\sum \varepsilon_\mu^{(r)*}\varepsilon_\nu^{(r)} = -\eta_{\mu\nu}$, because only o ne external gluon is involved. Evaluating then the trace $H^{\mu\nu}$ gives us

$$H_\mu^{\ \mu} = C_F N_C (e e_q g_s)^2 \left[\frac{S_{qq}}{(2q\cdot g)^2} + \frac{S_{\bar{q}\bar{q}}}{(2\bar{q}\cdot g)^2} - \frac{S_{qq}+S_{\bar{q}\bar{q}}}{(2q\cdot g)(2\bar{q}\cdot g)}\right]\,. \tag{17.62}$$

In this expression, IR divergences have shown up as poles for $q\cdot g \to 0$ and $\bar{q}\cdot g \to 0$. With $q\cdot g = E_q E_g(1-\cos\beta_q)$, where β_q is the quark velocity, we can identify them as a combination of soft ($E_g \to 0$) and collinear ($\beta_q \to 0$) singularities. Next we evaluate the spinor traces, finding

$$S_{qq} = S_{\bar{q}\bar{q}} = 32(1-\varepsilon)^2[(q\cdot g)(\bar{q}\cdot g)] \tag{17.63}$$

and

$$S_{q\bar{q}} + S_{\bar{q}q} = -32(1-\varepsilon)[(q\cdot q)Q^2 - 2\varepsilon(q\cdot g)(\bar{q}\cdot g)]\,. \tag{17.64}$$

As we have already seen discussing the splitting functions, the double pole from the propagators will be converted into a simple pole, cancelling against the factor $(q \cdot g)(\bar{q} \cdot g)$ in the traces.

Now we introduce the momentum fractions $x_i = 2p_i \cdot Q/Q^2$, which for massless quarks become e.g. $1 - x_q = 2\bar{q} \cdot g/Q^2$. Combining then H_μ^μ together with the three-particle phase space given in (17.83), we obtain

$$\begin{aligned} H_\mu^\mu = & \frac{C_F N_C e_q^2 \alpha \alpha_s Q^2}{\pi} \left(\frac{4\pi\mu^2}{Q^2} \right)^\varepsilon \frac{(1-\varepsilon)}{\Gamma(2-2\varepsilon)} \\ & \times \int_0^1 \mathrm{d}x_q \int_{1-x_{\bar{q}}}^1 \mathrm{d}x_{\bar{q}} \frac{1}{[(1-x_q)(1-x_{\bar{q}})(x_q + x_{\bar{q}} - 1)]^\varepsilon} \\ & \times \left[(1-\varepsilon) \left(\frac{1-x_q}{1-x_{\bar{q}}} + \frac{1-x_{\bar{q}}}{1-x_q} \right) + \frac{2(x_q + x_{\bar{q}} - 1)}{(1-x_q)(1-x_{\bar{q}})} - 2\varepsilon \right] . \end{aligned} \tag{17.65}$$

The nested integrals become easier substituting $x_q = x$ and $x_{\bar{q}} = 1 - vx$. Solving them using the definition (A.28) of Euler's Beta function, adding the 3-particle phase space (17.83) and expanding then the Gamma function around their poles, we end up with

$$\begin{aligned} \sigma &= \frac{2C_F N_C e_q^2 \alpha \alpha_s}{Q^2} \mu^\varepsilon \left(\frac{4\pi\mu^2}{Q^2} \right)^\varepsilon \frac{(1-\varepsilon)}{\Gamma(2-2\varepsilon)} \frac{\Gamma^3(1-\varepsilon)}{\Gamma(1-3\varepsilon)} \left(\frac{2}{\varepsilon^2} + \frac{3}{\varepsilon} + \frac{19}{2} + \mathcal{O}(\varepsilon) \right) \\ &= \sigma^{(0)} \frac{C_F \alpha_s}{2\pi} \left(\frac{4\pi\mu^2}{Q^2} \right)^\varepsilon \frac{\Gamma^2(1-\varepsilon)}{\Gamma(1-3\varepsilon)} \left(\frac{2}{\varepsilon^2} + \frac{3}{\varepsilon} + \frac{19}{2} + \mathcal{O}(\varepsilon) \right) . \end{aligned} \tag{17.66}$$

Here, the $1/\varepsilon^2$ term corresponds to a combined soft and collinear divergence, while the $1/\varepsilon$ term corresponds to soft singularities.

Virtual corrections The QCD one-loop contribution to $e^+e^- \to \bar{q}q$ consists of self-energy corrections to the quark lines and the vertex correction. The interference terms between these virtual corrections and the tree-level diagram are of the same order as the real emission, cf. Eq. (17.51a), and should cancel the IR divergences found above. Since we restrict ourselves to massless quarks, the self-energy diagram contains no scale and vanishes in DR, cf. the discussion of tadpoles in section 12.3.2. Therefore we have to consider only the vertex correction. Contracting the hadronic tensor of the interference term between the vertex correction evaluated in R_ξ gauge with $\xi = 1$ and the tree-level diagram gives

$$\begin{aligned} -\eta_{\mu\nu} H^{\mu\nu} = 2\mathrm{i}(eqg_s\mu^{2\varepsilon})^2 \mathrm{t}r\{T^b T^b\} \int \frac{\mathrm{d}^d k}{(2\pi)^d} \frac{1}{k^2(k+q)^2(k-\bar{q})^2} \\ \times \mathrm{t}r\{\not{q}\gamma_\sigma(\not{k} + \not{q})\gamma_\mu(\not{k} - \not{\bar{q}})\gamma^\sigma \not{q}\gamma^\mu\} . \end{aligned} \tag{17.67}$$

Combining the denominators using Feynman parameters results in

$$\begin{aligned} \frac{1}{k^2(k+q)^2(k-\bar{q})^2} &= \int_0^1 \mathrm{d}\alpha \int_0^{1-\alpha} \mathrm{d}\beta \frac{2}{[\alpha(k+q)^2 + \beta(k-\bar{q})^2 + (1-\alpha-\beta)k^2]^3} \\ &= \int_0^1 \mathrm{d}\alpha \int_0^{1-\alpha} \mathrm{d}\beta \frac{2}{[(k + \alpha q - \beta\bar{q})^2 + \alpha\beta Q^2]^3} . \end{aligned} \tag{17.68}$$

Now the change of variables $k \to k - \alpha q + \beta \bar{q}$ transforms the loop integral into standard form, $I(\omega, 3)$. Note that the loop integral in UV finite, since QCD corrections cannot renormalise the $\gamma\bar{q}q$ vertex. Next we work out the trace in (17.67), and shift then variables, obtaining

$$\begin{aligned} tr\{\cdots\} &= 8(1-\varepsilon)\{Q^4 - 4[(k\cdot q)(k\cdot\bar{q}) - 2k\cdot(q-\bar{q})Q^2 + \varepsilon k^2 Q^2\} \\ &= 8(1-\varepsilon)Q^2\{[1-\alpha-\beta+(1-\varepsilon)\alpha\beta]Q^2 - (1-\varepsilon)^2(2-\varepsilon)^{-1}k^2\}\,. \end{aligned} \tag{17.69}$$

Here, we have omitted all terms linear in k and replaced in the second line $k^\mu k^\nu$ by $k^2\eta^{\mu\nu}/d$. Combining then the trace and the loop integral, we obtain

$$\begin{aligned} \sigma_V &= \mathrm{i}\sigma_0^{(\varepsilon)} 4(g_s\mu^\varepsilon)^2 C_F \int_0^1 \mathrm{d}\alpha \int_0^{1-\alpha} \mathrm{d}\beta \int \frac{\mathrm{d}^d k}{(2\pi)^d} \frac{1}{[k^2+\alpha\beta Q^2]^3} \\ &\quad\times \left[(1-\alpha-\beta+(1-\varepsilon)\alpha\beta)Q^2 - (1-\varepsilon)^2/(2-\varepsilon)k^2\right] \\ &= -\sigma_0^{(\varepsilon)} \frac{\alpha_s}{2\pi} C_F \left(\frac{4\pi\mu^2}{-Q^2}\right)^\varepsilon \Gamma(1+\varepsilon)\, I(\varepsilon)\,, \end{aligned} \tag{17.70}$$

where we decouple the parameter integrals in $I(\varepsilon)$ by the substitution $\beta = (1-\alpha)v$,

$$\begin{aligned} I(\varepsilon) &= \int_0^1 \mathrm{d}\alpha \int_0^{1-\alpha} \mathrm{d}\beta \, \frac{1}{(\alpha\beta)^\varepsilon} \left[\frac{1-\alpha-\beta+(1-\varepsilon)\alpha\beta}{\alpha\beta} - \frac{(1-\varepsilon)^2}{\varepsilon}\right] \\ &= \int_0^1 \mathrm{d}\alpha \int_0^1 \mathrm{d}v \, \frac{1-\alpha}{[\alpha(1-\alpha)v]^\varepsilon} \left[\frac{(1-\alpha)(1-v)+(1-\varepsilon)\alpha(1-\alpha)v}{\alpha(1-\alpha)v} - \frac{(1-\varepsilon)^2}{\varepsilon}\right] \\ &= \frac{\Gamma^2(1-\varepsilon)}{\Gamma(2-2\varepsilon)} \left(\frac{1}{\varepsilon^2} + \frac{1}{2} + \frac{1-\varepsilon}{2\varepsilon}\right) = \frac{1}{2}\frac{\Gamma^2(1-\varepsilon)}{\Gamma(1-2\varepsilon)} \left(\frac{2}{\varepsilon^2} + \frac{3}{\varepsilon} + \frac{8}{1-2\varepsilon}\right)\,. \end{aligned} \tag{17.71}$$

Inserting the result in the formula for the cross section gives

$$\sigma_V = -\sigma_0^{(\varepsilon)} \frac{\alpha_s}{2\pi} C_F \left(\frac{4\pi\mu^2}{-Q^2}\right)^\varepsilon \frac{\Gamma(1+\varepsilon)\Gamma^2(1-\varepsilon)}{\Gamma(1-2\varepsilon)} \left(\frac{2}{\varepsilon^2} + \frac{3}{\varepsilon} + 8 + \mathcal{O}(\varepsilon)\right)\,. \tag{17.72}$$

Recall that only the real part of this expression enters the total cross section.

Total cross section Combining the result for the real and virtual correction, we arrive at

$$\begin{aligned} \sigma = \sigma_0^{(\varepsilon)} \Bigg\{ 1 + \frac{\alpha_s}{2\pi} C_F \left(\frac{4\pi\mu^2}{Q^2}\right)^\varepsilon \frac{\Gamma^2(1-\varepsilon)}{\Gamma(1-3\varepsilon)} \Bigg[\left(\frac{2}{\varepsilon^2} + \frac{3}{\varepsilon} + \frac{19}{2} + \mathcal{O}(\varepsilon)\right) \\ + \Re\{(-1)^\varepsilon\} \frac{\Gamma(1+\varepsilon)\Gamma(1-3\varepsilon)}{\Gamma(1-2\varepsilon)} \left(\frac{2}{\varepsilon^2} + \frac{3}{\varepsilon} - 8 + \mathcal{O}(\varepsilon)\right) \Bigg] \Bigg\}\,. \end{aligned} \tag{17.73}$$

A cancellation of the poles requires thus that the prefactor in the second line equals one up to $\mathcal{O}(\varepsilon^3)$ terms. Using

$$\Gamma(1+\varepsilon) = 1 - \varepsilon\gamma + \left(\frac{\pi^2}{12} + \frac{1}{2}\gamma^2\right)\varepsilon^2 + \mathcal{O}(\varepsilon^4) \tag{17.74}$$

we obtain

$$\frac{\Gamma(1+\varepsilon)\Gamma(1-3\varepsilon)}{\Gamma(1-2\varepsilon)} = 1 + \frac{1}{2}(\pi\varepsilon)^2 + \mathcal{O}(\varepsilon^3)\,. \tag{17.75}$$

Combined with $\Re\{(-1)^\varepsilon\} = \Re\{\exp(i\pi\varepsilon)\} = 1 - (\pi\varepsilon)^2/2 + \mathcal{O}(\varepsilon^4)$, we see that the $\mathcal{O}(\varepsilon^2)$ terms do cancel as required in the combined prefactor. The total cross section of e^+e^- annihilation into hadrons at $\mathcal{O}(\alpha\alpha_s)$ is therefore free of IR divergences and we can perform the limit $\varepsilon \to 0$, obtaining

$$\sigma = \sigma_0 \left\{ 1 + \frac{3}{2}\frac{\alpha_s}{2\pi} C_F + \mathcal{O}(\alpha_s^2) \right\} . \tag{17.76}$$

Two obvious questions arise: First, what should we use as argument of α_s? And second, which conditions characterise in general a quantity free of IR divergences? The answer to the first question follows from our general discussion of the RGE equation in section 12.4: We can absorb all scale dependence into the running coupling $\alpha_s(\mu^2)$ choosing as scale $\mu^2 = s$. As a bonus, we sum thereby the leading logarithmic corrections of all orders perturbation theory.

IR safe observables The second question is more complex, and we sketch only a few general results about so-called IR safe observables. From an experimental point of view, physical observables have to be measurable. However, any experimental device has a finite energy and angular resolution, ΔE and $\Delta\vartheta$, respectively. Any soft or collinear splitting below these resolution limits in practise cannot be measured and therefore the individual cross sections of these unresolved processes have to be summed up. Thus any observable like e.g. a n-jet rate should be defined such that it is invariant under the replacement $\boldsymbol{p}_i \to \boldsymbol{p}_j + \boldsymbol{p}_k$ whenever $\boldsymbol{p}_j$ or $\boldsymbol{p}_k$ are small or collinear. While the individual rates of unresolved processes may have IR divergences, they should disappear if we use sufficiently inclusive quantities. This idea is made precise in the Kinoshita-Lee-Nauenberg theorem which states that the total rate

$$\sum_{i,f\in D} |\langle f|\, S\, |i\rangle|^2 \tag{17.77}$$

in a general quantum field theory is IR finite. Here, $D \equiv D(p_i^\mu, \Delta E, \Delta\vartheta_i)$ denotes the set of states which are degenerated for a given energy and angular resolution with the initial and final states, respectively. In general, one has to sum thus also over all unresolved initial states, including e.g. initial bremsstrahlung. An exception is QED, where a summation over unresolved final states is sufficient to obtain IR finite quantities, because no massless particles with non-zero electric charged exist. This also the reason why it was sufficient to sum over final-states in our example, where only the final state contained strongly interacting particles.

17.A. Appendix: Phase space integrals in DR

The two particle phase space $\mathrm{d}\Phi_2$ was given in Eq. (9.120) for $d = 4$ dimensions. We rewrite first this expression for general d,

$$\mathrm{d}^{2d-2}\Phi_2 = (2\pi)^d\, \delta^{(d)}(q - p_1 - p_2)\, \frac{\mathrm{d}^{d-1}p_1}{2E_1(2\pi)^{d-1}}\, \frac{\mathrm{d}^{d-1}p_2}{2E_2(2\pi)^{d-1}}\,, \tag{17.78}$$

and integrate then over $\mathrm{d}^{d-1}p_2$,

$$\int \mathrm{d}^{d-1}p_2\, \delta^{(d)}(q - p_1 - p_2) = \delta(q - E_1 - E_2)\,. \tag{17.79}$$

The remaining integrals become

$$\frac{\mathrm{d}^{d-1}p_2}{2E_1} = \frac{1}{2}E_1^{d-3}\sin^{d-3}\vartheta_1\sin^{d-4}\vartheta_2\cdots\sin\vartheta_{d-3}\mathrm{d}\vartheta_1\mathrm{d}\vartheta_2\cdots\mathrm{d}\vartheta_{d-2}\mathrm{d}E_1\,, \tag{17.80}$$

where the ϑ_i are the angles to $d-2$ coordinate axes. The spin-averaged $|\mathcal{A}|^2$ does not depend on the angles ϑ_i, and thus the integrals can be performed with the help of Eq. (A.29), giving

$$\int\frac{\mathrm{d}^{d-1}p_2}{2E_1} = 2^{d-3}\,\pi^{(d-2)/2}\,\frac{\Gamma(d/2-1)}{\Gamma(d-2)}\,\mathrm{d}E_1\,. \tag{17.81}$$

Evaluating finally the energy integral as in $d=3$, we arrive at

$$\int\mathrm{d}^{2d-2}\Phi_2 = \frac{1}{4\pi}\frac{p}{\sqrt{s}}\left(\frac{\pi}{p^2}\right)^{\varepsilon}\frac{\Gamma(1-\varepsilon)}{\Gamma(2-2\varepsilon)}\,. \tag{17.82}$$

Following the same procedure, one obtains for the three-particle phase space

$$\int\mathrm{d}\Phi_3 = \frac{Q^2}{2(4\pi)^3}\left(\frac{4\pi}{Q^2}\right)^{2\varepsilon}\frac{1}{\Gamma(2-2\varepsilon)}\int_0^1\mathrm{d}x\int_{1-x}^1\mathrm{d}y\,\frac{1}{[(1-x)(1-y)(1-z)]^{\varepsilon}}\,. \tag{17.83}$$

Summary

Theories like QED or QCD with massless fermions contain no dimensionful parameters and are classically scale and conformal invariant. The scale invariance of these theories is broken by quantum corrections which unavoidably introduce a mass scale. This quantum effect is responsible for the bulk of hadron masses.

In the parton picture, we replace a hadron which is probed in a process with momentum transfer Q^2 by a collections of quarks and gluons. For $Q^2 \gg \Lambda^2_{QCD}$, we can treat them as "free" particles which are interacting independently. Measuring the parton distribution functions $f_i(x,Q^2)$ at one scale Q^2, perturbative QCD describes via the DGLAP equation their evolution to the new scale Q'^2. In the calculation of sufficiently exclusive quantities, infrared singularities due to massless gluons and quarks cancel.

Further reading

Two very useful references covering most aspects of perturbative QCD are [DKS09] and [ESW03]. Our presentation follows the one given in these references, where you can find also a discussion of coherence effects and the resulting angular-ordered cascade. Effective low-energy models for QCD as well as the trace anomaly are discussed by [DGH14].

18. Gravity as a gauge theory

We introduce in this chapter the action and the field equations of gravity, proceeding in a way which stresses the similarity of gravity and Yang-Mills theories: In particular, we determine the coupling between matter and gravity promoting the invariance of matter fields under global Lorentz transformations to a local symmetry. As a bonus, this approach allows us to describe also the gravitational interactions of fermions as well as to understand how gravity selects the connection among the many mathematically possible ones on a Riemannian manifold. We also derive the linearised Einstein equations which can be used to describe the emission and propagation of weak gravitational waves.

18.1. Vielbein formalism and the spin connection

The equivalence principle postulates that in a small enough region around the center of a freely falling coordinate system all physics is described by the laws of special relativity. On larger distances, gravity manifests itself as curvature of space-time. Thus physical laws involving only quantities transforming as tensors on Minkowski space are valid on a curved space-time performing the replacement

$$\{\partial_\mu, \eta_{\mu\nu}, \mathrm{d}^4x\} \to \{\nabla_\mu, g_{\mu\nu}, \mathrm{d}^4x\sqrt{|g|}\}\,. \tag{18.1}$$

Here, the covariant derivative ∇_μ was defined using as connection the Christoffel symbols (or Levi-Civita connection) from Eq. (6.43). We recall that the two requirements $\nabla_\rho g_{\mu\nu} = 0$ ("metric connection") and $\Gamma^\alpha{}_{\beta\gamma} = \Gamma^\alpha{}_{\gamma\beta}$ ("torsionless connection") select uniquely this connection. In the following, we want to understand if these conditions are a consequence of Einstein gravity or necessary additional constraints. A useful framework to address these questions is the vielbein formalism which is also necessary to include fermions into the framework of general relativity.

Vielbein formalism We apply the equivalence principle as physical guide line to obtain the physical laws including gravity: More precisely, we use that we can find at any point P a local inertial frame in which the physical laws become those known from Minkowski space. We demonstrate this first for the case of a scalar field ϕ. The usual Lagrange density without gravity,

$$\mathscr{L} = \frac{1}{2}\partial_\mu\phi\partial^\mu\phi - V(\phi)\,, \tag{18.2}$$

is still valid on a general manifold $\mathcal{M}(\{x^\mu\})$, if we use in each point P locally free-falling coordinates, $\xi^a(P)$. In order to distinguish these two sets of coordinates, we label inertial coordinates by Latin letters $a, b, \dots$ while we keep Greek indices $\alpha, \beta, \dots$ for arbitrary coordinates. We choose the locally free-falling coordinates ξ^a to be ortho-normal. Thus in these coordinates the metric is given by $\mathrm{d}s^2 = \eta_{ab}\mathrm{d}\xi^a\mathrm{d}\xi^b$ with $\eta = \mathrm{d}iag(1,-1,-1,-1)$. Then the action of a scalar field including gravity is

$$S[\phi]=\int \mathrm{d}^4\xi\left[\frac{1}{2}\eta_{ab}\partial^a\phi\partial^b\phi-V(\phi)\right] \tag{18.3}$$

with $\partial_a=\partial/\partial\xi^a$. This action looks formally exactly as the one without gravity – however we have to integrate over the manifold $\mathcal{M}(\{x^\mu\})$ and all effects of gravity are hidden in the dependence $\xi^a(x^\mu)$.

We introduce now the vielbein (or for $d=4$ tetrad) fields e^a_α by

$$\mathrm{d}\xi^m=\frac{\partial\xi^m}{\partial x^\mu}\,\mathrm{d}x^\mu\equiv\mathrm{e}^m_\mu(x)\,\mathrm{d}x^\mu\,. \tag{18.4}$$

Thus we can view the vielbein $\mathrm{e}^m_\mu(x)$ both as the transformation matrix between arbitrary coordinates x and inertial coordinates ξ or as a set of four vectors in T^*_xM. In the absence of gravity, we can find in the whole manifold coordinates such that $\mathrm{e}^m_\mu(x)=\delta^m_\mu$. The inverse vielbein e^μ_m is defined analogously by

$$\mathrm{d}x^\mu=\frac{\partial x^\mu}{\partial\xi^m}\,\mathrm{d}\xi^m\equiv\mathrm{e}^\mu_m(x)\,\mathrm{d}\xi^m\,. \tag{18.5}$$

The name is justified by

$$\mathrm{d}\xi^m=\mathrm{e}^m_\mu\,\mathrm{d}x^\mu=\mathrm{e}^m_\mu\mathrm{e}^\mu_n\,\mathrm{d}\xi^n \tag{18.6}$$

and thus $\mathrm{e}^m_\mu\mathrm{e}^\mu_n=\delta^m_n$. We can view the vielbein as a kind of square-root of the metric tensor, since

$$\mathrm{d}s^2=\eta_{mn}\mathrm{d}\xi^m\mathrm{d}\xi^n=\eta_{mn}\mathrm{e}^m_\mu\mathrm{e}^n_\nu\mathrm{d}x^\mu\mathrm{d}x^\nu=g_{\mu\nu}\mathrm{d}x^\mu\mathrm{d}x^\nu \tag{18.7}$$

and hence

$$g_{\mu\nu}=\eta_{mn}\mathrm{e}^m_\mu\mathrm{e}^n_\nu\,. \tag{18.8}$$

Taking the determinant, we see that the volume element is

$$\mathrm{d}^4\xi=\sqrt{|g|}\mathrm{d}^4x=\det(\mathrm{e}^m_\mu)\mathrm{d}^4x\equiv E\mathrm{d}^4x\,. \tag{18.9}$$

We can construct mixed tensors, having both Latin and Greek indices. Then Latin indices are raised and lowered by the flat metric, while Greek indices are raised and lowered by the curved metric. For instance, we can rewrite the energy-momentum tensor as $T_{\mu\nu}=\mathrm{e}^m_\mu T_{m\nu}=\mathrm{e}^m_\mu\mathrm{e}^n_\nu T_{mn}$.

We have now all the ingredients needed to express the action (18.3) in arbitrary coordinates x^μ of the manifold $\mathcal{M}$. We first change the derivatives,

$$\mathscr{L}=\frac{1}{2}\eta_{mn}\mathrm{e}^m_\mu\mathrm{e}^n_\nu\partial^\mu\phi\partial^\nu\phi-V(\phi)=\frac{1}{2}g_{\mu\nu}\partial^\mu\phi\partial^\nu\phi-V(\phi) \tag{18.10}$$

and then the volume element in the action,

$$S[\phi]=\int \mathrm{d}^4x\sqrt{|g|}\left[\frac{1}{2}g_{\mu\nu}\partial^\mu\phi\partial^\nu\phi-V(\phi)\right]\,. \tag{18.11}$$

As it should, we reproduced the usual action of a scalar field including gravity. Note that the sole effect of gravitational interactions is contained in the metric tensor and its determinant, while the connection plays no role since $\nabla_\mu\phi=\partial_\mu\phi$. Similarly, the connection drops out of the Lagrangian of a Yang-Mills field, since its field-strength tensor is antisymmetric.

Fermions and the spin connection We now proceed to the spin-1/2 case. In this case, the simple substitution rule $\partial_\mu \to \nabla_\mu$ can not be used, because the connections defined by Eq. (6.26) can be applied only to objects with tensorial indices. Applying instead the vielbein formalism, we recall first the Dirac Lagrangian without gravity,

$$\mathscr{L} = \bar{\psi}(\mathrm{i}\gamma^\mu \partial_\mu - m)\psi \tag{18.12}$$

where $\{\gamma^\mu, \gamma^\nu\} = 2\eta^{\mu\nu}$. Performing a Lorentz transformation, $\tilde{x}^\nu = \Lambda^\nu{}_\mu x^\mu$, the Dirac spinor ψ transforms as

$$\tilde{\psi}(\tilde{x}) = S(\Lambda)\psi(x) = \exp\left(-\frac{\mathrm{i}}{4}\omega^{\mu\nu}\sigma_{\mu\nu}\right)\psi(x) \tag{18.13}$$

with $\omega^{\mu\nu} = -\omega^{\nu\mu}$ as the six parameters and $\sigma^{\mu\nu} = \frac{\mathrm{i}}{2}[\gamma^\mu, \gamma^\nu]$ as the six infinitesimal generators of these transformations.

Switching on gravity, we replace $x^\mu \to \xi^m$ and $\gamma^\mu \to \gamma^m$. General covariance reduces now to the requirement that we have to allow in any local inertial system independent Lorentz transformations. In other words, we promote the invariance under global, space-time independent Lorentz transformations Λ to the invariance under local Lorentz transformations $\Lambda(x)$. This requirement allows us to derive the correct covariant derivative, in a manner completely analogous to the Yang-Mills case: We have to compensate the term introduced by the space-time dependence of $S(\Lambda(\xi))$ in

$$\partial_a \psi \to \tilde{\partial}_a \tilde{\psi}(\tilde{\xi}) = \Lambda^b{}_a \partial_b [S(\xi)\psi(\xi)] \tag{18.14}$$

by introducing a "Latin" covariant derivative,

$$\nabla_a = \mathrm{e}_a^\alpha \left(\partial_\alpha + \mathrm{i}\omega_\alpha\right) , \tag{18.15}$$

and requiring the inhomogeneous transformation law

$$\omega_\alpha \to \widetilde{\omega}_\alpha = S\omega_\alpha S^\dagger - \mathrm{i}S\partial_\alpha S^\dagger \tag{18.16}$$

for ω. As a result, the covariant derivative transforms as

$$\nabla_a \to \widetilde{\nabla}_a = \Lambda^b{}_a S \nabla_b S^\dagger \tag{18.17}$$

and the Dirac Lagrangian is invariant under local Lorentz transformations. The connection ω_α is a matrix in spinor space. Expanding it in the basis elements $\sigma^{\mu\nu}$, we find as more explicit expression for the covariant derivative

$$\nabla_a = \partial_a + \frac{\mathrm{i}}{2}\omega_a^{\mu\nu}\sigma_{\mu\nu} = \mathrm{e}_a^\alpha \left(\partial_\alpha + \frac{\mathrm{i}}{2}\omega_\alpha^{\mu\nu}\sigma_{\mu\nu}\right) = \mathrm{e}_a^\alpha \left(\partial_\alpha + \omega_\alpha^{\mu\nu} J_{\mu\nu}\right) . \tag{18.18}$$

In the last step we replaced the infinitesimal generators $\sigma^{\mu\nu}$ specific for the spinor representation by the general generators $J^{\mu\nu}$ of Lorentz transformations chosen appropriate for the representation the ∇_a act on. In this form, the covariant derivative can be applied to a field with arbitrary spin. The Lie algebra of the Lorentz group implies that the connection $\omega_a^{\mu\nu}$ is antisymmetric in its Greek indices, if they are both up or down, $\omega_a^{\mu\nu} = -\omega_a^{\nu\mu}$.

The transformation law (18.16) of the spin connection ω_a under a Lorentz transformation S is completely analogous to the transformation properties (10.16) of a Yang-Mills field A^μ under a gauge transformation U. One should keep in mind however two important differences: First, a vector lives in a tangent space which is naturally associated to a manifold: In particular, we can associate a vector in T_PM with a trajectory $x^\mu(\sigma)$ through P. Therefore we have the natural coordinate basis ∂_μ in T_PM and can introduce in a second step vielbein fields. In contrast, matter fields $\psi(x)$ live in their group manifold which is attached arbitrarily at each point of the manifold, and the gauge fields act as a connection telling us how we should transport $\psi(x)$ to $\tilde{\psi}(x')$. Second, we associate physical particles with spin s to irreducible representations of the Poincaré group: Thus we identify fluctuations of the gauge field A_μ as the quanta of the vector field, while we associate in the case of gravity not the fluctuations of the connection but of the metric tensor $g_{\mu\nu}$ with particles.

Transition to the standard notation We now establish the connection between the vielbein and the standard formalism, using that for tensor fields the two formalisms have to agree. Inserting into the definition of the covariant derivative for the components A^a of a vector $\boldsymbol{A}$,

$$\nabla_\mu A^a = \partial_\mu A^a + \omega_\mu{}^a{}_b A^b \,, \tag{18.19}$$

the decomposition $A^a = \mathrm{e}^a_\nu A^\nu$ and requiring the validity of the Leibniz rule gives

$$\nabla_\mu A^a = (\nabla_\mu \mathrm{e}^a_\nu) A^\nu + \mathrm{e}^a_\nu (\nabla_\mu A^\nu) \tag{18.20a}$$

$$= (\nabla_\mu \mathrm{e}^a_\nu) A^\nu + \mathrm{e}^a_\nu \left(\partial_\mu A^\nu + \Gamma^\nu{}_{\mu\lambda} A^\lambda \right) . \tag{18.20b}$$

Using

$$\partial_\mu A^a = \partial_\mu (\mathrm{e}^a_\nu A^\nu) = \mathrm{e}^a_\nu (\partial_\mu A^\nu) + (\partial_\mu \mathrm{e}^a_\nu) A^\nu \tag{18.21}$$

to eliminate the second term in (18.20b), we obtain

$$\nabla_\mu A^a = (\nabla_\mu \mathrm{e}^a_\nu) A^\nu + \partial_\mu A^a - (\partial_\mu \mathrm{e}^a_\nu) A^\nu + \mathrm{e}^a_\nu \Gamma^\nu{}_{\mu\lambda} A^\lambda \,. \tag{18.22}$$

Comparing this to Eq. (18.19) we can read off how the covariant derivative acts on an object with mixed indices,

$$\nabla_\mu \mathrm{e}^a_\nu = \partial_\mu \mathrm{e}^a_\nu - \Gamma^\lambda{}_{\nu\mu} \mathrm{e}^a_\lambda + \omega_\mu{}^a{}_b \mathrm{e}^b_\nu \,. \tag{18.23}$$

More generally, covariant indices are contracted with the usual connection $\Gamma^\lambda{}_{\nu\mu}$ while vielbein indices are contracted with $\omega_\mu{}^a{}_b$. In a moment, we will show that the covariant derivative of the vielbein field is zero, $\nabla_\mu \mathrm{e}^a_\nu = 0$. Sometimes this property is called "tetrade postulate", but in fact it follows naturally from the definition of the vielbein field.

In order to derive an explicit formula for the spin connection $\omega_\mu{}^a{}_b$ we compare now the covariant derivative of a vector in the two formalisms. First, we write in a coordinate basis

$$\boldsymbol{\nabla A} = (\nabla_\mu A^\nu) \mathrm{d}x^\mu \otimes \partial_\nu = (\partial_\mu A^\nu + \Gamma^\nu{}_{\mu\lambda} A^\lambda) \mathrm{d}x^\mu \otimes \partial_\nu \,. \tag{18.24}$$

Next we compare this expression to the one using a mixed basis,

$$\boldsymbol{\nabla A} = (\nabla_\mu A^m)\, \mathrm{d}x^\mu \otimes e_m = (\partial_\mu A^m + \omega_\mu{}^m{}_n A^n)\, \mathrm{d}x^\mu \otimes e_m \tag{18.25a}$$

$$= [\partial_\mu (e^m_\nu A^\nu) + \omega_\mu{}^m{}_n e^n_\lambda A^\lambda]\, \mathrm{d}x^\mu \otimes (e^\sigma_m \partial_\sigma) \,. \tag{18.25b}$$

Moving e^σ_m to the left and using the Leibniz rule as well as $e^\sigma_m e^m_\nu = \delta^\sigma_\nu$, it follows

$$\boldsymbol{\nabla A} = e^\sigma_m [e^m_\nu \partial_\mu A^\nu + A^\nu \partial_\mu e^m_\nu + \omega_\mu{}^m{}_n e^n_\lambda A^\lambda]\, \mathrm{d}x^\mu \otimes \partial_\sigma \tag{18.26a}$$

$$= [\partial_\mu A^\sigma + e^\sigma_m (\partial_\mu e^m_\nu) A^\nu + e^\sigma_m e^n_\lambda \omega_\mu{}^m{}_n A^\lambda]\, \mathrm{d}x^\mu \otimes \partial_\sigma \,. \tag{18.26b}$$

Relabelling indices and comparing to Eq. (18.24), we find as relation between the two connections

$$\Gamma^{\nu}{}_{\mu\lambda} = e^{\nu}_{m}\partial_{\mu}e^{m}_{\lambda} + e^{\nu}_{m}e^{n}_{\lambda}\omega_{\mu}{}^{m}{}_{n}\,. \tag{18.27}$$

Clearly, the connection $\Gamma^{\nu}{}_{\mu\lambda}$ is in general not symmetric, $\Gamma^{\nu}{}_{\mu\lambda} \neq \Gamma^{\nu}{}_{\lambda\mu}$. Now we are also in the position to show that the covariant derivative of the vielbein field vanishes. Multiplying Eq. (18.27) with two vielbein fields, we arrive first at

$$\omega_{\mu}{}^{m}{}_{n} = e^{m}_{\nu}e^{\lambda}_{n}\Gamma^{\nu}{}_{\mu\lambda} - e^{\lambda}_{n}\partial_{\mu}e^{m}_{\lambda}\,. \tag{18.28}$$

Multiplying once again with a vielbein field, we obtain

$$e^{n}_{\nu}\omega_{\mu}{}^{m}{}_{n} = e^{n}_{\nu}(e^{m}_{\sigma}e^{\lambda}_{n}\Gamma^{\sigma}{}_{\mu\lambda} - e^{\lambda}_{n}\partial_{\mu}e^{m}_{\lambda}) \tag{18.29a}$$

$$= e^{m}_{\sigma}\delta^{\lambda}_{\nu}\Gamma^{\sigma}{}_{\mu\lambda} - \delta^{\lambda}_{\nu}\partial_{\mu}e^{m}_{\lambda} = e^{m}_{\sigma}\Gamma^{\sigma}{}_{\mu\nu} - \partial_{\mu}e^{m}_{\nu}\,. \tag{18.29b}$$

Inserted into Eq. (18.23), the “tetrade postulate” $\nabla_{\mu}e^{a}_{\nu} = 0$ follows.

18.2. Action of gravity

Einstein-Hilbert action The analogy between the Yang-Mills field-strength and the Riemann tensor shown by Eq. (10.49) suggests as Lagrange density for the gravitational field

$$\mathscr{L} = -\sqrt{|g|}R_{\mu\nu\sigma\tau}R^{\mu\nu\sigma\tau}\,. \tag{18.30}$$

This Lagrange density has mass dimension 4 and would thus lead to a dimensionless gravitational coupling constant and a renormalisable theory of gravity. However, such a theory would be in contradiction to Newton’s law. Hilbert chose instead the curvature scalar which has the required mass dimension $d = 6$,

$$\mathscr{L}_{\mathrm{E}H} = -\sqrt{|g|}R\,. \tag{18.31}$$

As we know, we can always add a constant term to the Lagrangian, $R \to R + 2\Lambda$. Such a term would imply that classically empty space has a constant vacuum energy which gravitates. The Lagrangian is a function of the metric, its first and second derivatives[1], $\mathscr{L}_{\mathrm{E}H}(g_{\mu\nu}, \partial_{\rho}g_{\mu\nu}, \partial_{\rho}\partial_{\sigma}g_{\mu\nu})$. The resulting action

$$S_{\mathrm{E}H}[g_{\mu\nu}] = -\int_{\Omega}\mathrm{d}^4x\sqrt{|g|}\,\{R + 2\Lambda\} \tag{18.32}$$

is a functional of the metric tensor $g_{\mu\nu}$, and a variation of the action with respect to the metric gives the field equations for the gravitational field. If we consider gravity coupled to fermions, we have to use the spin connection ω_{μ} in the matter Lagrangian as well as expressing $\sqrt{|g|}$ and R through Latin quantities, $\sqrt{|g|} \to E$ and $R = R_{\mu\nu}g^{\mu\nu} \to R_{mn}\eta^{mn}$.

We derive the resulting field equations for the metric tensor $g_{\mu\nu}$ directly from the action principle

$$\delta S_{\mathrm{E}H} = -\delta\int_{\Omega}\mathrm{d}^4x\sqrt{|g|}(R + 2\Lambda) = -\delta\int_{\Omega}\mathrm{d}^4x\sqrt{|g|}\,(g^{\mu\nu}R_{\mu\nu} + 2\Lambda) = 0\,. \tag{18.33}$$

[1] Recall that the Lagrange equations are modified in the case of higher derivatives which is one reason why we directly vary the action in order to obtain the field equations.

We allow for variations of the metric $g_{\mu\nu}$ restricted by the condition that the variation of $g_{\mu\nu}$ and its first derivatives vanish on the boundary $\partial\Omega$,

$$\delta S_{\mathrm{E}H} = -\int_\Omega \mathrm{d}^4x \left\{ \sqrt{|g|}\, g^{\mu\nu}\delta R_{\mu\nu} + \sqrt{|g|} R_{\mu\nu}\delta g^{\mu\nu} + (R+2\Lambda)\,\delta\sqrt{|g|} \right\}. \tag{18.34}$$

Our task is to rewrite the first and third term as variations of $\delta g^{\mu\nu}$ or to show that they are equivalent to boundary terms. Let us start with the first term. Choosing inertial coordinates, the Ricci tensor at the considered point P becomes

$$R_{\mu\nu} = \partial_\rho \Gamma^\rho{}_{\mu\nu} - \partial_\nu \Gamma^\rho{}_{\mu\rho}\,. \tag{18.35}$$

Hence

$$g^{\mu\nu}\delta R_{\mu\nu} = g^{\mu\nu}(\partial_\rho \delta\Gamma^\rho{}_{\mu\nu} - \partial_\nu \delta\Gamma^\rho{}_{\mu\rho}) = g^{\mu\nu}\partial_\rho\delta\Gamma^\rho{}_{\mu\nu} - g^{\mu\rho}\partial_\rho\delta\Gamma^\nu{}_{\mu\nu}\,, \tag{18.36}$$

where we exchanged the indices ν and ρ in the last term. Since $\partial_\rho g_{\mu\nu} = 0$ at P, we can rewrite the expression as

$$g^{\mu\nu}\delta R_{\mu\nu} = \partial_\rho(g^{\mu\nu}\delta\Gamma^\rho{}_{\mu\nu} - g^{\mu\rho}\delta\Gamma^\nu{}_{\mu\nu}) = \partial_\rho X^\rho\,. \tag{18.37}$$

The quantity X^ρ is a vector, since the difference of two connection coefficients transforms as a tensor. Replacing in Eq. (18.37) the partial derivative by a covariant one promotes it therefore in a valid tensor equation,

$$g^{\mu\nu}\delta R_{\mu\nu} = \nabla_\mu V^\mu = \frac{1}{\sqrt{|g|}}\partial_\mu(\sqrt{|g|}V^\mu)\,. \tag{18.38}$$

Thus this term corresponds to a surface term which we assume to vanish. Next we rewrite the third term using

$$\delta\sqrt{|g|} = \frac{1}{2\sqrt{|g|}}\,\delta|g| = \frac{1}{2}\sqrt{|g|}\,g^{\mu\nu}\delta g_{\mu\nu} = -\frac{1}{2}\sqrt{|g|}\,g_{\mu\nu}\delta g^{\mu\nu} \tag{18.39}$$

and obtain

$$\delta S_{\mathrm{E}H} = -\int_\Omega \mathrm{d}^4x \sqrt{|g|}\left\{ R_{\mu\nu} - \frac{1}{2}\,g_{\mu\nu}R - \Lambda\, g_{\mu\nu} \right\}\delta g^{\mu\nu} = 0\,. \tag{18.40}$$

Hence the metric fulfils in vacuum the equation

$$-\frac{1}{\sqrt{|g|}}\,\frac{\delta S_{\mathrm{E}H}}{\delta g^{\mu\nu}} = R_{\mu\nu} - \frac{1}{2}\,R\,g_{\mu\nu} - \Lambda g_{\mu\nu} \equiv G_{\mu\nu} - \Lambda g_{\mu\nu} = 0\,, \tag{18.41}$$

where we introduced the Einstein tensor $G_{\mu\nu}$. The constant Λ is called the cosmological constant.

We consider now the combined action of gravity and matter, as the sum of the Einstein-Hilbert Lagrangian $\mathscr{L}_{\mathrm{E}H}/2\kappa$ and the Lagrangian $\mathscr{L}_{\mathrm{m}}$ including all relevant matter fields,

$$\mathscr{L} = \frac{1}{2\kappa}\mathscr{L}_{\mathrm{E}H} + \mathscr{L}_{\mathrm{m}} = -\frac{1}{2\kappa}\sqrt{|g|}(R+2\Lambda) + \mathscr{L}_{\mathrm{m}}\,. \tag{18.42}$$

We will determine the value of the coupling constant κ in the next section, demanding that we reproduce Newtonian dynamics in the weak-field limit. We have already argued that the source of the gravitational field is the energy-momentum stress tensor which lead to the definition

$$\frac{2}{\sqrt{|g|}}\frac{\delta S_{\rm m}}{\delta g^{\mu\nu}}=T_{\mu\nu}\,. \tag{18.43}$$

Einstein's field equation follows then as

$$G_{\mu\nu}-\Lambda g_{\mu\nu}=\kappa T_{\mu\nu}\,. \tag{18.44}$$

Two remarks are in order: First, the cosmological constant Λ measures the curvature of an empty classical space-time. Moving Λ to the RHS of the Einstein equations, we see that its effect is equivalent to the stress density $T_{\mu\nu}=\kappa\Lambda g_{\mu\nu}$ or a vacuum energy density $\rho_\Lambda=\kappa\Lambda$ with EoS $w=P_\Lambda/\rho_\Lambda=-1$, see problem 19.??. Second, we assumed that the boundary term arising from $g^{\mu\nu}\delta R_{\mu\nu}=\nabla_\mu V^\mu$ vanishes, which correspond to a zero flux of V^μ through $\partial\Omega$. In general, this flux is zero only for a compact manifold Ω (since then $\partial\Omega=0$) or if we require that not only $\delta g_{\mu\nu}$ but also $\delta\Gamma^\kappa{}_{\mu\nu}$ vanishes on the boundary $\partial\Omega$. The second option is naturally implemented in the Palatini action principle which we consider next. Using the Einstein-Hilbert action, one should add instead a boundary term which cancels the variation of $g^{\mu\nu}\delta R_{\mu\nu}=\nabla_\mu V^\mu$, for details see [Poi07].

Palatini action We start from the Einstein-Hilbert Lagrangian (18.31), but consider it now as a function of the metric tensor, the connection and its first derivatives, $\mathscr{L}_{\rm E}H(g_{\mu\nu},\Gamma^\sigma{}_{\mu\nu},\partial_\tau\Gamma^\sigma{}_{\mu\nu})$, while we allow an independent variation of the metric tensor and the connection in the action. We obtain the desired dependence of the Lagrangian expressing the Ricci tensor through the connection and its derivatives,

$$\mathscr{L}_{\rm E}H=-\sqrt{|g|}g^{\mu\nu}R_{\mu\nu}=\sqrt{|g|}g^{\mu\nu}\left(\partial_\nu\Gamma^\sigma{}_{\mu\sigma}-\partial_\sigma\Gamma^\sigma{}_{\mu\nu}+\Gamma^\tau{}_{\mu\sigma}\Gamma^\sigma{}_{\tau\nu}-\Gamma^\tau{}_{\mu\nu}\Gamma^\sigma{}_{\tau\sigma}\right)\,. \tag{18.45}$$

For simplicity, we set also $\Lambda=0$. Then the variation with respect to the metric,

$$\delta_g S_{\rm E}H=-\int_\Omega {\rm d}^4x\,\delta\left\{\sqrt{|g|}\,g^{\mu\nu}\right\}R_{\mu\nu}=0 \tag{18.46}$$

gives $R_{\mu\nu}=0$. As we will show in (18.66), this condition is equivalent with the usual Einstein equations in vacuum. For the variation with respect to the connection we use first the Palatini equation,

$$\delta_\Gamma S_{\rm E}H=-\int_\Omega {\rm d}^4x\sqrt{|g|}\,g^{\mu\nu}\delta R_{\mu\nu}=-\int_\Omega {\rm d}^4x\sqrt{|g|}\,g^{\mu\nu}\left[\nabla_\nu(\delta\Gamma^\sigma{}_{\mu\sigma})-\nabla_\sigma(\delta\Gamma^\sigma{}_{\mu\nu})\right]\,. \tag{18.47}$$

Applying then the Leibniz rule and relabelling some indices, we find

$$\begin{aligned}\delta_\Gamma S_{\rm E}H=&-\int_\Omega {\rm d}^4x\sqrt{|g|}\nabla_\nu\left[g^{\mu\nu}\delta\Gamma^\sigma{}_{\mu\sigma}-g^{\mu\sigma}\delta\Gamma^\sigma{}_{\mu\sigma}\right]\\&-\int_\Omega {\rm d}^4x\sqrt{|g|}\left[(\nabla_\rho g^{\mu\nu})\delta\Gamma^\rho{}_{\mu\nu}-(\nabla_\nu g^{\mu\nu})\delta\Gamma^\rho{}_{\mu\rho}\right]\,.\end{aligned} \tag{18.48}$$

We kept the second line, because we we do not know yet if the covariant derivative of the metric vanishes for an arbitrary connection. Next we perform a partial integration of the first two terms, converting it into a surface term which we can drop. In the remaining part we relabel indices so that we can factor out the variation of the connection,

$$\delta_\Gamma S_{\mathrm{E}H} = \int_\Omega \mathrm{d}^4x\sqrt{|g|}\,\left[\delta^\nu_\rho \nabla_\sigma g^{\mu\sigma} - \nabla_\rho g^{\mu\nu}\right]\delta\Gamma^\rho{}_{\mu\nu}\,. \tag{18.49}$$

We use now that the connection is symmetric in the absence of fermion. Then also the variation $\delta\Gamma^\rho{}_{\mu\nu}$ is symmetric and the antisymmetric part in the square bracket drops out. Asking that $\delta_\Gamma S_{\mathrm{E}H} = 0$ gives therefore

$$\frac{1}{2}\delta^\nu_\rho \nabla_\sigma g^{\mu\sigma} + \frac{1}{2}\delta^\mu_\rho \nabla_\sigma g^{\nu\sigma} - \nabla_\rho g^{\mu\nu} = 0 \tag{18.50}$$

or $\nabla_\sigma g^{\mu\nu} = \nabla_\sigma g_{\mu\nu} = 0$. Thus the Einstein-Hilbert action *implies* the metric compatibility of the connection.

Performing the same exercise with the Einstein-Hilbert plus the matter action considered as functional of the vielbein e^μ_m and the connection ω_μ, one finds the following: From the variation $\delta_\omega S_{\mathrm{E}H}$ one obtains automatically a metric connection which is however in general not symmetric. The torsion is sourced by the spin-density of fermions. The variation $\delta_e S_{\mathrm{E}H}$ gives the usual Einstein equation. This result justifies the usual choice of a torsionless connection which is metric compatible: Although e.g. a star consists of a collection of individual particles carrying spin $\boldsymbol{s}_i$, its total spin sums up to zero, $\sum_i \boldsymbol{s}_i \simeq 0$, because they are uncorrelated. Thus we can describe macroscopic matter in general relativity as a classical spinless[2] point particle (or a fluid if its extension is important) leading to a symmetric connection.

18.3. Linearised gravity

We are looking for small perturbations $h_{\mu\nu}$ around the Minkowski metric $\eta_{\mu\nu}$,

$$g_{\mu\nu} = \eta_{\mu\nu} + h_{\mu\nu}\,, \qquad h_{\mu\nu} \ll 1\,. \tag{18.51}$$

These perturbations may be caused either by the propagation of gravitational waves or by the gravitational potential of a star. In the first case, experimental limits showed that one should not hope for h larger than $\mathcal{O}(h) \sim 10^{-21}$. Keeping only terms linear in h is therefore an excellent approximation. Choosing in the second case as application e.g. the spiral-in of a binary system, deviations from the Newtonian limit can become arbitrarily large. Hence one needs a systematic "post-Newtonian" expansion or has to perform a full numerical analysis to describe properly such cases.

Linearised Einstein equations in vacuum From $\partial_\mu \eta_{\nu\rho} = 0$ and the definition

$$\Gamma^\mu{}_{\nu\lambda} = \frac{1}{2}g^{\mu\kappa}(\partial_\nu g_{\kappa\lambda} + \partial_\lambda g_{\nu\kappa} - \partial_\kappa g_{\nu\lambda}) \tag{18.52}$$

we find for the change of the connection linear in h

$$\delta\Gamma^\mu{}_{\nu\lambda} = \frac{1}{2}\eta^{\mu\kappa}(\partial_\nu h_{\kappa\lambda} + \partial_\lambda h_{\nu\kappa} - \partial_\kappa h_{\nu\lambda}) = \frac{1}{2}(\partial_\nu h^\mu_\lambda + \partial_\lambda h^\mu_\nu - \partial^\mu h_{\nu\lambda})\,. \tag{18.53}$$

[2] The curious reader may wonder if orbital angular momentum leads to torsion: One way to see that the answer is no is to realise that one cannot define an orbital angular momentum *density* which transforms properly as a tensor, cf. Eq. (5.26f).

Here we used η to raise indices which is allowed in linear approximation. Remembering the definition of the Riemann tensor,

$$R^{\mu}{}_{\nu\lambda\kappa} = \partial_\lambda \Gamma^{\mu}{}_{\nu\kappa} - \partial_\kappa \Gamma^{\mu}{}_{\nu\lambda} + \Gamma^{\mu}{}_{\rho\lambda}\Gamma^{\rho}{}_{\nu\kappa} - \Gamma^{\mu}{}_{\rho\kappa}\Gamma^{\rho}{}_{\nu\lambda} \,, \tag{18.54}$$

we see that we can neglect the terms quadratic in the connection terms. Thus we find for the change

$$\begin{aligned} \delta R^{\mu}{}_{\nu\lambda\kappa} &= \partial_\lambda \delta\Gamma^{\mu}{}_{\nu\kappa} - \partial_\kappa \delta\Gamma^{\mu}{}_{\nu\lambda} \\ &= \frac{1}{2}\left\{\partial_\lambda\partial_\nu h^\mu_\kappa + \partial_\lambda\partial_\kappa h^\mu_\nu - \partial_\lambda\partial^\mu h_{\nu\kappa} - (\partial_\kappa\partial_\nu h^\mu_\lambda + \partial_\kappa\partial_\lambda h^\mu_\nu - \partial_\kappa\partial^\mu h_{\nu\lambda})\right\} \\ &= \frac{1}{2}\left\{\partial_\lambda\partial_\nu h^\mu_\kappa + \partial_\kappa\partial^\mu h_{\nu\lambda} - \partial_\lambda\partial^\mu h_{\nu\kappa} - \partial_\kappa\partial_\nu h^\mu_\lambda\right\} \,. \end{aligned} \tag{18.55}$$

The change in the Ricci tensor follows by contracting μ and λ,

$$\delta R^{\lambda}{}_{\nu\lambda\kappa} = \frac{1}{2}\left\{\partial_\lambda\partial_\nu h^\lambda_\kappa + \partial_\kappa\partial^\lambda h_{\nu\lambda}) - \partial_\lambda\partial^\lambda h_{\nu\kappa} - \partial_\kappa\partial_\nu h^\lambda_\lambda\right\} \,. \tag{18.56}$$

Next we introduce $h \equiv h^\mu_\mu$, $\Box = \partial_\mu\partial^\mu$, and relabel the indices,

$$\delta R_{\mu\nu} = \frac{1}{2}\left\{\partial_\mu\partial_\rho h^\rho_\nu + \partial_\nu\partial_\rho h^\rho_\mu - \Box h_{\mu\nu} - \partial_\mu\partial_\nu h\right\} \,. \tag{18.57}$$

We now rewrite all terms apart from $\Box h_{\mu\nu}$ as derivatives of the vector

$$\xi_\mu = \partial_\nu h^\nu_\mu - \frac{1}{2}\partial_\mu h \,, \tag{18.58}$$

obtaining

$$\delta R_{\mu\nu} = \frac{1}{2}\left\{-\Box h_{\mu\nu} + \partial_\mu \xi_\nu + \partial_\nu \xi_\mu\right\} \,. \tag{18.59}$$

Looking back at the properties of $h_{\mu\nu}$ under gauge transformations, Eq. (7.41), we see that we can gauge away the second and third term. Thus the linearised Einstein equation in vacuum, $\delta R_{\mu\nu} = 0$, becomes simply

$$\Box h_{\mu\nu} = 0 \,, \tag{18.60}$$

if the harmonic gauge

$$\xi_\mu = \partial_\nu h^\nu_\mu - \frac{1}{2}\partial_\mu h = 0 \tag{18.61}$$

is chosen. Hence the familiar wave equation holds for all ten independent components of $h_{\mu\nu}$, and the perturbations propagate with the speed of light. Inserting plane waves $h_{\mu\nu} = \varepsilon_{\mu\nu}\exp(-\mathrm{i}kx)$ into the wave equation, one finds immediately that k is a null vector.

TT gauge We want to re-derive our old result for the polarisation tensor describing the physical states contained in a gravitational perturbation. We consider a plane wave $h_{\mu\nu} = \varepsilon_{\mu\nu}\exp(-\mathrm{i}kx)$. After choosing the harmonic gauge (18.61), we can still perform a gauge transformation using four functions ξ_μ satisfying $\Box\xi_\mu = 0$. We can choose them such that four components of $h_{\mu\nu}$ vanish. In the TT gauge, we set ($i = 1, 2, 3$)

$$h_{0i} = 0, \qquad h = 0 \,. \tag{18.62}$$

The harmonic gauge condition becomes $\xi_\alpha = \partial_\beta h^\beta_\alpha$ or

$$\xi_0 = \partial_\beta h^\beta_0 = \partial_0 h^0_0 = -\mathrm{i}\omega\varepsilon_{00}\mathrm{e}^{-\mathrm{i}kx} = 0\,, \tag{18.63a}$$

$$\xi_a = \partial_\beta h^\beta_a = \partial_b h^b_a = \mathrm{i}k^b\varepsilon_{ab}\mathrm{e}^{\mathrm{i}kx} = 0\,. \tag{18.63b}$$

Thus $\varepsilon_{00} = 0$ and the polarisation tensor is transverse, $k^b\varepsilon_{ab} = 0$. If we choose the plane wave propagating in z direction, $\boldsymbol{k} = k\boldsymbol{e}_z$, the z raw and column of the polarisation tensor vanishes too. Accounting for $h = 0$ and $\varepsilon_{\alpha\beta} = \varepsilon_{\beta\alpha}$, only two independent elements are left, and we recover our old result,

$$\varepsilon_{\alpha\beta} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & \varepsilon_{11} & \varepsilon_{12} & 0 \\ 0 & \varepsilon_{12} & -\varepsilon_{11} & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}. \tag{18.64}$$

In general, one can construct the polarisation tensor in TT gauge by setting first the non-transverse part to zero and then subtracting the trace. The resulting two independent elements are (again for $\boldsymbol{k} = k\boldsymbol{e}_z$) then $\varepsilon_{11} = 1/2(\varepsilon_{xx} - \varepsilon_{yy})$ and ε_{12}.

Linearised Einstein equations with sources We rewrite first the Einstein equation in an alternative form where the only geometrical term on the LHS is the Ricci tensor. Because of

$$R^{\ \mu}_\mu - \frac{1}{2}\delta^{\ \mu}_\mu(R + 2\Lambda) = R - 2(R + 2\Lambda) = -R - 4\Lambda = \kappa T^{\ \mu}_\mu\,, \tag{18.65}$$

we can perform with $T \equiv T^\mu_\mu$ the replacement $R = -\kappa T - 4\Lambda$ in the Einstein equation and obtain

$$R_{\mu\nu} = \kappa(T_{\mu\nu} - \frac{1}{2}g_{\mu\nu}T) - g_{\mu\nu}\Lambda\,. \tag{18.66}$$

This form of the Einstein equations is often useful, when it is easier to calculate T that R. Note also that (18.66) informs us that an empty universe with $\Lambda = 0$ has a vanishing Ricci tensor $R_{\mu\nu} = 0$.

Now we move on to the determination of the linearised Einstein equations with sources. We found $2\delta R_{\mu\nu} = -\Box h_{\mu\nu}$. By contraction follows $2\delta R = -\Box h$. Combining both terms gives

$$\Box\left(h_{\mu\nu} - \frac{1}{2}\eta_{\mu\nu}h\right) = -2(\delta R_{\mu\nu} - \frac{1}{2}\eta_{\mu\nu}\delta R) = -2\kappa\delta T_{\mu\nu}\,. \tag{18.67}$$

Since we assumed an empty universe in zeroth order, $\delta T_{\mu\nu}$ is the complete contribution to the stress tensor. We omit therefore in the following the δ in $\delta T_{\mu\nu}$. Next we introduce as useful short-hand notation the "trace-reversed" amplitude as

$$\bar{h}_{\mu\nu} \equiv h_{\mu\nu} - \frac{1}{2}\eta_{\mu\nu}h\,. \tag{18.68}$$

The harmonic gauge condition becomes then

$$\partial_\mu\bar{h}_{\mu\nu} = 0 \tag{18.69}$$

and the linearised Einstein equation in the harmonic gauge follow as

$$\Box\bar{h}_{\mu\nu} = -2\kappa T_{\mu\nu}\,. \tag{18.70}$$

Because of $\bar{\bar{h}}_{\mu\nu} = h_{\mu\nu}$ and Eq. (18.66), we can rewrite this wave equation also as

$$\Box h_{\mu\nu} = -2\kappa\bar{T}_{\mu\nu} \tag{18.71}$$

with the trace-reversed stress tensor $\bar{T}_{\mu\nu} \equiv T_{\mu\nu} - \frac{1}{2}\eta_{\mu\nu}T$.

Newtonian limit The Newtonian limit corresponds to $v/c \to 0$ and thus the only non-zero element of the energy-momentum tensor becomes $T^{00} = \rho$. We will show later that the metric around a point mass can be written in the weak-field limit as

$$ds^2 = (1 + 2\Phi)dt^2 - (1 - 2\Phi)\left(dx^2 + dy^2 + dz^2\right) \tag{18.72}$$

with Φ as the Newtonian gravitational potential. Comparing this metric to Eq. (18.51), we find as metric perturbations

$$h_{00} = 2\Phi \qquad h_{ij} = 2\delta_{ij}\Phi \qquad h_{0i} = 0\,. \tag{18.73}$$

In the static limit $\Box \to -\Delta$ and $v = 0$, and thus

$$-\Delta\left(h_{00} - \frac{1}{2}\eta_{00}h\right) = -4\Delta\Phi = -2\kappa\rho\,. \tag{18.74}$$

Hence the linearised Einstein equation has the same form as the Newtonian Poisson equation, and the constant κ equals $\kappa = 8\pi G$.

Detection principle of gravitational waves Consider the effect of a gravitational wave on a free test particle that is initially at rest, $u^\alpha = (1, 0, 0, 0)$. Then the geodesic equation simplifies to $\dot{u}^\alpha = \Gamma^\alpha{}_{00}$. The four relevant Christoffel symbols are in the linearised approximation, cf. Eq. (18.53),

$$\Gamma^\alpha{}_{00} = \frac{1}{2}(\partial_0 h_0^\alpha + \partial_0 h_0^\alpha - \partial^\alpha h_{00})\,. \tag{18.75}$$

We are free to choose the TT gauge in which all component of $h_{\alpha\beta}$ appearing on the RHS are zero. Hence the acceleration of the test particle is zero and its coordinate position is unaffected by the gravitational wave: The TT gauge defines a "comoving" coordinate system.

The physical distance l between two test-particles is given by integrating

$$dl^2 = g_{ab}d\xi^a d\xi^b = (h_{ab} - \delta_{ab})d\xi^a d\xi^b\,, \tag{18.76}$$

where g_{ab} is the spatial part of the metric and $d\xi$ the spatial coordinate distance between infinitesimal separated test particles. Hence the passage of a gravitational wave, $h_{\alpha\beta} \propto \varepsilon_{\alpha\beta}\cos(\omega t)$, results in a periodic change of the separation of freely moving test particles. Figure 18.1 shows that a gravitational wave exerts tidal forces, stretching and squashing test particles in the transverse plane. The relative size of the change, $\Delta L/L$, is given by the amplitude h of the gravitational wave. It is this tiny periodic change, $\Delta L/L \lesssim 10^{-21}\cos(\omega t)$, which gravitational wave experiments aim to detect. There are two basic types of gravitational wave experiments: In the first one, one uses that the tidal forces of a passing gravitational wave excite lattice vibrations in a solid state. If the wave frequency is resonant with a lattice mode, the vibrations might be amplified to detectable levels. In the second type of experiment, the free test particles are replaced by mirrors. Between the mirrors, a laser beam is many times reflected, increasing thereby the effective length L and thus ΔL, before two beams at 90° are brought to interference. As the most promising gravitational wave source the inspiral of binary systems composed of neutron stars or black holes have been suggested. In September 2015, the Advanced Laser Interferometer Gravitational-Wave Observatory (Advanced LIGO) detected for the first time such a signal [A+16, CM16]. Additionally, a stochastic background of gravitational waves might be produced during inflation and phase transitions in the early universe.

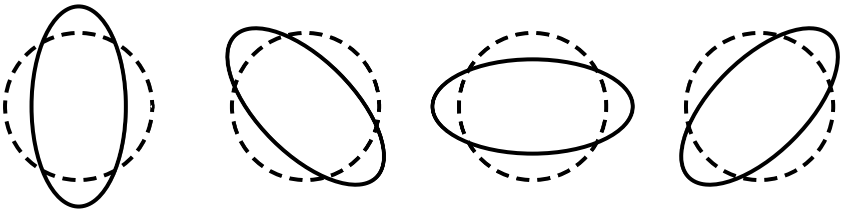


Figure 18.1.: The effect of a right-handed polarised gravitational wave on a ring of transverse test particles as function of time; the dashed line shows the state without gravitational wave.

Linearised action of gravity We have derived the linearised wave equation describing gravitational waves directly from the Einstein equation. Now we want to obtain the linearised action of gravity. A straightforward but lengthy approach would be to expand the Einstein-Hilbert action to $\mathcal{O}(h^3)$. Instead, we profit from our knowledge of the graviton propagator, which we derived in chapter 4 in Eq. (7.46),

$$D_F^{\mu\nu;\rho\sigma}(k) = \frac{\frac{1}{2}(-\eta^{\mu\nu}\eta^{\rho\sigma} + \eta^{\mu\rho}\eta^{\nu\sigma} + \eta^{\mu\sigma}\eta^{\nu\rho})}{k^2 + \mathrm{i}\varepsilon} = \frac{P^{\mu\nu;\rho\sigma}(k)}{k^2 + \mathrm{i}\varepsilon}\,. \tag{18.77}$$

The corresponding Lagrangian is as usually quadratic in the fields, $\frac{1}{2}h^{\mu\nu}P_{\mu\nu;\rho\sigma}\Box h^{\rho\sigma}$. Performing partial integrations and the contractions with the metric tensors gives us as the corresponding action

$$S = -\frac{1}{32\pi G}\int \mathrm{d}^4x \left[\frac{1}{2}\left(\partial_\alpha h_{\beta\gamma}\right)^2 - \frac{1}{4}\left(\partial_\alpha h\right)^2\right]. \tag{18.78}$$

We know that the propagator of a massless particles with helicity $h > 1/2$ can be inverted only, if either the gauge freedom is completely fixed or a gauge-fixing term is added. In contrast to the TT gauge, the harmonic gauge contains still unphysical degrees of freedom. Thus the expression (18.78) equals the quadratic Einstein-Hilbert Lagrangian in the harmonic gauge plus a gauge-fixing term. The propagator derived in Eq. (7.46) corresponds to the R_ξ-gauge with $\xi = 1$, and thus the effective Lagrangian is

$$\mathscr{L}_{\mathrm{E}H}^{(2)} + \mathscr{L}_{\mathrm{g}f} = -\frac{1}{32\pi G}\left[\frac{1}{2}\left(\partial_\alpha h_{\beta\gamma}\right)^2 - \frac{1}{4}\left(\partial_\alpha h\right)^2\right] = \mathscr{L}_{\mathrm{E}H}^{(2)} - \left(\partial^\mu h_{\mu\nu} - \frac{1}{2}\partial_\mu h\right)^2. \tag{18.79}$$

Specialising Eq. (18.78) to the TT gauge, we obtain

$$S_{\mathrm{E}H} = -\frac{1}{32\pi G}\int \mathrm{d}^4x\, \frac{1}{2}\left(\partial_\mu h_{ij}\right)^2. \tag{18.80}$$

We can express an arbitrary polarisation state as the sum over the polarisation tensors for circular polarised waves,

$$h_{\mu\nu} = \sum_{a=+,-} h^{(a)}\varepsilon_{\mu\nu}^{(a)}\,. \tag{18.81}$$

Inserting this decomposition into (18.80) and using $\varepsilon_{\mu\nu}^{(a)}\varepsilon^{\mu\nu(b)} = \delta^{ab}$, the action becomes

$$S_{\mathrm{E}H}^{TT} = -\frac{1}{32\pi G}\sum_a \int \mathrm{d}^4x\,\frac{1}{2}\left(\partial_\mu h^{(a)}\right)^2 . \tag{18.82}$$

Thus the gravitational action in the TT gauge consists of two scalar degrees of freedom, h^+ and h^-, which determine the contribution of left- and right-circular polarised waves. Apart from the prefactor, the action is the same as the one of two scalar fields. This means that we can short-cut many calculations involving gravitational waves by using simply the corresponding results for scalar fields. We can understand this equivalence by recalling that the part of the action action quadratic in the fields just enforces the relativistic energy-momentum relation via a Klein-Gordon equation for each field component. The remaining content of (18.78) is just the rule how the unphysical components in $h^{\mu\nu}$ have to be eliminated. In the TT gauge, we have applied already this information, and thus the two scalar wave equations for $h^{(\pm)}$ summarise the Einstein equations at $\mathcal{O}(h^2)$.

Before summarising, we mention a recently found connection between gravity and gauge theories. First we note that we can express the polarisation tensor $\varepsilon_{\mu\nu}$ of a graviton as the tensor product of two polarisation vectors ε_μ of a gauge boson, $\varepsilon_{\mu\nu}^{(a)} = \varepsilon_\mu^{(a)}\varepsilon_\nu^{(a)}$. Motivated by this relation, one may wonder if one can connect amplitudes containing gravitons to those containing e.g. gluons. In general, this cannot be true since loop graphs of renormalisable and non-renormalisable have very different properties. However, one can connect tree-level amplitudes in gravity to the "square" of the corresponding amplitude in a non-abelian gauge theory: For instance, the vertex $V^{\mu_1\mu_2\mu_3\nu_1\nu_2\nu_3}(p_1,p_2,p_3)$ connecting three on-shell gravitons can be expressed as

$$V^{\mu_1\mu_2\mu_3\nu_1\nu_2\nu_3}(p_1,p_2,p_3) = \frac{\mathrm{i}\kappa}{4}\,V^{\mu_1\mu_2\mu_3}(p_1,p_2,p_3)V^{\nu_1\nu_2\nu_3}(p_1,p_2,p_3)\,, \tag{18.83}$$

with $V^{\nu_1\nu_2\nu_3} = \eta^{\nu_1\nu_2}(p_1-p_2)^{\nu_3} + \eta^{\nu_2\nu_3}(p_2-p_3)^{\nu_1} + \eta^{\nu_3\nu_1}(p_3-p_1)^{\nu_2}$. But $V^{\nu_1\nu_2\nu_3}$ agrees with the three-gluon vertex (10.82) after stripping off its colour factor.

Summary

The vielbein formalism allows us not only to couple fermions to gravity but clarifies also that the metric compatibility of the connection is a consequence of Einstein's general relativity. In contrast, the connection is in general not symmetric, although deviations can be neglected for macroscopic sources of gravity. The linearised Einstein equations are often sufficient to describe the propagation of gravitational waves. In the TT gauge, the Einstein-Hilbert action is proportional to the action of a massless scalar field.

Further reading

[Car03] presents a clear discussion of the vielbein formalism and Cartan's structure equations. The Einstein-Hilbert action including boundary term is discussed by [Poi07]. For an introduction to gravitational waves see e.g. [HEL06]. The connection of tree-level amplitudes in gauge and gravity theories is discussed by [Wei16].

19. Cosmological models for an homogeneous, isotropic universe

The most important observational property of our Universe is its homogeneity and isotropy on sufficiently large scales: Only this feature allows us to deduce from observations performed at a single point generic properties of the Universe. While the homogeneity and isotropy of the Universe was merely a postulate—dubbed cosmological principle— at the beginning of the 20th century, observations of the cosmic microwave background (CMB) have shown that the early Universe was isotropic at the level of 10^{-4} around us. Note that a space isotropic around at least two points is also homogeneous, while a homogeneous space is not necessarily isotropic. Baring suggestions that we live at a unique place, the early Universe was therefore also homogeneous on large scales. The homogeneity of the Universe implies moreover that beyond a sufficiently large scale no new types of gravitationally bound structures appear: In particular, the sequence of hierarchical structures → star clusters → galaxies → galaxy clusters → supercluster of galaxies stops and superclusters with sizes of order 100 Mpc are the largest bound systems found in the Universe. For comparison, the size of the Universe visible at the present epoch is of order 3000 Mpc, i.e. a factor 30 larger. In this chapter, we will discuss the evolution of perfectly homogeneous and isotropic cosmological models. The question how the inhomogeneities measured in the CMB were generated and how they have evolved into the very inhomogeneous present Universe will be addressed in chapter 23.

19.1. FLRW metric

Weyl's postulate In 1923, Hermann Weyl suggested the existence of a privileged class of observers in the universe, namely those following the "average" motion of galaxies. He postulated that these observers follow time-like geodesics that never intersect. They may however diverge from a point in the (finite or infinite) past or converge towards such a point in the future. Weyl's postulate implies that we can find coordinates such that galaxies—and the observers residing in them—are at rest. These coordinates are called *comoving coordinates* and can be constructed as follows: One chooses first a space-like hypersurface. Through each point P in this hypersurface passes a unique worldline of a comoving observer. We choose the coordinate time such that it agrees with the proper-time of all observers, $g_{00} = 1$, and the spatial coordinate vectors such that they are constant and lie in the tangent space T_P at this point. Then $u^\alpha = \delta^\alpha_0$ and for $n^\alpha \in T_P$ it follows $n^\alpha = (0, n^i)$ and

$$0 = u_\alpha n^\alpha = g_{\alpha\beta} u^\alpha n^\beta = g_{0b} n^b \,. \tag{19.1}$$

Since n^α is arbitrary, the elements g_{0b} of the metric tensor have to vanish, $g_{0b} = 0$. Hence as a consequence of Weyl's postulate we may choose the metric as

$$\mathrm{d}s^2 = \mathrm{d}t^2 - \mathrm{d}l^2 = \mathrm{d}t^2 - g_{ij}\mathrm{d}x^i \mathrm{d}x^j \,. \tag{19.2}$$

The cosmological principle constrains further the form of $\mathrm{d}l^2$: Homogeneity requires that the g_{ij} can depend on time only via a common factor $S(t)$, while isotropy requires that only the scalars $\boldsymbol{x}\cdot\boldsymbol{x}\equiv r^2$, $\boldsymbol{x}\cdot\mathbf{d}\boldsymbol{x}$ and $\mathbf{d}\boldsymbol{x}\cdot\mathbf{d}\boldsymbol{x}$ enter $\mathrm{d}l^2$. Hence the spatial part of the metric has the form

$$\mathrm{d}l^2 = C(r)(\boldsymbol{x}\cdot\mathbf{d}\boldsymbol{x})^2 + D(r)\mathbf{d}\boldsymbol{x}\cdot\mathbf{d}\boldsymbol{x} \tag{19.3a}$$

$$= C(r)r^2\mathrm{d}r^2 + D(r)[\mathrm{d}r^2 + r^2\mathrm{d}\vartheta^2 + r^2\sin^2\vartheta\mathrm{d}\phi^2]\,. \tag{19.3b}$$

We can eliminate the function $D(r)$ by the rescaling $r^2 \to Dr^2$. Then the line-element becomes

$$\mathrm{d}l^2 = S(t)\left[B(r)\mathrm{d}r^2 + r^2\mathrm{d}\Omega\right] \tag{19.4}$$

with $\mathrm{d}\Omega = \mathrm{d}\vartheta^2 + \sin^2\vartheta\mathrm{d}\phi^2$. While the dynamical function $S(t)$ has to be determined from Einstein equations for a given matter content of the Universe, we can constrain the function $B(r)$ further using the cosmological principle.

Maximally symmetric spaces This are spaces with constant curvature. Hence the Riemann tensor of such spaces can depend only on the metric tensor and a constant K specifying the curvature. The only form that respects the (anti-)symmetries of the Riemann tensor is

$$R_{\alpha\beta\gamma\delta} = K(g_{\alpha\gamma}g_{\beta\delta} - g_{\alpha\delta}g_{\beta\gamma})\,. \tag{19.5}$$

Then we obtain for the Ricci tensor in three spatial dimensions

$$R_{bd} = g^{ac}R_{abcd} = Kg^{ac}(g_{ac}g_{bd} - g_{ad}g_{bc}) = K(3g_{bd} - g_{bd}) = 2Kg_{bd}\,. \tag{19.6}$$

A final contraction gives as curvature R of a three-dimensional maximally symmetric space

$$R = g^{ab}R_{ab} = 2K\delta^a_a = 6K\,. \tag{19.7}$$

A comparison of Eq. (19.6) with the Ricci tensor for the metric (19.4) will fix the still unknown function $B(r)$. We proceed in the standard way: We calculate the Christoffel symbols with the help of the geodesic equation (or alternatively directly from Eq. (6.43)) and use then the definition of the Ricci tensor.

Example 19.1: Find the Christoffel symbols and the Ricci tensor for the metric (19.4).
We solve the Lagrange equations for a test particle moving in (19.4),

$$L = B(r)\dot{r}^2 + r^2(\dot{\vartheta}^2 + \sin^2\vartheta\dot{\phi}^2)\,,$$

where we neglected the overall factor $S(t)$. Comparing the Lagrange equations

$$\ddot{r} - \frac{B'}{2B}\dot{r}^2 - (\dot{\vartheta}^2 + \sin^2\vartheta\dot{\phi}^2)\frac{r}{B} = 0\,, \qquad \ddot{\phi} + \frac{2}{r}\dot{r}\dot{\phi} + 2\cot\vartheta\dot{\vartheta}\dot{\phi} = 0\,, \qquad \ddot{\vartheta} + \frac{2\dot{r}\dot{\vartheta}}{r} - \cos\vartheta\sin\vartheta\dot{\phi}^2 = 0$$

with the geodesic equation $\ddot{x}^\kappa + \Gamma^\kappa{}_{\mu\nu}\dot{x}^\mu\dot{x}^\nu = 0$, we can read off the non-vanishing Christoffel symbols as

$$\Gamma^r{}_{rr} = -B'/(2B)\,, \qquad \Gamma^r{}_{\phi\phi} = -r\sin^2\vartheta/B \quad \textit{and} \quad \Gamma^r{}_{\vartheta\vartheta} = -r/B$$

$$\Gamma^\phi{}_{r\phi} = \Gamma^\vartheta{}_{r\vartheta} = 1/r\,, \qquad \Gamma^\phi{}_{\phi\vartheta} = \cot\vartheta \quad \textit{and} \quad \Gamma^\vartheta{}_{\phi\phi} = -\cos\vartheta\sin\vartheta\,.$$

Since the metric is diagonal, the non-diagonal elements of the Ricci tensor are zero too. We calculate with

$$R_{ab} = R^c{}_{acb} = \partial_c\Gamma^c{}_{ab} - \partial_b\Gamma^c{}_{ac} + \Gamma^c{}_{ab}\Gamma^d{}_{cd} - \Gamma^d{}_{bc}\Gamma^c{}_{ad}$$

for instance the rr component as

$$R_{rr} = 0 - 0 + \Gamma^c{}_{rr}\Gamma^d{}_{cd} - \Gamma^d{}_{rc}\Gamma^c{}_{rd} = \Gamma^r{}_{rr}(\Gamma^\phi{}_{r\phi} + \Gamma^\vartheta{}_{r\vartheta}) = -\frac{B'}{rB}\,. \tag{19.8}$$

Similarly, we find $R_{\vartheta\vartheta} = 1 + \frac{r}{2B^2}\frac{\mathrm{d}B}{\mathrm{d}r} - \frac{1}{B}$ and $R_{\phi\phi} = \sin^2\vartheta R_{\vartheta\vartheta}$.

Inserting our result for the spatial Ricci tensor into Eq. (19.6), we obtain

$$R_{rr} = \frac{1}{rB}\frac{\mathrm{d}B}{\mathrm{d}r} = 2Kg_{rr} = 2KB \tag{19.9a}$$

$$R_{\vartheta\vartheta} = 1 + \frac{r}{2B^2}\frac{\mathrm{d}B}{\mathrm{d}r} - \frac{1}{B} = 2Kg_{\vartheta\vartheta} = 2Kr^2\,, \tag{19.9b}$$

while the $\phi\phi$ equation contains no additional information. Integration of Eq. (19.9a) gives

$$B = \frac{1}{A - Kr^2} \tag{19.10}$$

with A as integration constant. Inserting then the result into Eq. (19.9b) fixes the constant as $A = 1$. Thus we have determined the line-element of a maximally symmetric 3-space with curvature K as

$$\mathrm{d}l^2 = \frac{\mathrm{d}r^2}{1 - Kr^2} + r^2(\sin^2\vartheta\mathrm{d}\phi^2 + \mathrm{d}\vartheta^2)\,. \tag{19.11}$$

Going over to the full four-dimensional line-element, we rescale for $K \neq 0$ the r coordinate by $r \to |K|^{1/2}r$. Then we absorb the factor $1/|K|$ in front of $\mathrm{d}l^2$ by defining the scale factor $a(t)$ as

$$a(t) = \begin{cases} S(t)/|K|^{1/2}, & K \neq 0 \\ S(t), & K = 0\,. \end{cases} \tag{19.12}$$

As result we obtain the Friedmann-Lemaître-Robertson-Walker (FLRW) metric for an homogeneous, isotropic universe,

$$\mathrm{d}s^2 = \mathrm{d}t^2 - a^2(t)\left[\frac{\mathrm{d}r^2}{1 - kr^2} + r^2(\sin^2\vartheta\mathrm{d}\phi^2 + \mathrm{d}\vartheta^2)\right]\,. \tag{19.13}$$

The constant k distinguishes the cases where the three-dimensional space has negative ($k = -1$), positive curvature ($k = 1$) or is flat ($k = 0$). Finally, we give two alternatives forms of the FLRW metric that are often more useful. In the first one, we introduce a new radial variable by $r = \sin\chi$ for $k = 1$. Then $\mathrm{d}r = \cos\chi\mathrm{d}\chi = (1 - r^2)^{1/2}\mathrm{d}\chi$ and the line-element follows as

$$\mathrm{d}s^2 = \mathrm{d}t^2 - a^2(t)\left[\mathrm{d}\chi^2 + S^2(\chi)(\sin^2\vartheta\mathrm{d}\phi^2 + \mathrm{d}\vartheta^2)\right] \tag{19.14}$$

with $S(\chi) = \sin\chi = r$. Defining

$$r = S(\chi) = \begin{cases} \sin\chi & \text{for} \quad k = 1\,, \\ \chi & \text{for} \quad k = 0\,, \\ \sinh\chi & \text{for} \quad k = -1\,, \end{cases} \tag{19.15}$$

the metric (19.14) is valid for all three values of k.

In the second alternative, we replace the comoving time t by the conformal time $\mathrm{d}\eta = \mathrm{d}t/a$,

$$\mathrm{d}s^2 = a^2(\eta)\left[\mathrm{d}\eta^2 - \frac{\mathrm{d}r^2}{1-kr^2} - r^2(\sin^2\vartheta\mathrm{d}\phi^2 - \mathrm{d}\vartheta^2)\right] \tag{19.16a}$$

$$= a^2(\eta)\left[\mathrm{d}\eta^2 - \mathrm{d}\chi^2 - S^2(\chi)(\sin^2\vartheta\mathrm{d}\phi^2 + \mathrm{d}\vartheta^2)\right] . \tag{19.16b}$$

Then the metric has for $k = 0$ the same form as the one of a uniformly expanding Minkowski space. The description of light propagation becomes therefore particularly simple using the coordinates (19.16b): Since $\mathrm{d}s^2 = 0$, a radial light-ray satisfies $\mathrm{d}\eta = \pm\mathrm{d}\chi$ and light-rays are straight lines at ±45 degrees in the $\eta - \chi$ plane.

Geometry of the Friedmann-Lemaître-Robertson-Walker metric Let us consider a sphere of fixed radius at fixed time, $\mathrm{d}r = \mathrm{d}t = 0$. The line-element $\mathrm{d}s^2$ simplifies then to $a^2(t)r^2(\sin^2\vartheta\mathrm{d}\phi^2 + \mathrm{d}\vartheta^2)$, which is the usual line-element of a sphere S^2 with radius $ra(t)$. The area of a sphere is $A = 4\pi(ra(t))^2 = 4\pi[S(\chi)a(t)]^2$ and the circumference of a circle is $L = 2\pi ra(t)$. Hence these quantities agree for $a = 1$ with the ones in an Euclidean space. By contrast, the radial distance between two points (r,ϑ,ϕ) and $(r+\mathrm{d}r,\vartheta,\phi)$ is $\mathrm{d}l = a(t)\mathrm{d}r/\sqrt{1-kr^2}$. Thus the radius of a sphere centered at $r = 0$ is

$$l = a(t)\int_0^r \frac{\mathrm{d}r'}{\sqrt{1-kr'^2}} = a(t)\times\begin{cases} \arcsin(r) & for \quad k = 1\,, \\ r & for \quad k = 0\,, \\ arcsinh\,(r) & for \quad k = -1\,. \end{cases} \tag{19.17}$$

Using χ as coordinate, the same result follows immediately,

$$l = a(t)\int_0^{\chi(r)} \mathrm{d}\chi = a(t)\chi\,. \tag{19.18}$$

Hence for $k = 0$, i.e. a flat space, one obtains the usual result $L/l = 2\pi$, while for $k = 1$ (spherical geometry) it is $L/l = 2\pi r/\arcsin(r) < 2\pi$ and for $k = -1$ (hyperbolic geometry) $L/l = 2\pi r/arcsinh\,(r) > 2\pi$. For $k = 0$ and $k = -1$, l is unbounded, while for $k = +1$ there exists a maximal distance $l_{\max}(t)$. Hence the first two case correspond to open spaces with an infinite volume, while the latter is a closed space with finite volume.

Hubble's law Lemaître, Hubble and other astronomers found empirically that the spectral lines of "distant" galaxies are redshifted, $z = \Delta\lambda/\lambda_0 > 1$, proportional to their distance d,

$$z = H_0 d\,. \tag{19.19}$$

If this redshift is interpreted as Doppler effect, $z = \Delta\lambda/\lambda_0 = v_r$, then the recession velocity of galaxies follows as

$$v = H_0 d\,. \tag{19.20}$$

The restriction "distant galaxies" means more precisely that the peculiar motion of galaxies caused by the gravitational attraction of nearby galaxy clusters should be small compared to the Hubble flow, $H_0 d \gg v_{\mathrm{pec}} \approx few \times 100\,\mathrm{km/s}$. Note that the interpretation of v as recession velocity is problematic. The validity of such an interpretation is certainly limited to $v \ll 1$. The parameter H_0 is called Hubble constant and has the value $H_0 \simeq (67.7 \pm 0.8)\,\mathrm{km/s/Mpc}$.

We will see soon that the Hubble law Eq. (19.20) is an approximation valid for $z \ll 1$. In general, the Hubble constant is not constant but depends on time, $H = H(t)$, and we will call it therefore Hubble parameter for $t \neq t_0$.

We can derive Hubble's law as consequence of the homogeneity of space: Consider Hubble's law as a vector equation with us at the center of the coordinate system,

$$\boldsymbol{v} = H\boldsymbol{d}. \tag{19.21}$$

What sees a different observer at position $\boldsymbol{d}'$? She has the velocity $\boldsymbol{v}' = H\boldsymbol{d}'$ relative to us. We are assuming that velocities are small and thus

$$\boldsymbol{v}'' \equiv \boldsymbol{v} - \boldsymbol{v}' = H(\boldsymbol{d} - \boldsymbol{d}') = H\boldsymbol{d}'' , \tag{19.22}$$

where $\boldsymbol{v}''$ and $\boldsymbol{d}''$ denote the position relative to the new observer. Hence a linear relation between v and d as Hubble law is the only relation compatible with homogeneity and thus the "cosmological principle" for $v \ll 1$.

Lemaître's redshift formula A non-static universe posses no time-like Killing vector field and thus we expect that the energy of a particle is affected by the expansion of space. In order to derive this change, we consider a massive particle moving along a geodesic $x(\tau)$ with four-velocity u^α. An observer with four-velocity U^α will measure as energy E of the massive particle $E = mu^\beta U_\beta$. Along the trajectory $x(\tau)$, this energy E changes as

$$\frac{\mathrm{d}}{\mathrm{d}\tau} E = u^\alpha \nabla_\alpha E = mu^\alpha \nabla_\alpha \left(u^\beta U_\beta\right) = mu^\alpha u^\beta \nabla_\alpha U_\beta , \tag{19.23}$$

where we used the geodesic equation $u^\alpha \nabla_\alpha u^\beta = 0$ in the last step. Evaluating the derivative using the Christoffel symbols for the FLRW metric and for comoving observers with $U^\alpha = (1, \mathbf{0})$ gives (cf. problem 20.**??**)

$$\nabla_\alpha U_\beta = \frac{\dot{a}}{a}(g_{\alpha\beta} - U_\alpha U_\beta) . \tag{19.24}$$

Inserting the derivative, we find with $u_\alpha u^\alpha = 1$

$$\frac{\mathrm{d}}{\mathrm{d}\tau} E = m\frac{\dot{a}}{a}\left(1 - E^2/m^2\right) = -\frac{\dot{a}}{a}\left(\frac{E^2 - m^2}{m}\right) , \tag{19.25}$$

or

$$-\frac{\mathrm{d}a}{a} = m\frac{\mathrm{d}t}{\mathrm{d}\tau}\frac{\mathrm{d}E}{E^2 - m^2} = \frac{E\mathrm{d}E}{E^2 - m^2} = \frac{1}{2}\,\frac{\mathrm{d}E(E^2 - m^2)}{E^2 - m^2} = \frac{\mathrm{d}p}{p} . \tag{19.26}$$

Thus the momentum of a particle is inverse proportionally to the scale factor of the universe, $p \propto 1/a$. An intuitive explanation of this result is that the expansion of the universe stretches all length scales of unbound systems, including the wave-length of free particles. As a consequence, the kinetic energy of non-relativistic particles goes quadratically to zero, and hence peculiar velocities relative to the Hubble flow are strongly damped by the expansion of the universe.

We can derive now Hubble's law by a Taylor expansion of the scale factor $a(t)$,

$$a(t) = a(t_0) + (t - t_0)\dot{a}(t_0) + \frac{1}{2}(t - t_0)^2\ddot{a}(t_0) + \ldots \tag{19.27a}$$

$$= a(t_0)\left[1 + (t - t_0)H_0 - \frac{1}{2}(t - t_0)^2 q_0 H_0^2 + \ldots\right] , \tag{19.27b}$$

where we introduced the Hubble constant

$$H_0 \equiv \frac{\dot{a}(t_0)}{a(t_0)} \qquad \text{and} \qquad q_0 \equiv -\frac{\ddot{a}(t_0)a(t_0)}{\dot{a}^2(t_0)}\,, \tag{19.28}$$

the so called deceleration parameter: If the expansion is slowing down, $\ddot{a} < 0$ and $q_0 > 0$. Hubble's law follows as an approximation for small redshift: For not too large time-differences, we can use the expansion Eq. (19.27a) and write

$$1 - z \simeq \frac{1}{1+z} = \frac{a(t)}{a_0} \simeq 1 + (t - t_0)H_0\,. \tag{19.29}$$

Hence Hubble's law, $z = (t_0 - t)H_0 = d/cH_0$, is valid as long as $z \simeq H_0(t_0 - t) \ll 1$. Deviations from its linear form arises for $z \gtrsim 1$ and can be used to determine q_0.

Proper distance In an expanding universe, the distance to an object depends on the expansion history, i.e. the behaviour of the scale factor $a(t)$ between the time of emission t of the observed light signal and its reception at t_0. From the metric (19.14) we can define the (radial) coordinate distance

$$\chi = \int_t^{t_0} \frac{\mathrm{d}t}{a(t)} \tag{19.30}$$

as well as the proper distance $d = g_{\chi\chi}\chi = a(t)\chi$. The proper distance d corresponds to the physical distance between two points on a hypersurface $t = const$. However, it is only for a static metric a directly measurable quantity and cosmologists use therefore other, operationally defined measures for the distance. The two most important examples are the luminosity and the angular diameter distances.

Luminosity distance The luminosity distance d_L is defined such that the inverse-square law between the luminosity L of a source at the distance d and the received energy flux $\mathcal{F}$ is valid,

$$d_L = \left(\frac{L}{4\pi\mathcal{F}}\right)^{1/2}\,. \tag{19.31}$$

Assume now that a (isotropically emitting) source with luminosity $L(t)$ and comoving coordinate χ is observed at t_0 by an observer at O. The cut at O through the forward light cone emitted at t_e by the source defines a sphere S^2 with proper area

$$A = 4\pi a^2(t_0)S^2(\chi)\,. \tag{19.32}$$

Additionally, we have to take into account that the frequency of a single photon is redshifted, $\nu_0 = \nu_e/(1 + z)$, and that the arrival rate of photons is reduced by the same factor due to time-dilation. Hence the received flux is

$$\mathcal{F}(t_0) = \frac{1}{(1+z)^2}\,\frac{L(t_e)}{4\pi a_0^2 S^2(\chi)} \tag{19.33}$$

and the luminosity distance in a FLRW universe follows as

$$d_L = (1 + z)\, a_0 S(\chi)\,. \tag{19.34}$$

Note that d_L depends via χ on the expansion history of the universe between t_e and t_0.

Angular diameter distance Instead of basing a distance measurement on standard candles, one may use standard rods with known proper length l whose angular diameter $\Delta\vartheta$ can be observed. Then we define the angular diameter distance as

$$d_A = \frac{l}{\Delta\vartheta}\,. \tag{19.35}$$

From the angular part of the FLRW metric it follows that $l = a(t_e)S(\chi)\Delta\vartheta$ and thus

$$d_A = a(t_e)S(\chi) = a(t_0)\frac{a(t_e)}{a(t_0)}S(\chi) = \frac{a_0 S(\chi)}{1+z}\,. \tag{19.36}$$

At small distances, $z \ll 1$, the angular diameter and the luminosity distance agree by construction, while for large redshift the difference increases as $(1+z)^2$.

Observable are not the coordinates χ or r, but the redshift z of a galaxy. Differentiating $1+z = a_0/a(t)$, we obtain

$$\mathrm{d}z = -\frac{a_0}{a^2}\,\mathrm{d}a = -\frac{a_0}{a^2}\frac{\mathrm{d}a}{\mathrm{d}t}\,\mathrm{d}t = -(1+z)H\mathrm{d}t\,. \tag{19.37}$$

Denoting by t_0 the present age of the Universe, we can find its age t at redshift z for a given $H(z)$ as

$$t_0 - t = \int_t^{t_0} \mathrm{d}t = \int_z^0 \frac{\mathrm{d}z}{H(z)(1+z)}\,. \tag{19.38}$$

Inserting the relation (19.37) into Eq. (19.30), we find the coordinate χ of a galaxy at redshift z as

$$\chi = \int_t^{t_0} \frac{\mathrm{d}t}{a(t)} = \frac{1}{a_0}\int_0^z \frac{\mathrm{d}z}{H(z)}\,. \tag{19.39}$$

For small redshift $z \ll 1$, we can use the expansion (19.27b) to approximate

$$\chi = \int_t^{t_0} \frac{\mathrm{d}t}{a_0}[1 - (t-t_0)H_0 + \ldots]^{-1} \tag{19.40a}$$

$$\approx \frac{1}{a_0}[(t-t_0) + \frac{1}{2}(t-t_0)^2 H_0 + \ldots] = \frac{1}{a_0 H_0}[z - \frac{1}{2}(1+q_0)z^2 + \ldots] \tag{19.40b}$$

In practise, one observes only the luminosity within a certain frequency range instead of the total (or bolometric) luminosity. A correction for this effect requires the knowledge of the intrinsic source spectrum.

19.2. Friedmann equations

We could determine the metric of a homogeneous and isotropic universe by purely geometrical considerations except for the value of $k \in \{-1, 0, +1\}$ and the unknown function $a(t)$: Since the scale factor is a dynamical quantity, it has to be determined by the Einstein equations. In order to write down these equations, we have to fix a model for the stress tensor $T_{\mu\nu}$ of the universe. A surprisingly generic ansatz for the matter content of the Universe which is consistent with Weyl's postulate is an ideal fluid: Although an ideal fluid is characterised by only one free parameter—the equation of state (EoS) $w = P/\rho$ fixing the ratio of its pressure

P and its energy density ρ—it can describe[1] both macrophysical systems (a fluid of galaxies) and elementary particles (e.g. a gas of photon or a fluid of cold dark matter particles). The various components of an ideal fluid evolve independently and thus the total stress tensor is simply the sum of the individual contributions.

In problem 5.**??** you derived as stress tensor of an ideal fluid in Minkowski space

$$T_{\mu\nu} = (\rho + P)u_\mu u_\nu - P\eta_{\mu\nu} \,. \tag{19.41}$$

Clearly, the homogeneity and isotropy of the FLRW metric imply that the stress tensor has the same symmetries and thus ρ and P can be only functions of time. Replacing $\eta_{\mu\nu}$ by $g_{\mu\nu}$ and choosing the comoving coordinates of Eq. (19.13) gives

$$g_{00} = 1 \,, \quad g_{11} = -\frac{a^2}{1-kr^2} \,, \quad g_{22} = -a^2r^2\sin^2\vartheta \quad and \quad g_{33} = -a^2r^2 \,. \tag{19.42}$$

Since the FLRW metric is diagonal, the elements of the inverse metric are simply given by $g^{\mu\mu} = 1/g_{\mu\mu}$. Thus the stress tensor follows for a comoving observer as

$$T_{00} = \rho g_{00} \quad and \quad T_{ij} = P\delta_{ij} \,. \tag{19.43}$$

Next we have to determine the still missing Christoffel symbols ${\Gamma^\sigma}_{\mu\nu}$, see problem 20.**??**. The result is

$${\Gamma^i}_{0i} = \frac{\dot a}{a} \quad and \quad {\Gamma^0}_{ij} = -\frac{\dot a}{a}\,\delta_{ij} \tag{19.44}$$

from which the non-zero components of the Ricci tensor follow as

$$R_{00} = -3\frac{\ddot a}{a}g_{00} \quad and \quad R_{ij} = \frac{a\ddot a + 2\dot a^2 + 2k}{a^2}\,\delta_{ij} \,. \tag{19.45}$$

The last ingredient needed for the Einstein equations is the curvature scalar,

$$R = g^{\mu\nu}R_{\mu\nu} = -6\,\frac{a\ddot a + \dot a^2 + k}{a^2} \,. \tag{19.46}$$

All the quantities appearing in the gravitational field equation are proportional to the metric tensor and thus the symmetries of the FLRW metric lead to only two independent equation of motions. The time-time component of the Einstein equation gives

$$3\frac{\dot a^2 + k}{a^2} = \kappa\rho + \Lambda \,, \tag{19.47}$$

while the space-space part results in

$$\frac{2a\ddot a + \dot a^2 + 2k}{a^2} = -\kappa P + \Lambda \,. \tag{19.48}$$

Using $\kappa = 8\pi G$, we obtain from Eq. (19.47) the Friedmann equation,

$$H^2 \equiv \left(\frac{\dot a}{a}\right)^2 = \frac{8\pi}{3}G\rho - \frac{k}{a^2} + \frac{\Lambda}{3} \,, \tag{19.49}$$

[1] Although deviations from this idealisation are small and happen only in specific phases during the evolution of the Universe, they are crucial to explain the observed amount of relic particles as baryons and dark matter: This topic will be introduced in the next chapter.

while the "acceleration equation" follows from (19.48) after eliminating $\ddot{a}$ with (19.47) as

$$\frac{\ddot{a}}{a} = \frac{\Lambda}{3} - \frac{4\pi G}{3}(\rho + 3P)\,. \tag{19.50}$$

This equation determines the (de-) acceleration of the Universe as function of its matter and energy content. "Normal" matter is characterised by $\rho > 0$ and $P \geq 0$. Thus a static solution is impossible for a universe with $\Lambda = 0$. Such a universe is decelerating and since today $\dot{a} > 0$, $\ddot{a}$ was always negative and there was a "big bang".

We define the *critical density* ρ_{cr} as the density for which the spatial geometry of the universe is flat. From $k = 0$, it follows

$$\rho_{cr} = \frac{3H_0^2}{8\pi G} \tag{19.51}$$

and thus ρ_{cr} is uniquely fixed by the value of H_0. One "hides" this dependence by introducing h,

$$H_0 = 100\, h\, \mathrm{k}m/(s\, Mpc)\,.$$

Then one can express the critical density as function of h,

$$\rho_{cr} = 2.77 \times 10^{11} h^2 M_\odot/\mathrm{M}pc^3 = 1.88 \times 10^{-29} h^2 \mathrm{g}/cm^3 = 1.05 \times 10^{-5} h^2 \mathrm{G}eV/cm^3\,.$$

Thus a flat universe with $h = 1$ requires an energy density of ~ 10 protons per cubic meter. We define the energy fraction Ω_i of the different players in cosmology as their energy density relative to the critical density, $\Omega_i = \rho/\rho_{cr}$. We will often include the cosmological constant Λ as another contribution to the energy density ρ via

$$\frac{8\pi}{3} G\rho_\Lambda = \frac{\Lambda}{3}\,. \tag{19.52}$$

Thereby one recognises also that the cosmological constant acts as a constant energy density with magnitude

$$\rho_\Lambda = \frac{\Lambda}{8\pi G} \qquad \text{or} \qquad \Omega_\Lambda = \frac{\Lambda}{3H_0^2}\,. \tag{19.53}$$

We can understand better the consequences of the cosmological constant by replacing Λ by $(8\pi G)\rho_\Lambda$ in the acceleration equation. Comparing then the effect of normal matter and of the Λ term on the acceleration,

$$\frac{\ddot{a}}{a} = \frac{8\pi G}{3}\rho_\Lambda - \frac{4\pi G}{3}(\rho + 3P)\,. \tag{19.54}$$

we recognise that Λ is equivalent to matter with an EoS $w_\Lambda = P/\rho = -1$, as we showed already in problem 19.??. Using ρ_Λ instead of Λ correspond on the level of the Einstein equations to a reshuffling of the cosmological constant from the geometry to the matter side, $\kappa T_{\mu\nu} + g_{\mu\nu}\Lambda \to \kappa\tilde{T}_{\mu\nu}$. Thus the observed value of the cosmological constant ρ_{obs} includes additionally to ρ_Λ both the classical contribution $V(\phi_0) \neq 0$ of all scalar potentials whose minima are not at zero and the (renormalised) quantum vacuum fluctuations of all matter fields. The borderline between an accelerating and decelerating universe is given by $\rho = -3P$ or $w = -1/3$. The condition $\rho < -3P$ violates the so-called strong energy condition for "normal" matter in equilibrium. An accelerating universe requires therefore a positive cosmological constant or a dominating form of matter that is not in equilibrium. As a phenomenological generalisation of the cosmological constant, one calls any form of matter *dark energy* which leads in the present epoch to an EoS parameter $w \approx -1$.

Relativistic species today are photons and possibly one of the three SM neutrinos. Their energy contribution is much smaller than the one of non-relativistic matter (stars, gas and cold dark matter). Thus the pressure term in the acceleration equation can be neglected at the present epoch. Measuring ρ, $\dot{a}/a$ and $\ddot{a}/a$ fixes therefore the geometry of the Universe: For a long time, cosmology was described therefore as the quest for two numbers, H_0 and q_0.

Thermodynamics The first law of thermodynamics becomes with $\mathrm{d}S = 0$ for a ideal fluid simply $\mathrm{d}U = T\mathrm{d}S - P\mathrm{d}V = -P\mathrm{d}V$. Considering a fixed comoving volume,

$$\mathrm{d}(\rho a^3) = -P\mathrm{d}(a^3)\,, \tag{19.55}$$

and dividing by $\mathrm{d}t$,

$$a\dot{\rho} + 3(\rho + P)\dot{a} = 0\,, \tag{19.56}$$

we obtain

$$\dot{\rho} = -3(\rho + P)H\,. \tag{19.57}$$

This result could be also derived from $\nabla_\alpha T^{\alpha\beta} = 0$, cf. problem **??**. Since $\nabla_\alpha T^{\alpha\beta} = 0$ is built-in in the Einstein equations, the three equations (19.49), (19.50), and (19.57) are not independent.

Scale-dependence of different energy forms The dependence of different energy forms as function of the scale factor a can derived from energy conservation, $\mathrm{d}U = -P\mathrm{d}V$, if an EoS $P = P(\rho) = w\rho$ is specified. For $w = const.$, it follows

$$\mathrm{d}(\rho a^3) = -3Pa^2\mathrm{d}a \tag{19.58}$$

or eliminating P

$$\frac{\mathrm{d}\rho}{\mathrm{d}a}a^3 + 3\rho a^2 = -3w\rho a^2\,. \tag{19.59}$$

Separating the variables,

$$-\,3(1+w)\frac{\mathrm{d}a}{a} = \frac{\mathrm{d}\rho}{\rho}\,, \tag{19.60}$$

we can integrate and obtain

$$\rho \propto a^{-3(1+w)} = \begin{cases} a^{-3} & \text{for matter} \quad (w=0)\,, \\ a^{-4} & \text{for radiation} \quad (w=1/3)\,, \\ \text{const.} & \text{for } \Lambda \quad (w=-1)\,. \end{cases} \tag{19.61}$$

The obtained scaling can be understood from heuristic arguments: The kinetic energy of non-relativistic matter is negligible, $kT \ll m$. Thus $\rho = nm \gg nT = P$ and non-relativistic matter is pressure-less, $w = 0$. The mass m is constant and $n \propto 1/a^3$, hence ρ is just diluted by the expansion of the universe, $\rho \propto 1/a^3$. Radiation is not only diluted but the energy of a single particle is additionally redshifted, $E \propto 1/a$. Thus the energy density of radiation scales as $\propto 1/a^4$. Alternatively, one can use that $\rho \propto T^4$ and $T \propto \langle E \rangle \propto 1/a$. Finally, the cosmological constant Λ acts by definition as an energy density $\rho_\lambda = \Lambda/(8\pi G)$ that is constant in time, independent from a possible expansion or contraction of the universe. Note also that cosmologists call relativistic particles radiation, while they include into matter only non-relativistic particles.

Let us rewrite the Friedmann equation for the present epoch as

$$\frac{k}{a_0^2} = H_0^2 \left(\frac{8\pi G}{3H_0^2}\rho_0 + \frac{\Lambda}{3H_0^2} - 1 \right) = H_0^2 \left(\Omega_{\mathrm{tot},0} - 1 \right) . \tag{19.62}$$

We express the curvature term for arbitrary times through $\Omega_{\mathrm{t}ot,0}$ and the redshift z as

$$\frac{k}{a^2} = \frac{k}{a_0^2}(1+z)^2 = H_0^2(\Omega_{\mathrm{t}ot,0} - 1)(1+z)^2 . \tag{19.63}$$

Note that the curvature term $(\Omega_{\mathrm{t}ot,0} - 1)(1+z)^2$ decreases slower than radiation or matter. Observations indicate that the universe is close to flat, and thus the curvature term can be safely neglected in the early universe. This behaviour poses also the question why the Universe is (nearly) flat, since for generic initial conditions one would expect $|\Omega_{\mathrm{t}ot,0} - 1|/\Omega_{\mathrm{t}ot} \gg 1$. Dividing the Friedmann equation (19.49) by $H_0^2 = 8\pi G\rho_{\mathrm{c}r}/3$, we obtain

$$\frac{H^2(z)}{H_0^2} = \sum_i \Omega_i(z) - (\Omega_{\mathrm{t}ot,0} - 1)(1+z)^2 \tag{19.64a}$$

$$= \Omega_{\mathrm{r}ad,0}(1+z)^4 + \Omega_{\mathrm{m},0}(1+z)^3 + \Omega_\Lambda - (\Omega_{\mathrm{t}ot,0} - 1)(1+z)^2 . \tag{19.64b}$$

This expression allows us to calculate the age of the universe (19.38), distances (19.34), etc. for a given cosmological model, i.e. specifying the energy fractions $\Omega_{i,0}$ and the Hubble parameter H_0 at the present epoch.

19.3. Evolution of simple cosmological models

Cosmological models with one energy component We consider a flat universe, $k = 0$, with one dominating energy component with EoS $w = P/\rho = const.$. With $\rho = \rho_{\mathrm{c}r}\,(a/a_0)^{-3(1+w)}$, the Friedmann equation becomes

$$\dot{a}^2 = \frac{8\pi}{3} G\rho a^2 = H_0^2 a_0^{3+3w}\, a^{-(1+3w)} , \tag{19.65}$$

where we inserted the definition of $\rho_{\mathrm{c}r} = 3H_0^2/(8\pi G)$. Separating variables we obtain

$$a_0^{-(3+3w)/2} \int_0^{a_0} \mathrm{d}a\, a^{(1+3w)/2} = H_0 \int_0^{t_0} \mathrm{d}t = t_0 H_0 \tag{19.66}$$

and hence the age of the Universe follows as

$$t_0 H_0 = \frac{2}{3+3w} = \begin{cases} 2/3 & \text{for matter} \quad (w=0) , \\ 1/2 & \text{for radiation} \quad (w = 1/3) , \\ \to \infty & \text{for } \Lambda \quad (w=-1) . \end{cases} \tag{19.67}$$

While a Universe with only a cosmological constant Λ has no "beginning", models with $w > -1$ needed a finite time to expand from the initial singularity $a(t = 0) = 0$ to the current size a_0. Since for $t \to 0$ the temperature and density formally diverge, one calls these cases often (hot) big bang models. We should expect that classical gravity breaks down when $\rho \sim M_{\text{Pl}}^4$. As long as $a \propto t^\alpha$ with $\alpha < 1$, most time elapsed during the last fractions of $t_0 H_0$. Hence our result for the age of the universe does not depend on unknown physics close to the big bang as long as $w > -1/3$.

If we integrate Eq. (19.66) to the arbitrary time t, we obtain as time-dependence of the scale factor

$$a(t) \propto t^{2/(3+3w)} = \begin{cases} t^{2/3} & \text{for matter} \quad (w = 0)\,, \\ t^{1/2} & \text{for radiation} \quad (w = 1/3)\,, \\ \exp(t) & \text{for } \Lambda \quad (w = -1)\,. \end{cases} \tag{19.68}$$

The special case with an exponential growing scale factor for $w = -1$ is called de Sitter space-time.

Age problem of the universe The age of a matter-dominated universe expanded around $\Omega_0 = 1$ is given by (problem 20.??)

$$t_0 = \frac{2}{3H_0}\left[1 - \frac{1}{5}(\Omega_0 - 1) + \ldots\right]. \tag{19.69}$$

Globular cluster ages require $t_0 \geq 13$ Gyr. Using $\Omega_0 = 1$ leads to $h \leq 0.50$. Thus a flat universe with $t_0 = 13\,\text{Gyr}$ without cosmological constant requires a value of H_0 which is too small compared with observations. Choosing $\Omega_{\text{m}} = 0.3$ increases the age by just 14%, cf. also with the left panel of Fig. 19.1.

We derive the age t_0 of a flat Universe with $\Omega_{\text{m}} + \Omega_\Lambda = 1$ in the next section as

$$\frac{3t_0 H_0}{2} = \frac{1}{\sqrt{\Omega_\Lambda}} \ln \frac{1 + \sqrt{\Omega_\Lambda}}{\sqrt{1 - \Omega_\Lambda}}\,. \tag{19.70}$$

Requiring $H_0 \geq 65$ km/s/Mpc and $t_0 \geq 13$ Gyr means that the function on the RHS should be larger than $\simeq 1.3$ or $\Omega_\Lambda \geq 0.55$. Thus the lower limit $t_0 \geq 13$ Gyr on the age of the Universe together with a lower limit on the Hubble constant is sufficient to deduce a non-zero value of the cosmological constant.

The ΛCDM model We consider as approximation to the late Universe a flat model containing as its only two components pressure-less matter and a cosmological constant, $\Omega_{\text{m}} + \Omega_\Lambda = 1$. Thus the curvature term in the Friedmann equation and the pressure term in the deceleration equation play no role and we can hope to solve these equations for $a(t)$. Multiplying the deceleration equation (19.50) by two and adding it to the Friedmann equation (19.49), we eliminate ρ_m,

$$2\,\frac{\ddot{a}}{a} + \left(\frac{\dot{a}}{a}\right)^2 = \Lambda\,. \tag{19.71}$$

Next we rewrite first the LHS and then the RHS as total time derivatives: With

$$\frac{\mathrm{d}}{\mathrm{d}t}(a\dot{a}^2) = \dot{a}^3 + 2a\dot{a}\ddot{a} = \dot{a}a^2\left[\left(\frac{\dot{a}}{a}\right)^2 + 2\,\frac{\ddot{a}}{a}\right], \tag{19.72}$$

we obtain

$$\frac{\mathrm{d}}{\mathrm{d}t}(a\dot{a}^2) = \dot{a}a^2\Lambda = \frac{1}{3}\,\frac{\mathrm{d}}{\mathrm{d}t}(a^3)\Lambda\,. \tag{19.73}$$

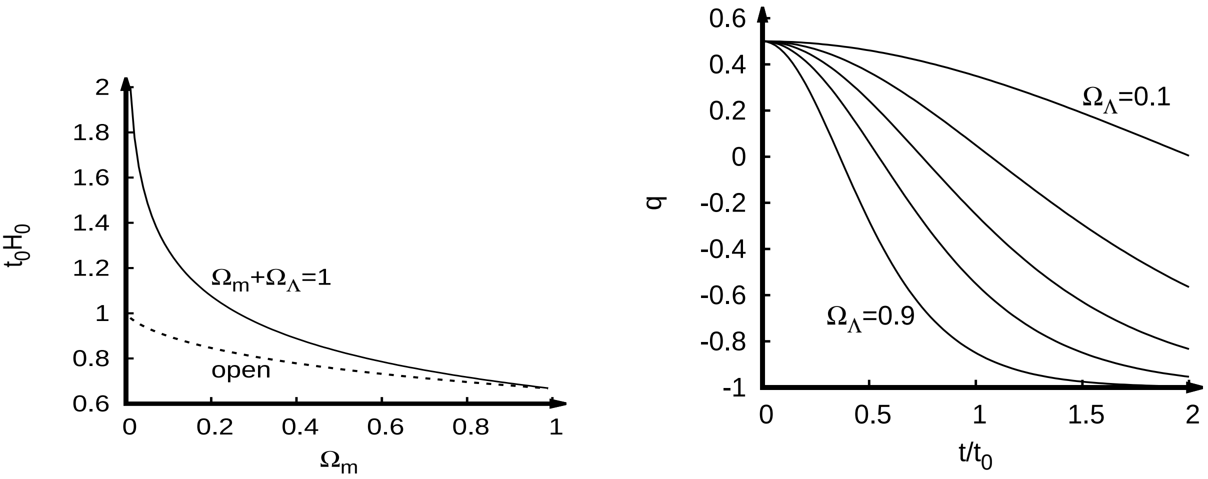


Figure 19.1.: Left: The product $t_0 H_0$ for an open universe containing only matter (dotted line) and for a flat cosmological model with $\Omega_\Lambda + \Omega_m = 1$ (solid line). Right: The deceleration parameter q as function of t/t_0 for the ΛCDM model and various values for Ω_Λ (0.1, 0.3, 0.5, 0.7 and 0.9 from the top to the bottom).

Integrating is now trivial,

$$a\dot{a}^2 = \frac{\Lambda}{3}\,a^3 + C\,. \tag{19.74}$$

The integration constant can be determined most easily setting $a(t_0) = 1$ and comparing the Friedmann equation (19.49) with (19.74) for $t = t_0$ as $C = 8\pi G\rho_{\mathrm{m},0}/3$. Now we introduce the new variable $x = a^{3/2}$. Then

$$\frac{\mathrm{d}a}{\mathrm{d}t} = \frac{\mathrm{d}x}{\mathrm{d}t}\,\frac{\mathrm{d}a}{\mathrm{d}x} = \frac{\mathrm{d}x}{\mathrm{d}t}\,\frac{2x^{-1/3}}{3}\,, \tag{19.75}$$

and we obtain as new differential equation

$$\dot{x}^2 - \Lambda x^2/4 + C/3 = 0\,. \tag{19.76}$$

Inserting the solution $x(t) = A\sinh(\sqrt{\Lambda}t/2)$ of the homogeneous equation fixes the constant A as $A = \sqrt{3C/\Lambda}$. We can express A also by the current values of Ω_i as $A = \Omega_m/\Omega_\Lambda = (1-\Omega_\Lambda)/\Omega_\Lambda$. Hence the time-dependence of the scale factor is

$$a(t) = A^{1/3}\sinh^{2/3}(\sqrt{3\Lambda}t/2)\,. \tag{19.77}$$

The time-scale of the expansion is set by $t_\Lambda = 2/\sqrt{3\Lambda}$. The present age t_0 of the universe follows by setting $a(t_0) = 1$ as

$$t_0 = t_\Lambda \mathrm{ar}\mathit{ctanh}\,(\sqrt{\Omega_\Lambda})\,. \tag{19.78}$$

The deceleration parameter $q = -\ddot{a}/aH^2$ is an important quantity for observational tests of the ΛCDM model. We calculate first the Hubble parameter

$$H(t) = \frac{\dot{a}}{a} = \frac{2}{3t_\Lambda}\coth(t/t_\Lambda) \tag{19.79}$$

and find then

$$q(t) = \frac{1}{2}[1 - 3\tanh^2(t/t_\Lambda)]\,. \tag{19.80}$$

The limiting behaviour of q corresponds with $q = 1/2$ for $t \to 0$ and $q = -1$ for $t \to \infty$ as expected to the one of a flat $\Omega_m = 1$ and a $\Omega_\Lambda = 1$ universe. More interesting is the transition region and, as shown in the right panel of Fig. 19.1, the transition from a decelerating to an accelerating universe happens for $\Omega_\Lambda = 0.7$ at $t \approx 0.55t_0$. This can be easily converted to redshift, $z_* = a(t_0)/a(t_*) - 1 \approx 0.7$, that is directly measured in supernova observations. The fact that the acceleration has set in recently, i.e. that z_* is of order one, has been coined the coincidence problem.

Remark 19.1: Topological defects as an energy component:
We have argued that topological defects are produced during cosmological phase transition via the Kibble mechanism. In order to understand the impact of topological defects on the evolution of the universe, one has to determine their EoS. The simplest case are magnetic monopoles which are created with negligible velocities. Hence their energy density behave as $\rho \sim 1/a^3$, and their abundance Ω increases during the radiation dominated phase. As a result, they tend to dominate the energy density of the universe, leading to its early collapse, cf. with exercise 20.??. Next we consider how the energy density of a network of domain walls evolves in an expanding universe. If we model them for simplicity as a set of static infinitely extended domain walls parallel to the yz plane, then the expansion of the universe along the y and z directions does not change the number density of domain walls. Therefore, the energy density of a network of domain walls evolves as $\rho \propto 1/a$. Simulations of dynamical networks of domain walls confirm approximately this behaviour. Interestingly, the equation of state of domain walls is thus negative, $w = -2/3$, leading to an accelerated expansion of the Universe. However, even a single wall inside the Hubble radius containing e.g. the electroweak vacuum would overclose the universe. Thus broken discrete symmetries are in general in conflict with cosmological observations. The remaining option, a network of cosmic strings, is viable if the string scale is below $\mu \lesssim (10^{-4}M_{\rm Pl})^2$.

19.4. Horizons

An important consequence of the finite speed of light and the expansion of the universe is the possibility that regions of space-time may be inaccessible for a given observer, either at the present time, or perhaps for all time. The borderline between the accessible and the inaccessible parts of the universe is termed a horizon.

Particle horizon We define the particle horizon d_p of a comoving observer O as the surface of the region in the past causally connected to O. Because of the homogeneity of the FLRW metric, it is sufficient to consider an observer at $\chi = 0$. The causally connected region is limited by light-rays, and thus bounded by

$$\chi_p(t_*) = \eta(t) - \eta(t_*) = \int_{t_*}^{t} \frac{\mathrm{d}t}{a(t)} = \int_0^{a(t)} \frac{\mathrm{d}a}{a\dot{a}}\,. \tag{19.81}$$

Here, the lower integration limit t_* is zero for models starting with a big bang. and $t_* \to -\infty$ otherwise. If the integral diverges, then the observer O can receive light-signals from an arbitrary spatial point in the past for sufficiently early t_*. If however the integral converges,

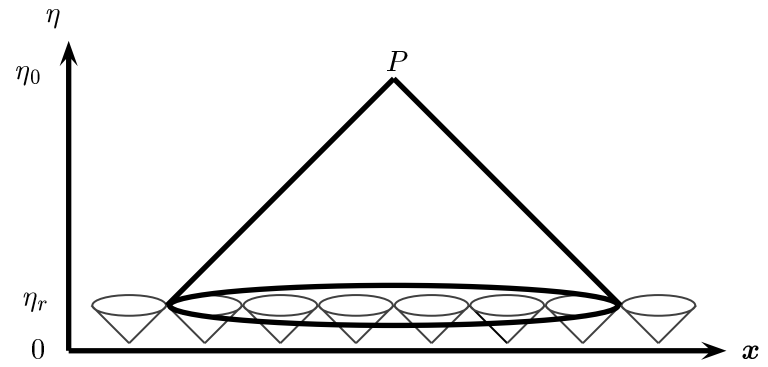


Figure 19.2.: Causal structure of a space-time with big bang: Particle horizons at the time η_r are shown in grey together with the event horizon of point P (black cone).

then $\chi_p(t_*)$ defines a particle horizon. The physical size of the horizon is given by $d_p(t) = a(t)\chi_p(t_*)$. If the model has a big bang origin, $a(0) = 0$, then also $z \to \infty$ for $t_* \to 0$ and thus the particle horizon corresponds to a surface of infinite redshift.

The second form of the integral shows that a finite particle horizon exists if $\ddot{a} < 0$, i.e. if the model was decelerating for $t \to t_*$. We can show this explicitly, if one form of matter dominates: Then $a(t) = a_0(t/t_0)^\alpha$, and thus

$$\chi_p(t_0) = \frac{1}{a_0}\int_0^t \mathrm{d}t \left(\frac{t}{t_0}\right)^{-\alpha} = \frac{t_0}{a_0(1-\alpha)} \tag{19.82}$$

for $\alpha < 1$. The physical size of the horizon follows as $d_p(t_0) = t_0/(1-\alpha)$. A particle horizon exists if $\alpha = 2/[3(1+w)] > 1$ or $w > -1/3$ and $q > 0$, i.e. for an decelerated expansion of the universe.

Differentiating the definition (19.81) gives $\mathrm{d}\chi_p/\mathrm{d}t = 1/a(t)$, and thus the particle horizon grows in an expanding universe: As time goes by, new objects become visible[2] after they cross the horizon and their redshift decreases from $z = \infty$. Note that in a universe with $w > -1/3$, these objects were never before in causal contact with the observer O: In such models, the observed homogeneity and isotropy of the Universe on large scales is actually a very puzzling fact called the horizon problem. The best observational probe of anisotropies are the photons of the cosmic microwave background (CMB) which scattered at redshift $z \simeq 1300$ the last time. Their intensity is the one of a black-body, with the same temperature in all directions except for fluctuations $\delta T/T \sim 10^{-5}$.

Hubble scale The Hubble scale (which is often called the Hubble horizon) is defined by $d_H(t) = H^{-1}(t)$. Since H^{-1} is of the same order as the curvature scalar R, the Hubble scale is a useful measure on which scales space-time curvature can be neglected and a simpler Newtonian analysis using an inertial coordinate system is possible. Moreover, the Hubble scale equals approximately the particle horizon for decelerating cosmological models, with $H^{-1} = (3+3w)/2t$. It is therefore often used as a substitute for the particle horizon. Note however that the difference between the two scales become very important for $w \to -1$.

[2] Since photons can interact, the boundary of the visible Universe is given more precisely by the "surface of last scattering", which is however comparable, see problem 20.??.

Event horizon Although the particle horizon grows in an expanding universe, there may be galaxies which may be never be seen by an observer. This is the case if the integral in Eq. (19.81) converges for $t \to t_{\max}$. We define the event horizon by

$$\chi_p(t_*) = \int_{t_*}^{t_{\max}} \frac{\mathrm{d}t}{a(t)} = \int_0^{a(t)} \frac{\mathrm{d}a}{a\dot{a}}\,, \tag{19.83}$$

where $t_{\max}$ is either infinite or corresponds to the big crunch, $a(t_{\max}) = 0$, i.e. the end of the universe in a future singularity. The existence of an event horizon in the limit $t_{\max} \to \infty$ suggests that space expands faster than the speed of light, which is often termed as "superluminal expansion". Note that this notion is misleading, since the definition of the relative velocity of two observers O and O' relies on the existence of an (approximate) inertial system connecting the two. For distances $\gtrsim 1/H$, when the "relative velocity" becomes superluminal, such a connection is not possible and thus their relative velocity is not defined.

In summary, the particle horizon is the maximal distance from which we can receive signals, while the event horizon defines the maximal distance to which we can send signals. The horizon problem of a radiation or matter dominated universe is illustrated schematically in Fig. 19.2: Particle horizons, i.e. causally connected regions, at the time of recombination η_r are shown in red. The event horizon (black cone) of the observer at the point P is much larger at the time of recombination and contains therefore many apparently causally disconnected patches.

Summary

A homogeneous and isotropic universe is described by a FLRW metric. This family of space-times is completely determined by the scale factor $a(t)$ and the parameter k distinguishing between hyperbolic, flat, or spherical 3-spaces. The time evolution of these models is specified, if at a given time t the Hubble parameter $H(t)$, the abundances Ω_i of all relevant matter forms and their equation of state $w_i = P_i/\rho_i$ is known. Observational data are well described by the ΛCDM model with $\Omega_\Lambda \simeq 0.7$, $\Omega_m \simeq 0.3$ and $H_0 \simeq 70$km/s/Mpc.

Further reading

For a discussion of important observations in cosmology and their interpretation see e.g. [CL02]. The properties of simple cosmological models are discussed in more detail by [HEL06].

20. Thermal relics

At large redshifts, we can treat the different contributions to the total energy density of the Universe as nearly homogeneous fluids whose momenta are increasing as $p \propto 1/a$ for $a \to 0$. This energy increase has three main effects: Firstly, bound states like atoms, nuclei and hadrons are dissolved when the temperature reaches their binding energy, $T \gtrsim E_b$. Secondly, particles with mass m_X can be produced in reactions like $\gamma\gamma \to \bar{X}X$, when $T \gtrsim 2m_X$. Finally, most scattering rates $\Gamma = n\sigma v$ increase faster than the expansion rate of the universe for $t \to 0$, since the number density n scales as $n \propto T^3$ for relativistic particles, while the expansion rate goes as $H \propto \rho_{rad}^{1/2} \propto T^2$. Therefore, reactions that have become "frozen in" today may have been important in the early Universe. As a result, the early Universe consisted of a plasma containing particles with $m_X \lesssim T$ in a state of thermal equilibrium—except those particles that are very weakly interacting as e.g. axions.

In the previous chapter, we have assumed that the evolution of the Universe proceeds adiabatically. While this is an excellent approximation during most of its evolution, deviations are crucial for the explanation of the present Universe: Without such deviations, the abundance of all particles with masses smaller than the present temperature of the CMB, $m \gg T_{\rm CMB} \simeq 2 \times 10^{-4}\,\text{eV}$ would be exponentially suppressed—or a non-zero chemical potential should ensure their survival. In particular, the abundance of baryons, of light elements from helium up to lithium, and likely of dark matter are connected to deviations from thermal equilibrium. Their present abundance can be calculated knowing their reaction rates and the expansion history of the universe, and we will develop the required formalism in this chapter.

20.1. Boltzmann equation

The evolution of the phase space distribution function of a thermal relic in the expanding universe can be modelled by a Boltzmann equation. As first step in its derivation, we recall Liouville's theorem and the resulting collisionless Boltzmann equation in Minkowski space, before we generalise this equation to a space-time described by a FLRW metric.

Liouville's theorem We consider first the evolution of a classical many-particle system in $6n$-dimensional phase space $\omega_j \equiv (q^j, p_j)$ with $j = 1, \ldots, 3n$. The phase space density $f(q^j, p_j, t)$ determines the probability to find the system in the state $\omega = (q^j, p_j)$ at time t. Conservation of the particle number leads to a conservation law for f,

$$\frac{\partial f}{\partial t} + \frac{\partial}{\partial \omega_i}(f\dot{\omega}_i) = 0\,. \tag{20.1}$$

We first use Hamilton's equations (1.22) to replace $\dot{\omega}$ with ω,

$$\frac{\partial}{\partial \omega_i}(f\dot{\omega}_i) = \frac{\partial}{\partial q^i}\left(f\dot{q}^i\right) + \frac{\partial}{\partial p_i}(f\dot{p}_i) = \frac{\partial}{\partial q^i}\left(f\frac{\partial H}{\partial p_i}\right) - \frac{\partial}{\partial p_i}\left(f\frac{\partial H}{\partial q^i}\right)\,, \tag{20.2}$$

and then that the mixed second derivatives of H commute, $\partial q^i \partial p_i = \partial p_i \partial q^i$,

$$\frac{\partial}{\partial \omega_i}(f\dot{\omega}_i) = \frac{\partial f}{\partial q^i}\frac{\partial H}{\partial p_i} - \frac{\partial f}{\partial p_i}\frac{\partial H}{\partial q^i} = \{f, H\} = \dot{q}^i \frac{\partial f}{\partial q^i} + \dot{p}_i \frac{\partial f}{\partial p_i}. \tag{20.3}$$

Inserting these results into the conservation law (20.1), Liouville's theorem follows

$$\frac{\mathrm{d}f}{\mathrm{d}t} = \frac{\partial f}{\partial t} + \{f, H\} = \frac{\partial f}{\partial t} + \frac{\partial f}{\partial q^i}\dot{q}^i + \frac{\partial f}{\partial p_i}\dot{p}_i = 0. \tag{20.4}$$

Thus the phase space density $f(q^j, p_j, t)$ of a Hamiltonian system stays constant along a trajectory in phase space.

Boltzmann equation in Minkowski space Particles in a thermal plasma can exchange energy via elastic collisions, and can be created and destroyed in inelastic collisions. We can split the time-evolution of such a system into the free movement of particles along space-time geodesics interrupted by collisions, if these scatterings are caused by short-range interactions. Without scatterings, the $6n$ dimensional phase-space function is separable, i.e. it is the n-dimensional product of single-particle distribution functions. Therefore it is sufficient to describe the time-evolution of a collisionless many-particle system in the simpler six-dimensional phase space $\omega \equiv (\boldsymbol{q}, \boldsymbol{p})$. In a second step, we will account for interactions by adding a collision term.

Now the six-dimensional phase space density $f(\boldsymbol{q}, \boldsymbol{p}, t)$ determines the density $\mathrm{d}N = f(\boldsymbol{x}, \boldsymbol{p})\mathrm{d}^3x\mathrm{d}^3p$ of particles in the state $\omega = (\boldsymbol{q}, \boldsymbol{p})$ at time t. We can view also this system as a fluid with $\dot{\omega}$ as the fluid velocity and f as its density. If the number of particles is not changed by interactions, then again a conservation law for f is valid,

$$\frac{\partial f}{\partial t} + \frac{\partial}{\partial w_i}(f\dot{w}_i) = \frac{\partial f}{\partial t} + \frac{\partial f}{\partial q^i}\dot{q}^i + \frac{\partial f}{\partial p_i}\dot{p}_i = 0. \tag{20.5}$$

This is the collisionless Boltzmann equation in Minkowski space.

Boltzmann equation for a FLRW metric Next we apply the Boltzmann equation to the case of a universe described by a FLRW metric. Particles move along geodesics which we parametrise by the parameter λ. Then the collisionless Boltzmann equation states that the phase space density $f(\boldsymbol{x}, \boldsymbol{p}, t)$ stays constant along all trajectories, $\mathrm{d}f/\mathrm{d}\lambda = 0$. In order to evaluate the total derivative $\mathrm{d}f/\mathrm{d}t = 0$, we have to fix the dependence $\lambda(t)$. Using the choice $\mathrm{d}\lambda = \mathrm{d}t/E$ as affine parameter (such that $\mathrm{d}x^\mu/\mathrm{d}\lambda = p^\mu$) and assuming an isotropic and uniform distribution of matter in space, $f(\boldsymbol{x}, \boldsymbol{p}, t) = f(p, t)$, we find the collisionless Boltzmann equation for a FLRW metric as

$$\frac{\mathrm{d}f}{\mathrm{d}\lambda} = \left(\frac{\mathrm{d}t}{\mathrm{d}\lambda}\frac{\partial}{\partial t} + \frac{\mathrm{d}\boldsymbol{p}}{\mathrm{d}\lambda}\frac{\partial}{\partial \boldsymbol{p}}\right) f(p,t) = E\left(\frac{\partial}{\partial t} - H\boldsymbol{p}\frac{\partial}{\partial \boldsymbol{p}}\right) f(p,t) = 0. \tag{20.6}$$

In the second step we used also the redshift formula for a free particle,

$$\frac{\mathrm{d}\boldsymbol{p}}{\mathrm{d}\lambda} = \frac{\mathrm{d}\boldsymbol{p}}{\mathrm{d}t}\frac{\mathrm{d}t}{\mathrm{d}\lambda} = -H\boldsymbol{p}E. \tag{20.7}$$

Adding the effect of collisions, the total time derivative of the phase space density f becomes non-zero,

$$\frac{\mathrm{d}f}{\mathrm{d}\lambda} = E\left(\frac{\partial}{\partial t} - H\boldsymbol{p}\frac{\partial}{\partial \boldsymbol{p}}\right) f(p,t) = C(p,t), \tag{20.8}$$

and is determined by the collision term $C(p,t)$.

Most often, we are not interested in the momentum dependence of $f(p,t)$ but only in the total density $n(t)$ of the relic particles. Dividing by the energy E and taking the first moment of the Boltzmann equation, we arrive at

$$g\int\frac{\mathrm{d}^3p}{(2\pi)^3}\left(\frac{\partial}{\partial t}-H\boldsymbol{p}\frac{\partial}{\partial\boldsymbol{p}}\right)f(p,t)=g\int\frac{\mathrm{d}^3p}{(2\pi)^3}\,\frac{1}{E_{\boldsymbol{p}}}\,C(p,t)\,. \tag{20.9}$$

The first term on the LHS is the time derivative of the number density n of relic particles, as a comparison to the general definition in Eq. (14.49) shows. In the second term, we use the isotropy of the particle distribution and perform a partial integration. The boundary term $p^3f(p,t)|_0^\infty$ vanishes[1], and thus we obtain

$$\frac{\partial}{\partial t}n+3Hn=g\int\frac{\mathrm{d}^3p}{(2\pi)^3}\,\frac{1}{E_{\boldsymbol{p}}}\,C(p)\,. \tag{20.10}$$

We can express the LHS as the change of the comoving number density a^3n,

$$\frac{\mathrm{d}(a^3n)}{\mathrm{d}t}=a^3\,g\int\frac{\mathrm{d}^3p}{(2\pi)^3}\,\frac{1}{E_{\boldsymbol{p}}}\,C(p)\equiv a^3\,\frac{\mathrm{d}n_{col}}{\mathrm{d}t}\,, \tag{20.11}$$

and thus the RHS is the net change $a^3\dot{n}_{col}$ of the comoving particle density due to interactions.

Let us determine $\dot{n}_{col}$ for the case that the relic particle interacts via decays and inverse decays, $1\leftrightarrow 1'2'\cdots n'$, with other particles in the thermal plasma. Recall first Eq. (9.117), that gives the number of decays $1\to 1'2'\cdots n'$ per time of a given particle 1 with momentum $\boldsymbol{p}_1$ at temperature $T=0$. This rate should be multiplied with $f_1(\boldsymbol{p})$ to get the number of decays per time of all relic particles of type 1 with momentum $\boldsymbol{p}_1$. Moreover, in a thermal bath, we have to account for the Pauli blocking of fermions and the stimulated emission of bosons by adding the factors $1\pm f_{i'}(\boldsymbol{p}_{i'})$ to the final state phase space. The inverse reaction $1'2'\cdots n'\to 1$ is proportional to the number of plasma particles and contains a factor $1\pm f_1(\boldsymbol{p})$ in the final phase-space. Combining everything, we have

$$\begin{aligned}\frac{\mathrm{d}n_1^{col}(\boldsymbol{p}_1)}{\mathrm{d}t}=&-\frac{1}{2E_1}\int\prod_{f=1}^{n'}\frac{\mathrm{d}^3p_f}{2E_f(2\pi)^3}\,(2\pi)^4\delta^{(4)}(p_1-P_f)\\&\times\left[f_1(1\pm f_{1'})\cdots(1\pm f_{n'})|\widetilde{\mathcal{A}_{fi}}|^2-f_{1'}\cdots f_{n'}(1\pm f_1)|\widetilde{\mathcal{A}_{if}}|^2\right]\,,\end{aligned} \tag{20.12}$$

with $P_f=\sum_{i'}p_{i'}^\mu$, $f_i\equiv f(\boldsymbol{p}_i)$, while $|\widetilde{\mathcal{A}_{fi}}|^2=\sum_{s_1,\ldots,s_{n'}}|\mathcal{A}_{fi}|^2$ denotes the squared Feynman amplitude summed over both final and initial internal degrees of freedom. If necessary, symmetry factors $S=1/n!$ should be added for n identical particles in the initial or final state. As final step, we have to integrate the differential rate $\dot{n}_1^{col}(\boldsymbol{p}_1)$ over $\mathrm{d}^3p_1/(2\pi)^3$ to obtain the total collision rate $\dot{n}_1^{col}$ per physical volume, finding as Boltzmann equation

$$\begin{aligned}\frac{1}{a^3}\frac{\mathrm{d}(a^3n_1^{col})}{\mathrm{d}t}=&-\int\frac{\mathrm{d}^3p_1}{2E_1(2\pi)^3}\prod_{f=1}^{n}\frac{\mathrm{d}^3p_f}{2E_f(2\pi)^3}\,(2\pi)^4\delta^{(4)}(p_1-P_f)\\&\times\left[f_1(1\pm f_{1'})\cdots(1\pm f_{n'})|\widetilde{\mathcal{A}_{fi}}|^2-f_{1'}\cdots f_{n'}(1\pm f_1)|\widetilde{\mathcal{A}_{if}}|^2\right]\,.\end{aligned} \tag{20.13}$$

[1] A finite energy density requires that $f(p)$ decreases faster than $1/p^4$ for large p.

It is instructive to specialise this equation to the simplest case of a $1 \leftrightarrow 2+3$ processes. Then the decay widths scale as $\Gamma_{1\to 2+3} \propto g_1^{-1}\sum_{s_1,s_2,s_3}|\mathcal{A}_{1\to 2+3}|^2$ and $\Gamma_{2+3\to 3} \propto g_2^{-1}g_3^{-1}\sum_{s_1,s_2,s_3}|\mathcal{A}_{2+3\to 1}|^2$, respectively. Thus we can rewrite the Boltzmann equation as

$$\frac{1}{a^3}\frac{\mathrm{d}(a^3 n_1^{col})}{\mathrm{d}t} = -n_1\langle\Gamma_{1\to 2+3}\rangle + n_2 n_3\langle\Gamma_{2+3\to 1}\rangle\,, \tag{20.14}$$

where the factors g_i entered the densities n_i and $\langle\Gamma\rangle$ are the rates thermally averaged over the gamma factors m_i/E_i. Going over from a decay $1 \leftrightarrow 1'2'\cdots n'$ to a scattering process $12\cdots n \leftrightarrow 1'2'\cdots m'$, we should add to Eq. (20.13) the corresponding phase space factors

$$\int\frac{\mathrm{d}^3p_2}{2E_i(2\pi)^3}\,f_2(\boldsymbol{p}_2) \quad and \quad \int\frac{\mathrm{d}^3p_2}{2E_i(2\pi)^3}\,(1-f_2(\boldsymbol{p}_2))$$

to the first and second term in the square bracket of Eq. (20.13), respectively.

In most cases of interest, the collision rate can be simplified drastically: First, we can set $|\mathcal{A}_{fi}|^2 \simeq |\mathcal{A}_{if}|^2$, because CP violation is generically small[2]. At tree-level, this relation holds exactly. Second, kinetic decoupling happens typically much later than chemical decoupling: Consider e.g. a heavy relic particle X scattering with neutrinos as part of the thermal bath. Then, elastic collisions $X\nu \to X\nu$ may keep the distribution function of X still close to the one in thermal equilibrium, although inelastic collisions like $\bar{\nu}\nu \to \bar{X}X$ became ineffective below $T \sim m_X$. As a result, deviations of the relic density from a thermal equilibrium distribution in the time between chemical and kinetic decoupling can be simply parametrised by a time-dependent chemical potential $\mu(t)$. Thirdly, we can assume that the densities are small enough such that the distribution functions can be approximated by Maxwell-Boltzmann distributions,

$$n(t) = \mathrm{e}^{\beta(t)\mu(t)}n_{\rm eq}(t) = \mathrm{e}^{\beta(t)\mu(t)}\;g\int\frac{\mathrm{d}^3p}{2E(2\pi)^3}\exp\left(-\beta(t)E\right)\,. \tag{20.15}$$

Since this corresponds to the classical limit, we can also neglect the factors for Pauli blocking and stimulated emission, setting $1 \pm f_i \simeq 1$. Finally, we often (but not always) assume that in the early universe any asymmetry in the number of particles and antiparticles is zero. If $\mu(0) = \bar{\mu}(0) = 0$, then $\mu(t) = \bar{\mu}(t)$ also at later times.

We employ now the first three approximations, but keep $\mu_i \neq 0$. Specialising also to the case of $12 \to 34$ scatterings, the second line in (20.13) becomes

$$\left[f_1 f_2(1\pm f_3)(1\pm f_4)|\widetilde{\mathcal{A}_{fi}}|^2 - f_3 f_4(1\pm f_1)(1\pm f_2)|\widetilde{\mathcal{A}_{if}}|^2\right] \tag{20.16}$$

$$\simeq \mathrm{e}^{-\beta(E_1+E_2)}\left(\mathrm{e}^{\beta(\mu_1+\mu_2)} - \mathrm{e}^{\beta(\mu_3+\mu_4)}\right)|\widetilde{\mathcal{A}_{fi}}|^2 \tag{20.17}$$

$$= \mathrm{e}^{-\beta(E_1+E_2)}\left[\frac{n_1 n_2}{n_1^{\rm eq}n_2^{\rm eq}} - \frac{n_3 n_4}{n_3^{\rm eq}n_4^{\rm eq}}\right]|\widetilde{\mathcal{A}_{fi}}|^2\,. \tag{20.18}$$

Here we used also energy conservation, $E_1 + E_2 = E_3 + E_4$, before we eliminated in the final step the chemical potentials in favour of the number densities, using $n_i = \mathrm{e}^{\beta\mu_i}n_i^{\rm eq}$. Now we define the thermally averaged cross section $\langle\sigma v\rangle$ as

$$\langle\sigma v\rangle \equiv \frac{1}{n_1^{\rm eq}n_2^{\rm eq}}\prod_{i=1}^{4}\int\frac{\mathrm{d}^3p_i}{2E_i(2\pi)^3}\,\mathrm{e}^{-\beta(E_1+E_2)}\,(2\pi)^4\delta^{(4)}(P_i - P_f)|\widetilde{\mathcal{A}_{fi}}|^2\,. \tag{20.19}$$

[2] We postpone the effect of CP violation to chapter 21.

Here the velocity v in $\langle\sigma v\rangle$ is the relative or Møller velocity, $v \equiv v_{\mathrm{M}\o l}$. Its sole purpose is to cancel the corresponding factor $v_{\mathrm{M}\o l}$ in the flux factor I contained in the $T = 0$ cross section σ, compare with Eq. (9.149). As a result, the integration on the RHS is over Lorentz invariant factors that do contain no explicit factor $v_{\mathrm{M}\o l}$. Employing the definition (20.19) for $\langle\sigma v\rangle$, we arrive at a rather compact expression for the Boltzmann equation,

$$\frac{1}{a^3}\frac{\mathrm{d}(a^3 n_1)}{\mathrm{d}t} = -n_1^{\mathrm{eq}} n_2^{\mathrm{eq}}\langle\sigma v\rangle\left[\frac{n_1 n_2}{n_1^{\mathrm{eq}} n_2^{\mathrm{eq}}} - \frac{n_3 n_4}{n_3^{\mathrm{eq}} n_4^{\mathrm{eq}}}\right]. \tag{20.20}$$

Note that we managed to reduce the integro-differential equation (20.8) to an ordinary differential equation. This equation together with the initial condition $n_1 \simeq n_{\mathrm{eq}}$ for $T \to \infty$ determines $n_1(t)$ for a given thermal cross section $\langle\sigma v\rangle$.

Important applications of this equation are the freeze-out of the baryon asymmetry and the dark matter abundance, the formation of light elements in big bang nucleosynthesis (BBN), and the recombination of electrons and protons into hydrogen. While the structure of the Boltzmann equation is the same for all theses processes, the energy scales involved vary from few eV to scales possibly as high as 10^{16} GeV. Correspondingly, the microphysics required as input to describe recombination consists of atomic physics, while BBN depends on weak interactions between nucleons and neutrinos as well as on nuclear physics. By contrast, the freeze-out of DM is a problem involving mainly particle physics. Therefore we will pick out this problem for a discussion of the freeze-out mechanism which is at work in all the four processes. After that, we will only sketch the main points of BBN. We postpone baryogenesis to the next chapter.

20.2. Thermal relics as dark matter

A wealth of observational data suggests that a viable dark matter (DM) candidate has to be non-baryonic and should be non-relativistic, at least from the time of matter-radiation equilibrium on [Lis16]. As none of the particles in the SM has the required properties, DM has to belong to a new sector of particles beyond the SM. The various particles X proposed as DM candidates can be divided in two main sub-categories: Thermal relics were at least once during the history of the Universe in chemical equilibrium with the thermal plasma, while non-thermal relics have either sufficiently small interactions or a high enough mass m_X to be never produced efficiently by processes like $e^-e^+ \to XX$. We consider in the following thermal relics, while we come back to the topic of non-thermal production of relics later on.

20.2.1. Abundance of thermal relics

We assume that the thermal relic X and their annihilation products are symmetric, i.e. that no asymmetry in the number density of particles and antiparticles exist. Moreover, we assume that the annihilation products are in thermal equilibrium. Then the Boltzmann equation simplifies to,

$$\frac{1}{a^3}\frac{\mathrm{d}(a^3 n_X)}{\mathrm{d}t} = -\langle\sigma_{\mathrm{ann}} v\rangle\left[n_X^2 - (n_X^{\mathrm{eq}})^2\right]. \tag{20.21}$$

In the calculation of $\langle\sigma_{\mathrm{ann}} v\rangle$ a sum over all relevant final states should be included. Knowing $\langle\sigma_{\mathrm{ann}} v\rangle$ as function of temperature, this equation can be numerically integrated using as initial condition $n_X \simeq n_X^{eq}$ for a large enough initial temperature T. In order to obtain simple numerical estimates and gaining some insight which physical parameters determine the final abundance, we will develop instead the more intuitive *Gamov criterion*: It states that a reaction becomes ineffective, when its rate Γ drops below the expansion rate H of the universe. Note that H defines also the typical time-scale for changes in the temperature T: Thus $\Gamma \ll H \propto \dot{T}/T$ implies that the particle can not follow any longer an equilibrium distribution. For the important case of a power-law expansion, the condition $\Gamma \ll H$ means that the time $1/\Gamma$ between a reaction becomes larger than the age $t \propto 1/H(t)$ of the universe. In the case of annihilations, we have to determine therefore the "freeze-out" time $t_{\rm f}$ defined by $\Gamma_{\rm A}(t_{\rm f}) = H(t_{\rm f})$ with $\Gamma_{\rm A} = n_{\rm eq}\langle\sigma_{\rm ann} v\rangle$ as the annihilation rate. At later times, the comoving density of the relic particle is then constant.

Freeze-out of thermal relic particles Without interactions, the number density n_X of a particle species X is only diluted by the expansion of space, $n_X \propto a^{-3}$. It is convenient to account for this trivial expansion effect by dividing n_X through the entropy density[3] s, i.e. to use the dimensionless quantity $Y = n_X/s$. The equilibrium abundance Y_{eq} expressed as function of $x = T/m_X$ is given for vanishing chemical potential $\mu_X = 0$ by

$$Y_{\rm eq} = \frac{n_X}{s} = \begin{cases} \frac{45}{2\pi^4}\left(\frac{\pi}{8}\right)^{1/2} \frac{\varepsilon_1 g_X}{g_{*S}}\, x^{3/2}\exp(-x) = 0.145\,\frac{\varepsilon_1 g_X}{g_{*S}}\, x^{3/2}\exp(-x) & \text{for } x \gg 3, \\ \frac{45\zeta(3)}{2\pi^4}\,\frac{\varepsilon g_X}{g_{*S}} = 0.278\,\frac{\varepsilon g_X}{g_{*S}} & \text{for } x \ll 3 \end{cases} \tag{20.22}$$

with $\varepsilon = 3/4$ ($\varepsilon = 1$) for fermions (bosons). If the particle X is in chemical equilibrium, its abundance is determined for $T \gg m$ by its contribution to the total number of degrees of freedom of the plasma, while $Y_{\rm eq}$ is exponentially suppressed for $T \ll m$.

In an expanding universe, one may expect that the reaction rate Γ for processes like $e^+e^- \leftrightarrow \bar{X}X$ drops below the expansion rate H mainly for two reasons: First, cross sections may decrease with energy as, e.g., weak processes $\sigma \propto s \propto T^2$ for $s \lesssim m_W^2$. Second, the density n_X decreases at least as $n_X \propto T^3$. When a reaction rate drops below the expansion rate, it becomes ineffective, or "freezes out." Around this freeze-out time $x_{\rm f}$, the true abundance Y starts to deviate from the equilibrium abundance $Y_{\rm eq}$ and becomes constant, $Y(x) \simeq Y_{\rm eq}(x_{\rm f})$ for $x \gtrsim x_{\rm f}$. This behaviour is illustrated in Fig. 20.1.

Next we rewrite the evolution equation for $n_X(t)$ using the dimensionless variables Y and x. We assume that the freeze-out occurs during the radiation-dominated epoch. Thus $\rho_{rad} \propto 1/a^4$, $H = 1/(2t)$ and the curvature term k/R^2 can be neglected. Then the Friedmann equation simplifies to $H^2 = (8\pi/3)G\rho$ with $\rho = g_*\pi^2/30T^4$, or

$$\frac{1}{2t} = H = 1.66\sqrt{g_*}\,\frac{T^2}{M_{\rm Pl}} \propto x^{-2}\,. \tag{20.23}$$

Here we introduced also the Planck mass $M_{\rm Pl} = 1/\sqrt{G_N} \simeq 1.2 \times 10^{19}$ GeV. Changing then from $n_X = sY$ to Y, we eliminate the $3Hn_X$ term obtaining

$$\frac{\mathrm{d}Y}{\mathrm{d}x} = -\frac{sx}{H}\langle\sigma_{\rm ann}v\rangle\left(Y^2 - Y_{\rm eq}^2\right)\,. \tag{20.24}$$

Finally we recast the Boltzmann equation in a form that makes our intuitive Gamov criterion

[3] Some formulas from equilibrium statistics are collected in the Appendix 14.A.

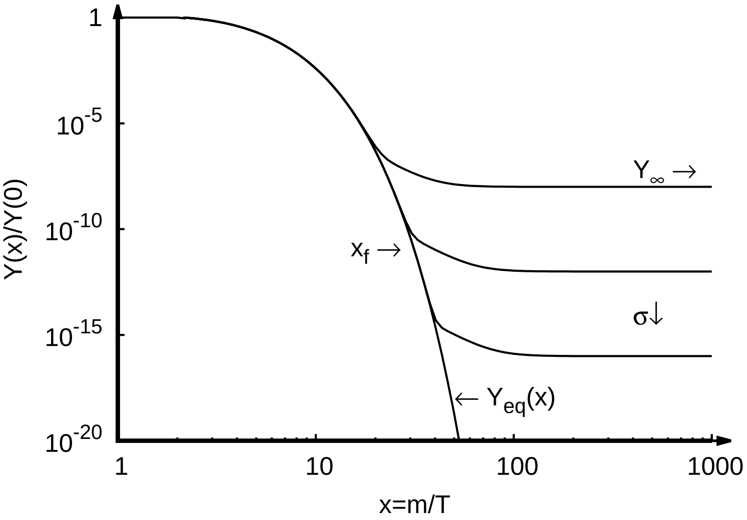


Figure 20.1.: Illustration of the freeze-out process: For $x \gtrsim x_\mathrm{f}$, the abundance Y becomes constant, $T \to Y_\infty$; for increasing $\langle\sigma_{ann}v\rangle$, the final abundance Y_∞ decreases.

explicit,

$$\frac{x}{Y_\mathrm{eq}}\frac{\mathrm{d}Y}{\mathrm{d}x} = -\frac{\Gamma_A}{H}\left[\left(\frac{Y}{Y_\mathrm{eq}}\right)^2 - 1\right] \tag{20.25}$$

with $\Gamma_A = n_\mathrm{eq}\langle\sigma_\mathrm{ann}v\rangle$: The relative change of Y is controlled by the factor Γ_A/H times the deviation from equilibrium. The evolution of $Y = n_X/s$ is shown schematically in Fig. 20.1: As the universe expands and cools down, n_X decreases at least as a^{-3}. Therefore, the annihilation rate quenches and is not longer sufficiently large to keep the particle in chemical equilibrium. As a result, the abundance freezes-out, i.e. the ratio n_X/s stays constant. For the discussion of approximate solutions to this equation, it is convenient to distinguish according to the freeze-out temperature hot dark matter (HDM) with $x_\mathrm{f} \ll 3$, cold dark matter (CDM) with $x_\mathrm{f} \gg 3$ and the intermediate case of warm dark matter with $x_\mathrm{f} \sim 3$.

Abundance of hot dark matter For $x_\mathrm{f} \ll 3$, freeze-out occurs when the particle is still relativistic and Y_{eq} is not changing with time. Thus the asymptotic value of the abundance, $Y(x \to \infty) \equiv Y_\infty$, is given by the equilibrium value at freeze-out,

$$Y_\infty = Y_{eq}(x_\mathrm{f}) = 0.278\,\frac{g_{eff}}{g_{*S}}\,, \tag{20.26}$$

and the only temperature-dependence is contained in g_{*S}. The number density today is then

$$n_0 = s_0 Y_\infty = 2970\,Y_\infty cm^{-3} = 825\,\frac{g_{eff}}{g_{*S}}\,cm^{-3}\,, \tag{20.27}$$

where we used for the present entropy density $s_0 \simeq 2891/\mathrm{cm}^3$ derived in problem 15.??. Although a HDM particle was relativistic at freeze-out, it is today non-relativistic if its mass is $m \gg 3\mathrm{K} \simeq 0.2\mathrm{meV}$. In this case its energy density is simply $\rho_0 = ms_0Y_\infty$ and its abundance $\Omega h^2 = \rho_0/\rho_{cr}$ is given by

$$\Omega h^2 = 7.8 \times 10^{-2}\,\frac{m}{\mathrm{e}V}\,\frac{g_{eff}}{g_{*S}}\,. \tag{20.28}$$

Hence the abundance of HDM particles with mass $m \gtrsim 10\,\mathrm{eV}$ exceeds the observed total abundance of matter, $\Omega_m h^2 \simeq 0.1$.

Abundance of cold dark matter For CDM with $x_\mathrm{f} \ll 3$, freeze-out occurs when the particles are already non-relativistic and Y_eq is exponentially changing with time. Thus the main problem is to find x_f, for late times we use $Y_\infty \approx Y(x_\mathrm{f})$, i.e. the equilibrium value at freeze-out. We parametrise the temperature-dependence of the annihilation cross section as $\langle\sigma_{\mathrm{ann}}v\rangle = \sigma_0(T/m)^n = \sigma_0/x^n$ which corresponds to an expansion in $v^{2n}_{\mathrm{M}\o l}$. For simplicity, we consider only the case of s-wave annihilation for CDM, $n = 0$. Then the Gamov criterion becomes with $H = 1.66\sqrt{g_*}\,T^2/M_{\mathrm{P}l}$ and $\Gamma_A = n_\mathrm{eq}\langle\sigma_{\mathrm{ann}}v\rangle$

$$g\left(\frac{mT_\mathrm{f}}{2\pi}\right)^{3/2}\exp(-m/T_\mathrm{f})\,\sigma_0 = 1.66\sqrt{g_*}\,\frac{T_\mathrm{f}^2}{M_{\mathrm{P}l}} \tag{20.29}$$

or

$$x_\mathrm{f}^{-1/2}\exp(x_\mathrm{f}) = 0.038\frac{g}{\sqrt{g_*}}M_{\mathrm{P}l}m\sigma_0 \equiv C\,. \tag{20.30}$$

To obtain an approximate solution, we neglect first in

$$\ln C = -\frac{1}{2}\ln x_\mathrm{f} + x_\mathrm{f} \tag{20.31}$$

the slowly varying term $\ln x_\mathrm{f}$. Inserting next $x_\mathrm{f} \approx \ln C$ into Eq. (20.31) to improve the approximation gives then

$$x_\mathrm{f} = \ln C + \frac{1}{2}\ln(\ln C)\,. \tag{20.32}$$

For a DM particle with thermal annihilation cross section $\sigma_0 = 3\times 10^{-26}cm^3/\mathrm{s}$ the freeze-out temperature changes slowly from $x_\mathrm{f} \simeq 23$ for $m = 10\,\mathrm{GeV}$ to $x_\mathrm{f} \simeq 28$ for $m = 1\,\mathrm{TeV}$. The relic abundance for CDM follows from $n(x_\mathrm{f}) = 1.66\sqrt{g_*}\,T_\mathrm{f}^2/(\sigma_0 M_{\mathrm{P}l})$ and $n_0 = n(x_\mathrm{f})[R(x_\mathrm{f})/R_0]^3 = n(x_\mathrm{f})[g_{*,f}/g_{*,0}][T_0/T(x_\mathrm{f})]^3$ as

$$\rho_0 = mn_0 \approx 10\frac{x_\mathrm{f}T_0^3}{\sqrt{g_{*,f}}\sigma_0 M_{\mathrm{P}l}} \tag{20.33}$$

or

$$\Omega_{CDM}h^2 = \frac{mn_0}{\rho_{cr}} \approx \frac{1\times 10^{-28}cm^3/s}{\sigma_0}\,x_\mathrm{f}\,. \tag{20.34}$$

Thus the abundance of a CDM particle is inverse proportionally to its annihilation cross section, since a more strongly interacting particle stays longer in equilibrium. Note that the explicit dependence on the freeze-out temperature T_f in the factors $n(x_\mathrm{f})$ and $[a(x_\mathrm{f})/a_0]^3$ cancelled, leading only to an implicit dependence of the abundance via $g_{*,\mathrm{f}}$ on T_f. Moreover, the abundance depends only logarithmically on the mass m via Eq. (20.32).

The observed value $\Omega_{CDM}h^2 \simeq 0.1$ implies $\sigma_0 \simeq 3\times 10^{-26}cm^3/\mathrm{s} = 1\times 10^{-36}cm^2$. Such cross sections are typical for a weakly interacting massive particles, and thus such CDM particles are called WIMPs. The numerical coincidence of the annihilation cross section of a thermal relic with a typical weak interaction cross section in the SM has been called by *aficionados* of this scenario WIMP miracle.

Unitarity limit for m_X We saw that the abundance of a CDM particle scales approximately inversely with its annihilation cross section. Therefore the observed abundance of DM implies a lower bound on $\langle\sigma v\rangle$. We show in this section that the unitarity of the S-matrix implies $\langle\sigma v\rangle \propto 1/m_X^2$ and derive thereby an upper bound on its mass m_X.

We repeat the steps from Eq. (**??**) to (**??**) which lead to the unitary bound on partial waves amplitudes, but take into account that the final states differs from the initial one, $p_{cms} \neq p'_{cms}$. Comparing

$$\sigma_{\rm tot} = \frac{16\pi}{s}\frac{p'_{cms}}{p_{cms}}\sum_{l=0}^{\infty}(2l+1)|T_l|^2 \equiv \sum_{l=0}^{\infty}\sigma_l\,, \tag{20.35}$$

to the result using the optical theorem (**??**) gives then the unitarity constraint

$$\frac{2p'_{cms}}{\sqrt{s}}\,|T_l| \leq 1\,. \tag{20.36}$$

Inserting this constraint for T_l bounds the contribution of the l.th partial wave to the total cross section σ_l as

$$\sigma_l = \frac{4\pi}{p_{cms}^2}(2l+1)\,. \tag{20.37}$$

For the non-relativistic scattering of two particles with mass m_X, we can use $p_{cms} \simeq v_{\rm M\o l}m_X/2$, giving as unitarity limit

$$\sigma_l v_{\rm M\o l} \leq (2l+1)\frac{16\pi}{v_{\rm M\o l}m_X^2}\,. \tag{20.38}$$

A bound stronger by a factor four applies to the annihilation cross section, $\sigma_{ann} = \sigma_{tot} - \sigma_{el}$. In the case of particles with spin, we have to divide this result by the number of spin degrees of freedom of the initial particles [GK90, Hui01]. For a CDM particle, only the $l = 0$ and $l = 1$ partial waves give a sizeable contribution to the total annihilation cross section. Using then $\Omega_X h^2 \leq \Omega_{CDM}h^2 = 0.11$ and $v_{\rm M\o l}^2 = x_{\rm f}/6$ and $x_{\rm f} \approx 30$ we obtain an upper limit of $m_X \lesssim 20\,\text{TeV}$ for any stable particle that was once in thermal equilibrium.

20.2.2. Annihilation cross section in the nonrelativistic limit

Since WIMPs freeze-out typically at $x_{\rm f} \sim 20$, it is useful to consider the WIMP annihilation cross section in the nonrelativistic limit $x \gg 1$. This is even more true for annihilations today in the Milky Way where WIMPs have velocities of order $v \sim 220\,\mathrm{k}m/s \sim 10^{-3}$. An expansion in $x \propto v^2$ corresponds to a partial wave expansion. Therefore we can either project out from the general Feynman amplitude the contribution to the first partial waves $l = 0, 1, \ldots$, or we can perform an expansion in v^2 of the thermally averaged cross section. Such an expansion contains even powers of the Møller velocity,

$$\langle\sigma_{\rm ann}v\rangle = \sigma_{\rm ann}^{(s)} + \sigma_{\rm ann}^{(l)}\langle v\rangle^2 + \sigma_{\rm ann}^{(p)}\langle v\rangle^4 + \ldots\,, \tag{20.39}$$

assuming that the Feynman amplitude is not singular in the limit $v \to 0$. The occurrence of singularities signals that the initial state can form a bound state: For instance, at low energies, the process $e^+e^- \to n\gamma$ should be not described as the annihilation of a free electron and a positron but as the annihilation of its bound-state positronium. More generally, the small relative velocities of CDM particles, $\beta = v \ll 1$, means that factors g^2/β or $\ln(g^2/\beta)$ can lead to a break-down of perturbation theory. This effect was first studied by Sommerfeld for Coulomb interactions, and the resulting boost of reaction rates is therefore called Sommerfeld enhancement.

The enhancement of the annihilation cross section can be calculated non-relativistically and is, neglecting bound-state effects, characterised by two parameters: The mass ratio $\varepsilon = M/m_X$ of the exchange and the DM particle determines, if the annihilation proceeds in the Coulomb ($\varepsilon \ll 1$) or in the Yukawa ($\varepsilon \gg 1$) regime, while the ratio $x = g_{\rm eff}^2/\beta$ of the squared effective coupling constant and the velocity determines, if factors $g_{\rm eff}^2/\beta$ or $\ln(g_{\rm eff}^2/\beta)$ lead to a break-down of perturbation theory. Here, the effective coupling constant g_{eff} includes all prefactors in front of the Yukawa potential, as e.g. mixing matrix elements. In the Coulomb case, the Sommerfeld factor $\mathcal{R}$ as ratio of the perturbative and non-perturbative annihilation cross section is given by

$$\mathcal{R} = \frac{\sigma_{\rm np}}{\sigma_{pert}} \sim \frac{\eta}{1-\exp(-\eta)} \tag{20.40}$$

with $\eta = \pm g_{\rm eff}^2/(2\beta)$ [LL81].

Projection operators and the nonrelativistic limit We give now an example how one can obtain the nonrelativistic expansion of a Feynman amplitude by projecting out the the partial waves $l = 0, 1, \ldots$. In many models, the DM particle is a Majorana fermion which leads to some particular features which the example will also illustrate.

We use spinors in the Dirac representation (8.51) to describe a Majorana fermion pair at rest. The adjoint anti-particle spinors are $\bar{v}(m,-) = \sqrt{2m}(0,0,0,-1)$ and $\bar{v}(m,+) = \sqrt{2m}(0,0,1,0)$. Since we have a pair of indistinguishable fermions in the initial state, we have to antisymmetrise the initial state. We compute the antisymmetrised two-particle state $u(m,s_1)\bar{v}(m,s_2) - u(m,s_2)\bar{v}(m,s_1)$ for different spin configurations: If the two spins are parallel, then the result is zero,

$$u(m,-)\bar{v}(m,-) - u(m,-)\bar{v}(m,-) = u(m,+)\bar{v}(m,+) - u(m,+)\bar{v}(m,+) = 0\,. \tag{20.41}$$

For anti-parallel spins, we obtain

$$u(m,+)\bar{v}(m,-) - u(m,-)\bar{v}(m,+) = 2m\begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \equiv \psi_1 \tag{20.42}$$

and

$$u(m,-)\bar{v}(m,+) - u(m,+)\bar{v}(m,-) = -2m\begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \equiv \psi_2\,. \tag{20.43}$$

The two states ψ_1 and ψ_2 are linearly dependent and we can combine them into

$$\Phi = \frac{1}{\sqrt{2}}(\psi_1 - \psi_2) = 2\sqrt{2}m\begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}\,. \tag{20.44}$$

Next we want to rewrite the expression for Φ which is valid in the rest-frame of the two Majorana fermions into a Lorentz invariant way. We express first $\left(\begin{smallmatrix} 0 & 1 \\ 0 & 0 \end{smallmatrix}\right)$ by gamma matrices,

$$\Phi = 2\sqrt{2}m\,\frac{1}{2}(1+\gamma^0)\gamma^5 = \sqrt{2}(m + \not{P}/2)\gamma^5\,. \tag{20.45}$$

In the second step we introduced the total momentum $P = (p_1 + p_2) = (2m, \mathbf{0})$ of the fermion pair and replaced $m\gamma^0$ by $\not{P}/2$.

Let us now illustrate the usefulness of this method with a concrete example. We consider the annihilation of two Majorana fermions χ with mass m_χ into a fermion pair $f\bar{f}$ with mass m_f via the exchange of a scalar $\tilde{f}_L$ with mass M. Their interaction is given by

$$\mathscr{L} = g_L \bar{\chi} P_L f \tilde{f}_L + \text{h.c.} = g_L \bar{\chi} P_L f \tilde{f}_L + g_L \bar{f} P_R \chi \tilde{f}_L \,. \tag{20.46}$$

For simplicity, we consider the limit that the mass M of the exchanged scalar is much larger than the DM mass, $M \gg m_\chi$. Then the interaction becomes an effective four-fermion interaction, and the Feynman amplitude simplifies to

$$\begin{aligned}\mathcal{A} &= \frac{g_L^2}{M^2} \bar{v}_{\bar{f}}(p_4) P_R [u_\chi(m_\chi, s_1)\bar{v}_\chi(m_\chi, s_2) - u_\chi(m_\chi, s_2)\bar{v}_\chi(m_\chi, s_1)] P_L u_f(p_3) &\text{(20.47a)}\\ &= \frac{\sqrt{2} g_L^2}{M^2} \bar{v}_{\bar{f}}(p_4) P_R (m_\chi + \not{P}/2)\gamma^5 P_L u_f(p_3) \,. &\text{(20.47b)}\end{aligned}$$

The m_χ term vanishes because of $P_L P_R = 0$. Using $P = p_3 + p_4$ and the Dirac equation, we obtain $\not{p}_3 + \not{p}_4 = -2m_f$ and thus

$$\mathcal{A} = -\frac{\sqrt{2} m_f g_L^2}{M^2} \bar{v}_{\bar{f}}(p_4)\gamma^5 P_L u_f(p_3) \,. \tag{20.48}$$

Thus the amplitude is proportional to m_f and the annihilation into light fermions is strongly suppressed. Additionally to getting this insight with little effort, the remaining amplitude is easier to evaluate and no expansion in $v_{\text{Møl}}$ has to be performed.

Let us now try to understand why the amplitude is proportional to m_f: The initial wave-function $|\Phi\rangle = |L, S\rangle$ of the identical fermion pair has to be antisymmetric. For zero relative velocity, the orbital angular momentum L is zero and thus $|L\rangle$ is symmetric. Therefore the spin wave-function $|S\rangle$ has to be anti-symmetric, $|S\rangle = |\uparrow, \downarrow\rangle$ or $|S\rangle = |\downarrow, \uparrow\rangle$. Thus the total spin is zero, and the pair of Majorana fermions is in a ${}^1S_0$ state. Consequently, $\Phi = \sqrt{2}(m + \not{P}/2)\gamma^5$ acts a projection operator which inserted between two arbitrary spinors extracts the ${}^1S_0$ state or the s-wave contribution to the annihilation amplitude. If the produced fermion pair were massless, the final state would be either $f_L \bar{f}_R$ or $f_R \bar{f}_L$: For instance, the Majorana fermion pair could annihilate into a left-handed electron and a right-handed positron. The electron-positron pair is produced back-to-back and therefore its total spin is $S = 1$. Since the Dirac mass term connects left- and right-chiral fields, the chirality flip required to create a $S = 0$ state leads to an amplitude proportional to m_f. Thus for massless fermions, the annihilation cross section has to vanish for $v_{\text{Møl}} \to 0$. Alternatively, we can allow for a non-zero relative velocity and consider that the orbital angular momentum is one. Thus $|L\rangle$ is anti-symmetric, and from the two states ${}^3P_0$ and ${}^3P_1$ only the latter is allowed. Now it is possible to produce the fermion pair without a helicity flip, but the amplitude will be proportional to $v_{\text{Møl}}$.

20.2.3. Detection of WIMPs

Searches for WIMPs rely on the assumption that they interact with SM particles. Such searches can be divided into three main categories: Indirect detection uses the annihilation process $X + X \to \text{S}M + SM$ into SM particles, while direct detection relies on the elastic scattering $X + \text{S}M \to X + \text{S}M$ of WIMPs on SM particles. Finally, one can search at accelerators for the production $\text{S}M + SM \to X + X$ of WIMPs by colliding normal matter.

Cosmology connects the measured CDM abundance with the required annihilation cross section of a thermal relic as $\langle\sigma_{ann}v\rangle \simeq 3\times10^{-26}\,\mathrm{cm}^3/\mathrm{s}$. Moreover, we know that the three processes used in the different search categories are related by crossing symmetry. This leads to the question how well we can constrain the possible signal strength to be expected in the WIMP scenario in these three channels. Let us consider first indirect detection where one can use the annihilation of WIMPs in the center of the Sun, the halo of the Milky Way or other galaxies. In all cases, typical WIMP velocities are much smaller than the ones during freeze-out. For instance, we can use as typical WIMP velocity relevant for the annihilation of DM in the Milky Way the rotation velocity of the Sun around the center of the Galaxy, $v\sim 220\,\mathrm{k}m/s\sim 10^{-3}$. As a result, the thermal annihilation cross section today is—in the absence of non-perturbative effects—only bounded from above by the one in the early universe: If annihilations are dominated by the p-wave contribution, the thermal cross section $\langle\sigma_{ann}v\rangle$ relevant for indirect searches could be six orders of magnitude smaller than in the early universe. On the other hand, the annihilation cross section at small velocities may be enhanced via the Sommerfeld effect compared to the cosmological one.

The connection is even less tight in the case of direct and accelerator searches: In these searches we test mainly the couplings of the WIMP to the first generation of quarks and leptons, while the annihilation cross section sums up all relevant channels. In many models, the main final states of WIMP annihilations are heavy fermions, gauge and higgs bosons. As a result, the elastic cross section on nucleons or the production via proton-proton scattering could be suppressed. Moreover, WIMPs are produced at accelerators as ultrarelativistic particles, probing again a different kinematical regime than annihilations in the early universe.

Direct detection A direct detection experiment aims to measure the nuclear recoil, when a WIMP scatters on a nucleus in a detector. Let us assume that the WIMP interacts via the exchange of a gauge or Higgs boson with the nucleus. The momentum transfer in such a reaction is small, $q^2\lesssim 100\,\mathrm{keV}^2$, cf. problem 21.??. Therefore, the exchanged virtual particles do not resolve the quark and gluon content of a nucleon, but interact with the whole nucleon[4]: Instead of the couplings to quarks, we have therefore to know the effective coupling of a gauge or Higgs boson to a nucleon. The small moment transfer implies also that one can integrate out the intermediate virtual particles, constructing an effective Lagrangian for the interactions between the DM particle and quarks and gluons. Within a given DM model, this step is lengthy but straight-forward and gives an effective Lagrangian containing higher-dimensional operators.

Let us illustrate this procedure with an example. We assume that the WIMP is a fermion and write down its interactions with light quarks and gluons. For our purposes, it is sufficient to consider as an example the contribution of scalar operators to the effective Lagrangian,

$$\mathscr{L}_{eff} = C_S^q\bar{X}Xm_q\bar{q}q + C_S^g\frac{\alpha_s}{\pi}\bar{X}XF^a_{\mu\nu}F^{a\mu\nu}\,. \tag{20.49}$$

All the information about the physics integrated out is contained in the coefficients C_S^N. We obtain the matrix elements of these effective operators between nucleon states following the same strategy as in section 17.1 discussing the trace anomaly: We use Eq. (17.18) at leading

[4] Since the moment transfer is comparable to the nuclear size, they interact (partly) coherently with the whole nucleus. Thus in an additional step, nuclear physics effects have to be incorporated.

order in α_s,

$$T_\mu^\mu = -\frac{9}{8}\frac{\alpha_s}{\pi}F^a_{\mu\nu}F^{a\mu\nu} + \sum_{q=u,d,s} m_q\bar{q}q\,, \tag{20.50}$$

together with the mass fractions $f_q^{(N)}$ calculated in lattice QCD. Evaluating then T_μ^μ between nucleon states $|N\rangle$ and using $\langle N|T^\mu_{\ \mu}|N\rangle = m_N$, we obtain

$$\langle N|\frac{\alpha_s}{\pi}F^a_{\mu\nu}F^{a\mu\nu}|N\rangle = -\frac{8}{9}m_N f_G^{(N)} \tag{20.51}$$

with $f_G^{(N)} \equiv 1 - \sum_{q=u,d,s} f_q^{(N)}$. This determines the effective operators at the scale $q^2 \simeq m_N$ and allows one to calculate scattering rates for given coefficients C_S^N. The effective Lagrangian (20.49) is however defined at the scale corresponding to the mass scale of the virtual particles integrated out. Thus as a final step, one has to derive the RGE of these effective operators and to evolve the coefficients C_S^N down to the scale m_N.

Indirect detection The average density of DM in the Galaxy is increased by a factor $\sim 10^5$ compared to the extragalactic space. Therefore the annihilation rate of DM can become again appreciable inside the Milky Way, and in particular in regions where DM is strongly accumulated. The secondaries of DM annihilations will be the stable particle of the standard model, i.e. photons, neutrinos, electrons and protons. The challenge for the indirect detection consists to disentangle these annihilation products from the background of high-energy particles produced by astrophysical sources. Two channels provide a rather unique signature: DM annihilations into two photons, or a photon and Z, lead to line features which cannot mimicked by a astrophysical background. Since a WIMP is by assumption neutral, this process can proceed however only via loop graphs and it thus suppressed by a factor $(\alpha/\pi)^2$.

Another unique signature are high-energy neutrinos produced by WIMPs that accumulate e.g. in the Sun. Here, the directional signal together with the fact that neutrinos produced by astrophysical processes have energies $\lesssim$ GeV provides the distinctive signature. In all other cases, a detailed knowledge of the spectral shape of anti-matter fluxes, both for the background produced e.g. in pp collisions and in DM annihilations is required. This requires the knowledge of strong interactions at small virtualities and relies on the use of Monte Carlo methods, as described in section 17.2.

20.3. Big bang nucleosynthesis

Big bang nucleosynthesis (BBN) is controlled mainly by two parameters: The mass difference between protons and neutrons, $\Delta \equiv m_n - m_p \simeq 1.29\,\mathrm{MeV}$, and the freeze-out temperature T_f of reactions converting protons into neutrons and vice versa. Since the binding energy per nucleon has a large peak for ^{4}He, essentially all free neutrons are bound into helium. Heavier elements (except ^{7}Li) are produced later by stellar fusion, where the densities are sufficiently high that e.g. the triple α process $3\,^4He \to^{12}$C can bridge the gap of missing tightly bound nuclei between ^{4}He and ^{12}C.

Equilibrium distributions In the non-relativistic limit $T \ll m$, the number density of a nuclear species with mass number A and charge Z is given by

$$n_A = g_A \left(\frac{m_A T}{2\pi}\right)^{3/2} \exp[\beta(\mu_A - m_A)] \,. \tag{20.52}$$

In chemical equilibrium, $\mu_A = Z\mu_p + (A - Z)\mu_n$ and we can eliminate μ_A by inserting the equivalent expression of (20.52) for protons and neutrons,

$$\mathrm{e}^{\beta\mu_A} = \exp[\beta(Z\mu_p + (A-Z)\mu_n)] = \frac{n_p^Z n_n^{A-Z}}{2^A}\left(\frac{2\pi}{m_N T}\right)^{3A/2} \exp[\beta(Zm_p + (A-Z)m_n)] \,. \tag{20.53}$$

Here and in the following we can set in the prefactors $m_p \simeq m_n \simeq m_N$ and $m_A \simeq Am_N$, keeping the exact masses only in the exponentials. Inserting this expression for $\exp(\beta\mu_A)$ together with the definition of the binding energy of a nucleus, $B_A = Zm_p + (A-Z)m_n - m_A$, we obtain

$$n_A = g_A \left(\frac{2\pi}{m_N T}\right)^{3(A-1)/2} \frac{A^{3/2}}{2^A}\, n_p^Z n_n^{A-Z} \exp(\beta B_A) \,. \tag{20.54}$$

The mass fraction X_A contributed by a nuclear species is

$$X_A = \frac{An_A}{n_B} \qquad \text{with} \qquad n_B = n_p + n_n + \sum_i A_i n_{A_i} \qquad \text{and} \qquad \sum_i X_i = 1 \,. \tag{20.55}$$

Next we introduce the baryon-photon ratio $\eta = n_B/n_\gamma$ as variable. With $n_p^Z n_n^{A-Z}/n_N = X_p^Z X_n^{A-Z} n_N^{A-1}$ and $\eta \propto T^3$ and thus $n_B^{A-1} \propto \eta^{A-1} T^{3(A-1)}$, we have

$$X_A \propto \left(\frac{T}{m_N}\right)^{3(A-1)/2} \eta^{A-1} X_p^Z X_n^{A-Z} \exp(\beta B_A) \,. \tag{20.56}$$

The fact that $\eta \simeq 6 \times 10^{-10} \ll 1$, i.e. that the number of photons per baryon is extremely large, means that nuclei with $A > 1$ are much less abundant and that nucleosynthesis takes place later than naively expected.

Let us consider the particular case of deuterium in Eq. (20.56),

$$\frac{X_D}{X_p X_n} = \frac{24\zeta(3)}{\sqrt{\pi}} \left(\frac{T}{m_N}\right)^{3/2} \eta \, \exp(\beta B_D) \tag{20.57}$$

with $B_D = 2.23$ MeV. The start of nucleosynthesis could be defined approximately by the condition $X_D/(X_p X_n) = 1$, or $T_{\rm NS} \simeq 0.07\,\text{MeV}$ according to the left panel in Fig. 20.2. The right panel of the same figure shows the results, if the equations (20.56) together with $\sum_i X_i = 1$ are solved for the lightest and stablest nuclei: In thermal equilibrium, essentially all free neutrons will bind to ^{4}He at temperatures $T \lesssim 0.2\,\text{MeV}$. The formation of heavier elements is strongly suppressed because of their much smaller binding energy per nucleon. Moreover, the Coulomb barrier will prevent the production of nuclei with $Z \gg 1$.

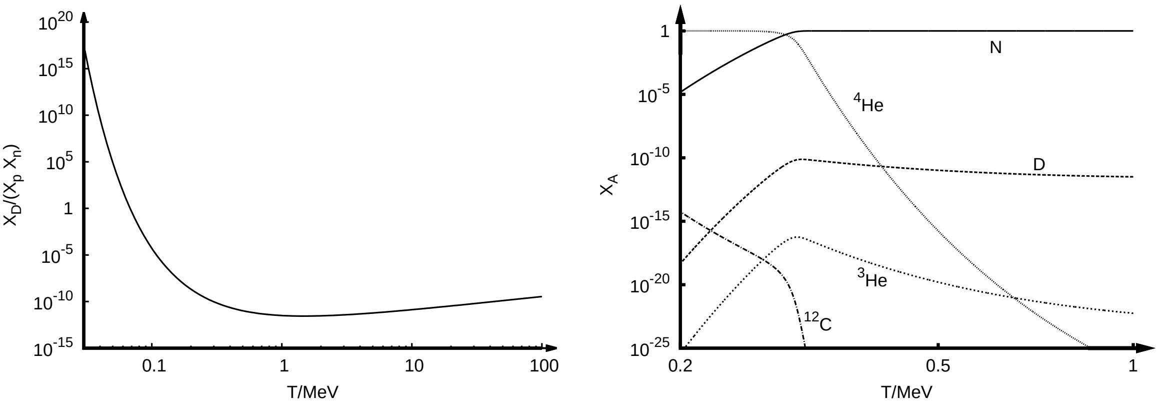


Figure 20.2.: Relative equilibrium abundance $X_D/(X_pX_n)$ of deuterium as function of temperature T (left) and equilibrium mass fractions of nucleons, D, ^{3}He, ^{2}He and ^{12}C (right).

Neutron abundance If we do not aim at the calculation of the (small) abundance of elements other than ^{4}He, we have to solve only the Boltzmann equation for the neutron abundance. Nucleons are inter-converted by weak processes as $n \leftrightarrow p + e^- + \nu_e$. For an estimate of the freeze-out temperature of weak interactions, we can use the Gamov criteria: The cross section of processes like $n \leftrightarrow p + e^- + \nu_e$ or $e^+e^- \leftrightarrow \bar{\nu}\nu$ is $\sigma \approx G_F^2 E^2$. If we approximate the energy of all particle species by their temperature T, their velocity by c and their density by $n \approx T^3$, then the interaction rate of weak processes is

$$\Gamma \approx \langle v\sigma n_\nu \rangle \approx G_F^2 T^5 \,. \tag{20.58}$$

In the radiation-dominated epoch, $\Gamma(T_{\rm fr}) = H(T_{\rm fr})$ gives as freeze-out temperature $T_{\rm fr}$ of weak processes

$$T_{\rm fr} \approx \left(\frac{1.66\sqrt{g_*}}{G_F^2 M_{\rm Pl}} \right)^{1/3} \approx 1{\rm M}eV \tag{20.59}$$

with $g_* = 10.75$. Thus weak interaction freeze out before nucleosynthesis starts. Because of $T_{\rm fr} \approx 1$ MeV, we can treat nucleons in the non-relativistic limit. Then their relative equilibrium abundance is given by the Boltzmann factor $\exp\left(-\Delta/T\right)$ for $T \gtrsim T_{\rm fr}$. Hence at freeze-out, the ratio of their equilibrium distributions is given by

$$\frac{n_n^{\rm eq}}{n_p^{\rm eq}} = \exp\left(-\frac{\Delta}{T_{\rm fr}}\right) \,. \tag{20.60}$$

We consider in Eq. (20.13) the reaction $n + e^- \leftrightarrow p + \nu_e$, using for the leptons equilibrium distributions,

$$\frac{1}{a^3}\frac{{\rm d}(a^3 n_n)}{{\rm d}t} = n_l^{\rm eq}\langle\sigma v\rangle \left[\frac{n_p n_n^{\rm eq}}{n_p^{\rm eq}} - n_n \right] \,. \tag{20.61}$$

Next we replace the ratio n_n^{eq}/n_p^{eq} by the Boltzmann factor $\exp(-\beta\Delta)$. Moreover, the loss term $n_l^{\rm eq} n_n \langle\sigma v\rangle$ equals the neutron-proton scattering rate Γ_{np}. Changing also from the variable n_n to the neutron mass fraction X_n, we arrive at

$$\frac{{\rm d}X_n}{{\rm d}x} = \Gamma_{np}\left[(1 - X_n){\rm e}^{-\beta\Delta} - X_n\right] \,. \tag{20.62}$$

Introducing again the dimensionless $x = \Delta/T$ as evolution variable, we obtain

$$\frac{\mathrm{d}X_n}{\mathrm{d}x} = \frac{x\Gamma_{np}}{H(1)}\left[\mathrm{e}^{-x} - X_n(1+\mathrm{e}^{-x})\right] \tag{20.63}$$

with $H(1) \simeq 1.13/\mathrm{s}$. The neutron-proton conversion rate can be connected to the neutron lifetime $\tau_n \simeq 886.7\mathrm{s}$ as (problem 21.**??**),

$$\Gamma_{np} = \frac{255}{\tau_n x^5}(12 + 6x + x^2)\,. \tag{20.64}$$

Now Eq. (20.63) can be integrated and the result shows that X_n freezes-out below 0.5 MeV, approaching the asymptotic value $X_n \simeq 0.15$.

Up-to now we have treated the neutron as stable. We can include into our simple picture the effect of neutron decays by adding the factor $\exp(-t/\tau_n)$ to the neutron abundance X_n. If we use that nucleosynthesis starts at $T_{\mathrm{NS}} \simeq 0.07\,\mathrm{MeV}$, $t_{\mathrm{NS}} \simeq 270\,\mathrm{s}$, then $\exp(-t/\tau_n) \simeq 0.74$. As result, the mass fraction of neutrons which can be fused into helium is $X_n = 0.11$ and thus the helium fraction $X_4 = 2X_n = 0.22$. Numerical calculations that include the full nuclear reaction network lead to $X_4 = 0.24$, and thus our simple estimate is only 10% away from the true value. A fit to these numerical results shows that the helium abundance depends logarithmically on η_b,

$$X_4 = 0.226 + 0.013\ln(\eta_b/10^{-10})\,. \tag{20.65}$$

Thus the helium abundance alone can be used to determine the baryon-photon ratio as $\eta_b \simeq few \times 10^{-10}$. A comparison of the predicted with the observed abundance of deuterium and lithium allows then for a consistency check of the BBN picture. An independent determination of Ω_b using CMB observations leads to $\eta_b \simeq 6.2 \times 10^{-10}$. The success of BBN can be used to limit e.g. the injection of high-energy through particle decays which could destroy light elements.

Summary

Boltzmann equations are an important tool to describe processes as diverse as the evolution of the DM density, BBN nucleosynthesis or recombination. The Gamov criterion states that processes freeze out when their rate becomes smaller than the Hubble rate. The mass of any thermal relic is bounded by $\lesssim 20\,\mathrm{TeV}$. The abundance of a CDM particle with $\langle\sigma v\rangle \simeq 3\times 10^{-26}\mathrm{cm}^3/\mathrm{s}$ corresponds to the observed one, $\Omega_{CDM} = 0.2$. BBN explains successfully the abundance of light elements like D and ^{4}He, and fixes thereby also η_b.

Further reading

A classic presentation of the freeze-out mechanism is given in [KT94]. A more detailed analytical treatment of BBN is presented by [Muk05] and [GR11b], who also discuss recombination. Dropping the assumption of uniformity, Boltzmann type equations can be also used to describe the evolution of perturbations, a topic which is treated in detail e.g. by [Dod03].

21. Baryogenesis

We have seen that BBN determines the baryon-photon ratio as $\eta = n_B/n_\gamma \simeq 6 \times 10^{-10}$. In problem 21.**??**, we calculated the baryon abundance in the usual freeze-out formalism for zero chemical potential. Since the $\bar{p}p$ annihilation cross section is large, nucleons freeze-out very late ($x_\mathrm{f} \sim 44$) when their density is already strongly suppressed. As a result, the baryon-photon ratio in a baryon symmetric world would be $\eta \simeq 7 \times 10^{-20}$, i.e. much smaller than the observed value. This implies that at temperatures above the freeze-out a tiny surplus of one quark per 10^{10} quarks and antiquarks existed. Therefore the proper definition of the baryon-photon ratio is $\eta = (n_b - n_{\bar{b}})/n_\gamma$ and the challenge is to explain the origin of the asymmetry between baryons and antibaryons. Astronomers do not observe the photons from the processes $e^+e^- \to 2\gamma$ and $\bar{p}p \to X\gamma$ that would occur at the boundaries of matter-antimatter domains. Using these observational limits, one can conclude that the whole observable universe consists of matter. Moreover, an inflationary period in the early universe eliminates any pre-existing baryon asymmetry, forcing us to explain the observed baryon asymmetry dynamically. We will see that such explanations require necessarily physics beyond the SM.

21.1. Sakharov conditions and the SM

Sakharov conditions for baryogenesis Sakharov developed in 1967 the first model which contained the three ingredients necessary for the dynamically generation of a non-zero baryon number. These so-called Sakharov conditions for baryogenesis are

1. violation of baryon number B,
2. violation of the discrete symmetries C and CP, and
3. departure from thermal equilibrium.

The first condition, the non-conservation of baryon number B, is obviously necessary, if the universe should evolve from a state with $n_B = 0$ to $n_B > 0$. We can understand the second condition from the transformation properties of the baryon number operator B under C and CP: Because of $\mathrm{C}B\mathrm{C}^{-1} = -B$ and $\mathrm{CP}B(\mathrm{CP})^{-1} = -B$, the (thermal) expectation value of B has to vanish, if C and CP are symmetries of the model considered. More explicitly, we can express the expectation value of B in this case with $A = \{\mathrm{C}, \mathrm{CP}\}$ as

$$\langle B \rangle = Z^{-1}\mathrm{Tr}\left[\mathrm{e}^{-\beta H} B\right] = Z^{-1}\mathrm{Tr}\left[AA^{-1}\mathrm{e}^{-\beta H} B\right] \tag{21.1a}$$

$$= Z^{-1}\mathrm{Tr}\left[\mathrm{e}^{-\beta H} A^{-1}BA\right] = -Z^{-1}\mathrm{Tr}\left[\mathrm{e}^{-\beta H} B\right] = -\langle B \rangle\,, \tag{21.1b}$$

where we used $[H, A] = 0$ going from the first to second line. Thus $\langle B \rangle = 0$, if $[H, \mathrm{C}] = 0$ or $[H, \mathrm{CP}] = 0$. By the same token, we can show that a departure from thermal equilibrium is required: Since any unitary Lorentz invariant quantum field theory is invariant under CPT, we have $[H, \mathrm{CPT}] = 0$ but also $\mathrm{CPT}B(\mathrm{CPT})^{-1} = -B$. We can avoid the conclusion $\langle B \rangle = 0$ only, if Eq. (21.1a), or in other words thermal equilibrium, does not hold.

CP violation Another consequence of CPT invariance is that CP violation is equivalent to T violation. The latter is an anti-unitary operator, $\mathrm{T}^{-1}\mathrm{iT} = -\mathrm{i}$, and therefore complex parameters in the Lagrangian lead to T and thus CP violation. The only complex parameters contained in the SM are the phases of the fermion mixing matrices. In order to see that such phases result indeed in CP violation, we express the charged current interaction for the example of leptons by mass eigenstates,

$$\mathscr{L}_{\mathrm{CC}} = \frac{g}{\sqrt{2}}\bar{\nu}_{L,i}\gamma^\mu U_{ij} e_{L,j} W^+_\mu + \text{h.c.} \tag{21.2}$$

$$= \frac{g}{\sqrt{2}}\left[\bar{\nu}_{L,i}\gamma^\mu U_{ij} e_{L,j} W^+_\mu + \bar{e}_{L,j}\gamma^\mu U^*_{ji}\nu_{L,i} W^-_\mu\right] . \tag{21.3}$$

We choose the rest-frame of the W-boson so that $W^\pm_0 = 0$. A CP transformation transforms the current $\bar{\nu}_{L,i}\gamma e_{L,j}$ into $\bar{e}_{L,j}\gamma\nu_{L,i}$ and the W^+_i into a W^-_i. It exchanges also the left- and right-circular polarisations of the W and transforms all arguments $x^\mu = (t, \boldsymbol{x})$ of the fields into $x^\mu = (t, -\boldsymbol{x})$. The latter two effects are harmless, since we integrate over x^μ and sum over the two polarisations of the W in the action. Thus the combined effect of CP on the first term in (21.3) is given by

$$\frac{g}{\sqrt{2}}\bar{\nu}_{L,i}\gamma^\mu U_{ij} e_{L,j} W^+_\mu \to \frac{g}{\sqrt{2}}\bar{e}_{L,j}\gamma^\mu U_{ji}\nu_{L,i} W^-_\mu . \tag{21.4}$$

The CP transformed term is identical to the Hermitian conjugated of the original term, if the mixing matrix is real, $\boldsymbol{U} = \boldsymbol{U}^*$. In this case, a CP transformation simply exchanges the first and the second term in (21.3). If however the mixing matrix is complex, the CP transformed Lagrangian differs from the original one, and CP is violated.

In general, the complex parameters required for CP violation can arise in two ways: First, interactions may contain physical phases, as in the case of the CKM and the MNSP matrices in the SM. Second, vacuum expectation values may contain physical phases. This option called spontaneous CP violation requires at least two Higgs doublets, as we discussed in remark 17.1. Enlarging the Higgs sector is therefore an efficient way to add more CP violation to the SM.

Sphaleron transitions and $\boldsymbol{B-L}$ violation Recall from section 16.2 the effect of an instanton transition on the divergence of the axial current,

$$\partial_\mu j^\mu_A = \partial_\mu (j^\mu_R - j^\mu_L) = \frac{g^2}{16\pi^2}\,tr(F_{\mu\nu}\tilde{F}^{\mu\nu}) . \tag{21.5}$$

Strong interactions couple equally to left- and right-chiral fermions, and therefore the sole effect of an instanton transition is a chirality flip of the fermion. In contrast, an electroweak instanton couples only to left-chiral fermions,

$$\sum_i \partial_\mu j^\mu_{L,i} = -n_g\frac{g^2}{16\pi^2}tr(W_{\mu\nu}\tilde{W}^{\mu\nu}) = \Delta\nu\,, \tag{21.6}$$

where n_g counts the number of generations, i stands for all components of the left-chiral SU(2) doublets, $L = (\nu_l, l)$ with $l = \{e, \mu\,\tau\}$ and the corresponding nine quark doublets. Now an instanton transition changes the fermion number of the left-chiral fields. More precisely, a transition which changes the winding number by one unit, $\Delta\nu = 1$, changes the fermion number in each doublet by one unit and thus the total fermion number by twelve. Since each

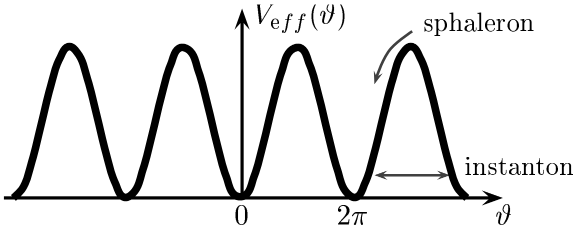


Figure 21.1.: Instanton versus sphaleron transition between different ϑ vacua.

quark carries baryon number 1/3, the changes in the lepton flavour and baryon number are connected by

$$\Delta L_e = \Delta L_\mu = \Delta L_\tau = \frac{1}{3}\Delta L = \frac{1}{3}\Delta B \tag{21.7}$$

or $\Delta L = \Delta B = n_g = 3$ for $\Delta\nu = 1$. Thus we see that the combination $B - L$ is conserved in the SM even in the presence of instanton transitions, while both the baryon and the lepton number are broken. Adding conservation of electric charge and colour, the states connected by an instanton are fixed. In particular, a transition with $\Delta\nu = 1$ creates the state

$$|0\rangle \to \left|u_L u_L d_L e_L^- + c_L c_L s_L \mu_L^- + t_L t_L b_L \tau_L^-\right\rangle ,$$

where the three quarks of each generations form a colour singlet. Since mass terms and Higgs interactions mix left- and right-chiral fields, the resulting change in the number of left-chiral fermions is transferred to the right-chiral fermions.

The differential probability p per time and volume of a tunnelling process connecting two vacua separated by one winding follows from (16.57) as

$$p = \xi m_W^{-4} \exp(-S) = \xi m_W^{-4} \exp\left(-\frac{8\pi^2}{g^2}\right) \approx \frac{10^{-160}}{m_W^4}\,, \tag{21.8}$$

where the prefactor ξ is a dimensionless function of order one. Thus at zero temperature, the effects of electroweak instantons are completely negligible. In the early Universe however, we should take into account not only quantum but also thermal fluctuations. We include the latter replacing the tunnelling factor $\exp(-S)$ by the Boltzmann factor $\exp(-E/T)$. The classical field configuration[1] connecting the top of the barrier with the vacuum which has the smallest free energy is called sphaleron, cf. Fig. 21.1. Its energy in the broken phase is $E_{sp} \sim m_W(T)/\alpha_W$, where $m_W(T)$ is the temperature dependent mass of the W. Thus the probability that a thermal fluctuation crosses the barrier is $\propto \exp(-2E_{\rm sp}(T)/T)$. At temperatures $T \gtrsim m_W$, thermal fluctuations are larger than the barrier height between the vacua and the baryon number violating processes should proceed unsuppressed. At these temperatures, the SM is in the unbroken phase and the only relevant length scale is the analogue of the Debye mass m_D for a gauge boson, $m_D \sim \alpha_W T$. Thus we expect that the sphaleron rate[2] is given in the high-temperature limit by

$$p \sim \xi(\alpha_W T)^4\,. \tag{21.9}$$

[1] The potential energy is a functional of the field configurations, $V_{eff}[W^\mu, \phi]$, and thus the figure represents a cut through an infinite dimensional field space; the maximum is in fact a saddle point.

[2] Our dimensional argument does not account for an additional factor α_W arising in perturbative calculations due to IR dynamics, but is nevertheless with $\xi \approx 0.1$ a good approximation for the sphaleron rate determined in lattice calculation.

Comparing the rate per particle, $\Gamma \sim \xi\alpha_W^4 T$, with the Hubble rate, wee see that the $B+L$ violating sphaleron transition are (with $\xi \sim 0.1$) up to $T_{sp} \sim 10^{12}$ GeV in equilibrium. As the universe cools down, sphaleron processes go out of equilibrium around $T \sim 100\,\text{GeV}$, when the rate becomes Boltzmann suppressed.

Connecting B and B-L In thermal equilibrium, a linear combination of $B + aL$ has to vanish. The exact value of a can be determined using that the universe is neutral, i.e. that the conserved charges should be zero. Before the electroweak phase transition, all the three gauge forces are of long-range. The corresponding charges have to be therefore zero (or better negligibly small), in order not to outweigh the gravitational forces. The only other conserved charge is $B-L$. It is sufficient to consider the hypercharge Y and $B-L$. Setting $\mu \equiv \mu_{B-L}$, a particle of type i has the chemical potential

$$\mu_i = \mu(B_i - L_i) + \mu_Y Y_i/2 = -\mu_{\bar{i}}\,. \tag{21.10}$$

Using the quantum number assignments of table **??**, it follows e.g.

$$\mu_{h^+} = \mu_{h^0} = \frac{1}{2}\mu_Y \quad and \quad \mu_{u_L} = \mu_{d_L} = \frac{\mu}{3} + \frac{\mu_Y}{6} \tag{21.11}$$

with analogous relations for the other members of the first fermion generation. Since the asymmetry is small, $\mu_i \ll T$, we can use Eq. (**??**) obtaining

$$\Delta n_i \equiv n_i - n_{\bar{i}} = \delta \sum_i g_i \mu_i q_i \frac{T^2}{3} \tag{21.12}$$

with $\delta = 2$ for bosons and $\delta = 1$ for fermions. The condition that the hypercharge of the plasma vanishes becomes

$$\sum_i \Delta n_i Y_i = n_g\left(\frac{5}{3}\mu_Y + \frac{4}{3}\mu\right) + \frac{1}{2}n_h\mu_Y = 0\,, \tag{21.13}$$

where we summed over n_g fermion generations and n_h Higgs doublets. Now we can solve for μ, eliminate then this variable in the relations for Δn_i, arriving at

$$n_B = \frac{1}{3}(\Delta n_{u_L} + \Delta n_{u_R} + \Delta n_{d_L} + \Delta n_{d_R}) = -\mu_Y\left(\frac{1}{2}n_g + \frac{1}{4}n_h\right)\frac{T^3}{3}\,, \tag{21.14}$$

$$n_L = \Delta n_{\nu_L} + \Delta n_{e_L} + \Delta n_{e_R} = \mu_Y\left(\frac{7}{8}n_g + \frac{9}{16}n_h\right)\frac{T^3}{3}\,. \tag{21.15}$$

Subtraction gives an expression for n_{B-L}. Eliminating T^3 with the help of (21.14), we find finally

$$n_B = \frac{24 + 4n_h}{66 + 13n_h} n_{B-L} = a n_{B-L} \tag{21.16}$$

with $a = 28/79$ in the SM. Because the $B+L$ violating sphaleron transitions are in thermal equilibrium above $T \sim 100\,\text{GeV}$, we have to modify therefore the first Sakharov condition: Any asymmetry proportional to $B+L$ will relax to zero, while the part proportional to $B-L$ will lead a final baryon asymmetry. In particular, a non-zero baryon number arises too, if we generate at $T \gg 100\,\text{GeV}$ an asymmetry only in the lepton number.

Toy model Let us consider a simple model, where we couple a heavy scalar X via Yukawa interactions to four SM fermions f_i,

$$\mathscr{L}_X = g_X^{12} X \bar{f}_2 f_1 + g_X^{34} X \bar{f}_4 f_3 + \text{h.c.} \tag{21.17}$$

To be concrete, we choose the scalar as a heavy leptoquark X with electric charge $q(X) = -4/3$ and the two decay modes as

$$X \to \bar{u}\bar{u}, \qquad r \tag{21.18}$$

$$X \to e^- d, \qquad 1 - r\,, \tag{21.19}$$

where r is the corresponding branching ratio. The baryon number of the first decay mode is $B = -2/3$, while the one of the second mode is $B = 1/3$. Therefore we can not assign a baryon number to X and thus baryon number is violated. Next we require that in these decays C is violated. Then the charge conjugated decay modes have different branching ratios $\bar{r} \neq r$,

$$\bar{X} \to uu, \qquad \bar{r} \tag{21.20}$$

$$\bar{X} \to e^+ \bar{d}, \qquad 1 - \bar{r}\,. \tag{21.21}$$

The resulting change ΔB of the baryon number B per decay of a $X, \bar{X}$ pair is thus

$$\Delta B = -\frac{2}{3} r + \frac{1}{3}(1 - r) + \frac{2}{3}\bar{r} - \frac{1}{3}(1 - \bar{r}) = \bar{r} - r\,, \tag{21.22}$$

i.e. proportional to the amount of C violation. If we consider these processes at tree-level, then the decay widths are $\Gamma(X \to \bar{f}_2 f_1) = |g_X^{12}|^2 I_X$ and $\Gamma(\bar{X} \to \bar{f}_2 f_1) = |g_X^{12*}|^2 I_{\bar{X}}$ where $I_X = I_{\bar{X}}$ is a real kinematical factor determined by the masses, $I = I(m_X, m_i, m_j)$. Thus at tree-level, CP violating effects are absent, $r = \bar{r}$, the second Sakharov condition is not satisfied, and thus no net baryon number is generated.

This example shows that for successful baryogenesis two additional requirements have to be satisfied: First, the kinematical terms I should be complex. The required imaginary part can be generated in a loop graph, if one of the virtual particles can become on-shell. This can happen, if the mass of the decaying particle is to be larger than the sum of the fermion masses in the loop. Second, at least two heavy particles have to interfere, such that the decay widths contain the complex combination $g_X g_Y^*$ instead of the real $|g_X|^2$. Adding therefore

$$\mathscr{L}_Y = g_Y^{12} Y \bar{f}_2 f_1 + g_Y^{34} Y \bar{f}_4 f_3 + \text{h.c.}$$

to the Lagrangian produces at the one-loop a potentially baryon number violating term,

$$\Delta B \propto \mathrm{Im}(g_Y^{12*} g_X^{12} g_Y^{34*} g_X^{34}) \mathrm{Im}(I_{XY})\,. \tag{21.23}$$

21.2. Baryogenesis in out-of equilibrium decays

Sakharov's third condition for baryogenesis, a departure from thermal equilibrium, can be satisfied during phase transitions or when a particle species is out of chemical equilibrium. In this section, we examine the conceptionally simpler second possibility, considering out-of equilibrium decays of unstable heavy particle X which violate baryon number.

Boltzmann equation for decays As first step, we simplify the Boltzmann equation for decays $X \to bb$ and its inverse reaction. Here, the unstable heavy X and Y particles may be out of chemical equilibrium, while the light decay products b are in equilibrium. In a GUT scenario for baryogenesis, we might identify the X particles with the lepto quark-like gauge bosons with electric charge $q = 4/3$. Finally, we assume $M_Y \gg M_X$ such that the heavier Y particle contributes only as virtual state in the loop correction. Then the decays $X \to bb$ corresponds to the annihilation term β and the inverse decays $bb \to X$ to the production term ψ in the Boltzmann equation. Using detailed balance to relate decays and inverse decays, we find

$$\frac{\mathrm{d}n_X}{\mathrm{d}t} + 3Hn_X = -\Gamma_{\mathrm{D}}(n_X - n_{X,\mathrm{eq}})\,. \tag{21.24}$$

Let us assume that the relevant decays of the X particle are to two bb particles and to two $\bar{b}\bar{b}$ antiparticles, with $B = +1/2$ and $B = -1/2$, respectively. We denote the squared Feynman amplitudes as

$$|\mathcal{A}(X \to bb)|^2 = |\mathcal{A}(\bar{b}\bar{b} \to X)|^2 = \frac{1}{2}(1+\varepsilon)|A_0|^2\,, \tag{21.25}$$

$$|\mathcal{A}(X \to \bar{b}\bar{b})|^2 = |\mathcal{A}(bb \to X)|^2 = \frac{1}{2}(1-\varepsilon)|A_0|^2\,. \tag{21.26}$$

Then the asymmetry per decay of one X particle is given by

$$\varepsilon_X = \sum_f B_f \,\frac{\Gamma(X \to f) - \Gamma(\bar{X} \to \bar{f})}{\Gamma_{tot}} = \frac{\frac{1}{2}(1+\varepsilon)|A_0|^2 - \frac{1}{2}(1-\varepsilon)|A_0|^2}{\frac{1}{2}(1+\varepsilon)|A_0|^2 + \frac{1}{2}(1-\varepsilon)|A_0|^2} \equiv \varepsilon\,. \tag{21.27}$$

We derive in the same way the Boltzmann equations for the number density of the b and $\bar{b}$ particles. However, we have to include additionally to the decays 2 → 2 scattering processes that may wash-out the created baryon asymmetry[3]. Subtracting the two Boltzmann equations and dividing by two, we obtain as equation for the baryon number density $n_B = n_b - n_{\bar{b}}$,

$$\frac{\mathrm{d}n_B}{\mathrm{d}t} + 3Hn_B = \varepsilon\Gamma_{\mathrm{D}}(n_X - n_{X,\mathrm{eq}}) - \Gamma_{\mathrm{D}} n_B \frac{n_{X,\mathrm{eq}}}{n_\gamma} - 2n_B n_n \langle\sigma v\rangle\,. \tag{21.28}$$

The only term which can be positive and can thus drive the baryon asymmetry, $\varepsilon\Gamma_{\mathrm{D}}(n_X - n_{X,\mathrm{eq}})$, shows clearly the three Sakharov conditions: It is zero in case of thermal equilibrium, $n_X = n_{X,\mathrm{eq}}$, or if C, CP, or B are not violated, i.e. if $\varepsilon = 0$. The second term accounts for the inverse decays of the X boson, where we used detailed balance to relate decays and inverse decays. Finally, the third part includes 2 → 2 baryon number violating scattering processes. Here, a subtle point arises: The imaginary part of this processes corresponds to the product of an inverse decay and a decay, which we have already taken into account. Thus we should include only the real, off-shell part of the scattering process.

Next we change to dimensionless variables, introducing $x = m_X/T$ and $Y_i = n_i/s$. We define as measure for the departure from equilibrium

$$K \equiv \left.\frac{\Gamma_D(x)}{2H(x)}\right|_{x=1} = \frac{\alpha M_{\mathrm{Pl}}}{3.3 g_*^{1/2} m_X}\,. \tag{21.29}$$

[3] The existence of such processes is implied by unitarity, cf. for an example Fig. 21.2

Inserting all this, the new Boltzmann equations can be written as

$$\frac{\mathrm{d}Y_X}{\mathrm{d}x} = -Kx\gamma_D \left(Y_X - Y_X^{\mathrm{e}q}\right) \tag{21.30a}$$

$$\frac{\mathrm{d}Y_B}{\mathrm{d}x} = \varepsilon Kx\gamma_D \left(Y_X - Y_X^{\mathrm{e}q}\right) - Kx\gamma_B Y_B \tag{21.30b}$$

with $\gamma_D = \Gamma_D(x)/\Gamma_D(1)$ and $\gamma_B = g_* Y_X^{\mathrm{e}q}\gamma_D + 2n_\gamma \langle\sigma v\rangle/\Gamma_D(1)$. The baryon asymmetry Y_B is driven by the departure from equilibrium, $\Delta = Y_X - Y_X^{\mathrm{e}q}$ and damped by inverse decays and $2 \to 2$ scatterings. Changing to Δ as variable, we can rewrite Eq. (21.30) as

$$\frac{\mathrm{d}\Delta}{\mathrm{d}x} = -\frac{\mathrm{d}Y_X^{\mathrm{e}q}}{\mathrm{d}x} - Kx\gamma_D\Delta\,, \tag{21.31a}$$

$$\frac{\mathrm{d}Y_B}{\mathrm{d}x} = \varepsilon Kx\gamma_D\Delta - Kx\gamma_B Y_B\,. \tag{21.31b}$$

Integrating these first-order equations results in

$$\Delta(x) = \Delta(x_0)\exp\left[-\int_{x_0}^{x}\mathrm{d}z\, zK\gamma_D(z)\right] - \int_{x_0}^{x}\mathrm{d}zX'_{\mathrm{e}q}(z)\exp\left[\int_{z}^{x}\mathrm{d}z'\, z'K\gamma_D(z')\right] \tag{21.32}$$

and

$$\begin{aligned} Y_B(x) &= Y_B(x_0)\exp\left[-\int_{x_0}^{x}\mathrm{d}z\, zK\gamma_B(z)\right] \\ &+ \varepsilon K\int_{x_0}^{x}\mathrm{d}z\gamma_D(z)\Delta(z)\exp\left[-\int_{z}^{x}\mathrm{d}z'\, z'K\gamma_B(z')\right]\,. \end{aligned} \tag{21.33}$$

The most interesting limiting case is $K \ll 1$, when the out-of-equilibrium condition is satisfied. Then the exponentials are of order one, and the abundances become with $\gamma_D(x) \simeq 1$ for $x \gg 1$

$$Y_X(x) \simeq Y_X(0)\exp\left[-Kx^2/2\right]\,, \tag{21.34a}$$

$$Y_B(x) = \varepsilon[X_X(0) - X(z)] \simeq \varepsilon X(0)\,. \tag{21.34b}$$

Thus the X particles decay around $x \approx K^{-1/2}$, resulting in a baryon asymmetry ε/g_*.

GUT baryogenesis Since GUT theories unify quarks and lepton, they contain also gauge bosons X_μ and Y_μ similar to the lepto-quark discussed in our toy model. In the simplest GUT theory, SU(5), $B - L$ is conserved and thus any non-zero B will be washed out by sphaleron processes. In GUT theories based on larger groups as e.g. SO(10), $B - L$ is broken and baryogenesis based on out-of equilibrium decays is in principle possible. However, the necessary temperatures, $T \gtrsim M_{GUT} \sim 10^{16}$ GeV, are larger than the maximal temperature the universe is reheated after inflation. A possible solution to this problem is sketched in section 23.3.

Leptogenesis This is an attractive model for baryogenesis which connects the smallness of neutrino masses with the creation of the baryon asymmetry. In a first step, L violating decays of heavy right-handed neutrinos generate $L \neq 0$. Latter, sphaleron processes that conserve only $B - L$ convert the lepton asymmetry into a baryon asymmetry.

In the seesaw model, we extend the SM by three right-chiral neutrinos $N_{R\alpha}$ with $\alpha = \{e, \mu, \tau\}$. Their interactions are described by

$$\mathscr{L} = M_\alpha \bar{N}^c_{R\alpha} N_{R\alpha} + (y_{\alpha\beta} \bar{l}_{L\alpha} \Phi N_{R\beta} + \text{h.c.}) \,. \tag{21.35}$$

Since we include Majorana masses, lepton number is violated in the decays of the right-chiral neutrinos,

$$N_{R\alpha} \to l_{L\beta} + \bar{\Phi} \quad \text{a}nd \quad N_{R\alpha} \to l^c_{L\beta} + \Phi \,. \tag{21.36}$$

If there is enough CP violation and the out-of-equilibrium condition is satisfied, these decays can result in a L asymmetry. We check first the latter condition, comparing the decay width $\Gamma \sim |y_{\alpha 1}|^2 M_1/(16\pi)$ of the lightest right-chiral neutrino N_{R1} with mass M_1 to the Hubble rate $H(T)$ at $T = M_1$. This results in the bound $M_1 \gtrsim 10^{14}\,\text{GeV}$ for a Yukawa coupling of $\mathcal{O}(y) = 0.1$. Numerical calculations show that the decoupling can happen later, what relaxes the bound to $M_1 \gtrsim 10^{12}\,\text{GeV}$.

As sources for CP violation, we have to consider the interference terms of the tree-level decays (21.36) with one loop corrections. We have to require $N_j \neq N_1$ as virtual particle, while the presence of the light particles l and Φ ensures that an imaginary part is created. Additionally to the vertex correction, these requirements are also satisfied by the self-energy insertion, cf. the top of Fig. 21.2. Cutting the loops as indicated leads automatically to scattering processes which violate lepton number by two units, $\Delta L = 2$. These scattering processes will wash out the generated L asymmetry, if they are in equilibrium. They are described by the effective interaction

$$\mathscr{L} = \frac{1}{M} ll\Phi\Phi \,, \tag{21.37}$$

which generates after electroweak symmetry breaking Majorana masses $m_\nu = v^2/M$ for the three light neutrinos via the seesaw mechanism. For dimensional reasons, the rate $\Gamma(\Delta L = 2)$ of these processes is proportional to

$$\Gamma(\Delta L = 2) \sim \frac{T^3}{M^2} \sim \frac{T^3 \sum_i m_i^2}{v^4} \,. \tag{21.38}$$

These processes are not effective, if their rate is smaller than the Hubble rate in the range $T_{ew} < T < M_1$. This implies a limit on the masses of the light neutrinos,

$$\sum_{i=e,\mu,\tau} m_i^2 \lesssim (0.2\,\text{e}V)^2 \,, \tag{21.39}$$

which is comparable to the upper limit from neutrinoless double beta decay and from cosmology.

21.3. Baryogenesis in phase transitions

The most economical model for baryogenesis is the attempt to use only the SM: It contains C, CP as well as B violation and hence the first two of the three Sakharov conditions are satisfied. Although electroweak interactions rates are fast compared to the Hubble rate at $T \sim m_W$ and we can thus not use out-of-equilibrium decays, a deviation from thermal equilibrium may occur during the electroweak phase transition. In practise, however, the amount of CP

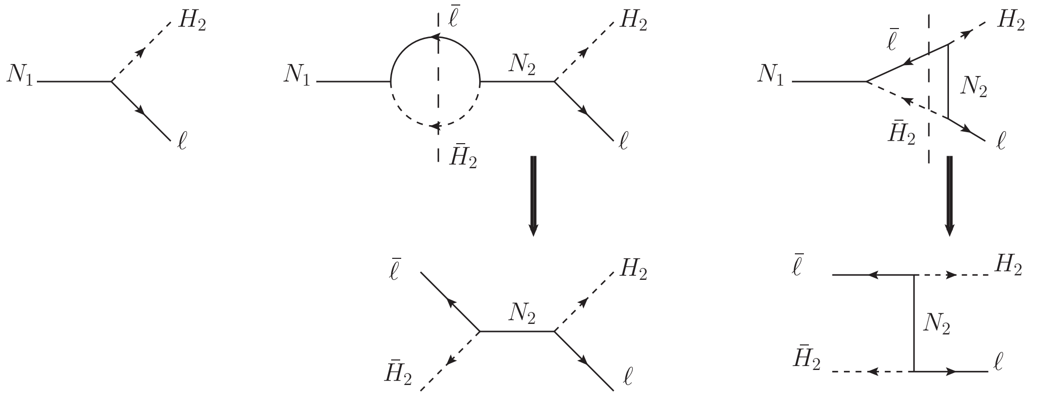


Figure 21.2.: Top: Feynman diagrams for the tree-level decay and the one loop corrections which source CP violation. Bottom: Scattering processes with $\Delta L = 2$; from [Rac16]

violation in the CKM matrix is too small and the phase transition is only a smooth cross-over. It is still tempting to ask, if electroweak baryogenesis is possible for a slightly modified SM, adding e.g. a second Higgs doublet, since such models are testable at the LHC.

Recall that in a first-order transition, the two minima at the critical temperature T_c are separated by a potential barrier. While the universe cools below T_c, it will be trapped for some time in the false minimum. As a result of tunnelling through the potential barrier, nucleation of bubbles containing the true vacuum starts. Initially, their surface tension is too large and the bubbles will collapse. When the temperature lowers further, the volume energy gain overweights the surface tension and the bubbles start to grow. As the bubbles expand, the expectation value $\langle\phi\rangle$ of the Higgs field changes from $\langle\phi\rangle = 0$ to $\langle\phi\rangle = v(T)$ when the bubble passes the considered point. It is this change of the order parameter $v(T)$ which provides the necessary departure from equilibrium.

Baryogenesis in a phase transitions is a complicated dynamical process and we give only a sketch of the basic physics involved. Let us estimate first the relative size of the relevant scales:

- The sphaleron rate changes from $\Gamma_{\mathrm{s}p} \sim \xi\alpha_W^4 T \sim 10^{-6}T$ outside the bubble to $\Gamma_{\mathrm{s}p} \sim \exp(-E_{\mathrm{s}p}/T) \sim \exp(-m_W(T)/\alpha_W T)$ in the broken phase contained in the expanding bubble. Requiring that the generated baryon number is not washed out in the broken phase, requires $m_W(T)/T = gv(T)/2T \gg 1$, or in other words a large jump in the order parameter $v(T)$ of the phase transition.

- The time scale for kinetic equilibration is given by strong or electroweak Coulomb scatterings with rate close to the plasma temperature, $\Gamma_{th} \sim 0.1T$.

- The time scale for the change of the Higgs vev when the bubble passes depends on the bubble speed v_b and the width of the bubble wall δ. Typically, one finds $\Gamma_{\mathrm{v}} \sim \dot{\phi}/\phi \sim v_b/\delta \sim (0.01 - 0.1)T$.

Hence the rate of baryon number violating processes is always out of equilibrium near the wall, $\Gamma_{\mathrm{s}p} \ll \Gamma_{\mathrm{v}}$. Depending on the relative size of Γ_{th} and Γ_{v} one distinguishes two regimes:

- The adiabatic thick wall regime, $\Gamma_{\mathrm{t}h} \gg \Gamma_{\mathrm{v}}$, where the plasma is in quasi-static equilibrium with time-dependent chemical potentials.

- The non-adiabatic thin wall regime, $\Gamma_{\mathrm{t}h} \ll \Gamma_{\mathrm{v}}$. Here, the individual CP violating transmission and reflection of particles at the bubble wall has to be calculated. These processes differ for L and R chiral fermions, because CP is violated: Therefore there is an excess of $f_l + \bar{f}_R$ compared to $f_R + \bar{f}_L$ in front of the bubble wall. Sphaleron interaction are effective only in the unbroken phase outside the bubble. Other processes exchange L to R. Since sphalerons react only with f_L, the rate for
$$e^-_L \mu^-_L \tau^-_L \to u_L u_L d_L + c_L c_L s_L + t_L t_L b_L \,,$$
is slower than for
$$e^+_R \mu^+_R \tau^+_R \to \bar{u}_R \bar{u}_R \bar{d}_R + \bar{c}_R \bar{c}_R \bar{s}_R + \bar{t}_R \bar{t}_R \bar{b}_R \,,$$
because there are more $\bar{f}_R$ than f_L. The change of the baryon number in the two reactions is opposite, and thus a difference in their rates results in the creation of a baryon asymmetry. Since the wall is moving, part of the created baryons will end up inside the broken phase where wash-out processes are not active. It is this part which survives and creates the final baryon asymmetry.

Adiabatic thick wall regime Let us consider now in a bit more detail the opposite regime. We use as a toy model the Yukawa interactions between one fermion doublet Q_L, a singlet q_R and two Higgs doublets. Their vev's are space-time dependent and contain one physical phase which we can choose as $\phi_1 = v_1 \mathrm{e}^{\mathrm{i}\delta}$ and $\phi_2 = v_2 \mathrm{e}^{-\mathrm{i}\delta}$. The Yukawa interactions contain terms of the type
$$\mathscr{L}_Y = -y_f v_1 \mathrm{e}^{\mathrm{i}\delta} \bar{d}_L d_R + \text{h.c.} \tag{21.40}$$
Both v_1 and δ vary across the bubble wall. Since the bubble is expanding, v_1 and δ and thus also the Lagrangian are time dependent. We can eliminate the time-dependence of $\mathscr{L}_Y$ caused by $\delta(t)$ performing the time dependent rotation $d_R \to \mathrm{e}^{-\mathrm{i}\delta(t)} d_R$. This induces an additional term in the kinetic energy of right-chiral quarks,
$$\mathrm{i}\bar{d}_R \not{\partial} d_R \to \mathrm{i}\bar{d}_R \not{\partial} d_R + \dot{\delta} \bar{d}_R \gamma^0 d_R \,. \tag{21.41}$$
As a result, the Hamiltonian changes as
$$H \to H - \dot{\delta} \int \mathrm{d}^3x \, \bar{d}_R \gamma^0 d_R = H - \dot{\delta} N_R \,. \tag{21.42}$$
Electroweak processes which interchange left- and right-chiral quarks are fast compared to $\dot{\delta}$. In contrast, the sphaleron rate is slower than $\dot{\delta}$ and we neglect in a first step this rate. In this approximation, the baryon number $B = N_R + N_L$ is zero and we have to add to (21.42) a chemical potential,
$$H \to H - \dot{\vartheta} N_R - \mu_B (N_R + N_L) \,. \tag{21.43}$$
The resulting effective chemical potentials for right- and left-chiral quarks differ, and using again (**??**), we obtain
$$\Delta n_B = n_{B,R} + n_{B,L} = \Big[(\mu_B + \dot{\delta}) + 2\mu_B \Big] \frac{T^2}{6} \,, \tag{21.44}$$

where the factor 2 accounts for the two components of the fermion doublet. Setting $\Delta n_B = 0$, we find

$$\mu_B = -\frac{1}{3}\dot{\delta}\,. \tag{21.45}$$

Next we include the effect of the non-zero sphaleron rate. Using $\Delta B = 1$ and $\Delta F = \mu_B \Delta B = \mu_B$ as change of the free energy F and the baryon number B per sphaleron process, we find applying detailed balance

$$\frac{\mathrm{d}n_B}{\mathrm{d}t} = -\frac{\Delta F\,\Delta B}{T}\,\Gamma_{\mathrm{sp}} = -\frac{\mu_B}{T}\,\Gamma_{\mathrm{sp}} = \frac{1}{3}\,\frac{\dot{\delta}}{T}\,\Gamma_{\mathrm{sp}}\,. \tag{21.46}$$

In the last step, we used also that the change of μ_B is mainly driven by $\dot{\delta}$, while Γ_{sp} is slow. Integrating this relation, we obtain the baryon number density generated by a passing bubble wall,

$$n_B = \frac{1}{3T}\int \mathrm{d}t\, \dot{\delta}\Gamma_{\mathrm{sp}}\,. \tag{21.47}$$

For an order of magnitude estimate, we can use that the sphaleron rate drops from $\Gamma_{sp} \approx \xi\alpha_W^4 T \sim 10^{-6}T$ to zero. Then we find for the ratio of baryon and entropy density

$$\frac{n_B}{s} \approx \frac{45\xi\alpha_W^4 T\Delta\vartheta}{6\pi^2 g_*} \sim 10^{-8}\Delta\vartheta\,. \tag{21.48}$$

This simple estimate shows that successful electroweak baryogenesis is possible, if the SM is extended such that there is sufficient CP violation and the phase transition is strong enough. While the first condition is relatively easy too satisfy, the second one requires additional particles which generically should be discovered at the LHC.

Summary

The dynamical generation of a baryon asymmetry is only possible, if the Sakharov conditions (violation of B, of C and CP, and departure from thermal equilibrium) are satisfied. The SM contains neither a sufficiently strong source of CP violation, nor leads to a departure from thermal equilibrium in the early Universe. Thus baryogenesis requires necessarily physics beyond the SM. Electroweak baryogenesis has the virtue of being testable at accelerators, while leptogenesis is supported by the observation that the light neutrino masses required to avoid wash-out of the baryon asymmetry fall in the experimentally allowed range.

Further reading

A more detailed discussion of baryogenesis which includes also the Affleck-Dine mechanism can be found in [GR11b]. The connection between leptogensis and neutrino masses is reviewed in [BDBP05]. Baryogenesis is an application where loop corrections involving thermal fields are essential. This requires to go beyond the Boltzmann equation, since its derivation is based on the distribution function in classical phase space; for some approaches in this direction see [ABDM11].

22. Quantum fields in curved space-time

Our treatment of quantum field theory has been restricted up to now to *inertial* observers and detectors in *Minkowski* space. This combination is very peculiar for two reasons: First, all Minkowski space is covered by a time-like Killing vector field. Second, no event horizons exist for inertial observers in this space-time. The existence of a unique time-like Killing vector field ∂_t which has as eigenfunctions the modes $\mathrm{e}^{-\mathrm{i}\omega t}$ implies that all inertial observers can agree how to split positive and negative frequency modes. This splitting selects in turn the standard Minkowski vacuum $|0\rangle_M$. As a result, the vacuum states defined by different inertial observers agree, and no inertial detector will register particles in the vacuum state $|0\rangle_M$. In this chapter we will consider more general situations. A case of special interest is the expanding universe described by the FLRW metric, where no time-like Killing vector field exists. Intuitively, we expect that a time-dependent gravitational field, in analogy to an electric field, can create particles. Our analysis confirms this expectation, but also teaches us that the concept of a "particle" becomes dubious in a non-stationary metric. Even more astonishingly, we will find that particle creation can occur also in the case of a *static* space-time, if an event horizon exists. Such a horizon obstructs the construction of an unique time-like Killing vector field and corresponds to a surface of infinite redshift. As a result, a *thermal* spectrum of particles is created close to the horizon.

22.1. Conformal invariance and scalar fields

Conformally flat space-times The problem of quantising field theories in curved space-times simplifies considerably, if we apply the following two restrictions: First, we consider only conformally flat space-times, i.e. those space-times which are connected by a conformal transformation to Minkowski space,

$$g_{\mu\nu}(x) = \Omega^2(x)\eta_{\mu\nu}(x) = \mathrm{e}^{2\omega(x)}\eta_{\mu\nu}(x)\,. \tag{22.1}$$

Note that conformal transformations $g_{\mu\nu}(x) \to \tilde{g}_{\mu\nu}(x) = \Omega^2(x)g_{\mu\nu}(x)$ of the *metric* are not equivalent to conformal transformations of the *coordinates*, $x \to \tilde{x} = \mathrm{e}^{\omega(x)}x$, which we considered in chapter 17.1. In the latter case, the argument of the metric tensor in the LHS and the RHS in (22.1) would differ, see also problem 23.??. Recall also that a coordinate transformation $x \to \tilde{x}(x)$ only relabels the space-time points, but does not affect physics, since we require that any action S is invariant under general coordinate transformations. By contrast, a conformal transformation of the metric shrinks and stretches the Riemannian manifold $\{\mathcal{M}, g_{\mu\nu}\}$ into another manifold $\{\tilde{\mathcal{M}}, \tilde{g}_{\mu\nu}\}$.

Conformal transformations change distances, but keep angles invariant. Thus the causal structure of two conformally related space-times is identical. In particular, light-rays propagate also in conformally flat space-times along straight lines at ±45 degrees to the time axis. Important examples for conformally flat space-times are the flat FLRW metric and all two-dimensional space-times.

Remark 22.1: Geometric quantities derived for the metric $g_{\mu\nu}(x)$ are connected to those of a conformally transformed metric $\tilde{g}_{\mu\nu}(x) = \Omega^2(x) g_{\mu\nu}(x) = \mathrm{e}^{2\omega(x)} g_{\mu\nu}(x)$ as

$$\tilde{\Gamma}^{\mu}{}_{\alpha\beta} = \Gamma^{\mu}{}_{\alpha\beta} + \Omega^{-1}\left[\delta^{\mu}_{\alpha}\partial_\beta\Omega + \delta^{\mu}_{\beta}\partial_\alpha\Omega - g_{\alpha\beta}g^{\mu\nu}\partial_\nu\Omega\right], \tag{22.2a}$$

$$\tilde{R}_{\mu\nu} = R_{\mu\nu} - g_{\mu\nu}\Box\omega - (d-2)\nabla_\mu\nabla_\nu\omega + (d-2)\nabla_\mu\omega\nabla_\nu\omega - (d-2)g_{\mu\nu}\nabla^\lambda\omega\nabla_\lambda\omega\,, \tag{22.2b}$$

$$\tilde{R} = \Omega^{-2}\left[R - 2(d-1)\Box\omega - (d-1)(d-2)\nabla^\mu\omega\nabla_\mu\omega\right], \tag{22.2c}$$

as one can check by direct (but tedious) computation.

Conformal invariance Another simplification occurs if we consider field theories which are conformally invariant in Minkowski space. Recall that such theories contain no dimensionsfull parameter and satisfy $T_\mu^{\ \mu} = 0$. In a curved space-time, we call such a theory conformally or Weyl invariant[1]. The action

$$S[\phi, g^{\mu\nu}] = \int \mathrm{d}^d x\sqrt{|g|}\mathscr{L}(\phi, \partial_\mu\phi, g^{\mu\nu}) \tag{22.3}$$

of such a theory is invariant under a conformal transformation of the metric,

$$S[\phi, g^{\mu\nu}] = S[\tilde{\phi}, \tilde{g}^{\mu\nu}] = \int \mathrm{d}^d x\sqrt{\tilde{g}}\mathscr{L}(\tilde{\phi}, \partial_\mu\tilde{\phi}, \tilde{g}^{\mu\nu})\,, \tag{22.4}$$

if we rescale the field according to its canonical dimension, i.e. $\phi(x) \to \tilde{\phi}(x) = \Omega^{(2-d)/2}(x)\phi(x)$ for a boson. Therefore the equations of motion for the field $\tilde{\phi}$ using the metric $\tilde{g}_{\mu\nu}$ are the same as those for the field ϕ using the metric $g_{\mu\nu}$. This allows us to relate the quantisation of a Weyl invariant theory in a conformally flat space to the known problem of quantising the field in Minkowski space.

Conformal invariance of a scalar field We have experienced that most calculations in Minkowski space for a scalar field are considerably less involved than for fields with non-zero spin. This holds true also in curved space-times, except for one aspect: While massless Dirac and Yang-Mills fields are classically conformal invariant, this is not the case for a scalar field minimally coupled to gravity: As discussed in section 17.1 the trace $T_\mu^{\ \mu}$ of the stress tensor for a massless scalar field is given by

$$T_\mu^{\ \mu} = -\left(1 - \frac{d}{2}\right)\Box\phi^2 \tag{22.5}$$

and vanishes only in $d = 2$, when the field ϕ is dimensionless.

However, we have the freedom to improve the action or the the stress tensor by appropriate terms which do not affect the equations of motion or the generators of the Poincaré algebra, respectively. We pursue the second approach way in problem 21.??. Here we ask if we can make the scalar action Weyl invariant, while retaining the usual equation of motions in Minkowski space.

The latter constraint is taken into account, if we modify only the coupling of the scalar field to gravity, i.e. if we abandon the substitution rule $\{\partial_\mu, \eta_{\mu\nu}, \mathrm{d}^4x\} \to \{\nabla_\mu, g_{\mu\nu}, \mathrm{d}^4x\sqrt{|g|}\}$.

[1] H. Weyl suggested first 1918 the combined scale transformation $\tilde{g}_{\mu\nu}(x) = \mathrm{e}^{2\omega(x)}g_{\mu\nu}(x)$ and $\tilde{A}_\mu(x) = \mathrm{e}^{\omega(x)}A_\mu(x)$ in an attempt to unify the gravitational and electromagnetic field.

While this rule is implied by the strong equivalence principle, there are two reasons to expect deviations: First, we know already that even within Einstein gravity non-zero torsion can exist. In this case, it is not possible to eliminate locally all effects of gravity by introducing a local inertial system. Second, we expect that quantum corrections will add all renormalisable coupling terms between the scalar and the gravitational field, even if we set them to zero at tree-level. Thus we should ask ourselves which additional renormalisable coupling terms between the scalar and the gravitational field exist. Since $[R] = m^2$ and thus also $[R_{\mu\nu}] = [R_{\mu\nu\rho\sigma}] = m^2$, the only dimensionless additional interaction term is a linear coupling of the curvature scalar R to ϕ^2. Such a coupling $R\phi^2/2$ acts as a curvature dependent, additional mass term for the scalar field. Now we ask how the action[2]

$$S = \int \mathrm{d}^d x \sqrt{|g|} \left[\frac{1}{2} g^{\alpha\beta} \partial_\alpha \phi \partial_\beta \phi - \frac{1}{2}(m^2 - \xi R)\phi^2 \right] , \tag{22.6}$$

where ξ parametrises the coupling to the curvature scalar, transforms under a conformal transformation of the metric,

$$g_{\mu\nu}(x) \to \tilde{g}_{\mu\nu}(x) = \Omega^2(x) g_{\mu\nu}(x) . \tag{22.7}$$

From our discussion of scale transformations in Minkowski space, we know that a bosonic field in d space-time dimensions scales as

$$\phi(x) \to \tilde{\phi}(x) = \Omega^{(2-d)/2}(x)\phi(x) = \Omega^D(x)\phi(x) . \tag{22.8}$$

This leads for $d > 2$ clearly to a non-trivial transformation of the kinetic term which has to be compensated by the non-trivial transformation of the scalar curvature term. The transformation (22.7) implies

$$g^{\mu\nu} \to \tilde{g}^{\mu\nu} = \Omega^{-2} g^{\mu\nu} \quad \text{and} \quad \sqrt{|g|} \to \sqrt{\tilde{g}} = \Omega^d \sqrt{|g|} . \tag{22.9}$$

As a result, the action S_0 obtained setting $\xi = 0$ changes as

$$S_0 \to \tilde{S}_0 = \frac{1}{2} \int \mathrm{d}^d x \sqrt{|g|}\, \Omega^d \left[\Omega^{-2} g^{\alpha\beta} \nabla_\alpha (\Omega^D \phi) \nabla_\beta (\Omega^D \phi) - m^2 \Omega^{2D} \phi^2 \right] \tag{22.10}$$

or

$$\begin{aligned} \tilde{S}_0 = \frac{1}{2} \int \mathrm{d}^d x \sqrt{|g|} \Big[& g^{\alpha\beta} \nabla_\alpha \phi \nabla_\beta \phi + 2\Omega^{d-2+D} g^{\alpha\beta} \nabla_\alpha \phi (\nabla_\beta \Omega^D) \phi \qquad (22.11a) \\ & + \Omega^{d-2} \phi^2 g^{\alpha\beta} (\nabla_\alpha \Omega^D)(\nabla_\beta \Omega^D) - m^2 \Omega^{2D} \phi^2 \Big] \\ = \frac{1}{2} \int \mathrm{d}^d x \sqrt{|g|} \Big[& g^{\alpha\beta} \partial_\alpha \phi \partial_\beta \phi - D(\Box\omega)\phi^2 + D^2 (\nabla\omega)^2 \phi^2 - m^2 \Omega^{2D} \phi^2 \Big] . \qquad (22.11b) \end{aligned}$$

Taking into account the transformation rule (22.2c) for the scalar curvature R,

$$\sqrt{|g|}\, R\phi^2 \to \sqrt{|g|}\, [R - 2(d-1)\Box\omega - (d-1)(d-2)\nabla^\mu \omega \nabla_\mu \omega]\phi^2 , \tag{22.12}$$

we find that the scalar action is invariant choosing $m = 0$ and

$$\xi = \xi_d \equiv \frac{d-2}{4(d-1)} . \tag{22.13}$$

A scalar field with $\xi = \xi_d$ is called conformally coupled to gravity, while the choice $\xi = 0$ is called minimally coupled to gravity. Note that for $d = 2$, i.e. when ϕ is dimensionless and does not scale under conformal transformations, minimal and conformal coupling agree.

[2] Since $g_{\alpha\beta}$ and $g^{\alpha\beta}$ transform inversely, we have to distinguish between $\partial_\mu \phi$ and $\partial^\mu \phi$. Convention is to use only $g^{\alpha\beta}$ and to write all fields and derivatives with lower indices.

Scalar field equation in a FLRW background Next we derive the equation of motion for a scalar field with arbitrary ξ in an expanding universe described by the flat FLRW metric. Choosing as coordinates $\{t, \boldsymbol{x}\}$, we have $g_{\mu\nu} = \mathrm{d}iag(1, -a^2, -a^2, -a^2)$, $g^{\mu\nu} = \mathrm{d}iag(1, -a^{-2}, -a^{-2}, -a^{-2})$, and $\sqrt{|g|} = a^3$. Varying the action

$$S = \int \mathrm{d}^4x\, a^3 \left\{ \frac{1}{2}\dot{\phi}^2 - \frac{1}{2a^2}(\boldsymbol{\nabla}\phi)^2 - V(\phi) \right\} \tag{22.14}$$

gives

$$\delta S = \int \mathrm{d}^4x\, a^3 \left\{ \dot{\phi}\delta\dot{\phi} - \frac{1}{a^2}(\boldsymbol{\nabla}\phi)\cdot\delta(\boldsymbol{\nabla}\phi) - V_{,\phi}\delta\phi \right\} \tag{22.15a}$$

$$= \int \mathrm{d}^4x \left\{ -\frac{\mathrm{d}}{\mathrm{d}t}(a^3\dot{\phi}) + a\boldsymbol{\nabla}^2\phi - a^3 V_{,\phi} \right\} \delta\phi \tag{22.15b}$$

$$= \int \mathrm{d}^4x\, a^3 \left\{ -\ddot{\phi} - 3H\dot{\phi} + \frac{1}{a^2}\boldsymbol{\nabla}^2\phi - V_{,\phi} \right\} \delta\phi\,, \tag{22.15c}$$

setting $V_{,\phi} \equiv \mathrm{d}V/\mathrm{d}\phi$. Thus the Klein-Gordon equation for a scalar field with the potential $V(\phi) = (m^2 + \xi R)\phi^2/2$ in a flat FLRW background is

$$\ddot{\phi} + 3H\dot{\phi} - \frac{1}{a^2}\boldsymbol{\nabla}^2\phi + (m^2 + \xi R)\phi = 0\,. \tag{22.16}$$

The term $3H\dot{\phi}$ acts in an expanding universe as a friction term for the oscillating ϕ field. Moreover, the gradient of ϕ is also suppressed for increasing a; this term can be therefore often neglected.

Next we want to rewrite this equation as the one for an harmonic oscillator with a time-dependent oscillation frequency. We introduce first the conformal time $\mathrm{d}\eta = \mathrm{d}t/a$.

Remark 22.2: For a power-law like behaviour of the scale factor, $a(t) \propto t^p$, the conformal time η evolves as $\eta = \int \mathrm{d}t\, t^{-p} \propto t^{1-p}$ and thus $a(\eta) \propto \eta^{\frac{p}{1-p}}$. In particular, η scales as $a(\eta) \propto \tau$ in the radiation dominated ($p = 1/2$) and as $a(\eta) \propto \tau^2$ in the matter dominated epoch ($p = 2/3$). For a de Sitter phase, $a(t) \propto \mathrm{e}^{Ht}$, we obtain

$$\eta = \int \mathrm{d}t\, \mathrm{e}^{-Ht} = -H^{-1}\mathrm{e}^{-Ht} + \eta_0 = -(aH)^{-1} + \eta_0\,.$$

Setting $\eta_0 = 0$ results in $a(\eta) = -1/(H\eta)$. Inflation is defined as the phase in the early universe with an accelerated expansion, $\ddot{a} > 0$. Using the convention $\eta_0 = 0$, inflation ends at $\eta = 0$, followed by the standard big bang evolution for $\eta > 0$.

Then we change the derivatives of the field,

$$\dot{\phi} = \frac{\mathrm{d}\phi}{\mathrm{d}t} = \frac{\mathrm{d}\phi}{\mathrm{d}\eta}\frac{\mathrm{d}\eta}{\mathrm{d}t} = \frac{1}{a}\phi'\,, \quad and \quad \ddot{\phi} = \frac{1}{a}\frac{\mathrm{d}}{\mathrm{d}\eta}\left(\frac{1}{a}\phi'\right) = \frac{1}{a^2}\phi'' - \frac{a'}{a^3}\phi'\,, \tag{22.17}$$

and express also the Hubble parameter as function of η,

$$H = \frac{\dot{a}}{a} = \frac{a'}{a^2} \equiv \frac{\mathscr{H}}{a}\,. \tag{22.18}$$

Inserting these expressions into Eq. (22.16) and multiplying with a^2 gives

$$\phi'' + 2\mathscr{H}\phi' - \nabla^2\phi + a^2 V_{,\phi} = 0\,. \tag{22.19}$$

Performing then a Fourier transformation, $\phi(\eta, \boldsymbol{x}) = \sum_k \phi_k(\eta) \mathrm{e}^{\mathrm{i}\boldsymbol{kx}}$, we obtain

$$\phi_k'' + 2H\phi_k' + [k^2 + (m^2 + \xi R)a^2]\phi_k = 0\,. \tag{22.20}$$

Note that $\boldsymbol{k}$ is the comoving wave-number. Since the proper distance varies as $\boldsymbol{x} \propto a$, the physical momentum is $\boldsymbol{p} = \boldsymbol{k}/a$.

Finally, we can eliminate the friction term $2H\phi_k'$ by introducing the auxiliary field $\chi_k(\eta) = a(\eta)\phi_k(\eta)$. Then we obtain a harmonic oscillator equation for χ_k,

$$\chi_k'' + \omega_k^2 \chi_k = 0\,, \tag{22.21}$$

with the time-dependent frequency

$$\omega_k^2(\eta) = k^2 + (m^2 + \xi R)a^2 - \frac{a''}{a}\,. \tag{22.22}$$

For a massless, conformally coupled ($\xi = 1/6$) scalar field, the frequency is independent of the expansion of the universe, $\omega_k^2(\eta) = k^2$, cf. problem 21.**??**.

Now we choose the special case of a de Sitter universe as approximation for the inflationary phase of the early universe. Moreover, we consider a minimally coupled scalar field with negligible mass. Combining then $a = -1/(H\eta)$ and $a'' = -2/(H\eta^3)$, or

$$\frac{a''}{a} = \frac{2}{\eta^2}\,, \tag{22.23}$$

the wave equation simplifies to

$$\chi_k'' + \left(k^2 - \frac{2}{\eta^2}\right)\chi_k = 0\,. \tag{22.24}$$

We examine first the short and the long-wavelength limit. In the first case, $k \gg |1/\eta|$, the field equation is conformally equivalent to the one in Minkowski space, with solution

$$\chi_k(\eta, \boldsymbol{x}) = \frac{1}{\sqrt{2k}}(A_k e^{-\mathrm{i}kx} + B_k \mathrm{e}^{\mathrm{i}kx})\,. \tag{22.25}$$

Here we factored out a normalisation factor $1/\sqrt{2k} \equiv 1/\sqrt{2\omega_k}$. With $\eta = -1/(aH)$, we can rewrite the short-wavelength condition as $|k|/a \gg H^{-1}$. Thus the comoving wavelength of the particle is much shorter than the comoving Hubble radius aH, or equivalently, its physical wavelength is much shorter than the Hubble radius. Therefore these solutions are called subhorizon modes.

In the opposite limit, we find $a''\chi_k = a\chi_k''$ which has as growing solution $\chi_k \propto a$ and thus $\phi_k = const.$ Thus modes with wavelengths larger than the horizon are “frozen in” and do not oscillate. They are called superhorizon modes. The complete solution is given by Hankel functions $H_{3/2}(\eta)$,

$$\chi_k(\eta, \boldsymbol{x}) = A_k \frac{\mathrm{e}^{-\mathrm{i}kx}}{\sqrt{2k}}\left(1 - \frac{\mathrm{i}}{k\eta}\right) + B_k \frac{\mathrm{e}^{\mathrm{i}kx}}{\sqrt{2k}}\left(1 + \frac{\mathrm{i}}{k\eta}\right)\,. \tag{22.26}$$

We could now set out for the quantisation of the scalar field χ. If we ignore the time-dependence and quantise the scalar field with the time-dependent mass term (22.22) in the standard way, we will obtain different vacua and different Fock spaces at different times t. As a result, a state which was empty at time t will contain in general particles at time t'. Thus the time-dependent gravitational field can excite modes χ_k, supplying energy and leading to particle production. We will postpone the quantisation of a scalar field in a FLRW metric to the next chapter, where this equation will play a prominent role. Before that we will introduce first some formalism needed and discuss two conceptionally simpler examples.

22.2. Quantisation in curved space-times

Similarly as in Minkowski space, we can use both canonical quantisation or the path integral approach to quantise classical field theories in curved space-times. The latter approach is particularly useful, if we are interested in quantum corrections to the stress tensor: Its expectation value for the quantum field ϕ in the background of a classical gravitational field $g^{\mu\nu}$ is

$$\langle T_{\mu\nu}\rangle = \frac{\int \mathcal{D}\phi\, T_{\mu\nu}\, \mathrm{e}^{\mathrm{i}S[\phi,g^{\mu\nu}]}}{\int \mathcal{D}\phi\, \mathrm{e}^{\mathrm{i}S[\phi,g^{\mu\nu}]}}\,. \tag{22.27}$$

Note that now the gravitational field $g^{\mu\nu}$ plays the usual role of a classical source. Inserting the definition (7.50) of the dynamical stress tensor and recalling that the denominator in (22.27) is the generating functional $Z = \exp(\mathrm{i}W)$ leads to

$$\langle T_{\mu\nu}\rangle = \frac{1}{Z[g^{\mu\nu}]}\,\frac{2}{\mathrm{i}\sqrt{|g|}}\,\frac{\delta}{\delta g^{\mu\nu}}\,Z[g^{\mu\nu}] = \frac{2}{\sqrt{|g|}}\,\frac{\delta W[g^{\mu\nu}]}{\delta g^{\mu\nu}}\,. \tag{22.28}$$

Having calculated $\langle T_{\mu\nu}\rangle$, one could aim at solving the Einstein equations in the semi-classical limit, replacing $T_{\mu\nu}$ by $\langle T_{\mu\nu}\rangle$. In this way, one takes into account two effects: First, the gravitational background can produce particles. Second, it changes the zero-point energies of the ϕ vacuum, analogous to the Casimir effect or vacuum polarisation.

One of the main advantages of this approach is that it is based on a local quantity, $\langle T_{\mu\nu}\rangle = \langle T_{\mu\nu}(x)\rangle$. Thus if we can show in a specific frame that e.g. $\langle T_{\mu\nu}(x)\rangle = 0$, then any observer will agree on that. By contrast, we will see that the expectation value $\langle\tilde{0}|N|\tilde{0}\rangle$ for the number of particles measured in a specific vacuum $|\tilde{0}\rangle$ depends on the trajectory of the considered detector. Therefore the concept of particle number and production is not a local one. This implies in particular that we can address the question where or when a particle is created only in an approximate way. An essential ingredient in both approaches is the vacuum state to be used in the calculation of expectation values. We will therefore concentrate first on this question, and then apply the simpler canonical quantisation formalism to specific examples.

22.2.1. Bogolyubov transformations

Ambiguity of the vacuum A basic ingredient of the canonical quantisation procedure is to split the fields into positive frequencies propagating forward in time, and negative frequencies propagating backwards. Since we associate annihilation operators with negative frequency modes and creation operators with positive frequency modes, this splitting defines the vacuum. As Minkowski space contains a time-like Killing vector field ∂_t which has as eigenfunctions the modes $\mathrm{e}^{-\mathrm{i}\omega t}$ with positive eigenvalues ω, the vacuum is invariant under

Lorentz transformation: All observers in inertial frames agree on the choice of the vacuum and thus also on one and many-particle states.

In curved space-times no inertial system can be globally extended to cover the whole manifold. No unique definition of the vacuum is possible and thus the notion of particle number becomes observer dependent which in turn implies the creation of particles. Using the correspondence $g^{\mu\nu}(x) \to J(t)$, we can illustrate this behaviour with the simple example of a harmonic oscillator which is driven during a finite time interval $0 < t < T$ by an external force $J(t)$.

Example 22.1: Excitation of a driven harmonic oscillator.
In the notation of chapter 2.4, the Hamiltonian of a driven harmonic oscillator is given by

$$H(\phi,\pi) = \frac{1}{2}\pi^2 + \frac{1}{2}\omega^2\phi^2 - J\phi\,. \tag{22.29}$$

We assume that the classical external source $J(t)$ acts only in the finite time interval $0 < t < T$. We keep the definition (2.58) of the annihilation and creation operators, from which we find now as equation of motion

$$\dot{a} = -\mathrm{i}\omega a + \frac{\mathrm{i}}{2\omega}\,J(t)\,. \tag{22.30}$$

Specifying the initial value before we apply the external force as $a(t=0) \equiv a_{\mathrm{i}n}$ results in

$$a(t) = a_{\mathrm{i}n}\mathrm{e}^{-\mathrm{i}\omega t} + \frac{\mathrm{i}}{2\omega}\int_0^t \mathrm{d}t'\,J(t')\mathrm{e}^{-\mathrm{i}\omega t'}\,. \tag{22.31}$$

For $t > T$, we set $a(t) \equiv \mathrm{e}^{-\mathrm{i}\omega t}a_{out} \equiv \mathrm{e}^{-\mathrm{i}\omega t}(a_{\mathrm{i}n} + J_0)$. Our aim is to express the in-states in term of the out-states. We set $|0_{\mathrm{i}n}\rangle = \sum_{n=0}^{\infty} c_n\,|n_{out}\rangle$, where the coefficients c_n have to be determined. Acting with a_{out} on the in-vacuum gives

$$a_{out}\,|0_{\mathrm{i}n}\rangle = \sum_{n=0}^{\infty} c_n a_{out}\,|n_{out}\rangle = \sum_{n=0}^{\infty}\sqrt{n}c_n\,|(n-1)_{out}\rangle = \sum_{n=0}^{\infty}\sqrt{n+1}c_{n+1}\,|n_{out}\rangle\ , \tag{22.32}$$

where we relabelled $n \to n+1$ in the last step. On the other hand, applying $a_{\mathrm{in}} + J_0 = a_{\mathrm{o}ut}$ on the in-vacuum results in

$$(a_{\mathrm{in}} + \tilde{J})\,|0_{\mathrm{in}}\rangle = J_0\,|0_{\mathrm{in}}\rangle = \sum_{n=0}^{\infty} J_0 c_n\,|n_{\mathrm{o}ut}\rangle\ . \tag{22.33}$$

Comparing the two expressions, we obtain $c_{n+1} = J_0 c_n/\sqrt{n+1}$ and thus $c_n = J_0^n\,c_0/\sqrt{n}!$. Requiring that the vacuum is normalised, we can determine c_0 as $|c_0| = \exp(-\frac{1}{2}|J_0|^2)$. Thus

$$|0_{\mathrm{in}}\rangle = \exp(-\frac{1}{2}|J_0|^2 + J_0 a_{\mathrm{o}ut}^\dagger)\,|0_{out}\rangle\ , \tag{22.34}$$

up to an undetermined phase: The driving force $J(t)$ converted the in-vacuum into a coherent state which contains all n-particle states with the amplitude c_n.

Our corresponding task in field theory is to find a mapping between field operators defined with respect to different vacua. The relation between the two sets of field operators is a special case of a Bogolyubov transformation. We will first discuss this transformation in general.

Bogolyubov transformation We defined a scalar product for solutions of the Klein-Gordon equation in Eq. (9.25). To simplify the notation, we normalise the solution in a box of finite volume, obtaining a discrete spectrum. Then the scalar product for plane-waves becomes

$$(\varphi_i, \varphi_j) = \delta_{ij}, \quad \text{a}nd \quad (\varphi_i^*, \varphi_j^*) = -\delta_{ij} \tag{22.35}$$

and zero otherwise,

$$(\varphi_i, \varphi_j^*) = (\varphi_i^*, \varphi_j) = 0\,. \tag{22.36}$$

If we quantise the scalar field (22.21) with its time-dependent frequency at different times t and $\tilde{t}$, we will obtain different vacua and different Fock spaces,

$$a_i\,|0\rangle = 0 \qquad \forall i\,, \qquad \tilde{a}_i\,|\tilde{0}\rangle = 0 \qquad \forall i\,.$$

We now search for the connection between the two sets of annihilation and creation operators. We can express the field at any time through the two sets of creation and annihilation operators,

$$\phi(x) = \sum_i \left[a_i\varphi_i(x) + a_i^\dagger\varphi_i^*(x)\right] = \sum_j \left[\tilde{a}_j\tilde{\varphi}_j(x) + \tilde{a}_j^\dagger\tilde{\varphi}_j^*(x)\right]\,. \tag{22.37}$$

Both sets of solutions, $\{\varphi_j(x), \varphi_j^*(x)\}$ and $\{\tilde{\varphi}_j(x), \tilde{\varphi}_j^*(x)\}$, form a complete basis. Thus we can decompose any basis vector $\tilde{\varphi}_j(x)$ as

$$\tilde{\varphi}_j(x) = \sum_i \left[\alpha_{ji}\varphi_i(x) + \beta_{ji}\varphi_i^*(x)\right]\,. \tag{22.38}$$

The unknown matrices α_{ij} and β_{ij} are called Bogolyubov coefficients. Using the orthogonality relations (22.35) and (22.36), we can determine the Bogolyubov coefficients as

$$(\varphi_k, \tilde{\varphi}_j) = \sum_i \left[\alpha_{ij}(\varphi_k, \varphi_i) + \beta_{ij}^\dagger(\varphi_k, \varphi_i^*)\right] = \alpha_{jk} \tag{22.39}$$

and $\beta_{jk} = -(\varphi_k^*, \tilde{\varphi}_j)$. In the reverse direction, we find in the same way

$$\varphi_i(x) = \sum_j \left[\alpha_{ji}^*\tilde{\varphi}_j(x) - \beta_{ji}^\dagger\tilde{\varphi}_j^*(x)\right]\,. \tag{22.40}$$

Rewriting Eq. (22.38) and its complex conjugated expression in matrix form gives

$$\begin{pmatrix} \tilde{\varphi} \\ \tilde{\varphi}^* \end{pmatrix} = \begin{pmatrix} \alpha & \beta \\ \beta^* & \alpha^* \end{pmatrix} \begin{pmatrix} \varphi \\ \varphi^* \end{pmatrix} = U \begin{pmatrix} \varphi \\ \varphi^* \end{pmatrix}\,. \tag{22.41}$$

Since both bases are orthonormal, the matrix U is unitary,

$$\alpha_{ij}\alpha_{jk}^* - \beta_{ij}\beta_{jk}^* = \delta_{ik}\,. \tag{22.42}$$

The transformation properties of the annihilation and creation operator follow as

$$\left(a, a^\dagger\right) = \left(\tilde{a}, \tilde{a}^\dagger\right) U\,. \tag{22.43}$$

If $\beta_{ij} \neq 0$, time-evolution mixes positive and negative frequency modes. As a result, the vacuum $|\tilde{0}\rangle$ evaluated with the number operator $N_i = a_i^\dagger a_i$ will contain particles,

$$\langle\tilde{0}|N_i|\tilde{0}\rangle = \sum_j |\beta_{ji}|^2 \,. \tag{22.44}$$

The corresponding energy density ρ in the continuum limit is

$$\rho_{\boldsymbol{k}} = \int \frac{\mathrm{d}^3 k'}{(2\pi)^3}\, \omega_{k'} |\beta_{\boldsymbol{k},\boldsymbol{k}'}|^2 \,. \tag{22.45}$$

The Bogolyubov coefficients β should therefore decrease faster than k^{-2} to ensure a finite energy density.

We will see that the presence of horizons lead to a thermal flux of particles. For a thermal spectrum with temperature T, the Bogolyubov coefficients have to satisfy the condition

$$|\alpha_{ji}|^2 = \mathrm{e}^{\omega_i/T} |\beta_{ji}|^2 \,. \tag{22.46}$$

Then the unitarity condition (22.42) of the Bogolyubov coefficients gives

$$\langle\tilde{0}|N_i|\tilde{0}\rangle = \sum_j |\beta_{ji}|^2 = \frac{1}{\mathrm{e}^{\omega_i/T} - 1} \,. \tag{22.47}$$

22.2.2. Choosing the vacuum state

Having set up the formalism of Bogolyubov transformations, we have to fill the formalism with physics: Since there is no unique vacuum, we have to decide case by case which is the physically relevant one. Additionally, we need a scheme for the calculation of the Bogolyubov coefficients. The first problem, the choice of a physically sensible vacuum, simplifies, if the space-time has "useful" symmetries. This includes in particular the case that the space-time is conformally flat. Space-times which approach asymptotically Minkowski space for $\eta \to \pm\infty$ are also useful toy-models: For such models, we can apply our standard formalism to construct the in and out Fock space. We study therefore first this case, using a model which has the further virtue of being analytically solvable.

Solvable model We illustrate now the production of particles in an expanding universe for a specific, exactly solvable model. In this two-dimensional model, the line-element is given by

$$\mathrm{d}s^2 = C(\eta)\left[\mathrm{d}\eta^2 - \mathrm{d}x^2\right] \tag{22.48}$$

and the scale factor changes as

$$C(\eta) \equiv a^2(\eta) = A + B\tanh(\rho\eta) \tag{22.49}$$

with $A > B$ and $\rho > 0$. For $\eta \to \pm\infty$, the scale factor approaches the constant value $C(\eta) \to A \pm B$ and thus the metric approaches asymptotically Minkowski space. In this limit we should be able to apply our standard approach to define vacua and one-particle states.

Using the separation ansatz $\phi_k(\eta, x) = \chi_k(\eta)\mathrm{e}^{\mathrm{i}kx}/(2\pi)^{1/2}$ in the scalar field equation with $\xi = 0$ gives

$$\chi_k'' + \left[k^2 + C(\eta)m^2\right]\chi_k = 0 \,. \tag{22.50}$$

Thus asymptotic in and out states have the frequency

$$\omega_{\rm in} = \sqrt{k^2 + m^2(A - B)} \tag{22.51}$$

$$\omega_{out} = \sqrt{k^2 + m^2(A + B)} \tag{22.52}$$

and approach

$$\chi_k(\eta) \to \frac{{\rm e}^{-{\rm i}(\omega_{\rm in}\eta - kx)}}{\sqrt{4\pi\omega_{\rm in}}} \quad and \quad \chi_k(\eta) \to \frac{{\rm e}^{-{\rm i}(\omega_{out}\eta - kx)}}{\sqrt{4\pi\omega_{out}}} \tag{22.53}$$

for $\eta \to \mp\infty$, respectively. The complete solutions are given by hypergeometric functions; we could find them either using a computer algebra program or performing the substitution $\xi = 1 + \tanh(\rho\eta)$. The important point is that one can relate the in and out solutions

$$\chi_k^{in}(\eta, x) = \alpha_k \chi_k^{out}(\eta, x) + \beta_k \chi_{-k}^{out*}(\eta, x) \tag{22.54}$$

using the linear transformation properties of these functions. The coefficients β_k determine the Bogolyubov coefficients, $\beta_{kk'} = \delta_{-k,k'}\beta_k$, and are given by

$$|\beta_k|^2 = \frac{\sinh^2(\pi\omega_-/\rho)}{\sinh(\pi\omega_{\rm in}/\rho)\,\sinh(\pi\omega_{out}/\rho)} \tag{22.55}$$

where $\omega_- = (\omega_{\rm in} - \omega_{out})/2$.

Using a variation of the Gamov criterion, particle creation should be controlled by the ratio of the expansion rate $H = \dot{C}/(2C) \sim \rho$ and ω: High-frequency modes (or subhorizon modes in the language of the previous section) with $\rho/\omega \ll 1$ should be not affected by the expansion and behave as in Minkowski space. Expanding $|\beta_k|^2$ for small ρ, we see that particle production is *exponentially* suppressed in this limit, $|\beta_k|^2 \to \exp(-2\pi\omega_-/\rho)$.

Adiabatic vacuum In cases of practical interest as the FLRW metric, the space-time does not approach asymptotically Minkowski space. In this case, we need an approximation scheme for the calculation of the Bogolyubov coefficients together with a suitable definition for the vacuum at an arbitrary intermediate time. Two particular choices for the vacuum state are the instantaneous and the adiabatic vacuum. The first one defines the vacuum as the state of lowest energy at each moment of time. This implies that one uses the time-dependent $\omega_k(\eta)$ in the usual Minkowski modes. This scheme over-predicts the effect of particle production, as the analogue of a (quantum) mechanical pendulum with variable length makes clear: The number of quanta E/ω is an invariant, if the length of the pendulum (i.e. ω) changes adiabatically [LL81]. This invariance is however not taken into account choosing the instantaneous vacuum.

The definition of the adiabatic vacuum is motivated by the requirement that high-frequency modes should not be affected by the expansion of the universe: In particular, the model (22.49) suggests that in this limit particle production is exponentially suppressed. Motivated by the WKB approximation, we express the positive mode functions as

$$\chi_k^+(\eta) = \frac{1}{\sqrt{W_k(\eta)}} \exp\left({\rm i}\int^\eta {\rm d}\eta\, W_k(\eta)\right), \tag{22.56}$$

while the negative modes functions are given by $\chi_k^-(\eta) = \chi_k^{+*}(\eta)$. We can implement the idea that high-frequency modes are unaffected by the expansion of the universe requiring the asymptotic expansion of W_k as

$$W_k(\eta) = \omega_k(\eta)\left[1 + \delta_2(\eta)\omega_k^{-2} + \delta_4(\eta)\omega_k^{-4} + \ldots\right]. \tag{22.57}$$

An adiabatic vacuum (of order n) is defined by choosing the initial conditions for the exact solution of $\chi_k'' + \omega_k^2\chi_k = 0$ such that only the positive modes (22.56) (with W_k evaluated at order δ_n) are present at the initial time η_0. This guaranties that particle production is exponentially suppressed in the adiabatic, high-frequency limit.

At lowest order in the asymptotic expansion, $W_k(\eta) = \omega_k(\eta)$, we can obtain a numerical approximation as follows: In general, the field modes can be expressed as

$$\chi_k(\eta) = \frac{\alpha_k(\eta)}{\sqrt{2\omega_k(\eta)}} \exp\left[-\mathrm{i}\int^\eta \mathrm{d}\eta\, \omega_k(\eta)\right] + \frac{\beta_k(\eta)}{\sqrt{2\omega_k(\eta)}} \exp\left[\mathrm{i}\int^\eta \mathrm{d}\eta\, \omega_k(\eta)\right], \tag{22.58}$$

where $\alpha_k(\eta)$ and $\beta_k(\eta)$ are the "instantaneous" Bogolyubov coefficient. The mode equation $\chi_k'' + \omega_k^2\chi_k = 0$ is satisfied, if the coefficients satisfy

$$\alpha_k' = \frac{\omega_k'}{2\omega_k} \exp\left(2\mathrm{i}\int \mathrm{d}\eta\, \omega_k\right)\beta_k \tag{22.59a}$$

$$\beta_k' = \frac{\omega_i'}{2\omega_k} \exp\left(-2\mathrm{i}\int \mathrm{d}\eta\, \omega_k\right)\alpha_k\,. \tag{22.59b}$$

Choosing the adiabatic vacuum at the time η_0 implies the initial conditions $\alpha_k(\eta_0) = 1$ and $\beta_k(\eta_0) = 0$. Neglecting a quadratic term, we obtain as closed expression for the Bogolyubov coefficients

$$\beta_k(\eta) \simeq \int^\eta \mathrm{d}\eta'\, \frac{\omega_k'}{2\omega_k} \exp\left(-2\mathrm{i}\int^{\eta'} \mathrm{d}\eta''\, \omega_k\right). \tag{22.60}$$

Conformal vacuum For a conformally flat space-time, we can connect the solutions of a conformally invariant theory to those of Minkowski space,

$$g_{\mu\nu}(x) = \Omega^{-2}(x)\eta_{\mu\nu}\,. \tag{22.61}$$

In the case of a conformally coupled scalar, the wave equation $(\Box - \xi_d R)\phi = 0$ becomes

$$\Box\tilde\phi = \eta_{\mu\nu}\partial_\mu\partial_\nu(\Omega^D\phi)\,, \tag{22.62}$$

where we used that in Minkowski space $\tilde R = 0$ and the scaling law (22.8) for bosonic fields. The field $\tilde\phi$ has the usual Minkowski Fourier modes,

$$\tilde\varphi_k(x) = [2\omega(2\pi)^{d-1}]^{-1/2}\mathrm{e}^{-\mathrm{i}kx} \tag{22.63}$$

which are eigenfunctions of the conformal Killing vector $\boldsymbol{\eta} = (1, \mathbf{0})$,

$$\frac{\partial}{\partial\eta}\tilde\varphi_k(x) = -\mathrm{i}\omega_{\boldsymbol{k}}\tilde\varphi_k(x) \tag{22.64}$$

with $\omega_{\boldsymbol{k}} > 0$. Reversing the scaling law (22.8), $\tilde\phi = \Omega^{-D}\phi$, we obtain the solution and the Green function in the space-time $(\mathcal{M}, g_{\mu\nu})$ as

$$\phi(x) = \Omega^{-D}(x)\sum_i \left[a_i\tilde\varphi_i(x) + a_i^\dagger\tilde\varphi_i^*(x)\right] \tag{22.65}$$

$$\Delta_F(x,x') = \Omega^{-D}(x)D_F(x,x')\Omega^{-D}(x')\,, \tag{22.66}$$

where the $\tilde{\varphi}_i$ are given by (22.63). The vacuum of the field ϕ defined by $a_i\,|0\rangle = 0$ is called conformal vacuum. The second relation for the propagator follows immediately from the definition of the Green function as time-ordered product of fields. Note also the correspondence to the variable substitution $u_k(\eta) = a(\eta)\phi_k(\eta) = \Omega(\eta)\phi_k(\eta)$ which transformed the scalar field equation in $d = 4$ into a Minkowskian form.

Since modes which are positive eigenmodes of Eq. (22.64) at one time remain positive for all times, no mixing between positive and negative frequency modes occurs. As result, particle creation of conformally invariant fields in a conformally flat space-time is absent. The exactly solvable model confirms this behaviour, since $m \to 0$ results in $\omega_- \to 0$ and no particle production for conformally coupled massless scalar occurs. Phenomenologically important cases of massless particles are the photon and the graviton. In the first case, the conformal invariance of the Maxwell equation implies that electromagnetic fields can not be generated during inflation, unless a mechanism which breaks gauge invariance and generates a mass term for the photon is invoked. In the case of the gravitational field $g_{\mu\nu}$, the scaling law $\tilde{g}_{\mu\nu} = \Omega^{(2-d)/2} g_{\mu\nu}$ of a bosonic field is in conflict with the conformal transformation law $\tilde{g}_{\mu\nu} = \Omega^2 g_{\mu\nu}$ for all d. Therefore gravitons are generated in an expanding universe.

22.3. Accelerated observers and the Unruh effect

Particle production can be divided in two different cases: In the first one, the space-time is time-dependent (e.g. via the scale factor $a(t)$) and can perform "work" and thus create particles. The second, perhaps more intriguing one, is the emission of a thermal spectrum of particles close to a horizon. We will consider in this section the second case, investigating the simplest case of an accelerated observer in Minkowski space.

Uniformly accelerated observer In the rest frame of an uniformly accelerated observer, the four-acceleration is given by $a^\alpha = \ddot{x}^\alpha = (0, \boldsymbol{a})$ with $|\boldsymbol{a}| = a = const.$ We can convert this condition into a covariant form, writing

$$\eta_{\alpha\beta}\ddot{x}^\alpha\ddot{x}^\beta = -a^2 \,. \tag{22.67}$$

In order to determine the trajectory $x^\alpha(\tau)$ of the accelerated observer, it is convenient to change to light-cone coordinates,

$$u = t - x \quad and \quad v = t + x \,. \tag{22.68}$$

Here, we assume that the trajectory is contained in the t-x plane; in the following we will suppress the transverse coordinates y and z. Forming the differentials $\mathrm{d}u$ and $\mathrm{d}t$, we see that the line-element in the new coordinates is $\mathrm{d}s^2 = \mathrm{d}u\mathrm{d}v$. The normalisation condition $\eta_{\alpha\beta}\dot{x}^\alpha\dot{x}^\beta = 1$ of the four-velocity becomes therefore $\dot{u}\dot{v} = 1$, while the acceleration equation (22.67) results in $\ddot{u}\ddot{v} = -a^2$. Differentiating then $\dot{u} = 1/\dot{v}$, we obtain $\ddot{u} = -\ddot{v}/\dot{v}^2$ or

$$\frac{\ddot{v}}{\dot{v}} = \pm a \,. \tag{22.69}$$

Integrating results in

$$v(\tau) = \frac{A}{a}\exp(a\tau) + C \tag{22.70}$$

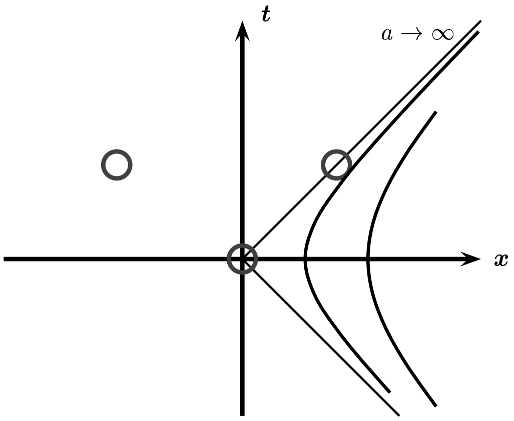


Figure 22.1.: Trajectories of uniformly accelerated observers together with the horizon and some vacuum fluctuations.

and, using $\dot{u} = 1/\dot{v}$, in

$$u(\tau) = -\frac{1}{Aa}\exp(-a\tau) + D\,. \tag{22.71}$$

Going back to the original Cartesian coordinates, we obtain

$$t(\tau) = \frac{1}{a}\sinh(a\tau) \quad \text{a}nd \quad x(\tau) = \frac{1}{a}\cosh(a\tau)\,, \tag{22.72}$$

where we set the integration constants $A = 1$ and $C = D = 0$ which selects the trajectory with $t(0) = 0$ and $x(0) = 1/a$. Two trajectories for finite acceleration a are shown together with the limiting curve $a \to \infty$ in Fig. 22.1.

Exponential redshift We will discuss later gravitational particle production as the effect of a non-trivial Bogolyubov transformation between different vacua. Before we apply this formalism, we will discuss the basis of this physical phenomenon in a classical picture. As a starter, we want to derive the formula for the relativistic Doppler effect. Consider an observer who is moving with constant velocity v relative to the Cartesian inertial system $x^\mu = (t, x)$ where we neglect the two transverse dimensions. We can parametrise the trajectory of the observer as

$$x^\mu(\tau) = (t(\tau), x(\tau)) = (\tau\gamma, \tau\gamma v)\,, \tag{22.73}$$

where γ denotes its Lorentz factor. A monochromatic wave of a scalar, massless field $\phi(k) \propto \exp[-\mathrm{i}\omega(t-x)]$ will be seen by the moving observer as

$$\phi(\tau) \equiv \phi(x^\mu(\tau)) \propto \exp\left[-\mathrm{i}\omega\tau\left(\gamma - \gamma v\right)\right] = \exp\left[-\mathrm{i}\omega\tau\sqrt{\frac{1-v}{1+v}}\right]\,. \tag{22.74}$$

Thus this simple calculation reproduces the usual Doppler formula, where the frequency ω of the scalar wave is shifted as

$$\omega' = \sqrt{\frac{1-v}{1+v}}\,\omega\,. \tag{22.75}$$

Next we apply the same method to the case of an accelerated observer. Then $t(\tau) = a^{-1}\sinh(a\tau)$ and $x(\tau) = a^{-1}\cosh(a\tau)$. Inserting this trajectory again into a monochromatic wave with $\phi(k) \propto \exp(-\mathrm{i}\omega(t-x)$ gives now

$$\phi(\tau) \propto \exp\left[-\frac{\mathrm{i}\omega}{a}\left[\sinh(a\tau) - \cosh(a\tau)\right]\right] = \exp\left[\frac{\mathrm{i}\omega}{a}\exp(-a\tau)\right] \equiv \mathrm{e}^{-\mathrm{i}\vartheta}\,. \tag{22.76}$$

Thus an accelerated observer does not see a monochromatic wave, but a superposition of plane waves with varying frequencies. Defining the instantaneous frequency by

$$\omega(\tau) = \frac{\mathrm{d}\vartheta}{\mathrm{d}\tau} = \omega\,\exp(-a\tau)\,, \tag{22.77}$$

we see that the phase measured by the accelerated observer is exponentially redshifted. As next step, we want to determine the power spectrum $P(\nu) = |\phi(\nu)|^2$ measured by the observer, for which we have to calculate the Fourier transform $\phi(\nu)$.

Example 22.2: Determine the Fourier transform of the wave $\phi(\tau)$.
Substituting $y = \exp(-a\tau)$ in

$$\phi(\nu) = \int_{-\infty}^{\infty} \mathrm{d}\tau \phi(\tau)\mathrm{e}^{\mathrm{i}\nu\tau} = \int_{-\infty}^{\infty} \mathrm{d}\tau \exp\left(\frac{\mathrm{i}\omega}{a}\exp(-a\tau)\right)\mathrm{e}^{\mathrm{i}\nu\tau} \tag{22.78}$$

gives

$$\phi(\nu) = \frac{1}{a}\int_0^{\infty} \mathrm{d}y\; y^{-\mathrm{i}\nu/a-1}\,\mathrm{e}^{\mathrm{i}(\omega/a)y}\,. \tag{22.79}$$

On the other hand, we can rewrite the integral representation (A.24) of the Gamma function as

$$\int_0^{\infty} \mathrm{d}t\, t^{z-1}\mathrm{e}^{-bt} = b^{-z}\,\Gamma(z) = \exp(-z\ln b)\,\Gamma(z) \tag{22.80}$$

for $\Re(z) > 0$ and $\Re(b) > 0$. Comparing these two expressions, we see that they agree setting $z = -\mathrm{i}\nu/a + \varepsilon$ and $b = -\mathrm{i}\omega/a + \varepsilon$. Here we added an infinitesimal positive real quantity $\varepsilon > 0$ to ensure the convergence of the integral. In order to determine the correct phase of b^{-z}, we have rewritten this factor as $\exp(-z\ln b)$ and have used

$$\ln b = \lim_{\varepsilon\to 0}\ln\left(-\frac{\mathrm{i}\omega}{a} + \varepsilon\right) = \ln\left|\frac{\omega}{a}\right| - \frac{\mathrm{i}\pi}{2}sign(\omega/a)\,. \tag{22.81}$$

Thus the Fourier transform $\phi(\nu)$ is given by

$$\phi(\nu) = \frac{1}{a}\left(\frac{\omega}{a}\right)^{\mathrm{i}\nu/a}\,\Gamma(-\mathrm{i}\nu/a)\mathrm{e}^{\pi\nu/(2a)}\,. \tag{22.82}$$

The Fourier transform $\phi(\nu)$ contains negative frequencies,

$$\phi(-\nu) = \phi(\nu)\mathrm{e}^{-\pi\nu/a} = \frac{1}{a}\left(\frac{\omega}{a}\right)^{\mathrm{i}\nu/a}\,\Gamma(-\mathrm{i}\nu/a)\mathrm{e}^{-\pi\nu/(2a)}\,. \tag{22.83}$$

Using the reflection formula of the Gamma function for imaginary arguments,

$$\Gamma(\mathrm{i}x)\Gamma(-\mathrm{i}x) = \frac{\pi}{x\sinh(\pi x)}\,, \tag{22.84}$$

we find the power spectrum at negative frequencies as

$$P(-\nu) = \frac{\pi}{a^2} \frac{\mathrm{e}^{-\pi\nu/a}}{(\nu/a)\sinh(\pi\nu/a)} = \frac{\beta}{\nu} \frac{1}{\mathrm{e}^{\beta\nu} - 1} \tag{22.85}$$

with $\beta = 2\pi/a$. Remarkably, the dependence on the frequency ω of the scalar wave—still present in the Fourier transform $\phi(\nu)$—has dropped from the negative frequency part of the power spectrum $P(-\nu)$ which corresponds to a thermal Planck law with temperature $T = 1/\beta = a/(2\pi)$.

The occurrence of negative frequencies is the classical analogue for the mixing of positive and negative frequencies in the Bogolyubov method. Therefore we expect that on the quantum level a uniformly accelerated detector will measure a thermal Planck spectrum with temperature $T = 1/\beta = a/(2\pi)$. This phenomenon is called Unruh effect and $T = a/(2\pi)$ the Unruh temperature.

Rindler space-time Recall that the trajectory of an accelerated observer is given by

$$t(\tau) = \frac{1}{a}\sinh(a\tau) \quad and \quad x(\tau) = \frac{1}{a}\cosh(a\tau)\,. \tag{22.86}$$

It describes one branch of the hyperbola $x^2 - t^2 = a^{-2}$, cf. Fig. 22.1.

Our aim is to determine the vacuum experienced by the uniformly accelerated observer. As first step, we have to find a frame $\{\xi, \chi\}$ comoving with the observer. In this frame, the observer is at rest, $\chi(\tau) = 0$, and the coordinate time ξ agrees with the proper time, $\xi = \tau$. Introducing comoving light-cone coordinates,

$$\tilde{u} = \xi - \chi \quad and \quad \tilde{v} = \xi + \chi\,, \tag{22.87}$$

these conditions become

$$\tilde{u}(\tau) = \tilde{v}(\tau) = \tau\,. \tag{22.88}$$

Moreover, we choose the comoving coordinates such that the metric is conformally flat,

$$\mathrm{d}s^2 = \Omega^2(\xi, \chi)(\mathrm{d}\xi^2 - \mathrm{d}\chi^2) = \Omega^2(\tilde{u}, \tilde{v})\mathrm{d}\tilde{u}\mathrm{d}\tilde{v}\,. \tag{22.89}$$

Next we have to relate the comoving coordinates $\{\tilde{u}, \tilde{v}\}$ to Minkowski coordinates $\{t, x\}$. Since $\mathrm{d}\tilde{u}^2$ and $\mathrm{d}\tilde{v}^2$ are missing in the line-element, the functions $u(\tilde{u}, \tilde{v})$ and $v(\tilde{u}, \tilde{v})$ can depend only on one of their two arguments. We can set therefore $u(\tilde{u})$ and $v(\tilde{v})$. Expressing the four-acceleration $\dot{u}$ as

$$\frac{\mathrm{d}u}{\mathrm{d}\tau} = \frac{\mathrm{d}u}{\mathrm{d}\tilde{u}}\frac{\mathrm{d}\tilde{u}}{\mathrm{d}\tau}\,, \tag{22.90}$$

inserting $\dot{u} = -au$ and $\dot{\tilde{u}} = 1$ we arrive at

$$-\,au = \frac{\mathrm{d}u}{\mathrm{d}\tilde{u}}\,. \tag{22.91}$$

Separating variables and integrating we end up with $u = C_1\mathrm{e}^{-a\tilde{u}}$. In the same way, we find $v = C_2\mathrm{e}^{a\tilde{v}}$. Since the line-element has to agree along the trajectory with the proper-time, $\mathrm{d}s^2 = \mathrm{d}\tau^2$, the two integration constants C_1 and C_2 have to satisfy the constraint

$-a^2C_1C_2 = 1$. Choosing $C_1 = -C_2$, the desired relation between the two sets of coordinates becomes

$$u = -\frac{1}{a}\,\mathrm{e}^{-a\tilde{u}} \quad and \quad v = \frac{1}{a}\,\mathrm{e}^{a\tilde{v}}\,, \tag{22.92}$$

or using Cartesian coordinates,

$$t = \frac{1}{a}\,\mathrm{e}^{a\xi}\sinh(a\xi) \quad and \quad x = \frac{1}{a}\,\mathrm{e}^{a\chi}\cosh(a\chi)\,. \tag{22.93}$$

The space-time described by the coordinates defining the comoving frame of the accelerated observer,

$$\mathrm{d}s^2 = \mathrm{e}^{2a\chi}(\mathrm{d}\xi^2 - \mathrm{d}\chi^2)\,, \tag{22.94}$$

is called Rindler space-time. It is locally equivalent to Minkowski space, but differs globally: If we vary the Rindler coordinates over their full range, $\xi \in \mathbb{R}$ and $\chi \in \mathbb{R}$, then we cover only the one fourth of Minkowski space with $x > |t|$. Thus for an accelerated observer an event horizon exist: Evaluating on a hypersurface of constant comoving time, $\xi = const.$, the physical distance from $\chi = -\infty$ to the observer placed at $\chi = 0$ gives

$$d = \int_{-\infty}^{0} \mathrm{d}\chi\sqrt{|g_{\chi\chi}|} = \frac{1}{a}\,. \tag{22.95}$$

This corresponds to the coordinate distance between the observer and the horizon in Minkowski coordinates.

Unruh effect We have found that the Fourier spectrum of a classical wave seen by an accelerated observer contains negative frequencies which exhibit a thermal spectrum. Now we want to discuss this phenomenon which is called the Unruh effect on the quantum level. In order to simplify the calculation, we consider the simplest case of a massless scalar field in 1+1 dimensions. Then the action is conformally invariant,

$$S = \frac{1}{2}\int \mathrm{d}t\mathrm{d}x\sqrt{|g|}\;\eta^{\alpha\beta}\nabla_\alpha\phi\nabla_\beta\phi = \frac{1}{2}\int \mathrm{d}\xi\mathrm{d}\chi\sqrt{|g|}\;g^{\alpha\beta}\nabla_\alpha\phi\nabla_\beta\phi \tag{22.96}$$

and the resulting wave equation has the same form for an an inertial observer using u, v and an accelerated observer using $\tilde{u}, \tilde{v}$ (light-cone) coordinates,

$$\frac{\partial^2\phi}{\partial u\partial v} = \frac{\partial^2\phi}{\partial\tilde{u}\partial\tilde{v}} = 0\,. \tag{22.97}$$

The corresponding solutions are

$$\phi(t,x) = f(u) + g(v) \qquad and \qquad \phi(\xi,\chi) = f(\tilde{u}) + g(\tilde{v})\,, \tag{22.98}$$

where f and g are arbitrary smooth functions specifying the wave form.

In the overlap region $x > |t|$, we can quantise the field using either set of coordinates,

$$\phi(x) = \int_0^\infty \frac{\mathrm{d}\omega}{\sqrt{(2\pi)2\omega}}\left[a_\omega\mathrm{e}^{-\mathrm{i}\omega u} + a_\omega^\dagger\mathrm{e}^{\mathrm{i}\omega u}\right] + \text{left-movers} \tag{22.99a}$$

$$= \int_0^\infty \frac{\mathrm{d}\Omega}{\sqrt{(2\pi)2\Omega}}\left[b_\Omega\mathrm{e}^{-\mathrm{i}\Omega\tilde{u}} + b_\Omega^\dagger\mathrm{e}^{\mathrm{i}\Omega\tilde{u}}\right] + \text{left-movers}\,. \tag{22.99b}$$

Here we wrote down explicitly only the right-moving modes: Because of $u(\tilde{u})$ and $v(\tilde{v})$, the two sets of modes propagate independently and we can consider them separately.

The vacuum defined by $b_i\,|0_{\rm R}\rangle = 0$ is called the Rindler vacuum, while the usual Minkowski vacuum is defined by $a_i\,|0_{\rm M}\rangle = 0$. One may wonder which one of the two vacua is the "better" one? First, the Rindler coordinates cover only part of the Minkowski space-time; they, and as a result also the Rindler vacuum $|0_{\rm R}\rangle$, are singular on the horizon: Using the two different representation for the field ϕ, its energy density follows as

$$\rho = \langle 0_M|\,(\partial_u\phi)^2\,|0_M\rangle = \langle 0_R|\,(\partial_{\tilde{u}}\phi)^2\,|0_R\rangle\ . \tag{22.100}$$

Now we compare the expectation value of $(\partial_u\phi)^2$ for the two vacua,

$$\rho = \langle 0_R|\,(\partial_u\phi)^2\,|0_R\rangle = \left(\frac{\partial\tilde{u}}{\partial u}\right)^2 \langle 0_R|\,(\partial_{\tilde{u}}\phi)^2\,|0_R\rangle = \frac{1}{(au)^2}\,\langle 0_M|\,(\partial_u\phi)^2\,|0_M\rangle\ . \tag{22.101}$$

Since the expectation value in the Minkowski vacuum is well-behaved, the Rindler vacuum diverges for $u \to 0$. More precisely, we see that the contribution of the left-movers to the energy density of the Rindler vacuum diverges at the future horizon $u = 0$. Similarly, the right-movers add an infinite energy density at the past horizon $v = 0$. While the Rindler vacuum is thus not able to describe physics close to the horizon, the corresponding set of field operators should be used to calculate the response of an uniformly accelerated detector to the Minkowski vacuum.

We express now in Eq. (22.99b) the operator b_Ω using the Bogolyubov relation, and compare then the coefficients of the positive frequency part,

$$\frac{1}{\sqrt{2\omega}}{\rm e}^{-{\rm i}\omega u} = \int_0^\infty \frac{{\rm d}\Omega'}{\sqrt{2\Omega'}}\left[{\rm e}^{-{\rm i}\Omega'\tilde{u}}\alpha_{\Omega'\omega} - {\rm e}^{-{\rm i}\Omega'\tilde{u}}\beta^*_{\Omega'\omega}\right]\ . \tag{22.102}$$

Next we multiply with ${\rm e}^{-{\rm i}\Omega\tilde{u}}$ and integrate over $\tilde{u}$. Performing then the trivial Ω' integral on the RHS, we arrive at

$$\int_0^\infty {\rm d}\tilde{u}{\rm e}^{\mp{\rm i}\omega u+{\rm i}\Omega\tilde{u}} = \begin{cases} \alpha_{\Omega\omega} \\ \beta_{\Omega\omega} \end{cases}, \tag{22.103}$$

which has the same form as (22.78). Hence the Bogolyubov coefficients satisfy the condition for a thermal spectrum,

$$|\alpha_{\Omega\omega}|^2 = \exp(2\pi\Omega/a)|\beta_{\Omega\omega}|^2\,. \tag{22.104}$$

The expectation value of the number operator valid for the accelerated observer in the Minkowski vacuum becomes

$$\left\langle \tilde{N}_\Omega \right\rangle = \langle 0_M|\,b^\dagger_\Omega b_\Omega\,|0_M\rangle = \int {\rm d}\omega|\beta_{\omega\Omega}|^2\,. \tag{22.105}$$

For $\omega = \Omega$, the normalisation condition (22.42) becomes in the continuum limit $\int_0^\infty {\rm d}\omega\left(|\alpha_{\Omega\omega}|^2 - |\beta_{\Omega\omega'}|^2\right) = \delta(0)$. Using also (22.104) we arrive at

$$\left\langle \tilde{N}_\Omega \right\rangle = \delta(0)\,\frac{1}{\exp(2\pi\Omega/a) - 1}\,. \tag{22.106}$$

Identifying the factor $\delta(0)$ as usually with the volume, we obtain for the number density of scalar particles detected by an accelerated observer in the Minkowski vacuum

$$\langle \tilde{n}_\Omega \rangle = \frac{1}{\exp(2\pi\Omega/a) - 1}\,. \tag{22.107}$$

Since energy is conserved in Minkowski space, you should be worried about this thermal flux measured in an accelerated detector: A tempting answer is that this energy is delivered by the external agent which accelerates the detector. Figure 22.1 suggests however a different interpretation: Minkowski space vacuum fluctuations that are crossing the horizon of an accelerated observer are experienced by the observer as real, thermal fluctuations: For instance, the vacuum fluctuation at $(t, x) = 0$ is seen for an accelerated observer as a real particle existing from $\xi = -\infty$ to $\xi = \infty$. Similarly, all other vacuum fluctuation crossing the horizon are interpreted as the creation and annihilation of real particles. Thus the horizon seems to be equipped with a thermal atmosphere, which temperature increases the closer an accelerated observer approaches it. In contrast, for any Minkowski observer these fluctuations are the usual “harmless” vacuum fluctuations.

We close this section with a remark on the topology of the Rindler and Minkowski space-times: Both space-times are flat and agree locally in their overlapping region. They differ only from a global point of view, and thus their topology should disagree. Performing a Wick rotation, the hyperbola $x^2 - t^2 = a^{-2}$ defining the horizon becomes a circle. Hence the Rindler space-time has the topology $S^1 \times \mathbb{R}^3$. We can view this periodicity as another indication for the presence of a thermal spectrum. Its temperature T equals the inverse of the circumference $2\pi/a$ of the event horizon in the compactified dimension.

Summary

The definition of the vacuum and of the number of particles in a space-time without time-like Killing vector field is ambiguous and dependent on the observer. Field operators defined with respect to different vacua are related by a Bogolyubov transformation; the coefficients relating positive and negative frequencies in two different vacua determine the amount of particle production. In space-times without time-like Killing vector field, all SM particles except photons and gluons are produced by the changing metric. In a space-time with event horizon, a thermal spectrum of particles is created close to the horizon.

Further reading

[MW07] give a pedagogical introduction of quantum effects in gravity, including path integral methods and the calculation of the conformal anomaly. Two exhaustive discussions of quantum field theory in curved space-time are the books of [BD82] and of [PT09].

23. Inflation

Observational data show that we live in a universe which is nearly flat, $|\Omega_k| \equiv |\Omega_{tot} - 1| \lesssim 0.1\%$. A flat universe however is an unstable fixed point of the time evolution of a radiation or matter dominated universe. This raises the question about the "naturalness" of the initial conditions of our universe: Is there any hope to explain why the universe is flat? A similar problem is posed by the formation of structures: Gravitational collapse can enhance initial fluctuations, but is not able to create them. Thus in an initially uniform FLRW universe, no structures like galaxies would evolve. A successful cosmological theory should answer therefore how the primordial fluctuations were created which have served as seeds for the observed structures today.

An inflationary phase in the early universe is a very successful attempt to explain these and other observations which are puzzling in the traditional big bang picture. Inflation is however not a specific theory, but resembles from the point of view of particle physics more a paradigm: Very different models can lead to an observationally indistinguishable inflationary phase in the early universe. In order to be specific, we will use in most of our discussion as prototype for inflation a single scalar field, the *inflaton*, equipped with one of our standard potentials, $V(\phi) = m^2\phi^2/2$ or $V(\phi) = \lambda\phi^4/4$.

23.1. Motivation for inflation

According to the standard big bang scenario, the universe evolved looking backwards to $t \to 0$ from a matter dominated into an radiation dominated epoch until it reached the initial singularity at $t = 0$. This standard big bang model leads to several shortcomings:

- *Causality* or *horizon problem:* Why are causally disconnected regions of the universe homogeneous, as the isotropy of the CMB shows? This problem arises because the (particle) horizon grows like the cosmic time t, but the scale factor a increases in the radiation or matter dominated epoch only as $t^{2/3}$ or $t^{1/2}$, respectively. Thus for any scale L contained today completely inside the horizon, there exists a time $t < t_0$ where L crossed the horizon. A solution to the horizon problem requires that a grows faster than t. Since $a \propto t^{2/[3(1+w)]}$, this demands $w < -1/3$ or $\ddot{a} > 0$, i.e. an accelerated expansion of the universe.

- *Flatness problem:* The curvature term in the Friedmann equation scales as k/a^2 and decreases thus slower than the matter ($\propto 1/a^3$) and radiation ($1/a^4$) terms. Let us rewrite the Friedmann equation as

$$\frac{k}{a^2} = H^2\left(\frac{8\pi G}{3H^2}\rho - 1\right) = H^2\left(\Omega_{\rm tot} - 1\right) . \tag{23.1}$$

The LHS scales as $(1 + z)^2$, the squared Hubble parameter in the matter dominated epoch as $(1+z)^3$, and in the radiation dominated epoch as $(1+z)^4$. Classical gravity is supposed to be valid until the Planck scale $M_{\mathrm{P}l}$. Most of time the universe was radiation dominated, so we can estimate the redshift at the Planck time as $1+z_{\mathrm{P}l} = (t_0/t_{\mathrm{P}l})^{1/2} \approx 10^{30}$. Thus if today the deviation from flatness is $|\Omega_{tot}-1| \lesssim 1\%$, then it had too be extremely small at the Planck time, $|\Omega_{tot}-1| \lesssim 10^{-2}/(1+z_{\mathrm{P}l})^2 \approx 10^{-62}$. Taking the time-derivative of

$$|\Omega_{tot}-1| = \frac{|k|}{H^2a^2} = \frac{|k|}{\dot{a}^2} \tag{23.2}$$

gives

$$\frac{\mathrm{d}}{\mathrm{d}t}\,|\Omega_{tot}-1| = \frac{\mathrm{d}}{\mathrm{d}t}\,\frac{|k|}{\dot{a}^2} = -\frac{2|k|\ddot{a}}{\dot{a}^3} < 0 \tag{23.3}$$

for $\ddot{a} > 0$. Hence $|\Omega_{tot}-1|$ increases, if the universe decelerates, i.e. $\dot{a}$ decreases, and decreases if the universe accelerates, i.e. $\dot{a}$ increases. Thus again a phase with $\ddot{a} > 0$ (or $w < -1/3$) may avoid this problem.

- *Magnetic monopole problem:* Grand unified theories (GUT) predict the existence of magnetic monopole with masses M_{GUT}/α_{GUT}. If they are produced via the Kibble mechanism during the GUT symmetry breaking, they would overclose the universe, cf. problem 20.**??**.

- The standard big bang model contains no source for the *initial fluctuations* required for structure formation.

Classical gravity breaks down around $t \approx t_{\mathrm{P}l}$ or $T \approx M_{\mathrm{P}l}$, and one may wonder if these problems can be avoided by an appropriate modification of gravity above the Planck scale. This possibility seems to be rather contrived, because for most predictions the time interval between the singularity and $t \sim t_{\mathrm{P}l}$ is negligible. Therefore setting $t_{\mathrm{P}l} = 0$ seems to be a good approximation, see also problem 24.**??**.

Solution by inflation Inflation is a modification of the standard big bang model where a phase of accelerated expansion in the very early universe is introduced. While the initial singularity in a big bang model happens at $t = 0$ or $\eta = 0$, an inflationary phase adds "additional" time at $\eta < 0$ (cf. the remark 23.2). During this phase, the universe behaves close to a de Sitter universe with $w = -1$, $H = const.$, and $a(t) = a_0 \exp(Ht)$. If this phase prevails long enough, it solves the horizon, flatness, and the monopole problem and generates as bonus density fluctuations:

- The exponential growth of the scale factor, $a(t_\mathrm{f})/a(t_\mathrm{i}) = \mathrm{e}^{(t_\mathrm{f}-t_\mathrm{i})H} \gg t_\mathrm{f}/t_\mathrm{i}$, blows up a small, at time t_i causally connected region to superhorizon scales.

- Similarly, the growth of the scale factor solves the flatness problem, driving $|\Omega_{tot}-1| \propto \mathrm{e}^{-2Ht}$ exponentially towards zero.

- Magnetic monopoles or other superheavy relics are diluted as $n \propto \exp(-3Ht)$. At the end of a sufficiently long inflationary phase, their density is therefore practically zero.

- Inflation blows-up quantum fluctuation to astronomical scales, generating initial fluctuation without scale, $P_0(k) \propto k^{-1+\varepsilon}$ with $\varepsilon \ll 1$, as required by observations.

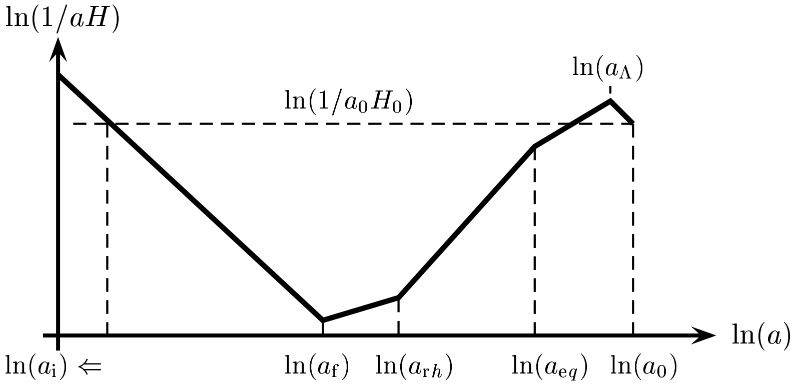


Figure 23.1.: Evolution of the present Hubble scale as function of time; the condition $(a_0H_0)^{-1} < (a_\mathrm{i}H_\mathrm{i})^{-1}$ sets a lower limit on $a_\mathrm{f}/a_\mathrm{i}$.

Figure 23.1 illustrates how the comoving Hubble scale $1/(aH)$ evolves in a universe with an inflationary phase. During inflation, $1/(aH)$ decreases, because $\ddot{a} > 0$ is equivalent to

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{1}{aH}\right) = \frac{\mathrm{d}}{\mathrm{d}t}(\dot{a})^{-1} = -\frac{\ddot{a}}{(\dot{a})^2}\,. \tag{23.4}$$

More precisely, $w \simeq -1$ and $H \simeq const.$ imply that a increases exponentially. Thus the comoving Hubble radius decreases as $\ln(1/aH) = -\ln(a) + const.$ during inflation: As a result, physical scales k which had been previously on superhorizon distance come in causal contact. Inflation ends at a_f, and the comoving Hubble scale starts to grow. During the intermediate phase $a_\mathrm{f} < a < a_{\mathrm{r}h}$ the universe goes through a phase of reheating before the standard big bang evolution starts at $a > a_{\mathrm{r}h}$. The value of $a_{\mathrm{r}h}$ is determined by the temperature to which the universe is reheated and is very model dependent.

Conditions for successful inflation We can rewrite the condition for accelerated expansion $\ddot{a} > 0$ as

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{1}{aH}\right) = -\frac{\dot{a}H + a\dot{H}}{(aH)^2} = -\frac{1}{a}(1-\varepsilon_H) \qquad \text{with} \qquad \varepsilon_H \equiv -\frac{\dot{H}}{H^2}\,. \tag{23.5}$$

The condition $\varepsilon_H \ll 1$ ensures that w is close to -1 (problem 24.**??**). Next we define as measure for the expansion the number N of e-foldings during inflation, $N = \ln(a_2/a_1)$. Using then $\mathrm{d}N = \mathrm{d}\ln(a) = H\mathrm{d}t$, we can express ε_H as

$$\varepsilon_H = -\frac{\dot{H}}{H^2} = -\frac{\mathrm{d}\ln H}{\mathrm{d}N}\,. \tag{23.6}$$

The relative change of ε_H should be small per Hubble time H^{-1} so that inflation can persist during a sufficiently large number of e-foldings. This motivates us to introduce as a second parameter

$$\eta_H \equiv \frac{\mathrm{d}\ln\varepsilon_H}{\mathrm{d}N} = \frac{1}{H}\frac{\dot{\varepsilon}_H}{\varepsilon_H}\,, \tag{23.7}$$

which should be also small, $|\eta_H| \ll 1$. Thus we can quantify the conditions for successful inflation by $\varepsilon_H \ll 1$ and $|\eta_H| \ll 1$: These two conditions guaranty that the time evolution of the scale factor is for a sufficiently long time close to the one in a de Sitter universe.

Example 23.1: How much inflation is needed?
We can find the minimal number of e-folding to solve the horizon problem by comparing the comoving Hubble radius today to the one at the beginning of inflation (cf. Fig. 23.1),

$$(a_0H_0)^{-1} < (a_\mathrm{i}H_\mathrm{i})^{-1} \,. \tag{23.8}$$

We neglect the reheating phase and approximate the complete evolution of the universe for $a > a_{rh}$ as radiation dominated. With $H \propto 1/a^2$, we find

$$\frac{a_0H_0}{a_\mathrm{f}H_\mathrm{f}} = \frac{a_0}{a_\mathrm{f}}\left(\frac{a_\mathrm{f}}{a_0}\right)^2 = \frac{a_\mathrm{f}}{a_0} = \frac{T_0}{T_\mathrm{f}} \approx 10^{-28} \,, \tag{23.9}$$

where we used $T_\mathrm{f} = 10^{15}\,\mathrm{GeV}$ as temperature of the universe at beginning of the standard big bang evolution for the numerical estimate. Thus

$$(a_\mathrm{i}H_\mathrm{i})^{-1} > (a_0H_0)^{-1} \sim 10^{28}(a_\mathrm{f}H_\mathrm{f})^{-1} \,. \tag{23.10}$$

With $H_i \approx H_f$, $N \approx 65$ e-foldings are required to solve the horizon problem for our choice of T_f.

23.2. Inflation in the homogeneous limit

We assume in the following that a single scalar field, the *inflaton*, is responsible for the inflationary phase in the early universe. We know that displacing the minimum of a scalar potential from zero generates a cosmological constant, which in turn leads to accelerated expansion. In a successful inflationary model, we have to convert this static picture into a dynamical process, since inflation has to start and to stop. In particular, the end of inflation has to be successfully connected to the smooth big bang picture, which is called often the "graceful exit problem."

Recall also our discussion of scalar fields with non-zero vev in chapter 13. There we split the scalar field $\phi(x)$ into a classical part $\langle\phi\rangle \equiv \phi_0$ and quantum fluctuations $\delta\phi(x)$ on top of it. Now we consider the more general case where also the classical field depends on time,

$$\phi(x) = \phi_0(t) + \delta\phi(x) \,, \tag{23.11}$$

and its dynamics is governed by the Einstein equations. Clearly, this problem can not be solved in general. Observations of the CMB show however that the early universe was very homogeneous, $\delta T/T \sim 10^{-5}$, and thus we can expect that lowest order perturbation theory in $\delta\phi(x)$ should work reliable. We considered already the evolution of $\phi_0(t)$ in a FLRW background. Now we analyse under which conditions this evolution leads successfully to inflation, before we calculate in the next section the fluctuations $\delta\phi(x)$.

Equation of state of a scalar field condensate A scalar field ϕ sitting at the minimum of its potential $V(\phi)$ has the desired EoS $w = -1$ to drive accelerated expansion. The simplest dynamical model for an inflationary phase is a single, scalar field which is initially displaced from its minimum. A necessary condition for accelerated expansion is $w < -1/3$, and thus we should find the EoS of a scalar field ϕ evolving in a FLRW background.

The stress tensor for a scalar field with $\mathscr{L} = \frac{1}{2}g^{\mu\nu}\nabla_\mu\phi\nabla_\nu\phi - V(\phi)$ follows from

$$T_{\mu\nu} = 2\frac{\partial\mathscr{L}}{\partial g^{\mu\nu}} - g_{\mu\nu}\mathscr{L} = \nabla_\mu\phi\nabla_\nu\phi - g_{\mu\nu}\left[\frac{1}{2}g^{\rho\sigma}\nabla_\rho\phi\nabla_\sigma\phi - V(\phi)\right] \,, \tag{23.12}$$

where we used the relation (??) derived in problem 19.??. We can describe the scalar field also as an ideal fluid. Equating the two expressions for the stress tensor gives

$$T_{\mu\nu} = \nabla_\mu\phi\nabla_\nu\phi - g_{\mu\nu}\mathscr{L} \overset{!}{=} (\rho + P)u_\mu u_\nu - Pg_{\mu\nu}\,. \tag{23.13}$$

Comparing the two independent tensor structures we can identify $P = \mathscr{L}$ and

$$\nabla_\mu\phi\nabla_\nu\phi = (\rho + P)u_\mu u_\nu\,. \tag{23.14}$$

Contracting the indices with $g^{\mu\nu}$, remembering $u_\mu u^\mu = 1$ and using $\nabla_\mu\phi\nabla^\mu\phi = 2L + 2V$ results in

$$\rho = P + 2V\,. \tag{23.15}$$

Now we have to calculate only the energy-density $\rho = T_{00}$ in order to determine the (isotropic) pressure P and the equation of state $w = P/\rho$. In a FLRW background, the energy-density of the field ϕ is given by

$$\rho = T_{00} = \dot{\phi}^2 - \left[\frac{1}{2}\dot{\phi}^2 - \frac{1}{2a^2}(\boldsymbol{\nabla}\phi)^2 - V(\phi)\right] = \frac{1}{2}\dot{\phi}^2 + \frac{1}{2a^2}(\boldsymbol{\nabla}\phi)^2 + V(\phi)\,. \tag{23.16}$$

Thus the pressure[1] follows as

$$P = \frac{1}{2}\dot{\phi}^2 + \frac{1}{2a^2}(\boldsymbol{\nabla}\phi)^2 - V(\phi)\,. \tag{23.17}$$

If we require that the field ϕ respects the symmetries of the FLRW background, then ϕ has to be homogeneous and the $(\boldsymbol{\nabla}\phi)^2$ term vanishes. As result, the equation of state simplifies to

$$w = \frac{P}{\rho} = \frac{\dot{\phi}^2 - 2V(\phi)}{\dot{\phi}^2 + 2V(\phi)} \in [-1 : 1]\,. \tag{23.18}$$

Thus a classical scalar field may act as dark energy, $w < -1/3$, leading to an accelerated expansion of the Universe. A necessary condition is that the field is "slowly rolling," i.e. that its kinetic energy is sufficiently smaller than its potential energy, $\dot{\phi}^2 < 2V/3$.

Slow-roll conditions for the potential We can integrate $\dot{a} = aH$ for an arbitrary time-evolution of H,

$$a(t) = a(t_0)\exp\left(\int \mathrm{d}tH(t)\right)\,. \tag{23.19}$$

The number N of e-foldings is connected to the evolution of H and ϕ as

$$N = \ln\frac{a_2}{a_1} = \int \mathrm{d}t\, H(t) = \int \frac{\mathrm{d}\phi}{\dot{\phi}} H(t)\,. \tag{23.20}$$

The potential energy can dominate only long enough, if $\ddot{\phi}$ is small. Therefore we can approximate the field equation $\ddot{\phi} + 3H\dot{\phi} + V' = 0$ as $\dot{\phi} \approx -V'/(3H)$. Using then also the Friedmann equation $H^2 = 8\pi GV/3$, it follows

$$N = -\int \mathrm{d}\phi\,\frac{3H^2}{V'} = -\int \mathrm{d}\phi\,\frac{8\pi GV}{V'} = -\int \frac{\mathrm{d}\phi}{\widetilde{M}_{\mathrm{Pl}}}\,\frac{V}{\widetilde{M}_{\mathrm{Pl}}V'}\,, \tag{23.21}$$

[1] For an alternative definition of the pressure see problem 24.??.

where we introduced the reduced Planck mass $\widetilde{M}_{\mathrm{P}l} \equiv (8\pi G)^{-1/2}$. Successful inflation requires $N \gtrsim 60 \gg 1$ and thus we require as slow-roll parameter for the potential

$$\varepsilon_V \equiv \frac{1}{2}\left(\frac{\widetilde{M}_{\mathrm{P}l}V'}{V}\right)^2 \ll 1\,. \tag{23.22}$$

Hence the inflaton potential should be flat and its value $V(\phi)$ should be large. An additional constraint on the curvature V'' follows by differentiating V'/V and using then $\varepsilon_V \ll 1$ as

$$\eta_V \equiv \widetilde{M}^2_{\mathrm{P}l}\frac{V''}{V} \ll 1\,. \tag{23.23}$$

This defines the second slow-roll condition which requires that the curvature V'' of the potential measured in Planck units is small compared to the value of the potential $V(\phi)$.

Example 23.2: The arguable least exotic model for inflation uses a single scalar field ϕ with potential $V = \frac{1}{2}m^2\phi^2$. Then the two slow-roll parameters coincide and are given by

$$\varepsilon_V = \eta_V = \frac{2\widetilde{M}^2_{\mathrm{P}l}}{\phi^2}\,. \tag{23.24}$$

Thus we see that inflation in this model requires trans-Planckian field values, $\phi_{\mathrm{i}} \gg \sqrt{2}\widetilde{M}_{\mathrm{P}l}$. (With a certain understatement, one calls in this context field values "large" for $\phi \gtrsim \widetilde{M}_{\mathrm{P}l}$ and small otherwise.) Combining (23.21) and (23.22), we can express the number of e-foldings as

$$N = \int_{\phi_{\mathrm{i}}}^{\phi_{\mathrm{f}}} \frac{\mathrm{d}\phi}{\widetilde{M}_{\mathrm{P}l}}\,(2\varepsilon_V)^{-1/2} = \frac{\phi_{\mathrm{i}}^2}{4\widetilde{M}^2_{\mathrm{P}l}} - \frac{1}{2}\,, \tag{23.25}$$

where we used $\max\{\varepsilon_V, \eta_V\} = 1$ as condition for the end of inflation. Solving the horizon and flatness problem with inflation requires $N \sim 60$ efoldings. Thus the largest scales observed in the CMB, $N \sim 60$, correspond to field values $\phi = 2\sqrt{N_{CMB}} \sim 15 M_{\mathrm{P}l}$.
Classical general relativity should be valid as long as the energy density ρ satisfies $\rho \sim V \ll M^4_{\mathrm{p}l}$. This allows field values as large as $\phi_{\mathrm{i}} \sim 10^3$ for $m = 10^{15}\,\mathrm{GeV}$, and as maximal number of e-folding $N \sim 10^6$. Therefore it is natural to expect that the true number of e-foldings is much larger than 60. In this case, deviations from flatness would be extremely small.

Hamilton-Jacobi equations and phase portraits The evolution of the scale factor a and the field ϕ in single-field inflation is determined by

$$H^2 = \frac{8\pi G}{3}\rho - \frac{k}{a^2} = \frac{8\pi G}{3}\left(\frac{1}{2}\dot{\phi}^2 + V\right) - \frac{k}{a^2}\,, \tag{23.26a}$$

$$\frac{\ddot{a}}{a} = \frac{8\pi G}{3}\left(V - \dot{\phi}^2\right)\,, \tag{23.26b}$$

$$\ddot{\phi} + 3H\dot{\phi} + V_{,\phi} = 0\,. \tag{23.26c}$$

If the field $\phi(t)$ is a monotonic function of the time t, then we can replace t by $\phi(t)$ as evolution variable: We differentiate first $H = \dot{a}/a$, obtaining $\dot{H} = \ddot{a}/a - H^2$. Inserting then (23.26a) and (23.26b) as well as setting $k = 0$, we arrive at

$$\dot{H} = -4\pi G\dot{\phi}^2 \tag{23.27}$$

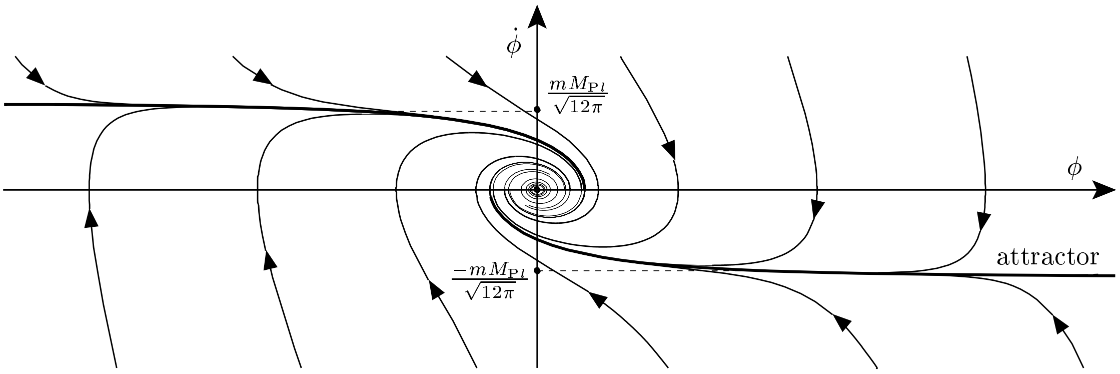

$\dot\phi$
$\frac{mM_{\mathrm{P}l}}{\sqrt{12\pi}}$
ϕ
attractor
$\frac{-mM_{\mathrm{P}l}}{\sqrt{12\pi}}$

Figure 23.2.: Phase portrait for the $V = m^2\phi^2/2$ potential, from [Muk05].

or

$$H_{,\phi} = -4\pi G\dot\phi\,. \tag{23.28}$$

Now we can eliminate $\dot\phi$ in (23.26a), obtaining

$$H_{,\phi}^2 - 12\pi G H^2 = -32\pi^2 G^2 V\,. \tag{23.29}$$

The last two equations are equivalent to the usual two Friedmann equations. They have the virtue to show that choosing the potential $V(\phi)$ according to (23.29), single-field inflation is flexible enough to describe an arbitrary evolution of the Hubble parameter H. Moreover, we can use them to connect the two sets of slow-roll parameters, see problem 24.??.

Next, we want to check how generic the slow-roll conditions in the specific case of a $V = \frac{1}{2}m^2\phi^2$ potential occur. Inserting the Friedmann equation (23.26a) and the explicit form of the potential into the Klein-Gordon equation (23.26c) gives

$$\ddot\phi + [12\pi G(\dot\phi + m^2\phi^2)]^{1/2}\dot\phi + m^2\phi = 0\,. \tag{23.30}$$

Since this second-order differential equation contains no explicit time-dependence, it can be reduced to a first-order equation eliminating $\ddot\phi$ with the help of

$$\ddot\phi = \dot\phi\,\frac{\mathrm{d}\dot\phi}{\mathrm{d}\phi}\,. \tag{23.31}$$

The result,

$$\frac{\mathrm{d}\dot\phi}{\mathrm{d}\phi} = -\frac{[12\pi G(\dot\phi + m^2\phi^2)]^{1/2}\dot\phi + m^2\phi}{\dot\phi}\,, \tag{23.32}$$

allows to plot the phase portrait shown in Fig. 23.2. We observe two lines at $\dot\phi_a = \pm mM_{\mathrm{P}l}/\sqrt{12\pi}$ that attract all trajectories starting at sufficiently large field values ϕ: A field that starts its evolution with $|\dot\phi| \gg \dot\phi_a$ loses fast its kinetic energy $\dot\phi$, moving to the attractor line. Close to the attractor line, it evolves with $|\dot\phi| \sim 0$, i.e. it satisfies the slow-roll condition and can drive inflation. Around $\phi_f \approx \sqrt{2}\widetilde{M}_{\mathrm{P}l}$, this condition is violated: The trajectory leaves the attractor line and spirals towards the origin. This final stage corresponds to coherent oscillations of the inflaton around its minimum and leads to the reheating of the universe. Thus the phase portrait in Fig. 23.2 illustrates that the potential $V = m^2\phi^2/2$ implies an inflationary phase for sufficiently large initial field values ϕ.

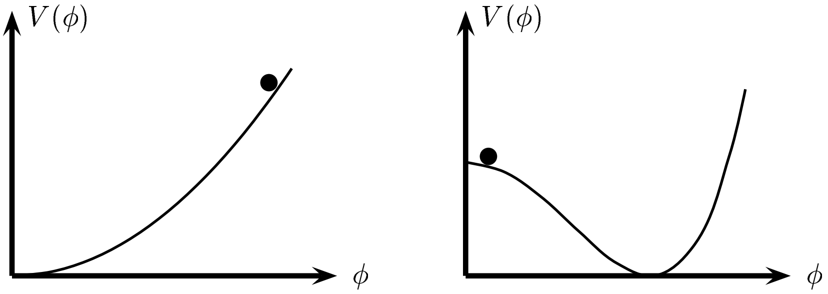


Figure 23.3.: Typical potential of a large-field model (left) and of a small-field model (right) for inflation.

Models for inflation Up-to now we have discussed the—for a particle physicist perhaps most natural—option that a scalar field drives inflation. Moreover, we have restricted our attention to the case of a *single* scalar field. Single field models can be characterised by two parameters, e.g. the width and the height of the potential. Generically, we can divide these models into large and small field models as shown in Fig. 23.3. In the first class, inflation requires trans-Planckian field values as e.g. for the $m^2\phi^2$ or the $\lambda\phi^4$ models. The potential of these models has positive curvature, $V'' > 0$, and thus $\varepsilon_V > 0$.

The trans-Planckian field values required to start inflation lead clearly to the question how the inflaton was displaced from its equilibrium position. A suggestion by Linde is "chaotic inflation": The inflaton field ϕ acquires random values due to quantum fluctuations. In a region of initial size $1/M_{\mathrm{P}l}^4$ with $\phi \gg M_{\mathrm{P}l}$, inflation starts and produces a homogeneous patch inside the universe. Other regions do not undergo inflation at all, or with only a few e-foldings. Thus on scales much larger than our successfully inflated patch, the universe is very inhomogeneous. In a variant, called "stochastic inflation", quantum fluctuations disturb the classical slow-roll trajectory so strongly that the volume filled with large quantum fluctuations $\phi \gg M_{\mathrm{P}l}$ grows exponentially. As a result, new patches of inflating "miniverses" are generated continuously, leading to an eternal self-reproduction of the inflationary universe. A controversial question in these types of models is, how generic the observed universe is, and how such a statement can be made precise.

Typical examples for a small field model are potentials like $V(\phi) = \lambda(\eta^2 - \phi^2)^2/4$ or of the Coleman-Weinberg type. Such potentials are generically much flatter than those of large field models. They often are connected to SSB, and the inflaton sits initially at the unstable equilibrium position $\phi = 0$. The potential has negative curvature, $V'' < 0$, and thus $\varepsilon_V < 0$. The idea of small field models implies that we choose parameters such that inflation starts at sub-Planckian field values, and thus e.g. $\lambda\phi_{\mathrm{f}}^4 \ll V_0 \lesssim M_{\mathrm{P}l}^4$ for $V(\phi) = V_0 - \lambda\phi^4/4$. Inflation may be realised in this class of models as follows: Before the start of inflation, the universe is at high temperature in a potentially inhomogeneous state. The symmetry of the temperature-dependent effective potential $V_{eff}(\phi, T)$ is restored. Thus the field ϕ sits initially at $\phi = 0$. As the universe expands, it cools down and the size of the temperature dependent corrections becomes smaller, $V(\phi, T) \simeq V(\phi, 0)$. Finally, the symmetry is broken, the field starts to roll down the potential and inflation starts.

Using a scalar field as the driving force of inflation, a natural question to address is if we can identify the inflaton with a Higgs field. During the 1980s, one tried to connect the

GUT phase transition and GUT Higgs fields to inflation. However, combining the slow-roll conditions and the size of density fluctuations (that will be discussed in the after next section) restricts the potential severely: Generically, loop corrections destroy the flatness of the potential, if the inflation is not extremely weakly coupled to the SM fields. Therefore, it seems natural to consider the inflaton as a gauge singlet. The discovery of the SM Higgs has created nevertheless interest in the question if the SM Higgs can act as inflaton: First, we know that the Higgs potential flattens for large values of the renormalisation scale. Second, its coupling $\xi\phi^2 R$ to the curvature scalar is unconstrained. A large enough number of e-foldings can be achieved, if the coupling ξ is large, $\xi \approx 50\,000$: Such a term flattens the Higgs potential below $M_{\mathrm{Pl}}/\sqrt{\xi}$ sufficiently to lead to slow-roll inflation. This scenario faces however two problems: First, perturbative unitarity is violated below M_{Pl}, requiring the existence of new degrees of freedom. Thus predictions in “Higgs inflation” depend on the unknown UV completion. More severely, we have seen that the SM Higgs potential (for the values of m_h and m_t currently favoured) develops an instability below M_{Pl}. Thus the SM Higgs cannot be the main agent of inflation, but may play some rôle during inflation.

The range of options widens drastically as soon as one uses several inflaton fields, reducing at the same time however also the predictive power of the models. We comment here only briefly on another option, namely abandoning the idea that the inflaton is a fundamental field. In particular, during the very early universe higher derivative terms in the gravitational action may have played an important role. As a specific possibility, one can modify gravity by generalising the Einstein-Hilbert action as $\mathscr{L}_{\mathrm{E}H} = R \to f(R)$. Here, the function $f(R)$ should be chosen such that observational constraints are obeyed in the $R \to 0$ limit, while for large R modified gravity may lead to inflation. An example for this approach is the Starobinsky model proposed in 1979 that uses $f(R) = R - R^2/(6M^2)$. It represents the first working theory of inflation and is still in excellent agreement with data. In vacuum, this theory is equivalent to standard gravity with a scalar field: Changing first the metric as $g_{\mu\nu} \to g_{\mu\nu}/\chi$ with $\chi \equiv \partial f(R)/\partial R$, and using then $\chi = \exp[4\sqrt{\pi}\phi/(\sqrt{3}M_{\mathrm{P}l})]$ in order to obtain a canonically normalised kinetic term gives the scalar potential

$$V(\phi) = \frac{3M^2 M_{\mathrm{P}l}^2}{32\pi^2}\,(1-1/\chi)^2 = \frac{3M^2 M_{\mathrm{P}l}^2}{32\pi^2}\left[1-\exp\left(\frac{4\sqrt{\pi}\phi}{\sqrt{3}M_{\mathrm{P}l}}\right)\right]^2 . \tag{23.33}$$

Thus one can analyse the Starobinsky model using $V(\phi)$ and standard gravity. The transformation $g_{\mu\nu} \to g_{\mu\nu}/\chi$ induces however couplings of gravitational strength between ϕ and all other SM fields. These additional gravitational couplings indicate that it is more natural to see this class of models as a modification of gravity.

In order to distinguish between these various possibilities, we have to work out the fluctuations predicted by these models. Before that, we consider first the transition from the inflationary phase to the standard radiation dominated universe.

23.3. Reheating and preheating

The inflationary period ends when the EoS becomes larger than $w = -1/3$ and the expansion slows down. In practise, we can define as the end of inflation the time when one of the slow-roll parameters becomes of order one. At this point, the universe became empty and cold; all its energy was contained in the inflaton field. Reheating is a collective term for all the mechanisms by which this energy is transferred to a thermal state of ordinary matter, initiating the usual radiation dominated epoch of the universe. One distinguishes between "perturbative reheating" where the inflaton transfers its energy via perturbative two- or three-particle decays and "preheating" where matter fields coupled to the inflaton develop exponentially growing instabilities.

Perturbative reheating We determine first the EoS valid when the energy density of the universe is dominated by the oscillating inflaton. Assuming a polynomial as inflaton potential, $V \propto \phi^n$, we can use the virial theorem $\langle T\rangle = -n/2\langle V\rangle$ to obtain the EoS averaged over several oscillations,

$$w = \frac{P}{\rho} = \frac{\langle T\rangle - \langle V\rangle}{\langle T\rangle + \langle V\rangle} = \frac{n-2}{n+2}\,. \tag{23.34}$$

Here we neglected the friction factor $3H$ which is by assumption small. Thus during reheating, the universe expands as matter-dominated for $n = 2$ and as radiation-dominated for $n = 4$.

Initially, the reheating process of the universe was modelled as perturbative two-particle decays of the inflaton ϕ. For instance, we can consider decays into bosons χ and fermions ψ via the interaction terms $\mathscr{L}_{int} = -g_1 v\phi\chi^2 - g_2\phi\bar{\psi}\psi$. The self-energy of the inflaton obtains an imaginary part for $q^2 \geq \min\{4m_\chi^2, 4m_\psi^2\}$ which is connected to the total decay width Γ_{tot} via the optical theorem (cf. problem 9.??) by

$$\Im[\Sigma(m)] = m\Gamma_{tot}(\phi)\,. \tag{23.35}$$

We can include this imaginary part of the self-energy into the Klein-Gordon equation, setting $m^2 \to m^2 + \mathrm{i}m\Gamma_{tot}(\phi)$,

$$\ddot{\phi} + 3H\dot{\phi} + (m^2 + \mathrm{i}m\Gamma_{tot})\phi = 0\,. \tag{23.36}$$

The inflaton has to be very weakly coupled to other fields, $\Gamma_{tot} \ll \phi$, and the friction term has to be small too, $H \ll m$. Therefore we can approximate the inflaton evolution as

$$\phi(t) = \Phi(t)\sin(mt) \tag{23.37}$$

with the time-dependent amplitude

$$\Phi(t) = \phi_{\mathrm{i}}\exp\left[-\frac{1}{2}\int \mathrm{d}t\,(3H + \Gamma_{tot})\right]\,. \tag{23.38}$$

We can view the oscillating classical field $\phi(t)$ as a coherent wave of ϕ particles with zero momentum and number density $n(t) = \rho/m \sim m\phi^2(t)/2$.

To be concrete, we fix the potential as $V = m^2\phi^2/2$. Then we find in the limit $\Gamma_{tot} \ll H$ with $H = 2/(3t)$

$$\Phi(t) = \phi_{\mathrm{f}}\frac{t_{\mathrm{f}}}{t} = \frac{1}{\sqrt{3\pi}t}\frac{M_{\mathrm{P}l}}{m}\,, \tag{23.39}$$

where we used $\phi_{\mathrm{f}} = \sqrt{2}\widetilde{M}_{\mathrm{P}l} = M_{\mathrm{P}l}/\sqrt{4\pi}$ as the inflaton value at the end of inflation, $t_{\mathrm{f}} = 2H_{\mathrm{f}}/3$ and $H_{\mathrm{f}} = V(\phi_f)$. This implies $\rho \sim 1/a^3$ as expected for $n = 2$ (problem 24.??). In the opposite limit, $\Gamma_{tot} \gg H$, the amplitude (23.38) follows the usual decay law,

$$\Phi(t) = \phi_{\mathrm{i}}\exp(-\Gamma_{tot}t/2)\,. \tag{23.40}$$

Defining the reheating time t_{rh} by $3\Gamma_{tot} = H$, the inflaton energy density at that time is

$$\rho_{\mathrm{r}h} = \frac{\Gamma_{tot}^2 M_{\mathrm{P}l}^2}{24\pi} \overset{!}{=} \frac{\pi^2}{30}g_* T_{\mathrm{r}h}^4\,. \tag{23.41}$$

In the second step, we assumed that reheating occurs instantaneously. In this approximation, the reheating temperature $T_{\mathrm{r}h}$ is therefore

$$\begin{aligned} T_{\mathrm{r}h} &= \left(\frac{5}{4\pi^3 g_*}\right)^{1/4} (\Gamma_{tot} M_{\mathrm{Pl}})^{1/2} \simeq 0.14 \left(\frac{100}{g_*}\right)^{1/4} (\Gamma_{tot} M_{\mathrm{Pl}})^{1/2} \\ &\ll (m M_{\mathrm{Pl}})^{1/2} \simeq 10^{15}\,\mathrm{G}eV\,, \end{aligned} \tag{23.42}$$

where we used $\Gamma_{tot} \ll m$ and as inflaton mass $m \approx 10^{-6} M_{\mathrm{P}l}$.

Using instead the estimate from problem 24.**??** for the maximal possible decay width of the inflaton, $\Gamma_{tot} \lesssim 10^8\,\mathrm{GeV}$, the instantaneous reheating temperature is of order $10^{13}\,\mathrm{GeV}$. It is thus considerably smaller than the GUT energy and thus GUT baryogenesis seems to be impossible in this picture. Moreover, in any realistic model, reheating will be delayed and thus the maximal temperature the universe reaches after inflation will be below $10^{13}\,\mathrm{GeV}$. In this case, even leptogenesis seems to be only marginally possible. A way out is the non-perturbative scenario of preheating we discuss next, where in a first stage particles are resonantly produced before they thermalise by perturbative processes.

Preheating We consider again a light scalar field χ coupled to the inflaton ϕ via the interaction $\mathscr{L}_{\mathrm{i}nt} = -g\phi^2\chi^2$, but treat now ϕ as a classical field and study the evolution of the quantum field χ in this background. Then the field modes χ_k satisfy

$$\ddot{\chi}_k + 3H\dot{\chi}_k + \left(\frac{k^2}{a^2} + g^2\phi^2\right)\chi_k = 0\,. \tag{23.43}$$

When the inflaton ϕ starts to oscillate, it can transfer energy to the χ field: In particular, modes χ_k which are resonant with the ϕ oscillations can became unstable, leading to an exponential growth of their occupation number. This phenomenon is called parametric resonance.

In order to illustrate the basis of the mechanism, we transform (23.43) into an equation resembling the Schrödinger equation for an electron in a periodic potential. Rescaling $u_k = a^{3/2}\chi_k$ gives the oscillator equation $\ddot{u}_k + \omega_k^2 u_k = 0$ with energy

$$\omega_k^2 = \frac{k^2}{a^2} + g^2\phi^2 - \frac{3}{4}(2\dot{H} + 3H^2)\,. \tag{23.44}$$

Now we introduce as new time variable $z = mt$ as well as $q^2 = g^2\Phi^2/(4m^2)$ and $A_k = 2q + k^2/(ma)^2$, where $\Phi(t)$ satisfies (23.38). Then we arrive at

$$\frac{\mathrm{d}^2 u_k}{\mathrm{d}z^2} + (A_k - 2q\cos(2z) + \Delta)u_k = 0\,, \tag{23.45}$$

where we lumped the unimportant factors into

$$\Delta = \frac{m_\chi^2}{m^2} - \frac{3}{4m^2}(2\dot{H} + 3H^2)\,. \tag{23.46}$$

Using $H = 2/(3t)$, and thus $2\dot{H} + 3H^2 = 0$, we see that the second term in Δ vanishes. Moreover, we can assume that the χ field is light relative to the inflaton, $m_\chi/m \ll 1$, and thus Δ can be neglected all-together. If the fields evolve fast compared to the expansion of the universe, one may neglect also the a dependence in A_k and q. The differential equation (23.45) has then constant coefficients and becomes a Mathieu equation. Its solutions have the form $u_k \propto \exp(\mathrm{i}v_k z)$, and are therefore unstable for $\Im(v_k) < 0$. The wave-bands with $\Im(v_k) < 0$ correspond to forbidden energy bands of the equivalent Schrödinger equation for an electron in a periodic potential. For $q \gtrsim 2A_k$, the bands of unstable wave-numbers k become large, and occupy most of k space. The production of particles is efficient, if ω_k changes non-adiabatically, i.e. if $\dot{\omega}_k \gg \omega_k^2$. This happens when the effective mass of χ, $m_\chi(t) = g\phi(t)$, becomes zero for $\phi(t) = 0$.

Example 23.3: Broad resonance as scattering:
We can break the evolution of $u_k(t)$ into adiabatic pieces where the modes evolve as $u_k \propto \exp(\pm\mathrm{i}\int \mathrm{d}t\omega_k)$ and the particle number is conserved, and "scattering events" when the inflaton crosses zero, $\phi(t_i) = 0$. Thus the field modes satisfy for $t \neq t_i$ Eq. (22.58) with constant Bogolyubov coefficients and energies,

$$u_k(t) = \frac{\alpha_k}{\sqrt{2\omega_k}} \exp\left[-\mathrm{i}\int^t \mathrm{d}t'\,\omega_k\right] + \frac{\beta_k}{\sqrt{2\omega_k}} \exp\left[\mathrm{i}\int^t \mathrm{d}t'\,\omega_k\right] . \tag{23.47}$$

At t_i, the incoming field is scattered into an outgoing field with new Bogolyubov coefficients. Let us assume that the incoming state $t < t_i$ contained no particles. Then we know that the rate of particle production is given by $|\beta_k|^2$. Close to $\phi = 0$, we can approximate $g^2\phi^2(t) \approx g^2\Phi^2 m^2(t-t_j)^2 \equiv k_*^4(t-t_j)^2$ and thus the oscillator equation $\ddot{u}_k + \omega_k^2 u_k = 0$ becomes

$$\ddot{u}_k + \left[k^2/a^2 + k_*^4(t-t_j)^2\right] u_k = 0 . \tag{23.48}$$

Now we rescale momenta, $\kappa = k/a$, and time, $\tau = k_*(t-t_j)$, mapping thereby our problem on the calculation of the tunnelling probability through the one-dimensional potential $V = -\tau^2$,

$$\frac{\mathrm{d}^2 u_k}{\mathrm{d}\tau^2} + \left(\kappa^2 + \tau^2\right) u_k = 0 . \tag{23.49}$$

The analytical solution to this problem is known. Alternatively, one can employ the WKB approximation to find $n_k = |\beta_k|^2 = \exp(-\pi\kappa^2)$, cf. with problem 24.??.

Thus particle production is efficient, if $\kappa^2 \lesssim 1$ or $(k/a)^2 \lesssim k_*^2 = gm\Phi$. The analysis for a non-empty initial state shows that generally particles can be created or destroyed, depending on the value of the relative phase between initial and final state. If one includes the expansion of the universe, the relative phases behave nearly as random variables. In this limit, one finds that in 75% of scatterings particle are created. The number of created particles grows therefore exponentially with time.

At some point, this growth has to slow down and finally stop. Generically, this happens already when still most energy is contained in the inflaton field. One possible reason is the back-reaction of the produced particles on the inflaton field: The term $\langle\chi^2\rangle\phi^2$ acts as additional mass term for the inflaton, changes for $\langle\chi^2\rangle \sim m^2$ the dynamics and shuts off particle production via preheating. Thus the initial phase of preheating has to be followed by perturbative reheating and thermalisation. Note also that the back-reaction of the particle production on the evolution of the scale factor couples the evolution of χ, ϕ and the Hubble parameter H.

Remark 23.1: Superheavy dark matter
Particles with masses up to $\sim 100T_{rh}$ could be produced during preheating. Another option is their

gravitational production during the end of the inflation, as discussed in section 22.2. While the first possibility is strongly model-dependent, the second one relies only gravitational interactions and is therefore universal. If such particles are stable (or have life-times large compared to the age of the universe), this opens the possibility of superheavy dark matter (SHDM). They should have weak enough interactions with SM particles not to thermalise, in order to avoid the unitarity limit (20.38). This ensures also that SHDM behaves as cold dark matter.
The energy density of the particle type X with mass M_X at the end of inflation can be calculated using Eq. (22.60). The numerical results are well described by

$$\rho_X \simeq 10^{-3} M_X^4 \left(\frac{M_X}{H(t_\mathrm{f})}\right)^{-3/2} \exp\left(-2M_X/H(t_\mathrm{f})\right) , \tag{23.50}$$

where $H(t_\mathrm{f})$ denotes the Hubble parameter at the end of inflation. Since the modes generated are non-relativistic, the relative abundance of SHDM has grown as $T_0/T_{\mathrm{r}h}$ since reheating. The present abundance follows therefore setting also $H(t_\mathrm{f})$ equal to the inflaton mass m as

$$\Omega_{X,0} \simeq 3\times 10^{-7} \left(\frac{T_{\mathrm{r}h}}{T_0}\right) \left(\frac{m}{M_{\mathrm{P}l}}\right)^2 \left(\frac{M_X}{m}\right)^{5/2} \exp\left(-2M_X/m\right) . \tag{23.51}$$

Since $T_{\mathrm{r}h}/T_0$ is large, the initial abundance of SHDM has to be tiny and a stable particle with mass $M_X \sim 10^{13}$ GeV would have today an abundance of order one.

23.4. Generation of perturbations

We move now on to the study of the fluctuations $\delta\phi(x)$. We neglect first that they back-react via the Einstein equations on the space-time and calculate their evolution in a fixed FLRW metric. Allowing in the next step for deviations from the FLRW metric means that no preferred foliation of the space-time in time and space exists. This arbitrariness in the coordinate choice requires a careful definition of observables, so that an unphysical gauge dependence of observables is avoided. Having defined suitable gauge invariant variables, we then solve the coupled equations for perturbation in the linear theory.

23.4.1. Fluctuations in a fixed FLRW background

We consider fluctuations of the inflaton field ϕ around its classical, uniform but time-dependent average value,

$$\phi(\boldsymbol{x},t) = \phi_0(t) + \delta\phi(\boldsymbol{x},t) . \tag{23.52}$$

Inserting this splitting into the Klein-Gordon equation (22.16) with $\xi = 0$ and a general potential $V(\phi)$ gives six terms[2],

$$\ddot{\phi} + \partial_t^2\delta\phi + 3H(\dot{\phi} + \dot{\delta\phi}) - \frac{1}{a^2}\boldsymbol{\nabla}^2\delta\phi + V'(\phi_0+\delta\phi) = 0 . \tag{23.53}$$

We evaluate first the potential term V', expanding it around the classical field ϕ_0,

$$V'(\phi_0+\delta\phi) = V'(\phi_0) + V''(\phi_0)\delta\phi = V'(\phi_0) + m_{eff}^2\delta\phi . \tag{23.54}$$

[2] Here, a prime on the potential, V', denotes its derivative with respect to its argument, not to conformal time.

Here we introduced also $m_{eff}^2 = V''(\phi_0)$ as an effective mass term for the fluctuations $\delta\phi$. Taking into account that the classical field ϕ_0 satisfies separately the field equation (22.16) gives as equation for the fluctuations

$$\left[\frac{\partial^2}{\partial t^2} - \frac{1}{a^2}\nabla^2 + 3H\frac{\partial}{\partial t} + m_{eff}^2\right]\delta\phi = 0\,. \tag{23.55}$$

We perform next a Fourier expansion of the fluctuations, $\delta\phi(\boldsymbol{x},t) = V^{-1}\sum_k \phi_k(t)\mathrm{e}^{\mathrm{i}\boldsymbol{kx}}$, with k as comoving wave-number. Since the proper distance varies as $a\boldsymbol{x}$, the physical momentum is then given by $\boldsymbol{p} = \boldsymbol{k}/a$. Inserting the expansion into (23.55), we obtain

$$\ddot{\phi}_k + 3H\dot{\phi}_k + \left(\frac{k^2}{a^2} + m_{eff}^2\right)\phi_k = 0\,. \tag{23.56}$$

Thus the fluctuations obey basically the same equation as the average field, with the effective mass term as the only difference. Introducing again the auxiliary field $\chi_k(\eta) = a(\eta)\phi_k(\eta)$, we arrive at

$$\chi_k'' + \left(k^2 - \frac{a''}{a} + m_{eff}^2\right)\chi_k = 0\,. \tag{23.57}$$

The effective mass term, $m_{eff}^2 = V_{,\phi\phi}$ is much smaller than the squared Hubble parameter H during slow-roll, since

$$\frac{V_{,\phi\phi}}{H^2} \approx \frac{3\widetilde{M}_{\mathrm{Pl}}V_{,\phi\phi}}{V} = 3\eta_V \ll 1\,. \tag{23.58}$$

This implies that we can neglect $V_{,\phi\phi}$ relative to a''/a, because

$$\frac{a''}{a} = \frac{2}{\eta^2} = 2a^2H^2 \ll a^2V_{,\phi\phi} \tag{23.59}$$

is valid in the slow-roll regime. Hence the modes $\chi_k(\eta)$ are also described by

$$\chi_k(\eta) = A_k\,\frac{\mathrm{e}^{-\mathrm{i}\omega_k\eta}}{\sqrt{2\omega_k}}\left(1 - \frac{\mathrm{i}}{k\eta}\right) + B_k\,\frac{\mathrm{e}^{\mathrm{i}\omega_k\eta}}{\sqrt{2\omega_k}}\left(1 + \frac{\mathrm{i}}{k\eta}\right)\,. \tag{23.60}$$

In order to find the spectrum of fluctuations, we should quantise χ and determine its two-point function. In the limit $\eta \to -\infty$, the auxiliary field χ resembles a Minkowski field, and thus we can write down immediately the field operator in this limit,

$$\chi(\eta,\boldsymbol{x}) = \int \frac{\mathrm{d}^3k}{\sqrt{2\omega_k(2\pi)^3}}\left[A_k a_k \mathrm{e}^{-\mathrm{i}(\omega_k\eta - \boldsymbol{kx})} + B_k a_k^\dagger \mathrm{e}^{\mathrm{i}(\omega_k\eta - \boldsymbol{kx})}\right]\,. \tag{23.61}$$

The annihilation and creation operators a_k and $a_k^\dagger$ satisfy the usual commutation relations, with the physical momenta replaced by the conformal momenta. However, we have still to choose the initial vacuum state, what amounts to fixing the coefficients A_k and B_k.

We observe modes which exited the horizon $\Delta N \sim 60$ e-folding before the end of inflation. If there is no special choice of the initial conditions for inflation, we expect that the total number N_{tot} of e-folding is much larger. Thus the physical momentum of these modes at the beginning of inflation, $p(t_i) \sim H\mathrm{e}^{N_{\mathrm{tot}}-\Delta N}$, is extremely high. A natural assumption is therefore that these modes at the start of inflation were empty. Thus we should require that for early times, $\eta \to -\infty$, only positive frequencies survive, what implies $B_k = 0$ and $A_k = 1$.

This choice is called the *Bunch-Davies vacuum.* Inserting this choice into (23.60), we find for the fluctuations inside the horizon, $k \gg |1/\eta| = aH$,

$$|\delta\phi_k| = \left|\frac{\chi_k}{a}\right| = \frac{H\eta}{\sqrt{2k}}\,. \tag{23.62}$$

Modes outside the horizon, $k \ll aH$, are frozen in with amplitude

$$|\delta\phi_k| = \left|\frac{\chi_k}{a}\right| = \frac{H}{\sqrt{2k^3}}\,. \tag{23.63}$$

Power spectrum of perturbations The two-point correlation function for the scalar field fluctuation $\delta\phi$ is given by

$$\langle\delta\phi(\boldsymbol{x}',t')\delta\phi(\boldsymbol{x},t)\rangle = \int \frac{\mathrm{d}^3k}{(2\pi)^3}|\delta\phi_k|^2\,\mathrm{e}^{-\mathrm{i}k(x'-x)}\,. \tag{23.64}$$

Introducing spherical coordinates in Fourier space and choosing $x = x'$ results in

$$\langle\delta\phi^2(\boldsymbol{x},t)\rangle = \int \frac{4\pi k^2\mathrm{d}k}{(2\pi)^3}\,|\delta\phi_k|^2 = \int \mathrm{d}k\,\underbrace{\frac{k^2}{2\pi^2}\,|\delta\phi_k|^2}_{\equiv P(k)} = \int \frac{\mathrm{d}k}{k}\,\Delta^2_\phi(k)\,. \tag{23.65}$$

The functions $P(k)$ and $\Delta^2_\phi(k)$ are the linear and logarithmic *power spectrum* of the fluctuations, respectively. The spectrum $\Delta^2_\phi(k)$ of fluctuations outside of the horizon is given by

$$\Delta^2_\phi(k) = \frac{k^3}{2\pi^2}|\delta\phi_k|^2 = \frac{H^2}{4\pi^2}\,. \tag{23.66}$$

Hence, the power spectrum of superhorizon fluctuations is independent of the wave-number in the approximation that H is constant during inflation. The total area below the function $\Delta^2_\phi(k)$ plotted versus $\ln(k)$ gives $\langle\phi^2(\boldsymbol{x},t)\rangle$. Therefore a spectrum with $\Delta^2_\phi(k) = const.$ contains the same amount of fluctuations in each decade of k. Such a spectrum of fluctuations is called a *Harisson-Zel'dovich* spectrum, and is produced by inflation in the limit of an infinitely slow-rolling inflaton.

Remark 23.2: Let us compare the power-spectrum of fluctuations in inflation to those of normal Minkowski space. Recalling (3.57), the latter are given by

$$\langle 0|\phi^2(x)|0\rangle = \int \frac{\mathrm{d}^3k}{(2\pi)^3 2\omega_k} = \int_0^\infty \frac{\mathrm{d}k}{k}\,\frac{k^2}{(2\pi)^2}\,. \tag{23.67}$$

Thus the amplitude of fluctuations in Minkowski space is given by $\Delta_\phi(k) \equiv (\Delta^2_\phi(k))^{1/2} = k/(2\pi)$, and their relative size $\Delta_\phi(k)/k = 1/(2\pi)$ is constant: Going to smaller and smaller scales, fluctuations keep being important – this is another point of view to understand the problem of UV divergences. Comparing now the relative size of fluctuations at the end of inflation and in Minkowski space, we find

$$\frac{\Delta\phi(k)|_{\mathrm{d}S}}{\Delta_\phi(k)|_{\mathrm{M}}} = H/k \gtrsim \mathrm{e}^{N_\mathrm{f}} \sim \mathrm{e}^{60}\,.$$

This explains why quantum fluctuations become macroscopically important by inflation.

The fluctuations in the inflaton field, $\phi = \phi_0 + \delta\phi$, lead to fluctuations in the stress tensor $T^{\mu\nu} = T_0^{\mu\nu} + \delta T^{\mu\nu}$, and thus to metric perturbations $g^{\mu\nu} = g_0^{\mu\nu} + \delta g^{\mu\nu}$. Thus as next step we have to examine metric perturbations.

23.4.2. Gauge invariant variables for perturbations

Metric scalar, vector and tensor perturbations The spatial uniformity and isotropy of the FLRW background metric suggests that we decompose perturbations of the metric tensor into irreducible components under spatial rotations. Thus we split the full metric tensor $g_{\mu\nu}$ into its background part $g^0_{\mu\nu}$ and scalar, vector and tensor perturbations,

$$g_{\mu\nu} = g^0_{\mu\nu} + \delta g^{\rm s}_{\mu\nu} + \delta g^{\rm v}_{\mu\nu} + \delta g^{\rm t}_{\mu\nu} \,. \tag{23.68}$$

This decomposition is useful, since perturbations with different helicities develop independently in the linear approximation (problem 24.**??**).

The line-element of the flat FLRW metric using conformal time is

$$\mathrm{d}s^2 = a^2[\mathrm{d}\eta^2 - \delta_{ij}\mathrm{d}x^i\mathrm{d}x^j] \tag{23.69}$$

and thus

$$g_{\mu\nu} = a^2(\eta_{\mu\nu} + h_{\mu\nu}) \quad and \quad g^{\mu\nu} = \frac{1}{a^2}(\eta^{\mu\nu} + h^{\mu\nu}) \,. \tag{23.70}$$

We break up the perturbation $h_{\mu\nu}$ in a first step into

$$h_{\mu\nu} = \begin{pmatrix} 2A & B_i \\ B_i & -C_{ij} \end{pmatrix} , \tag{23.71}$$

which gives as line-element

$$\mathrm{d}s^2 = a^2 \left[(1-2A)\mathrm{d}\eta^2 + 2B_i\mathrm{d}x^i\mathrm{d}\eta - (\delta_{ij} + C_{ij})\mathrm{d}x^i\mathrm{d}x^j\right] \,. \tag{23.72}$$

The function A is already a scalar, while we have to find the irreducible components of B_i and C_{ij}. Any vector in $\mathbb{R}^3$ can be written as the sum of a divergence-free and a rotation-free vector; the latter is the gradient of a scalar. Thus we can perform the replacements

$$B_i = -\partial_i B + V_i \tag{23.73}$$

and

$$C_{ij} = 2D\delta_{ij} + 2\partial_i\partial_j E + (\partial_i E_j + \partial_j E_i) + h_{ij} \,. \tag{23.74}$$

The six degrees of freedom of the reducible tensor C_{ij} are decomposed into two scalar (D, E), two vector (E_i with the constraint $\partial_i E_i = 0$) and two tensor (h_{ij}) degrees of freedom. The tensor h_{ij} corresponds to gravitational waves in the TT gauge, with $h_{ii} = 0$ and $\partial_i h_{ij} = 0$. The three degrees of freedom of the reducible vector B_i are decomposed into one scalar (B) and two vector (B_i with the constraint $\partial_i B_i = 0$) degrees of freedom. Thus all-together, $\delta g_{\mu\nu}$ contains four scalar, four vector and two tensor degrees of freedom. Now we can split the line-element into scalar, vector and tensor perturbations around the uniform FLRW background,

$$\mathrm{d}s^2 = a^2 \left\{(1+2A)\mathrm{d}\eta^2 + 2\partial_i B\mathrm{d}x^i\mathrm{d}\eta - \left[(1-2D)\delta_{ij} + \partial_i\partial_j E\right]\mathrm{d}x^i\mathrm{d}x^j\right\} \tag{23.75a}$$

$$\mathrm{d}s^2 = a^2 \left[\mathrm{d}\eta^2 + 2V_i\mathrm{d}x^i\mathrm{d}\eta - (\delta_{ij} + (\partial_i E_j + \partial_j E_i))\mathrm{d}x^i\mathrm{d}x^j\right] \tag{23.75b}$$

$$\mathrm{d}s^2 = a^2 \left[\mathrm{d}\eta^2 - (\delta_{ij} + h_{ij})\mathrm{d}x^i\mathrm{d}x^j\right] \,. \tag{23.75c}$$

As next step, we have to determine the source terms for the different perturbations.

Perturbations of a (non-) ideal fluid The perturbations of the stress tensor serve as source for the perturbations of the metric. If we model the energy content of the universe as an ideal fluid,

$$T_{\mu\nu} = (\rho + P)u_\mu u_\nu - P g_{\mu\nu} \tag{23.76}$$

with $u^\alpha = (1,0,0,0)$ for a comoving observer, then $\delta T_{\mu\nu}$ is parametrised by the perturbations $\delta\rho$, δP and velocities v^i. Thus the perturbed stress tensor $\delta T_\mu^{\ \nu}$ of an ideal fluid contains five degrees of freedom,

$$\delta T_0^{\ 0} = \delta\rho\,, \qquad \delta T_i^{\ j} = \delta_i^j \delta P \quad \mathit{and} \quad \delta T_0^{\ j} = (\rho + P)v^j\,. \tag{23.77}$$

We split the vector v_i again in its irreducible components, $v_i = \tilde{v}_i + \partial_i v$. Thus the perturbed stress tensor of an ideal fluid contains three scalars ($\delta\rho$, δP, v) and one vector ($\tilde{v}_i$), summing up to five degrees of freedom. The remaining five degrees of freedom of a general stress tensor $\delta T_{\mu\nu}$ are contained in the anisotropic pressure tensor Π_{ij}. The presence of this term is characteristic for a non-ideal fluid, i.e. a fluid with viscosity. Since we extracted the isotropic pressure $P\delta_{ij}$, the anisotropic pressure tensor is traceless, $\Pi_{ii} = 0$. Looking back at the decomposition of C_{ij}, we see that the anisotropic pressure contains two tensor ($\propto h_{ij}$), one vector ($\propto (\partial_i E_j + \partial_j E_i)$ and one scalar ($\propto \partial_i\partial_j E$) degrees of freedom.

We can now associate the metric perturbations with the various contribution to the perturbation of the stress tensor: Tensor perturbation h_{ij} have as source only the anisotropic pressure Π_{ij}. This term is generated e.g. by freely streaming neutrinos after weak decoupling, and its effects are always subleading. Therefore, we will set $\Pi_{ij} = 0$ in the following and treat gravitational waves as freely propagating. Vector perturbations correspond to rotational flows of matter—these perturbations are without practical interest for two reasons: First, it is unlikely that they are generated during inflation and, second, they have only decaying solutions. Finally, scalar perturbations are sourced by $\delta\rho$, δP and v_i. We will see that they contain a growing solution and are connected to the inhomogeneities of matter in the universe.

Gauge invariant variables for scalar perturbations In a second step, we have to identify how the perturbations are connected to the physical degrees of freedom. We know already that tensor perturbations, i.e. gravitational waves, contain only two physical degrees of freedom. Similarly, the line-element for scalar perturbations agrees with the Newtonian weak-field limit (18.72), if one sets $A = D = \Phi$ and $E_{,ij} = B_{,i} = 0$. Therefore, one may suspect that not all the four scalar variables describing scalar perturbations are physical.

In order to identify the physical degrees of freedom, we examine how the splitting into scalar perturbations changes under a finite gauge transformation,

$$\tilde{h}_{\mu\nu} = h_{\mu\nu} - \nabla_\mu \xi_\nu - \nabla_\nu \xi_\mu\,, \tag{23.78}$$

of the metric. We decompose the gauge vector ξ_μ as usually in its irreducible components,

$$\xi^\mu = (\xi^0, \xi^i) = (\xi^0, \partial_i \xi + \xi_\perp)\,. \tag{23.79}$$

Consider now e.g. the 00 component of the metric tensor, $\delta g_{00}^{(s)} = h_{00} = 2Aa^2$, from which we can read off the transformation law of A. With $\xi_0 = a^2\xi^0$, $\nabla_0\xi_0 = (a^2\xi_0)' - \Gamma^\mu{}_{00}\xi_\mu$ and $\Gamma^0{}_{00} = \mathscr{H}$, we find

$$\delta g_{00}^{(s)} = 2Aa^2 \to 2\tilde{A}a^2 = 2Aa^2 - 2\nabla_0\xi_0 = 2Aa^2 - 2(a^2\xi^0)' - 2\mathscr{H}a^2\xi^0\,. \tag{23.80}$$

Thus the scalar perturbation A changes under a gauge transformation as

$$\tilde{A} = A - \mathscr{H}\xi^0 - \xi^{0\prime} . \tag{23.81}$$

Proceeding in the same way for the other perturbations, we obtain

$$\tilde{B} = B + \xi^0 - \xi' , \tag{23.82}$$

$$\tilde{D} = D + \mathscr{H}\xi^0 , \tag{23.83}$$

$$\tilde{E} = E - \xi . \tag{23.84}$$

The transformations of the potentials are parametrised by only two arbitrary parameters, ξ^0 and ξ', since $\xi_\perp$ influences only vector perturbations. Moreover, the gauge vector (23.79) contains no tensor component, which expresses the fact that $h_{\mu\nu}$ in the TT gauge contains only physical degrees of freedom. As a result, we can eliminate two variables: For instance, choosing $\xi^0 = -D/\mathscr{H}$ and $\xi = E$ we can set $\tilde{D} = \tilde{E} = 0$.

There are two ways to eliminate the gauge ambiguity in the scalar perturbations: In the first one, we combine the gauge dependent variables into invariant combinations. Consider e.g. the quantity

$$\Psi = D - \mathscr{H}(B - E') \tag{23.85}$$

which is invariant under gauge transformations,

$$\Psi \to \tilde{\Psi} = \tilde{D} - \tilde{\mathscr{H}}(\tilde{B} - \tilde{E}') = D - \mathscr{H}(B - E') . \tag{23.86}$$

Similarly, the combination

$$\Phi = A + \mathscr{H}(B - E') + (B - E')' \tag{23.87}$$

is shown to be invariant. This method was suggested by Bardeen and therefore Ψ and Φ are called Bardeen potentials. An alternative way is to choose a gauge condition which eliminates the gauge freedom partly or completely, analogously to the Coulomb gauge in electrodynamics.

The conformal Newtonian gauge has the virtue that the metric for scalar perturbations is diagonal, since one sets $B = E = 0$. Then the Bardeen potentials become $\Psi = D$ and $\Phi = A$ and we can write the scalar part of metric as

$$\mathrm{d}s^2 = a^2 \left\{(1 + 2\Phi)\mathrm{d}\eta^2 - (1 - 2\Psi)\delta_{ij}\mathrm{d}x^i\mathrm{d}x^j\right\} . \tag{23.88}$$

We show in the appendix that the combination $\Phi - \Psi$ is only sourced by the anisotropic pressure. In the absence of anisotropic pressure, $\Psi = \Phi$, and thus the line-element (except for the conformal factor a^2) coincides with the Schwarzschild metric in the Newtonian limit. This makes an understanding of the perturbation, especially on sub-horizon scales, easier.

Gauge invariant curvature perturbation In the full theory, gauge transformations couple matter and curvature fluctuations. Thus we have to find as final step a gauge invariant combination of both fluctuations. We determine first how the matter perturbation $\delta\phi$ transforms under a gauge transformation $\tilde{x}^\mu = x^\mu + \xi^\mu(x^\nu)$,

$$\widetilde{\delta\phi}(\tilde{x}) = \tilde{\phi}(\tilde{x}) - \phi_0(\tilde{x}) = \phi(x) - \phi_0(x + \xi) = \underbrace{\phi(x) - \phi_0(x)}_{\delta\phi(x)} - \xi^0\partial_0\phi_0(x) . \tag{23.89}$$

Here we used that ϕ is a scalar field, $\tilde{\phi}(\tilde{x}) = \phi(x)$, and that ϕ_0 is uniform. Thus the perturbation transforms as

$$\widetilde{\delta\phi}(\tilde{x}) = \delta\phi(x) - \xi^0\phi_0' \simeq \delta\phi(x) - \xi^0\phi' , \tag{23.90}$$

where we could replace also $\phi_0' \sim \phi'$ neglecting a quadratic term. We know already that D transforms as $\tilde{D} = D + \mathscr{H}\xi^0$. Thus the quantity

$$\mathcal{R} \equiv D + \mathscr{H}\frac{\delta\phi}{\phi_0'} \to \tilde{\mathcal{R}} = D + \mathscr{H}\xi^0 + \mathscr{H}\frac{\delta\phi - \xi^0\phi'}{\phi_0'} = \mathcal{R} \tag{23.91}$$

called the curvature perturbation is a gauge invariant combination of the metric perturbation D and the matter perturbation $\delta\phi$.

23.4.3. Fluctuations in the full linear theory

Up to now, we have considered the evolution of the inflaton field ϕ in a fixed FLRW background, neglecting the back-reaction of the fluctuation $\delta\phi$ on the metric $g_{\mu\nu}$. The Einstein equations couple however the inflaton field ϕ and the metric $g_{\mu\nu}$ already at the linear level. Having identified gauge invariant variables for the perturbations, it remains to perform the straightforward but tedious linearisation of the Einstein equations.

Perturbed stress tensor of the inflaton Inserting into the stress tensor (23.12) of a scalar field,

$$T_\mu^{\ \nu} = g^{\nu\lambda}\partial_\lambda\phi\partial_\mu\phi - \delta_\mu^\nu\left[\frac{1}{2}g^{\rho\sigma}\partial_\rho\phi\partial_\sigma\phi - V(\phi)\right] , \tag{23.92}$$

the decomposition $\phi(x) = \phi_0(t) + \delta\phi(x)$ into a homogeneous background field and fluctuation, we find

$$\delta T_0^0 = \frac{1}{a^2}\left(-\Phi\phi_0'^2 + \phi_0'\delta\phi + V_{,\phi_0}\delta\phi\right) \tag{23.93a}$$

$$= \frac{1}{a^2}\left[-\Phi\phi_0'^2 + \phi_0'\delta\phi - \left(\phi_0'' + 2\frac{a'}{a}\phi_0'\right)\delta\phi\right] , \tag{23.93b}$$

$$\delta T_i^0 = \frac{1}{a^2}\phi_0'\partial_i\delta\phi . \tag{23.93c}$$

Here we used the Klein-Gordon equation for ϕ_0 to replace $V_{,\phi_0}$. The only information we will need on the spatial components δT_i^j of the perturbed stress tensor of the inflaton is that the uniformity of the background requires that $\delta T_i^j \propto \delta_i^j$.

Next we observe that $\rho = \dot{\phi}^2/2 + V(\phi)$ and $P = \dot{\phi}^2/2 - V(\phi)$ imply for the background field ϕ_0 that

$$\rho + P = \dot{\phi}_0^2 = \frac{\phi_0'^2}{a^2} . \tag{23.94}$$

Comparing $\delta T_i^0 = a^{-2}\phi_0'\partial_i\delta\phi$ to an ideal fluid, $\delta T_i^0 = -(\rho + P)\partial_i v = -(\phi_0'^2/a^2)\partial_i v$, we have thus

$$v = -\frac{\delta\phi}{\phi_0'} . \tag{23.95}$$

The choice of a co-moving gauge, where $v = 0$, leads hence to $\delta\phi = 0$: The inflaton fluctuations are zero in the frame where the observer is at rest relative to the inflaton. Thus the curvature perturbation $\mathcal{R}$ becomes $\mathcal{R} = D$ on the hypersurfaces defined by $v = 0$. Evaluating then the three-dimensional curvature $R^{(3)}$ using the spatial part of the metric (23.75a), one finds $R^{(3)} = 4/a^2\Delta\mathcal{R}$. This explains the name curvature perturbation for $\mathcal{R}$.

Perturbed Einstein equations As next step, we need the linearised Einstein tensor for perturbations around the FLRW metric. The result of the straight-forward but lengthy calculations are given in the appendix 23.A. We use the conformal Newtonian gauge where the metric contains the two Bardeen potentials Φ and Ψ. Since the anisotropic pressure is zero, it follows $\Phi = \Psi$, and thus the Einstein equations contain only two free variables, which we choose as Φ and $\delta\phi$. Therefore we can select out of three Einstein equations (00, $0i$ and ij) the two most convenient ones. Choosing the time-time and the time-space equations, we avoid second-order time derivatives and find

$$\Delta\Phi - 3\frac{a'}{a}\Phi' - 3\frac{a'^2}{a^2}\Phi = 4\pi G\left[-\Phi\phi_0'^2 + \phi_0'\delta\phi' - \left(\phi_0'' + 2\frac{a'}{a}\phi_0'\right)\delta\phi\right], \tag{23.96a}$$

$$\Phi' + \frac{a'}{a}\Phi = 4\pi G\phi_0'\delta\phi\,. \tag{23.96b}$$

Here we could integrate the last equation immediately, since H and ϕ_0 are uniform.

Our aim is to combine these two first-order equations into a single second order equation for the gauge-invariant variable $\mathcal{R}$. An often employed strategy is to simplify first Eq. (23.96a), neglecting terms which are small in the slow-roll approximation. We prefer to derive the exact equation for single field inflation, massaging the complete Eq. (23.96a) into a suitable form. We use that the background field ϕ_0 and the scale factor a are connected by the unperturbed Einstein equations. Changing Eqs. (19.48) and (19.49) to conformal time, we have

$$\frac{2a''}{a^3} + \frac{a'^2}{a^4} = -8\pi GP \quad and \quad 3\frac{a'^2}{a^4} = 8\pi G\rho\,. \tag{23.97}$$

Subtracting then these two equations, we can write

$$\frac{a''}{a} - 2\frac{a'^2}{a^2} = -4\pi G(\rho + P)a^2 = -4\pi G\phi_0'^2\,, \tag{23.98}$$

where we used also (23.94). This allows us to eliminate the Φ term on the RHS of (23.96a),

$$\Delta\Phi - 3\frac{a'}{a}\Phi' - \left(\frac{a'^2}{a^2} + \frac{a''}{a}\right)\Phi = 4\pi G\left[\phi_0'\delta\phi' - \left(\phi_0'' + 2\frac{a'}{a}\phi_0'\right)\delta\phi\right]. \tag{23.99}$$

Next we eliminate the Φ term on the LHS: We substitute Φ by $\delta\phi$ and Φ' with the help of (23.96b), and use then again (23.98). The LHS of (23.99) becomes thereby

$$\Delta\Phi - 4\pi G(\phi_0'^2\Phi' + \phi_0'\delta\phi)\,.$$

Then we bring all terms linear in G to the RHS and combine them as

$$\Delta\Phi = 4\pi G\frac{a}{a'}\phi_0'^2\frac{\mathrm{d}}{\mathrm{d}\eta}\left(\Phi + \frac{a'}{a\phi_0'}\delta\phi\right). \tag{23.100}$$

Now we recognise the term in the parenthesis as the gauge-invariant variable $\mathcal{R}$ we were out for. We can combine the dependence of $\mathcal{R}$ on Φ and $\delta\phi$ introducing as new variable

$$\tilde{\phi} = \delta\phi + \frac{a\phi_0'}{a'}\Phi\,. \tag{23.101}$$

Analogous to the uniform case, we replace next $\tilde{\phi}$ by the auxiliary field $u = a\tilde{\phi}$, which is often called the Mukhanov-Sasaki variable. It is connected to $\mathcal{R}$ by

$$u = -\frac{a^2\phi_0'}{a'}\mathcal{R} \equiv z\mathcal{R}\,. \tag{23.102}$$

Expressed by the Mukhanov-Sasaki variable, Eq. (23.100) becomes

$$\Delta\Phi = 4\pi G\frac{z}{a}\phi_0'^2\frac{\mathrm{d}}{\mathrm{d}\eta}\left(\frac{u}{z}\right)\,. \tag{23.103}$$

Then we rewrite the not yet used Eq. (23.96b) as function of u,

$$\frac{a'}{a^2}\frac{\mathrm{d}}{\mathrm{d}\eta}\left(\frac{a^3}{a'}\Phi\right) = 4\pi G\phi_0' u\,. \tag{23.104}$$

The remaining task is to combine these two equations. Applying Δ on (23.100), replacing $\Delta\Phi$ via (23.103) and inserting $\phi_0' = za'/a^2$, we arrive at

$$\frac{\mathrm{d}}{\mathrm{d}\eta}\left(z^2\left(\frac{\mathrm{d}}{\mathrm{d}\eta}\frac{u}{z}\right)\right) = z\Delta u \tag{23.105}$$

or

$$u'' - \frac{z''}{z}u^2 - \Delta u = 0\,. \tag{23.106}$$

This equation describes linear perturbations in single-field inflation exactly, i.e. without implying the slow-roll approximation or any specific shape of the inflation potential.

For superhorizon modes, $\Delta u \approx 0$, and the growing modes satisfies $u \propto z$. The definition of the Mukhanov-Sasaki variable implies then that $\mathcal{R}$ is conserved during inflation, i.e. does not depend on time for superhorizon modes. More generally, one can show that $\mathcal{R}$ is conserved on superhorizon scales also after inflation, if the perturbations are adiabatic.

In order to quantise u, we need to find its action $S[u]$. Formally, we could derive $S[u]$ by integrating out the the gravitational potential Φ from the combined action $S[g_{\mu\nu}, \phi]$. We use instead that the equation of motion (23.106) fixes the action up to an unknown constant A as

$$S[u] = A\int \mathrm{d}^4x\left[u'^2 - (\partial_i u)^2 + \frac{z''}{z}u^2\right]\,. \tag{23.107}$$

Then we determine A by requiring that the term χ'^2 which comes form the scalar action, has the correct coefficient 1/2. With $u = \chi + z\Phi$, we obtain $A = 1/2$. The effective mass of the u field is now

$$m_{\mathrm{e}ff}^2 = -\frac{z''}{z} = -\frac{H}{a\dot{\phi}}\frac{\partial^2}{\partial\eta^2}\left(\frac{a\dot{\phi}}{H}\right)\,, \tag{23.108}$$

where we used

$$z = -\frac{a^2\phi_0'}{a'} = -\frac{a\dot{\phi}_0}{H} = -\sqrt{2}a\widetilde{M}_{\mathrm{P}l}\varepsilon\,. \tag{23.109}$$

The back-reaction between the scalar and the gravitational field is encoded in the time behaviour of $\dot{\phi}$ and H. The exact solutions of Eqs. (23.106) and (23.107) have to be found by numerical integration. To proceed analytically, we employ instead the slow-roll approximation: Then the Hubble parameter is close to constant, $H \approx const.$, and the kinetic energy

$\dot{\phi}_0$ should be small for a sufficiently long time, while a is increasing exponentially. Thus we can treat $\dot{\phi}$ and H as constant, obtaining in the slow-roll approximation

$$m_{\mathrm{e}ff}^2 = -\frac{z''}{z} \to -\frac{\eta''}{\eta}\,. \tag{23.110}$$

Therefore we can identify in the slow-roll approximation the Mukhanov-Sasaki variable u with χ, while $\tilde{\phi}$ coincides with χ/a. Hence, our results obtained for a fixed FLRW background remain valid in the lowest order of the slow-roll approximation, if we express $\mathcal{R}$ by $\tilde{\phi}$: Inserting (23.109) in $u = a\tilde{\phi} = z\mathcal{R}$, we obtain this relation as

$$\mathcal{R} = \frac{a}{z}\,\tilde{\phi} = \frac{H}{\dot{\phi}_0}\,\tilde{\phi}\,. \tag{23.111}$$

We can now use the connection between perturbations in $\mathcal{R}$ and $\tilde{\phi}$ using our result in a fixed background,

$$\Delta^2_{\mathcal{R}}(k) = \frac{H^2}{\dot{\phi}^2}\Delta^2_{\tilde{\phi}}(k) = \frac{H^2}{\dot{\phi}^2}\frac{H^2}{4\pi^2} = \left(\frac{H^2}{2\pi\dot{\phi}}\right)^2_{k=aH} = \frac{1}{8\pi^2\varepsilon}\left(\frac{H}{\widetilde{M}_{\mathrm{Pl}}}\right)^2\Bigg|_{k=aH}\,. \tag{23.112}$$

Here, we used also $\dot{H} = -4\pi G\dot{\phi}^2$ from Eq. (23.27) and $\varepsilon = -\dot{H}/H^2$. Moreover, we accounted for deviations from de Sitter (i.e. the time dependence of H) by evaluating the power spectrum for wave-vectors at horizon crossing, $k = aH$. Since during slow-roll inflation $H^2 \propto V$, we see that the normalisation of the power-spectrum informs us about the ratio V/ε.

Tensor perturbations While scalar perturbations are a gauge-dependent mixture of perturbations in $\delta g^{\mathrm{s}}_{\mu\nu}$ and $\delta\phi$, tensor perturbations $\delta g^{\mathrm{t}}_{\mu\nu}$ are fully fixed by the two physical degrees of freedom in the gravitational wave tensor $h_{\mu\nu}$ present in the TT gauge. Moreover, the source term for gravitational waves corresponds to anisotropic pressure, which is absent in the perturbed stress tensor of the inflaton. We have seen that the action (18.80) for gravitational waves in the TT gauge is identical to the one of a free minimally coupled scalar field, changing its normalisation by 2 × $32\pi G$. Hence the power spectrum $\Delta^2_T(k)$ of the metric tensor perturbations is connected to the scalar perturbations $\Delta^2_\phi(k)$ by

$$\Delta^2_{\mathrm{t}}(k) = 2\times 32\pi G\Delta^2_\phi(k)\big|_{k=aH} = \frac{2}{\pi^2}\left(\frac{H}{\widetilde{M}_{\mathrm{Pl}}}\right)^2\Bigg|_{k=aH}\,. \tag{23.113}$$

Measuring the amplitudes of scalar and tensor perturbations on the scale $k = aH$ determines thus both the slow-roll parameter ε and the Hubble parameter H at horizon crossing of this mode.

Deviations from a scale-invariant spectrum The spectrum of fluctuations is scale-invariant only for a de Sitter universe. Since inflation has to end, we expect deviations from the scale-invariant spectrum $P(k) \propto 1/k$ which we parametrise via the scalar spectral index n_s of the scalar fluctuations. Thus $\Delta^2_{\mathcal{R}}(k) \propto k^{-(1-n_s)}$ and

$$n_{\mathrm{s}} - 1 \equiv \frac{\ln\Delta^2_{\mathcal{R}}(k)}{\ln(k)} = \frac{\mathrm{d}\ln\Delta^2_{\mathcal{R}}(k)}{\mathrm{d}N}\frac{\mathrm{d}N}{\mathrm{d}\ln(k)}\,. \tag{23.114}$$

Since the slow-roll parameters determine the deviations from de Sitter expansion, we should be able to express n_s via ε and η. Using $\Delta^2_{\mathcal{R}}(k) = H^2/(8\pi^2\varepsilon)\big|_{k=aH}$, we find for the first factor

$$\frac{\mathrm{d}\ln\Delta^2_{\mathcal{R}}(k)}{\mathrm{d}N} = 2\frac{\mathrm{d}\ln H}{\mathrm{d}N} - \frac{\mathrm{d}\ln\varepsilon}{\mathrm{d}N} = -2\varepsilon - \eta\,, \tag{23.115}$$

where we used the definition of the slow-roll parameters. For the evaluation of the second factor, we use that modes crossing the horizon satisfy $k = aH$, or

$$\ln k = \ln a + \ln H\,. \tag{23.116}$$

Differentiating this expression and recalling $\mathrm{d}N = d\ln(a)$ gives

$$\frac{\mathrm{d}\ln(k)}{\mathrm{d}N} = 1 - \varepsilon\,. \tag{23.117}$$

Combining the results for the two factors and neglecting second order terms, we arrive at

$$n_\mathrm{s} - 1 \approx (-2\varepsilon - \eta)(1+\varepsilon) \approx -2\varepsilon - \eta\,. \tag{23.118}$$

Thus the slope of the scalar fluctuations informs us about deviations from a perfect de Sitter phase, or via ε_V and η_V, on the shape of the inflaton potential.

Next we consider tensor fluctuations, defining their spectral slope by

$$n_\mathrm{t} \equiv \frac{\ln\Delta^2_\mathrm{r}(k)}{\ln(k)}\,. \tag{23.119}$$

In contrast to the scalar slope, n_t contains no ε term and thus it is given simply by

$$n_\mathrm{t} = -2\varepsilon\,. \tag{23.120}$$

Tensor-scalar ratio The ratio r of perturbations in tensor and scalar modes is determined in single field inflation fully by the slow-roll parameter ε,

$$r \equiv \frac{\Delta^2_\mathrm{t}}{\Delta^2_{\mathcal{R}}} = \frac{2}{\pi^2}\left(\frac{H}{\widetilde{M}_{\mathrm{Pl}}}\right)^2 \Big/ \frac{H^2}{8\pi^2\varepsilon} = \frac{8\dot{\phi}^2}{\widetilde{M}^2_{\mathrm{P}l}H^2} = 16\varepsilon\,. \tag{23.121}$$

This result has two important consequences. First, we can derive an upper limit on $r|_{k=aH}$ as function of the inflaton field value at the time of horizon crossing of the scale k. This limit, often called the Lyth bound, can be derived using $\mathrm{d}\phi/(\mathrm{d}tH) = \mathrm{d}\phi/\mathrm{d}N$. Then the amount $\Delta\phi$ the inflaton evolved between the horizon exit of the CMB modes and the end of inflation is given by

$$\Delta\phi \geq \widetilde{M}_{\mathrm{P}l}\int_0^{N_{CMB}} \mathrm{d}N\sqrt{r/8} \approx \text{few} \times \widetilde{M}_{\mathrm{P}l}\left(\frac{r}{0.01}\right)^{1/2}\,. \tag{23.122}$$

Here, we used that r during slow-roll should be nearly constant. This relation connects the tensor-scalar ratio on CMB scales and the minimal initial value of the inflaton. Hence large, observable tensor perturbations require trans-Planckian initial values of the inflaton.

As second consequence, we can derive a so-called consistency relation. We can combine $r = 16\varepsilon$ and $n_\mathrm{t} = -2\varepsilon$ as follows

$$n_\mathrm{t} = -2\varepsilon = -\frac{r}{8}\,. \tag{23.123}$$

The slope of the tensor perturbations is fixed by the ratio of the amplitudes of scalar and tensor perturbations. Measuring n_t independently provides therefore a consistency check of single field inflation.

Emergence of classical fields Inflation amplifies the length-scales of perturbations which have been generated as quantum fluctuations. Since field operators as

$$\chi(\eta, \boldsymbol{x}) = \int \frac{\mathrm{d}^3 k}{(2\pi)^{3/2}} \left(\chi_k^{(+)}(\eta) a_{\boldsymbol{k}}^\dagger \mathrm{e}^{-\mathrm{i}\boldsymbol{kx}} + \chi_k^{(-)}(\eta) a_{\boldsymbol{k}} \mathrm{e}^{\mathrm{i}\boldsymbol{kx}} \right) \tag{23.124}$$

are complex, the expectation values of products of fields are complex too. Moreover, they are depending on their order. In contrast, the fluctuations of a fluid are classical and real. In general, we expect that the initial quantum fluctuations are converted into classical fluctuations by the phenomenon of decoherence, i.e. by the coupling to a thermal bath. In the case of superhorizon fluctuations, the coupling between the inflaton and the gravitational field is sufficient for this conversion process: For modes after horizon exit, $k_* \ll 1/|\eta|$, which are frozen in, the field operator simplifies to

$$\chi(\eta, \boldsymbol{x}) = \int_{k_*} \frac{\mathrm{d}^3 k}{\sqrt{(2\pi)^3 2\omega_k}} \left(\frac{-1}{k\eta} \right) \left(a_{\boldsymbol{k}}^\dagger \mathrm{e}^{-\mathrm{i}(\boldsymbol{kx}+\alpha_{\boldsymbol{k}})} + a_{\boldsymbol{k}} \mathrm{e}^{\mathrm{i}(\boldsymbol{kx}-\alpha_{\boldsymbol{k}})} \right) , \tag{23.125}$$

where $\alpha_{\boldsymbol{k}}$ are arbitrary phases. Thus the mode functions are real, $\chi_k^{(+)}(\eta) = \chi_k^{(-)}(\eta)$. As a result, also the expectation values of products of fields become real and do not depend on the order. This can be also seen calculating the canonically conjugated field operator on super-horizon scales,

$$\pi(\eta, \boldsymbol{x}) = \frac{\partial \mathscr{L}}{\partial \chi'} = -\frac{1}{\eta} \chi(\eta, \boldsymbol{x}) \tag{23.126}$$

which is proportional to $\chi(\eta, \boldsymbol{x})$. Therefore, the two operators commute and χ behaves as a classical field. Thus we can treat the perturbations as classical random fields and replace quantum averages by statistical averages. Since we included only the quadratic part of the potential, cf. Eq. (23.54), the quantum field χ is a free Gaussian field. As a result, the corresponding classical random field is Gaussian too, implying that all information is contained in the two-point correlation functions. Deviations from Gaussianity should be non-zero but tiny.

23.5. Outlook: Further evolution of fluctuations

We have found that inflation generates a nearly power-law like spectrum of curvature fluctuations and gravitational waves. Superhorizon modes are frozen-in and can be described by classical Gaussian random fields. These fluctuations become observable through the temperature fluctuations of the CMB and the large-scale structure of the universe. The evolution of these fluctuations after inflation follows classical physics, and thus this topic is outside the focus of this book. We give therefore in this section only a sketchy overview.

After reheating, the matter content of the universe consists of a plasma containing at least the SM degrees of freedom. Additionally, there should be new particles associated to a dark matter sector and to baryogenesis. Initially, all the species[3] are tightly coupled and can described therefore as an ideal fluid with a single mode of perturbation $\delta\rho$. This implies that the fluctuations are *adiabatic*, i.e. that they satisfy

$$\delta\rho_i = \dot{\rho}_i \delta(x) \quad and \quad \delta P_i = \dot{P}_i \delta(x) \tag{23.127}$$

[3] An exception might be e.g. non-thermal DM which presence would lead to isothermal fluctuations.

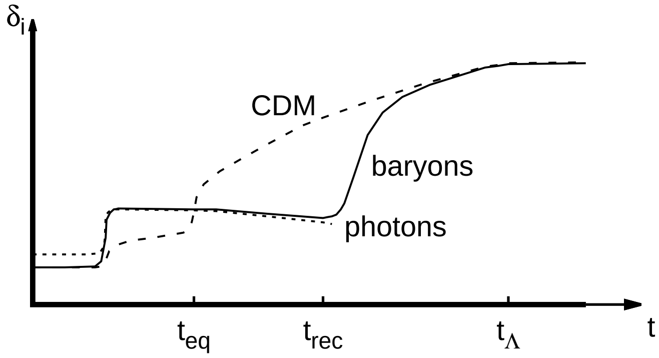


Figure 23.4.: Schematic evolution of perturbations in cold dark matter, photons, and baryons.

with a single function $\delta(x)$ describing the density and pressure fluctuations of all the components i. In this case, the conservation law for $\mathcal{R}$ continues to hold for superhorizon modes. As the universe cools down, some species go first out of chemical and later out of kinetic equilibrium. As a result, the evolution of these species has to be described either by a non-ideal fluid including anisotropic pressure or by a set of coupled Boltzmann equations. Examples where such a treatment is necessary are the free-streaming of neutrinos after weak decoupling and of photons after recombination.

During the evolution of the universe, at least two episodes happened where unknown physics beyond the SM can influence the evolution of fluctuations: These two episodes are the generation of the baryon asymmetry and of the dark matter abundance. If both are generated through the freeze-out mechanism discussed in chapter 20 and 21, then the same ratio n_B/s and $n_{\rm DM}/s$ is generated in the whole universe. As a result, the fluctuations remain adiabatic. Another option is that the ratios n_B/s or $n_{\rm DM}/s$ vary in space. Such fluctuations are called isocurvature perturbations. This may happen e.g. in axion models, if the Peceei-Quinn phase transition takes place at a smaller temperature than reheating. Then the observable universe today contains many patches with different values of the misalignment angle ϑ. Since ϑ is space dependent, the resulting axion dark matter density $n_{\rm DM}/s$ varies in space, too. Observations of the CMB are consistent with a purely adiabatic fluctuation spectrum and can be thus used to limit models predicting isocurvature perturbation.

Matter perturbations The evolution of the perturbations $\delta\rho_i$ and δP_i can be determined from the Einstein equations, following the same strategy as in the case of the inflaton but using the stress tensor for a sum of the various fluid components. The evolution of the density contrast in different components of the energy density of the universe is schematically shown in Fig. 23.4 assuming adiabatic fluctuations.

An important feature to note is that the density contrast of radiation is approximately constant. This can be understood from the fact that the scale on which radiation is gravitationally bound is comparable to the Hubble radius (problem 24.**??**). As a result, also the density contrast of baryons is approximately constant as long the photon-baryon fluid is tightly coupled. In contrast, the density contrast of CDM grows as $\delta_{CDM} \propto (1+z)$ starting from matter-radiation equilibrium z_{eq}. At recombination, baryons decouple from photons and start to fall into the potential walls formed by CDM. Note also that fluctuations of CDM on small scales k crossed earlier the horizon. They started therefore to grow and to form gravitationally bound systems earlier, leading to the hierarchical formation of structures in a CDM universe. Observations strongly favour this picture. Finally, the growth of the density contrast stops outside of already gravitationally bound structures, when the accelerated expansion of Universe starts at $z \simeq 0.5$.

Example 23.4: Derive the connection between $\mathcal{R}$ and the gravitational potentials.
We matched already the fluctuations of the inflaton and the plasma, see Eqs. (23.94) and (23.95), finding $v = -\delta\phi/\phi_0'$. Thus we can express the curvature perturbation in the conformal Newtonian gauge as

$$\mathcal{R} = \Psi - \mathscr{H} v\,. \tag{23.128}$$

Using now the (integrated) $0i$ component of the Einstein equation, $\Psi' + \mathscr{H}\Phi = 4\pi G a^2(\rho + P)v$ and $\mathscr{H}^2 = 8\pi G a^2 \rho/3$, we obtain

$$\mathcal{R} = \Psi + \frac{2}{3}\,\frac{\Phi + \Psi'/\mathscr{H}}{1+w}\,. \tag{23.129}$$

Here, $w = P/\rho$ is the effective EoS with $P = \sum_i P_i$ and $\rho = \sum_i \rho_i$ for the different components of the plasma. This equation allows us to deduce two important results, cf. problem 24.??: First, differentiating it for adiabatic perturbations shows that $\mathcal{R}$ is also after inflation conserved on superhorizon scales. Second, we can use it to deduce from $\mathcal{R}$ the gravitational potentials, i.e. we can connect the predictions of inflationary models with the input for the formation of large-scale structures. In particular, Eq. (23.129) simplifies neglecting anisotropic pressure and assuming $w = const.$ on superhorizon scales to

$$\Phi = \Psi = \frac{3+3w}{5+3w}\,\mathcal{R}\,. \tag{23.130}$$

This gives $\Phi = 2\mathcal{R}/3$ for the radiation and $\Phi = 3\mathcal{R}/5$ for the matter dominated epoch.

Angular power spectrum We will next examine in somewhat more detail the cosmic microwave background (CMB). Fluctuations in the CMB temperature are seen on the sphere of last scattering. Thus it is convenient to decompose a map of CMB temperatures $T(\vartheta,\phi)$ into spherical harmonics $Y_{lm}(\vartheta,\phi)$,

$$T = \sum_{l=0}^{\infty}\sum_{m=-l}^{l} a_{lm} Y_{lm}(\vartheta,\phi)\,. \tag{23.131}$$

The first two moments are usually considered separately: The monopole moment $l = 0$ of a CMB temperature map corresponds to the average temperature of the CMB, $T_0 = 2.725\,\mathrm{K}$. The relative motion of the Sun with respect to the CMB introduces (mainly) a dipole $l = 1$ anisotropy. More precisely, the temperature transforms as

$$T = \frac{T_0\sqrt{1-\beta^2}}{1-\beta\cos\vartheta} = T(1 + \cos\beta + \mathcal{O}(\beta^2))\,, \tag{23.132}$$

where we can neglect the higher order terms because the peculiar velocity of the Sun is small, $\beta = v/c \ll 1$: From the size of the dipole one deduces that the Sun moves with 370 km/h relative to the CMB.

If this dipole anisotropy is subtracted, temperature differences of the order $\delta T/T \sim 10^{-5}$ remain between different directions of the sky. However, in each direction the spectrum is the one of a perfect black-body. The moments $l \geq 2$ of the fluctuations $\delta T/T$,

$$\Theta(\vartheta, \phi) \equiv \frac{\delta T}{T} = \sum_{l=2}^{\infty} \sum_{m=-l}^{l} a_{lm} Y_{lm}(\vartheta, \phi) \tag{23.133}$$

are connected with the cosmological parameters and physical processes between recombination and today. For an isotropic universe, the m dependence of the coefficients a_{lm} contains no information and one defines therefore

$$C_l = \langle a_{lm} a_{lm}^* \rangle = \frac{1}{2l+1} \sum_{m=-l}^{l} a_{lm} a_{lm}^* \,. \tag{23.134}$$

Since a single spherical harmonic Y_{lm} corresponds roughly to angular variations of $\vartheta \sim \pi/l$, the coefficient C_l determines the power of fluctuation with the angular scale π/l. Note that this procedure allows one to replace an average over an ensemble of universes by an average over different patches of the observed Universe. For small ℓ, this introduces an irreducible statistical error $\delta C_\ell / C_\ell \sim 1/\sqrt{2\ell+1}$ which is called *cosmic variance.*

We are not interested in the temperature fluctuations $\Theta(\vartheta, \phi)$ themselves, but only in their statistical properties. For Gaussian initial quantum fluctuations, all the information is contained in the two-point correlation function,

$$C(\vartheta) = \big\langle \Theta(\vartheta, \phi) \Theta(\vartheta', \phi') \big\rangle \,. \tag{23.135}$$

(Since the initial curvature fluctuations are small, the perturbations of all other quantities are linearly related to the curvature fluctuations and therefore Gaussian too.) A Legendre transformation gives then

$$C_l = 2\pi \int_{-1}^{1} \mathrm{d}\cos\vartheta \; C(\vartheta) P_l(\cos\vartheta) \,, \tag{23.136}$$

which summarise in an efficient way the experimental data $\delta T(\vartheta, \phi)$ contained in the many pixels of an CMB experiment. In the next step, we have to connect the temperature fluctuations observed today to the fluctuations at recombination.

Temperature fluctuations Our aim is to connect the frequency of a photon emitted on the last scattering surface (LSS) to its frequency today. This calculation is simplified performing it in the frame which is conformally flat for zero perturbations, $\tilde{g}_{\mu\nu} = g_{\mu\nu}/a^2(\eta)$, using that light-like geodesics coincide in two conformally related frames.

Let us consider the covariant momentum $P^\mu = \mathrm{d}x^\mu/\mathrm{d}\lambda$ of a photon as function of conformal time $\eta = x^0(\lambda)$. Then we can rewrite the geodesic equation $\mathrm{d}P^\mu/\mathrm{d}\lambda + \Gamma^\mu{}_{\rho\sigma} P^\rho P^\sigma = 0$ using

$$\frac{\mathrm{d}P^\mu}{\mathrm{d}\lambda} = \frac{\mathrm{d}\eta}{\mathrm{d}\lambda} \frac{\mathrm{d}P^\mu}{\mathrm{d}\eta} = P^0 \frac{\mathrm{d}P^\mu}{\mathrm{d}\eta} \tag{23.137}$$

as

$$\frac{\mathrm{d}P^\mu}{\mathrm{d}\eta} + \Gamma^\mu{}_{\rho\sigma} \frac{P^\rho}{P^0} \frac{P^\sigma}{P^0} P^0 = 0 \,. \tag{23.138}$$

In the following, we will need only the component of this equation describing the evolution of P^0. Using the conformal Newtonian gauge and curvature perturbations, the Christoffel symbols related to the frame $\tilde{g}_{\mu\nu}$ are given by ${\Gamma^0}_{i0} = \partial_i\Phi$, and ${\Gamma^0}_{ij} = \Psi'\delta_{ij}$. Inserting them into the geodesic equation for P^0, it follows

$$\frac{\mathrm{d}P^0}{\mathrm{d}\eta} + \left(\Phi' + \Psi'\delta_{ij}\frac{P^i}{P^0}\frac{P^j}{P^0} + 2\partial_i\Phi\frac{P^i}{P^0}\right)P^0 = 0\,. \tag{23.139}$$

Without perturbations, $\mathrm{d}P^\mu/\mathrm{d}\eta = 0$ and thus the vector $n_i = n^i = P^i/P^0$ is a constant tangent vector along the photon trajectory with unit norm. Next we rewrite the parenthesis as

$$\Phi' + \Psi'\delta_{ij}n^in^j + 2\boldsymbol{n}\cdot\boldsymbol{\nabla}\Phi = -(\Phi' - \Psi') + 2(\phi' + \boldsymbol{n}\cdot\boldsymbol{\nabla}\Phi)\,. \tag{23.140}$$

Here, we combined the terms such that the second parenthesis on the RHS is a total derivative, since

$$\frac{\mathrm{d}\Phi(\eta,\boldsymbol{x}(\eta))}{\mathrm{d}\eta} = \Phi' + \frac{\partial\boldsymbol{x}}{\partial\eta}\cdot\boldsymbol{\nabla}\Phi = \Phi' + \frac{\partial\boldsymbol{x}/\partial\lambda}{\partial\eta/\partial\lambda}\cdot\boldsymbol{\nabla}\Phi = \Phi' + \frac{\boldsymbol{P}}{P^0}\cdot\boldsymbol{\nabla}\Phi = \phi' + \boldsymbol{n}\cdot\boldsymbol{\nabla}\Phi\,. \tag{23.141}$$

Integrating (23.140) along the photon trajectory and using that P^μ is constant at zero order, we obtain

$$\frac{P^0(\eta_0) - P^0(\eta_1)}{P^0(\eta_1)} = \int_{\eta_1}^{\eta_0}\mathrm{d}\eta(\Phi' - \Psi') - 2[\phi(\eta_0) - \phi(\eta_1)]\,. \tag{23.142}$$

Now we have to connect this expression to the observed photon frequency today in the frame $g_{\mu\nu}$. Let us denote the frequency in the rest-frame of the plasma of a photon emitted at the point x of the LSS with $\bar{\omega}$. In the frame connected to the $g_{\mu\nu}$ coordinates, the plasma moves with the four-velocity u^α, with the components $u^0 = (1-\Phi)$ and $u^i = v^i$ (recall that we use conformal quantities). Thus it follows

$$u_0 = (1+\Phi) \quad and \quad u^i = u_i = -v^i\,.$$

We choose in the rest-frame of the plasma at the position where the photon is emitted normal coordinates, $\bar{g}_{\mu\nu}(x) = \eta_{\mu\nu}$. Then $\bar{u}^\alpha = (1,\boldsymbol{0})$ and $\omega = \bar{u}_\alpha\bar{P}^\alpha = \bar{P}^0$. Since $\omega = u_\alpha P^\alpha = \bar{u}_\alpha\bar{P}^\alpha$, we can use

$$\omega = u_\alpha P^\alpha = (1+\Phi)P^0 - \boldsymbol{v}\cdot\boldsymbol{P} = (1+\Phi-\boldsymbol{n}\cdot\boldsymbol{v})P^0 \tag{23.143}$$

to calculate the frequency $\omega(\eta_1)$ of the photon emitted at LSS. The same formula applies to the frequency $\omega(\eta_0)$ measured by an observer at present time η_0 with velocity v^i in the conformal frame. Thus the relative frequency shift is

$$\frac{\omega(\eta_0) - \omega(\eta_1)}{\omega(\eta_1)} = \int_{\eta_1}^{\eta_0}\mathrm{d}\eta(\Phi' - \Psi') - [\phi(\eta_0) - \phi(\eta_1) + \boldsymbol{n}\cdot\boldsymbol{v}|_{\eta_0} - \boldsymbol{n}\cdot\boldsymbol{v}|_{\eta_1}]\,. \tag{23.144}$$

The RHS is independent of the photon frequency and thus an initially thermal photon spectrum remains thermal. Consequently, the same formula gives also the relative temperature change $\Theta(-\boldsymbol{n}) = \delta T/T$ of the CMB in the direction $\boldsymbol{k} = -k\boldsymbol{n}$. The monopole contribution $\Phi(\eta_0)$ is not observable, since it can be absorbed into the average CMB temperature. Similarly, the term $\boldsymbol{n}\cdot\boldsymbol{v}|_{\eta_0}$ corresponds to the dipole term due to the movement of the observer versus the CMB which we subtracted. Finally, we should add the temperature fluctuations on the LSS which are connected to the energy density fluctuation on the LSS via $\delta\rho_\gamma/\rho_\gamma = 4\delta T/T$. Combining everything, we obtain

$$\Theta(\boldsymbol{n},\eta_0) = \underbrace{\int_{\eta_r}^{\eta_0} \mathrm{d}\eta(\Phi' - \Psi')}_{ISW} + \underbrace{\frac{1}{4}\frac{\delta\rho_\gamma}{\rho_\gamma} + \Phi(\eta_r)}_{SW} + \underbrace{\boldsymbol{n}\cdot\boldsymbol{v}|_{\eta_r}}_{\mathrm{D}oppler} . \tag{23.145}$$

The last term accounts for the Doppler effect induced by the movement of the plasma relative to the coordinate frame. The two terms in the middle describe the Sachs-Wolfe effect, which has two contributions: An overdense spot emits photons with larger average energies but has also a larger (negative) gravitational potential which in turn leads to a larger redshift. Thus the two contributions cancel partly. Finally, the first terms called the integrated Sachs-Wolfe effect take into account the change of the gravitational potentials between LSS and today. Next we should stress the approximations we made: First, we have assumed that recombination happens instantaneously. Second, we have neglected that some of the photons may scatter on free electrons after the universe become reionised by the first stars at $z \lesssim 10$.

We omit from now on the argument η_0 and write simply $\Theta(\boldsymbol{n},\eta_0) \equiv \Theta(\boldsymbol{n})$. Once again, it is useful to move to Fourier space: The integral in $\Theta(\boldsymbol{n})$ receives contributions along the path $x(\eta) = \eta\boldsymbol{n}$ with $\eta \in [\eta_r, \eta_0]$. Thus the Fourier transformed $\Theta(\boldsymbol{k})$ is a function only of $\eta\boldsymbol{n}$ and the magnitude k, or $\Theta(\boldsymbol{k}) = \Theta(k\boldsymbol{n}, k) \equiv \Theta(k\cos\vartheta, k)$. Performing the Fourier transformation, we can expand thus the phase in Legendre polynomials,

$$\Theta(\boldsymbol{n}) = \frac{\delta T(-\boldsymbol{n})}{T} = \int \mathrm{d}^3k\, \Theta(\cos\vartheta, k)\mathrm{e}^{\mathrm{i}k\eta\cos\vartheta} \tag{23.146}$$

$$= \sum_l \mathrm{i}^l(2l+1)\int \mathrm{d}^3k\, \widetilde{\Theta}_l(k)P_l(\cos\vartheta) \tag{23.147}$$

We denoted the Legendre transformed by $\tilde{\Theta}_l(k)$, because one usually extracts the effect of primordial fluctuations in the gravitational potential setting

$$\widetilde{\Theta}_l(k) \equiv \Phi_i(k)\Theta_l(k) . \tag{23.148}$$

Inserting $\Theta(\boldsymbol{n})$ into (23.135) and the result then in (23.136), we arrive at our final formula for the angular power spectrum induced by scalar perturbations,

$$C_l = 4\pi \int \frac{\mathrm{d}k}{k}\Delta_{\Phi_i}(k)\Theta_l^2(k) . \tag{23.149}$$

Hence the CMB power spectrum is determined by the product of the power spectrum $\Delta_{\Phi_i}(k)$ of initial fluctuations and the functions $\Theta_l^2(k)$. The latter contain projection effects of the plane waves onto the sphere of last scattering, encode the physics of the cosmological fluid at recombination and the evolution of the gravitational potential between inflation and today.

As simplest application, we consider the Sachs-Wolfe effect which is relevant on large angular scales. From an analysis of the fluid equations one finds that on superhorizon scales $\Theta = \Phi/3 = 3\Phi_i/10 = \mathcal{R}/5$ holds (cf. example 24.4),

$$\Theta(\boldsymbol{n}) = \int \mathrm{d}^3k\Theta(\boldsymbol{n}\cdot\boldsymbol{n}, k) = \frac{1}{3}\int \mathrm{d}^3k\Theta(\boldsymbol{k})\mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{n}(\eta_0-\eta_r)} . \tag{23.150}$$

Expanding the plane wave in Legendre polynomials, we obtain as coefficients spherical Bessel functions,

$$\widetilde{\Theta}_l = \frac{1}{3}\Phi(\boldsymbol{k})j_l(k(\eta_0 - \eta_r)) = \frac{3}{10}\Phi_i(\boldsymbol{k})j_l(k(\eta_0 - \eta_r)), \tag{23.151}$$

or, if we extract the primordial spectrum and use $\eta_0 \gg \eta_r$,

$$\Theta_l = \frac{3}{10}j_l(k(\eta_0 - \eta_r)) \simeq \frac{3}{10}j_l(k\eta_0). \tag{23.152}$$

Inserting Θ_l into Eq. (23.149) results in

$$C_l = \frac{36\pi}{100}\int \frac{\mathrm{d}k}{k}\Delta_\Phi(k)j_l^2(k\eta_0). \tag{23.153}$$

Choosing a power-law $\Delta_\Phi(k) \simeq A_\Phi(k/k_0)^{n_s-1}$, we can evaluate the integral (problem 24.**??**) and obtain finally in the limit of a flat spectrum

$$C_l = \frac{18\pi}{100}\frac{A_\Phi}{l(l+1)}. \tag{23.154}$$

Thus a flat primordial spectrum results in a flat temperature power-spectrum (at $l \lesssim 100$). Comparing this result to observations determines the normalisation of the primordial fluctuations as $A_\Phi \simeq 2.6 \times 10^{-9}$, what in turn fixes $(V/\varepsilon)^{1/4} \sim 6 \times 10^{16}$ GeV using (23.112).

In the left panel of Fig. 23.5, numerical results for the temperature power spectra from scalar perturbations are shown as a black line. After the plateau up to $l \sim 100$, a series of peaks with declining amplitude is visible. These peaks are caused by the coherent oscillations of the baryon-photon fluid, with gravitation as the driving and photon pressure as the restoring force. The fundamental frequency of these sound waves corresponds to the sound horizon, and thus the first peak at $l \sim 200$ indicates the horizon scale at recombination. The fact that these peaks are visible in the power spectrum requires that the oscillations are in phase on the whole LSS—what is a natural prediction of inflation: Since Fourier modes δ_k are frozen-in outside the horizon, their initial condition is $\delta_k = const.$ and $\dot{\delta}_k = 0$ at horizon crossing. In other words, all modes $\delta_k \sim \cos(\boldsymbol{kx} + \boldsymbol{\phi_k})$ start with $\boldsymbol{\phi_k} = 0$ at horizon crossing. The relative size of these peaks depends on the cosmological parameters; and the precise measurement of the CMB power spectrum has led to the standard model of cosmology.

Metric perturbations and CMB polarisation Another important prediction of inflation are tensor perturbations which lead to additional temperature fluctuations. Their derivation proceeds analogous to the scalar case, but now the non-vanishing Christoffel symbols in the conformal Newtonian gauge are given by ${\Gamma^0}_{ij} = -h'_{ij}$. Thus the geodesic equation becomes

$$\frac{\mathrm{d}P^0}{\mathrm{d}\eta} - \frac{1}{2}P^0 h'_{ij}n_i n_j = 0. \tag{23.155}$$

In linear order, we can neglect again the time dependence of P^0 and obtain immediately

$$\frac{P^0(\eta_0) - P^0(\eta_1)}{P^0(\eta_1)} = \frac{1}{2}\int_{\eta_1}^{\eta_0}\mathrm{d}\eta\, h'_{ij}n_i n_j. \tag{23.156}$$

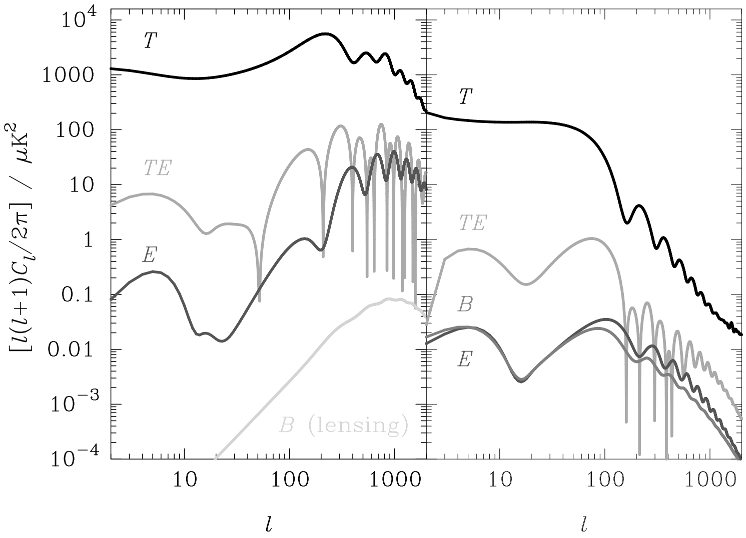


Figure 23.5.: CMB temperature and polarisation power spectra from scalar (left) and tensor perturbations (right) for a (unrealistic) tensor-to-scalar ratio $r = 0.38$, from [Cha06].

The perturbations of an ideal fluid contain no tensor component and thus the plasma four-velocity is undisturbed by metric perturbations, $u^\alpha = (1, 0)$. Therefore it is $\omega = P^0$, and only the ISW effect contributes to temperature fluctuations induced by tensor perturbations,

$$\Theta(\boldsymbol{n}, \eta_0) = \frac{1}{2} \int_{\eta_1}^{\eta_0} \mathrm{d}\eta \, h'_{ij} n_i n_j \, . \tag{23.157}$$

The connection between the angular power spectrum C_l and the primordial spectrum of tensor fluctuations is derived following the same logic as in the case of scalar perturbation, and we summarise therefore just the result: The temperature fluctuations induced by tensor perturbations have also a plateau up to $\lesssim 100$, and decay then faster than those of scalar perturbations, cf. with the black line in the right panel of Fig. 23.5. Since the relative size of tensor and scalar perturbations is bounded as $r < 0.1$, it seems thus hopeless to disentangle the two using only temperature fluctuations. What comes to our rescue is that the CMB is polarised and that one of the two polarisation states can be generated only by gravitational perturbations.

In order to understand how the CMB became polarised, we have to abandon the approximation of instantaneous recombination. Let us model instead the LSS as a layer of thickness $2\Delta\eta_r$: Then for $\eta_r - \Delta\eta_r$, photons are tightly coupled to the baryon fluid, while for $\eta_r + \Delta\eta_r$ they are free-streaming. In the intermediate region, they scatter on free electrons. The scattering is described by the non-relativistic Thomson cross section, $\sigma \propto |\boldsymbol{\varepsilon}' \cdot \boldsymbol{\varepsilon}|^2$, cf. with Eq. (9.69). If we choose linear polarisation vectors with $\boldsymbol{\varepsilon}_\parallel$ contained in the scattering plane spanned by $\boldsymbol{k}$ and $\boldsymbol{k}'$, and $\boldsymbol{\varepsilon}_\perp$ perpendicular to the plane, then $\sigma_\parallel \propto \cos^2\vartheta$ and $\sigma_\perp \propto 1$. Thus Thomson scattering generates linearly polarised photons. However, this is not sufficient for the generation of polarised radiation: If the initial photon intensity is isotropic, the polarisation is averaged out integrating over the directions of the initial photons. In the case of the CMB, the initial photon intensity is however anisotropic, because the intensity of black-body radiation is a function of T. Thus inhomogeneities $\delta T/T$ lead to fluctuations of the intensity which result in turn in fluctuations of the polarisation. This implies first that the fluctuations in temperatures and the degree of polarisation are correlated and second that the degree of polarisation is bounded by $\delta T/T$. Third, the CMB polarisation disappears in the limit of instantaneous recombination, $\Delta\eta_r \to 0$. More precisely, the polarisation is proportional to the ratio of the mean-free-path l of photons and the thickness $\Delta\eta_r$ of the LSS.

Next we discuss how we can describe the polarisation states of the CMB photons. The intensity I of a photon beam is determined by the square of the electric field strength vector, $I \propto \langle E^2 \rangle$, where the average $\langle \cdots \rangle$ is taken over one oscillation period. If we distinguish the two polarisation states of a photon, i.e. set $E \to E_a = \{E_+, E_-\}$, then the intensity becomes a tensor, I_{ab}. This tensor[4] is a Hermitian 2×2 matrix and has thus four real components, while the polarisation of a photon beam is fully described by two real parameters. We are only interested in the polarisation and we introduce therefore instead of I_{ab} the normalised, traceless 2×2 polarisation tensor

$$\mathcal{P}_{ab} = \frac{1}{\langle E_c E^c \rangle} \left(\langle E_a E_b \rangle - \frac{1}{2} \langle E_c E_c \rangle g_{ab} \right) . \tag{23.158}$$

Since we consider the polarisation on the LSS, we have to use for g_{ab} the two-dimensional metric tensor on S^2. We break the vector E_a into its rotation and divergence free part, which we call E and B, respectively,

$$\mathcal{P}_{ab} = \left(\nabla_a \nabla_b - \frac{1}{2} g_{ab} \right) E + \left(\varepsilon_{ac} \nabla_b \nabla_c + \varepsilon_{bc} \nabla_a \nabla_c \right) B . \tag{23.159}$$

From the three quantities T, E, and B, we can form six bilinear observables: the auto-correlations TT, EE, BB and the cross correlations TE, TB, EB. Since T and E are scalars, but B is a pseudo-scalar, the combinations TB and EB are zero (if parity is conserved). Moreover, the polarisation signal is weak and thus the cross correlation TE is (after TT) easiest to observe. Figure 23.5 shows additional to the temperature power spectra also the cross correlation and the E and B power spectrum.

Typical E and B modes are plotted in Fig. 23.6: The expression for $\mathcal{P}_{ab}$ shows that E polarisation transforms as a scalar, while B polarisation is a pseudo-scalar. As a result, E modes are symmetric with respect to reflections at a line through the center, while B modes change sign. This distinction makes B modes on super-horizon scales[5] to a "smoking gun" for the presence of gravitational waves during recombination: Since scalar perturbations are characterised by scalar functions, they can not lead to any B polarisation. On the other hand, gravitational waves consists of left- and right-circularly polarised waves, whose amplitudes $h^\pm$ are random variables. Thus in some directions left- and in other direction right-circular polarised waves dominate, leading in turn to a (locally) parity breaking B polarisation.

The detection of tensor perturbations would provide important information on the inflationary models. Such a measurement would inform us that *i)* B modes are correlated with temperature fluctuations. No causal mechanism can generate these correlations on superhorizon scales. *ii)* The tensor-scalar-ratio r measures the energy scale of inflation,

[4] Expanding I_{ab} on the sphere, we have to use spin-2 spherical harmonics, ${}^2Y_m^l(\vartheta, \phi)$. This indicates already that the (irreducible) spin-2 part of the polarisation signal is sourced by tensor perturbations of the metric.

[5] Weak gravitational lensing can transform E into B-modes, as a causal mechanism however only on subhorizon scales: this contribution is shown in the left panel of Fig. 23.5.

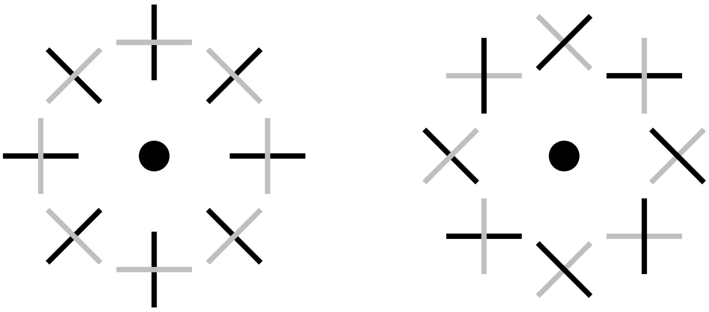


Figure 23.6.: The two E polarisation (left panel) and B polarisation (right panel) modes around a polarisation extremum shown by a circle.

$V^{1/4} \sim (r/0.01)^{1/4} 10^{16}$ GeV. *iii*) A large value $r > 0.01$ requires large-field models and thus trans-Planckian field values. *iv*) The observation of tensor perturbations would be a proof that gravity is a usual quantum theory, since they are sourced by quantum fluctuations.

Successes and problems of the inflationary paradigm Inflation has a set of successful predictions: First, the flatness of the universe, $\Omega \simeq 1$, which seemed in the 1980s and early 1990s at odds with observations. Second, inflation seeds scalar perturbations which are very close to Gaussian, nearly scale-invariant with a small red tilt, and lead, at least in the simplest models, to adiabatic temperature fluctuations of the CMB. Third, inflation predicts tensor perturbations which can be detected via B modes. Last but not least, the auto- and cross correlation of these fluctuations have fixed phase relations on superhorizon scales, which are difficult to understand in any causal mechanism. From these predictions, only the existence of tensor perturbations still awaits confirmation.

Having summarised the successes, we turn now to possible problems of the inflationary paradigm. First, we come back to the question if trans-Planckian field values of the inflaton are problematic. More precisely, we ask if, we can trust our analysis of large field models like e.g. $V = \frac{1}{2}m^2\phi^2$. From the point of view of general relativity, the answer is yes: If m is sufficiently small, the energy density satisfies $\rho \sim V \ll M_{\mathrm{pl}}^4$ and classical relativity holds. Moreover, the virtuality of the quantum fluctuations is close to zero and thus special relativity ensures that the large spatial momenta are not dangerous as long as their four-momentum squared is small. However, if we conceive a specific inflationary model $\mathscr{L}_{infl}$ as part of a complete effective field theory $\mathscr{L}_{\mathrm{B}SM}$ beyond the SM, then

$$\mathscr{L} = \mathscr{L}_{\mathrm{B}SM} + \mathscr{L}_{\mathrm{in}fl} + \sum_{n=1}^{\infty} \frac{c_n}{M_{\mathrm{P}l}^n} \mathcal{O}_{4+n} \,. \tag{23.160}$$

Usually, we can neglect the operators $\mathcal{O}_{4+n}$ of dimension five and higher, which are suppressed by powers of $(E/M_{\mathrm{P}l})^n$. However, in large field models the whole infinite tower of higher-dimensional operators should contribute during inflation with a priori equal weight. Thus there is no reason to trust the analysis of these models restricted to the operators with $d \leq 4$ contained in $\mathscr{L}_{BSM} + \mathscr{L}_{\mathrm{in}fl}$. If the upper limits on r continue to improve, this problem might become in the future however a purely academic one, Another potential problem becomes clear looking back at Eqs. (23.65) and (23.66). For a constant power spectrum, $\Delta^2_\phi(k) = H^2/(4\pi^2)$, the two-point function $\langle\delta\phi^2\rangle$ is logarithmically IR and UV divergent. The former divergence is caused by exponentially large wavelengths. These modes are homogeneous over the horizon scale both in the present and in the inflationary epoch. Therefore this IR divergence can be absorbed into a rescaling of the classical inflaton field. Considering the UV divergence, we have to recall that for field modes $k \gtrsim aH$ the spectrum is the one of a free field in Minkowski space. Thus the divergence of $\langle\delta\phi^2\rangle = \int \mathrm{d}k/k\,\Delta^2_\phi(k)$ is the same as in Minkowski space, and should be cured with our usual renormalisation procedure. Since the modes ϕ_k of a free field are independent, the subtraction of subhorizon modes should not affect the spectra of superhorizon modes in $\Delta^2_\phi(k)$. While thus the use of $\Delta^2_\phi(k) = H^2/(4\pi^2)$ seems to be in accord with our standard practise, other subtraction procedures are possible and would lead to different predictions. Moreover, we should keep in mind that our predictions depend on the choice of the vacuum state. Again, our selection of the Bunch-Davies vacuum is well-motivated, but cannot be derived from first principles.

Finally, let us comment on the naturalness of the initial conditions in inflation. We have seen that inflation in the homogeneous limit happens for a large set of initial values, considering e.g. the phase portrait for the classical $V = m^2\phi^2/2$ potential in Fig 23.2. Such large fluctuations should be rare, but the probability per volume to sit inside an inflationary patch should be large, precisely because these patches are exponentially inflated. In the "eternal inflation" scenario, this simple classical picture is changed since quantum fluctuations modify the classical inflation trajectories. It is unclear what measure for the initial conditions should be used and thus it is also disputed how natural the initial conditions for successful inflation are in this scenario.

23.A. Appendix: Perturbed Einstein equations

The derivation of the Einstein equations in the conformal Newtonian gauge (23.88) is straightforward but tedious. One may either calculate directly the linearised Einstein tensor for this metric, or one may use that the $k = 0$ FLRW metric is conformally flat. Therefore one can map the linearised Einstein equations for perturbations around Minkowski space derived in chapter 18.3 on the FLRW metric using the transformation rules (22.2) between conformally related space-times, see e.g. [GR11a]. Alternatively, one may use a program like differentialGeometry.py to calculate first the Einstein tensor $G_{\mu\nu}$ for the metric (23.88), expanding then $G_{\mu\nu}$ in Φ and Ψ. Either way, the Einstein tensor in the conformal Newtonian gauge at linear order in the perturbations Ψ and Φ follows as

$$\delta G^0_0 = \frac{2}{a^2}\left(\Delta\Psi - 3\frac{a'}{a}\Psi' - 3\frac{a'^2}{a^2}\Phi\right) \tag{23.161a}$$

$$\delta G^0_i = \frac{2}{a^2}\left(\partial_i\Psi' + \frac{a'}{a}\partial_i\Phi\right) \tag{23.161b}$$

$$\delta G^j_i = \frac{1}{a^2}\partial_i\partial^j(\Phi-\Psi) - \frac{2}{a^2}\delta^j_i\left(\Psi'' + \frac{1}{2}\Delta(\Phi-\Psi) + \frac{a'}{a}(\Phi'+2\Psi') + 2\frac{a''}{a}\Phi - \frac{a'^2}{a^2}\Phi\right). \tag{23.161c}$$

The spatial components δG^j_i of the Einstein tensor contain the two independent tensor structures $\partial_i\partial^j$ and δ^j_i. In the case of an ideal fluid or a scalar field, anisotropic pressure is absent,

$\delta T_i^j = \delta_i^j P$. As a result, the Einstein equation implies $\partial_i \partial^j(\Phi - \Psi) = 0$ for $i \neq j$. But $\Phi - \Psi$ cannot be constant, since the perturbed Einstein tensor should vanish in the limit $\Phi, \Psi \to 0$. Moreover, isotropy and homogeneity forbid a linear dependence of the potentials on the coordinates. Thus we found that

$$\Phi = \Phi \quad and \quad A = D \tag{23.162}$$

in the absence of anisotropic pressure.

Summary

Inflation denotes a phase of nearly exponential expansion of the early universe, which solves the horizon and flatness problem of the standard big bang model. It generates a nearly scale-invariant spectrum of Gaussian density fluctuations which is typically red-tilted, $n_s < 1$, and in the simplest models adiabatic. The fluctuations have fixed phase relations on superhorizon scales what results in characteristic oscillations of the CMB temperature fluctuations. The ratio r of power in gravitational waves and curvature perturbations is determined by the slow-roll parameter $\varepsilon = r/16$, which controls also the slope of the tensor perturbations $n_T = -2\varepsilon = -r/8$.

Further reading

Our derivation of the perturbation spectrum follows the one of [GR11a]. For a somewhat complementary approach see [LL09]. Both references as well as [Muk05] discuss also in much more detail the connection to CMB fluctuations. A more detailed review of preheating is given by [ABCRM10] and [GR11a].

24. Black holes

John Michell speculated already 1784 that the gravitational attraction of a mass concentrated inside a sufficiently small radius can become so strong that not even light escapes. The advent of general relativity put Mitchell's premise that energy is subject to gravity on a firm footing but nevertheless the idea of black holes was accepted only very slowly. In modern language, we call a black hole a solution of Einstein's equations containing a physical singularity which in turn is covered by an event horizon: The horizon acts as a perfect unidirectional membrane which any causal influence can cross only towards the singularity. After this picture became accepted in the 1960ies, the discovery by Hawking that quantum effects lead to the emission of thermal radiation by a black holes came as big surprise. Before we examine this process of Hawking radiation, we discuss the essential features of stationary and rotating black holes.

24.1. Schwarzschild black holes

Definitions Let us start by introducing few definitions: First, we need to distinguish between *physical* and *coordinate singularities*. The latter arise only for a specific coordinate choice and all measurable quantities remain finite at a coordinate singularity. By contrast, physical singularities can not be eliminated by a change of coordinates and physical quantities as the curvature or the stress tensor diverge. Next, we recall our definition of an *event horizon* as a three-dimensional hypersurface which limits a region of a space-time which can never influence an observer. The event horizon is formed by light-rays and is therefore a null surface. Hence we require that at each point of such a surface defined by $f(x^\mu) = 0$ a null tangent vector n^μ exists that is orthogonal to two space-like tangent vectors. The normal n^μ to this surface is parallel to the gradient along the surface, $n^\mu = h\nabla^\mu f = h\partial^\mu f$, where h is an arbitrary non-zero function. From

$$0 = n_\mu n^\mu = g_{\mu\nu} n^\mu n^\nu \tag{24.1}$$

we see that the line-element vanishes on the horizon, $\mathrm{d}s = 0$. Hence the (future) light cones at each point of an event horizon are tangential to the horizon.

We add two additional definitions for space-times with special symmetries. A *stationary* space-time has a time-like Killing vector field. In appropriate coordinates, the metric tensor is independent of the time coordinate,

$$\mathrm{d}s^2 = g_{00}(\boldsymbol{x})\mathrm{d}t^2 + 2g_{0i}(\boldsymbol{x})\mathrm{d}t\mathrm{d}x^i + g_{ij}(\boldsymbol{x})\mathrm{d}x^i\mathrm{d}x^j\,. \tag{24.2}$$

A stationary space-time is *static*, if it is invariant under time reversal. Thus the off-diagonal terms g_{0i} have to vanish, and the metric simplifies to

$$\mathrm{d}s^2 = g_{00}(\boldsymbol{x})\mathrm{d}t^2 + g_{ij}(\boldsymbol{x})\mathrm{d}x^i\mathrm{d}x^j\,. \tag{24.3}$$

An example for a stationary space-time is the metric around a spherically symmetric mass distribution which rotates with constant velocity. If the mass distribution is at rest, then the space-time becomes static.

Schwarzschild metric The Schwarzschild solution describes the static space-time outside a spherically symmetric mass distribution and is therefore of the form (24.3). Our discussion of symmetric spaces in section 19.1 implies that the spatial part g_{ij} of the metric tensor is given by Eq. (19.4) with $S(t) = const.$ Moreover, the metric tensor can depend only on the radial distance r to the center of the mass distribution. Thus the complete line-element is

$$\mathrm{d}s^2 = A(r)\mathrm{d}t^2 - B(r)\mathrm{d}r^2 - r^2(\mathrm{d}\vartheta^2 + \sin^2\vartheta\mathrm{d}\phi^2)\,, \tag{24.4}$$

where $A(r)$ and $B(r)$ are two arbitrary functions. Following the steps from Eq. (19.9a) to (19.13), but using now $T_{\mu\nu} = 0$ appropriate for the vacuum outside a spherical mass distribution, leads[1] to (problem 25.**??**)

$$\mathrm{d}s^2 = \left(1 - \frac{2M}{r}\right)\mathrm{d}t^2 - \left(1 - \frac{2M}{r}\right)^{-1}\mathrm{d}r^2 - r^2(\mathrm{d}\vartheta^2 + \sin^2\vartheta\mathrm{d}\phi^2)\,. \tag{24.5}$$

Here we required also that the metric is asymptotically flat, i.e. that we recover Minkowski space for $M/r \to 0$. The weak-field limit $r \gg 2M$ implies that M is the total mass as measured by an observer at infinity. The specific coordinates used in (24.5) which make the static property of the space-time manifest are called Schwarzschild coordinates. The Schwarzschild solution which is parametrised only by the mass M is the unique spherically symmetric vacuum solution of the Einstein equations: Allowing for time-dependent functions, $A(t,r)$ and $B(t,r)$ would result in the same *static* space-time, a result known as Birkhoff's theorem.

The main properties of the Schwarzschild solution can be summarised as follows: The time-independence and spherically symmetry of the metric imply the existence of four Killing vectors. If we order coordinates as $\{t, r, \phi, \vartheta\}$, then the two Killing vectors leading to the conservation of energy and z component of the angular momentum are $\xi^\mu \equiv (\xi^t, \xi^r, \xi^\phi, \xi^\vartheta) = (1,0,0,0)$ and $\eta^\mu = (0,0,0,1)$. The Schwarzschild coordinates have two singularities at $r = 2M$ and $r = 0$. The radius $2M$ is called Schwarzschild radius R_s and has numerically the value

$$R_s = 2M = \frac{2G_\mathrm{N}M}{c^2} \simeq 3\,\mathrm{k}m\,\frac{M}{M_\odot}\,, \tag{24.6}$$

where $M_\odot$ denotes the mass of the Sun. In a stationary, radial-symmetric space-time the general equation of a surface, $f(x^\mu) = 0$ simplifies to $f(r) = 0$. Then the condition defining a horizon becomes simply $g^{rr} = 0$ or $g_{rr} = 1/g^{rr} = \infty$. Thus $r = R_s$ is an event horizon.

Moreover, at R_s the coordinate t and r switch their character: for $r < R_s$, the "time coordinate" t becomes space-like, while r is time-like. In order to decide if $r = 2M$ and $r = 0$ are coordinate or physical singularities one can calculate the scalar invariants formed from the Riemann tensor. For instance, one finds $R_{\mu\nu\rho\sigma}R^{\mu\nu\rho\sigma} = 48M^2/r^6$, indicating that $r = 2M$ is a coordinate and $r = 0$ a physical singularity: Approaching $r = 0$, any macroscopic body would be destroyed by tidal forces. However, the question if at $r = 0$ a true singularity exists cannot be addressed within classical gravity which is expected to breakdown for curvatures larger than $R \sim M_{\mathrm{Pl}}^2$.

Gravitational redshift An observer with four-velocity $\boldsymbol{u}$ measures the frequency

$$\omega = \boldsymbol{k}\cdot\boldsymbol{u} \tag{24.7}$$

[1] We set $G_\mathrm{N} = 1$ in this chapter, if not otherwise stated.

of a photon with four-momentum $\boldsymbol{k}$. For an observer at rest,

$$\boldsymbol{u}\cdot\boldsymbol{u} = 1 = g_{tt}(u^t)^2\,, \tag{24.8}$$

and hence we can express $\boldsymbol{u}$ through the Killing vector $\boldsymbol{\xi}$ as

$$\boldsymbol{u} = (1-2M/r)^{-1/2}\boldsymbol{\xi}\,. \tag{24.9}$$

Inserting this expression into Eq. (24.7), we find for the frequency measured by an observer at the position r

$$\omega(r) = (1-2M/r)^{-1/2}\boldsymbol{\xi}\cdot\boldsymbol{k}\,. \tag{24.10}$$

Since $\boldsymbol{\xi}\cdot\boldsymbol{k}$ is conserved and $\omega_\infty = \boldsymbol{\xi}\cdot\boldsymbol{k}$, we obtain

$$\omega_\infty = \omega(r)\sqrt{1-\frac{2M}{r}}\,. \tag{24.11}$$

Thus a photon climbing out of the potential wall of the mass M looses energy, in agreement with the principle of equivalence. In the same way, a signal sent towards an observer at infinity by a spaceship falling towards $r=2M$ will be more and more redshifted, with $\omega_\infty \to 0$ for $r\to 2M$. Thus the event horizon at $r=2M$ is also an infinite redshift surface.

Radial infall into a black hole We investigate now the time dependence of a trajectory describing the infall of an object into a black hole. The geodesics of the Schwarzschild metric are easiest derived combining the normalisation condition of the four-velocity with the conserved quantities defined by the Killing vectors $\boldsymbol{\xi}$ and $\boldsymbol{\eta}$.

A test particle moving in the Schwarzschild metric initially in the radial direction will continue so, because then $\dot{u}^\phi = \dot{u}^\vartheta = 0$ (problem 25.??). The normalisation condition $\boldsymbol{u}\cdot\boldsymbol{u} = 1$ written out for a radial trajectory simplifies to

$$1 = A\left(\frac{\mathrm{d}t}{\mathrm{d}\tau}\right)^2 - A^{-1}\left(\frac{\mathrm{d}r}{\mathrm{d}\tau}\right)^2\,, \tag{24.12}$$

where we set also $A \equiv 1-2M/r$. Now we replace the velocity u^t by the conserved quantity

$$e \equiv \boldsymbol{\xi}\cdot\boldsymbol{u} = A\frac{\mathrm{d}t}{\mathrm{d}\tau}\,, \tag{24.13}$$

obtaining

$$1 = -\frac{e^2}{A} + \frac{1}{A}\left(\frac{\mathrm{d}r}{\mathrm{d}\tau}\right)^2\,. \tag{24.14}$$

We consider the free fall of a particle that was at rest at spatial infinity. Then the proper and coordinate time coincide for $r\to\infty$, $\mathrm{d}t/\mathrm{d}\tau = 1$, and thus $e^2 = 1$. Then the radial equation (24.14) simplifies to

$$\frac{1}{2}\left(\frac{\mathrm{d}r}{\mathrm{d}\tau}\right)^2 = -\frac{M}{r} \tag{24.15}$$

and can be integrated by separation of variables,

$$\int_r^0 \mathrm{d}r r^{1/2} = \sqrt{2M}\int_{\tau_*}^{\tau} \mathrm{d}\tau\,, \tag{24.16}$$

with the result

$$\frac{2}{3}r^{3/2} = \sqrt{2M}(\tau_* - \tau)\,. \tag{24.17}$$

Hence a freely falling particle needs only a finite proper time to fall from finite r to $r = 0$. In particular, it passes the Schwarzschild radius $2M$ in finite proper time, and no singular behaviour of the trajectory at $2M$ is apparent.

We can answer the same question using the coordinate time t by combining Eqs. (24.13) and (24.15),

$$\frac{\mathrm{d}t}{\mathrm{d}r} = \frac{\mathrm{d}t}{\mathrm{d}\tau}\frac{\mathrm{d}\tau}{\mathrm{d}r} = -\left(\frac{2M}{r}\right)^{-1/2}\left(1 - \frac{2M}{r}\right)^{-1}\,. \tag{24.18}$$

Integration gives

$$\begin{aligned} t - t_0 &= \int_{r_0}^{r} \mathrm{d}r' \left(\frac{2M}{r'}\right)^{-1/2}\left(1 - \frac{2M}{r'}\right)^{-1} = \qquad (24.19)\\ &= -2M\left[-\frac{2}{3}\left(\frac{r'}{2M}\right)^{3/2} - 2\left(\frac{r'}{2M}\right)^{1/2} + \ln\left|\frac{\sqrt{r'/2M}+1}{\sqrt{r'/2M}-1}\right|\right]_{r_0}^{r} \to \infty \quad \text{for} \quad r \to 2M. \end{aligned}$$

Since the coordinate time t is the proper time for an observer at infinity, a freely falling particle reaches the Schwarzschild radius $r = 2M$ only for $t \to \infty$ as seen from spatial infinity. The last result can be derived immediately for light-rays. Choosing a light-ray in radial direction with $\mathrm{d}\phi = \mathrm{d}\vartheta = 0$, the metric (24.5) simplifies with $\mathrm{d}s^2 = 0$ to

$$\frac{\mathrm{d}r}{\mathrm{d}t} = 1 - \frac{2M}{r}\,. \tag{24.20}$$

Thus light travelling towards the star, as seen from the outside, will travel slower and slower as it comes closer to the Schwarzschild radius $r = 2M$. The coordinate time is $\propto \ln|1 - 2M/r|$ and thus for an observer at infinity the signal will reach $r = 2M$ again only asymptotically for $t \to \infty$.

We noted already that at $R_s = 2M$ the coordinates t and r switch their character, r becoming time-like. Since the proper time τ of an observer has to increases continuously, the time-like coordinate has to change continuously too. Because of $\mathrm{d}r/\mathrm{d}t = 1 - 2M/r < 0$, we anticipate therefore that r has to decrease continuously for any particle that crossed the horizon until it hits the singularity at $r = 0$.

Eddington-Finkelstein coordinates We next try to find new coordinates which are regular at $r = 2M$ and valid in the whole range $0 < r < \infty$. Such a coordinate transformation has to be singular at $r = 2M$, otherwise we can not hope to cancel the singularity present in the Schwarzschild coordinates. We can eliminate the troublesome factor $g_{rr} = (1 - \frac{2M}{r})^{-1}$ introducing a new radial coordinate r^* defined by

$$\mathrm{d}r^* = \frac{\mathrm{d}r}{1 - \frac{2M}{r}}\,. \tag{24.21}$$

Integrating (24.21) results in

$$r^*(r) = r + 2M\ln\left|\frac{r}{2M} - 1\right| + A\,, \tag{24.22}$$

with $A \equiv -2Ma$ as integration constant. The coordinate $r^*(r)$ is often called tortoise coordinate, because $r^*(r)$ changes only logarithmically close to the horizon. This coordinate change maps the range $r \in [2M, \infty]$ of the radial coordinate onto $r^* \in [-\infty, \infty]$. A radial null geodesics satisfies $\mathrm{d}(t \pm r^*) = 0$, and thus in- and out-going light-rays are given by

$$\tilde{u} \equiv t - r^* = t - r - 2M \ln \left| \frac{r}{2M} - 1 \right| - A \,, \qquad \text{outgoing rays}\,, \tag{24.23}$$

$$\tilde{v} \equiv t + r^* = t + r + 2M \ln \left| \frac{r}{2M} - 1 \right| + A \,, \qquad \text{ingoing rays}\,. \tag{24.24}$$

For $r > 2M$, Eq. (24.20) implies that $\mathrm{d}r/\mathrm{d}t > 0$ so that r increases with t. Therefore (24.23) describes outgoing light-rays, while (24.24) corresponds to ingoing light-rays for $r > 2M$.

We can extend now the Schwarzschild metric using as coordinate the "advanced time parameter $\tilde{v}$" instead of t. Forming the differential,

$$\mathrm{d}\tilde{v} = \mathrm{d}t + \mathrm{d}r + \left(\frac{r}{2M} - 1\right)^{-1} \mathrm{d}r = \mathrm{d}t + \left(1 - \frac{2M}{r}\right)^{-1} \mathrm{d}r \,, \tag{24.25}$$

we can eliminate $\mathrm{d}t$ from the Schwarzschild metric and find

$$\mathrm{d}s^2 = \left(1 - \frac{2M}{r}\right) \mathrm{d}\tilde{v}^2 - 2\mathrm{d}\tilde{v}\mathrm{d}r - r^2 \mathrm{d}\Omega \,. \tag{24.26}$$

This metric was found first by Eddington and was later rediscovered by Finkelstein. Although $g_{\tilde{v}\tilde{v}}$ vanishes at $r = 2M$, the determinant $g = r^4 \sin^2 \vartheta$ is non-zero at the horizon and thus the metric is invertible. Moreover, r^* was defined by (24.22) initially only for $r > 2M$, but we can use this definition also for $r < 2M$, arriving at the same expression (24.26). Therefore, the metric using the advanced time parameter $\tilde{v}$ is regular at $2M$ and valid for all $r > 0$. We can view this metric hence as an extension of the $r > 2M$ part of the Schwarzschild solution, similar to the process of analytic continuation of complex functions. The price we have to pay for a non-zero determinant at $r = 2M$ are non-diagonal terms in the metric: As a result, the space-time described by (24.26) is not symmetric under the exchange $t \to -t$. We will see shortly the consequences of this asymmetry.

We now study the behaviour of radial light-rays, which are determined by $\mathrm{d}s^2 = 0$ and $\mathrm{d}\phi = \mathrm{d}\vartheta = 0$. Thus radial light-rays satisfy $A\mathrm{d}\tilde{v}^2 - 2\mathrm{d}\tilde{v}\mathrm{d}r = 0$, which is trivially solved by ingoing light-rays, $\mathrm{d}\tilde{v} = 0$ and thus $\tilde{v} = const$. The solutions for $\mathrm{d}\tilde{v} \neq 0$ are given by (24.23). Additionally, the horizon $r = 2M$ which is formed by stationary light-rays satisfies $\mathrm{d}s^2 = 0$. In order to draw a space-time diagram, it is more convenient to replace the light-like coordinate $\tilde{v}$ by a new time-like coordinate. We show in the left panel of Fig. 24.1 geodesics using as new time coordinate $\tilde{t} = \tilde{v} - r$. Then the ingoing light-rays are straight lines at 45 degrees with the r axis. Radial light-rays which are outgoing for $r > 2M$ and ingoing for $r < 2M$ follow Eq. (24.24). A few future light-cones are indicated: They are formed by the intersection of light-rays, and they tilt towards $r = 0$ as they approach the horizon. At $r = 2M$, one light-ray forming the light-cone becomes stationary and part of the horizon, while the remaining part of the cone lies completely inside the horizon.

Let us now discuss how Fig. 24.1 would like using the retarded Eddington-Finkelstein coordinate $\tilde{u}$. Now the outgoing radial null geodesics are straight lines at 45°. They start from the singularity, crossing smoothly $r = 2M$ and continue to spatial infinity. Such a situation, where the singularity is not covered by an event horizon is called a "white hole." The cosmic censorship hypothesis postulates that singularities formed in gravitational collapse

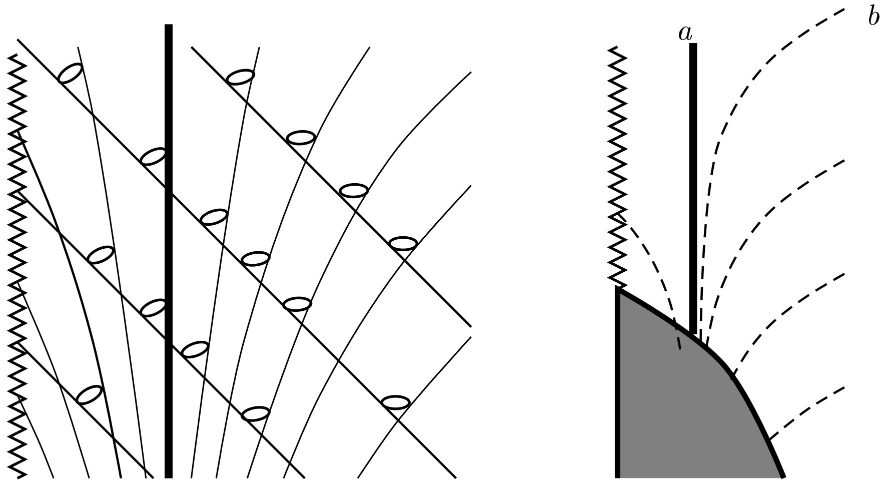


Figure 24.1.: Left: The Schwarzschild space-time using advanced Eddington-Finkelstein coordinates; the singularity is shown by a zigzag line, the horizon by a thick and geodesics by thin lines. Right: Collapse of a star modelled by pressureless matter; dashes lines show geodesics, the thin solid line encompasses the collapsing stellar surface.

are always covered by event horizons. This implies that the time-invariance of the Einstein equations is broken by its solutions. In particular, only the BH solution using the retarded Eddington-Finkelstein coordinates should be realised by nature—otherwise we should expect that causality is violated. This behaviour may be compared to classical electrodynamics, where all solutions are described by the retarded Green function, while the advanced Green function seems to have no relevance.

Collapse to a BH After a star has consumed its nuclear fuel, gravity can be balanced only by the Fermi degeneracy pressure of its constituents. Increasing the total mass of the star remnant, the stellar EoS is driven towards the relativistic regime until the star becomes unstable. As a result, the collapse of its core to a BH seems to be inevitable for a sufficiently heavy star.

Let us consider a toy model for such a gravitational collapse. We describe the star by a spherically symmetric cloud of pressureless matter. While the assumption of negligible pressure is unrealistic, it implies that particles at the surface of the star follow radial geodesics in the Schwarzschild space-time. Thus we do not have to bother about the interior solution of the star, where $T_{\mu\nu} \neq 0$ and our vacuum solution does not apply. In advanced Eddington-Finkelstein coordinates, the collapse is schematically shown in the right panel of Fig. 24.1. At the end of the collapse, a stationary Schwarzschild BH has formed. Note that in our toy model the event horizon forms before the singularity, as required by the cosmic censorship hypothesis. The horizon grows from $r = 0$ following the light-like geodesic a shown by the thin black line until it reaches its final size $R_s = 2M$. What happens, if we drop latter a lump of matter δM on a radial geodesics into the BH? Since we do not add angular momentum to the BH, the final stage is according to the Birkhoff's theorem still a Schwarzschild BH: All deviations from spherical symmetry corresponding to gradient energy in the intermediate regime are being radiated away as gravitational waves. Thus in the final stage, the only change is an increase of the horizon, size $R_s \to 2(M + \delta M)$. Therefore some light-rays (e.g. b) which we expected to escape to spatial infinity will be trapped. Similarly, light-ray a which we thought to form the horizon will be deflected by the increased gravitational attraction towards the singularity. In essence, knowing only the space-time up to a fixed time t we are not able to decide which light-rays form the horizon. The event horizon of a black hole is a global property of the space-time: It is not only independent of the observer, but influenced by the complete space-time.

How does the stellar collapse looks like for an observer at large distances? Let us assume that the observer uses a neutrino detector and is able to measure the neutrino luminosity $L_\nu(r) = \mathrm{d}E_\nu/\mathrm{d}t = N_\nu\omega_\nu/\mathrm{d}t$ emitted by a shell of stellar material at radius r. In order to determine the luminosity $L_\nu(r)$, we have to connect r and t. Linearising Eq. (24.19) around $r = 2M$ gives

$$\frac{r - 2M}{r_0 - 2M} = \mathrm{e}^{-(t-t_0)/2M} . \tag{24.27}$$

For an observer at large distance r_0, the time difference between two pulses send by a shell falling into a BH increases thus exponentially for $r \to 2M$. As a result the energy ω_ν of an individual neutrino is also exponentially redshifted

$$\omega_\nu(r) = \omega_\nu(r_0)\mathrm{e}^{-(t-t_0)/2M} . \tag{24.28}$$

A more detailed analysis confirms the expectation that then also the luminosity decreases exponentially. Thus an observer at infinity will not see shells which slow logarithmically down as they fall towards $r \to 2M$, as it is suggested by Eq. (24.19). Instead the signal emitted by the shell will fade away exponentially, with the short characteristic time scale of $M = Mt_{\mathrm{Pl}}/M_{\mathrm{Pl}} \approx 10^{-5}\,$s for a stellar-size BH.

Kruskal coordinates We have been able to extend the Schwarzschild solution into two different branches, a BH solution using the advanced time parameter $\tilde{v}$ and a white hole solution using the retarded time parameter $\tilde{u}$. The analogy with the analytic continuation of complex functions leads naturally to the question, if we can combine these two branches into one common solution. Moreover, our experience with the Rindler metric suggests that an event horizon where energies are exponentially redshifted implies the emission of a thermal spectrum: If true, our BH would not be black after all. One way to test this suggestion is to relate the vacua as defined by different observers via a Bogolyubov transformation. In order to simplify this process, we would like to find new coordinates for which the Schwarzschild space-time is conformally flat.

An obvious try to proceed is to use both the advanced and the retarded time parameters. For most of our discussion, it is sufficient to concentrate on the t, r coordinates in the line-element $\mathrm{d}s^2 = \mathrm{d}\bar{s}^2 + r^2\mathrm{d}\Omega$, and to neglect the angular dependence from the $r^2\mathrm{d}\Omega$ part. We start by eliminating r in favour of r^*,

$$\mathrm{d}\bar{s}^2 = \left(1 - \frac{2M}{r(r^*)}\right)(\mathrm{d}t^2 - \mathrm{d}r^{*2}) , \tag{24.29}$$

where r has to be expressed through r^*. This metric is conformally flat, but the definition of $r(r^*)$ on the horizon contains the ill-defined factor $\ln(2m/r - 1)$. Clearly, a new set of coordinates where this factor is exponentiated is what we are looking for.

This is achieved introducing both Eddington-Finkelstein parameters,

$$\tilde{u} = t - r^* , \qquad \tilde{v} = t + r^* , \tag{24.30}$$

for which the metric simplifies to

$$d\bar{s}^2 = \left(1 - \frac{2M}{r(\tilde{u},\tilde{v})}\right) d\tilde{u}d\tilde{v} . \tag{24.31}$$

From (24.22) and (24.30), it follows

$$\frac{\tilde{v}-\tilde{u}}{2} = r^*(r) = r + 2M \ln\left|\frac{r}{2M} - 1\right| - 2Ma , \tag{24.32}$$

or

$$1 - \frac{2M}{r} = \frac{2M}{r} \exp\left(\frac{\tilde{v}-\tilde{u}}{4M}\right) \exp\left(a - \frac{r}{2M}\right) . \tag{24.33}$$

This allows us to eliminate the singular factor $1 - 2M/r$ in (24.31), obtaining

$$d\bar{s}^2 = \frac{2M}{r} \exp\left(a - \frac{r}{2M}\right) \exp\left(-\frac{\tilde{u}}{4M}\right) d\tilde{u} \exp\left(\frac{\tilde{v}}{4M}\right) d\tilde{v} . \tag{24.34}$$

Finally, we change to Kruskal light-cone coordinates u and v defined by

$$u = -4M \exp\left(-\frac{\tilde{u}}{4M}\right) \quad and \quad v = 4M \exp\left(\frac{\tilde{v}}{4M}\right) , \tag{24.35}$$

arriving at

$$ds^2 = \frac{2M}{r} \exp\left(a - \frac{r}{2M}\right) dudv + r^2 d\Omega . \tag{24.36}$$

Kruskal diagram The coordinates $\tilde{u}, \tilde{v}$ cover only the exterior $r > 2M$ of the Schwarzschild space-time, and thus u, v are initially only defined for $r > 2M$. Since they are regular at the Schwarzschild radius, we can extend these coordinates towards $r = 0$. In order to draw the space-time diagram of the full Schwarzschild space-time shown in Fig. 24.2, it is useful to go back to time- and space-like coordinates via

$$u = T - R \quad and \quad v = T + R . \tag{24.37}$$

Then the connection between the pair of coordinates $\{T, R\}$, $\{u, v\}$ and $\{t, r\}$ is given by

$$uv = T^2 - R^2 = -16M^2 \exp\left(\frac{r^*}{2M}\right) = -16M^2 \left(\frac{r}{2M} - 1\right) \exp\left(\frac{r}{2M} - a\right) , \tag{24.38a}$$

$$\frac{u}{v} = \frac{T-R}{T+R} = \exp\left[-t/(2M)\right] . \tag{24.38b}$$

Lines with $r = const.$ are given by $uv = T^2 - R^2 = const.$ They are thus parabola shown as dotted lines in Fig. 24.2. Lines with $t = const.$ are determined by $u/v = const.$ and are thus given by straight (solid) lines through zero. In particular, null geodesics correspond to straight lines with angle 45° in the $R - T$ diagram. The horizon $r = 2M$ is given by to $u = 0$ or $v = 0$. Hence two separate horizons exist: A past horizon at $t = -\infty$ (for $v = 0$) and

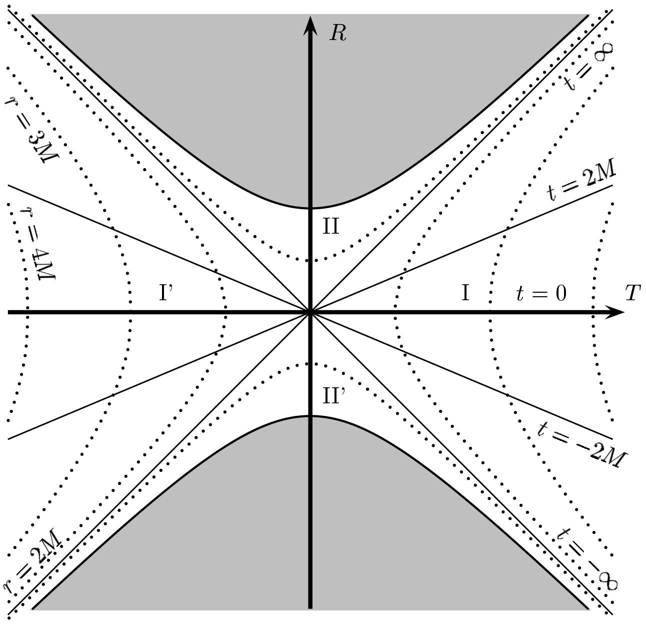


Figure 24.2.: Space-time diagram for the Kruskal coordinates T and R.

a future horizon at $t = +\infty$ (for $u = 0$). Also the singularity at $r = 0$ corresponds to two separate lines in the $R - T$ Kruskal diagram[2] and is given by

$$T = \pm\sqrt{16M^2 + R^2}\,. \tag{24.39}$$

The horizon lines $\{t = -\infty, r = 2M\}$ and $\{t = \infty, r = 2M\}$ divide the space-time in four parts: The future singularity is unavoidable in part II, while in region II' all trajectories start at the past singularity. Region I corresponds to the original Schwarzschild solution outside the horizon $r > 2M$, while region I and II encompass the advanced Eddington-Finkelstein solution. The regions I' and II' represent the retarded Eddington-Finkelstein solution, where II' corresponds to a white hole. Note that I' represents a new asymptotically flat Schwarzschild exterior solution.

The presence of a past horizon $v = 0$ at $t = -\infty$ makes the complete BH solutions time-symmetric and corresponds to an eternal BH. If we model a realistic BH, i.e. one that was created at finite t by a collapsing mass distribution, with Kruskal coordinates, then any effect induced by the past horizon should be considered as unphysical.

24.2. Kerr black holes

The stationary space-time outside a rotating mass distribution can be derived by symmetry arguments similarly to the case of the Schwarzschild metric, but it was found first accidentally by R. Kerr in 1963. The black hole solutions of this space-time is fully characterised by two quantities, the mass M and the angular momentum L of the Kerr BH. Both parameters can be

[2] Recall that we suppress two space dimension: Thus a point in the $R - T$ Kruskal diagram correspond to a sphere S^2, and a line to $\mathbb{R} \times S^2$.

manipulated, at least in a Gedankenexperiment, dropping material into the BH: Examining the response of a Kerr black hole to such changes was crucial for the discovery of "black hole thermodynamics."

In Boyer-Lindquist coordinates, the metric outside of a rotating mass distribution is given by

$$\begin{aligned}\mathrm{d}s^2 = &\left(1-\frac{2Mr}{\rho^2}\right)\mathrm{d}t^2 + \frac{4Mar\sin^2\vartheta}{\rho^2}\mathrm{d}\phi\mathrm{d}t - \frac{\rho^2}{\Delta}\mathrm{d}r^2 - \rho^2\mathrm{d}\vartheta^2 \\ &-\left(r^2+a^2+\frac{2Mra^2\sin^2\vartheta}{\rho^2}\right)\sin^2\vartheta\mathrm{d}\phi^2\,,\end{aligned} \tag{24.40}$$

with the abbreviations

$$a = L/M\,, \qquad \rho^2 = r^2 + a^2\cos^2\vartheta\,, \qquad \Delta = r^2 - 2Mr + a^2\,. \tag{24.41}$$

The metric is time-independent and axially symmetric. Hence two obvious Killing vectors are as in the Schwarzschild case $\boldsymbol{\xi} = (1,0,0,0)$ and $\boldsymbol{\eta} = (0,0,0,1)$, where we order again coordinates as $\{t,r,\vartheta,\phi\}$. The presence of the mixed term $g_{t\phi}$ means that the metric is stationary, but not static – as one expects for a star or BH rotating with constant rotation velocity. Finally, the metric is asymptotically flat and the weak-field limit shows that L is the angular momentum of the rotating black hole.

Singularity We examine first the potential singularities at $\rho = 0$ and $\Delta = 0$. The calculation of the scalar invariants formed from the Riemann tensor shows that only $\rho = 0$ is a physical singularity, while $\Delta = 0$ corresponds to a coordinate singularity. The physical singularity at $\rho^2 = 0 = r^2 + a^2\cos\vartheta^2$ corresponds to $r = 0$ and $\vartheta = \pi/2$. Thus the value $r = 0$ is surprisingly not compatible with all ϑ values. To understand this point, we consider the $M \to 0$ limit of the Kerr metric (24.40) keeping $a = L/M$ fixed,

$$\mathrm{d}s^2 = \mathrm{d}t^2 - \frac{\rho}{r^2+a^2}\mathrm{d}r^2 - r^2\mathrm{d}\vartheta^2 - (r^2+a^2)\sin^2\vartheta\mathrm{d}\phi^2\,. \tag{24.42}$$

The comparison with the Minkowski metric shows that

$$\begin{aligned} x &= \sqrt{r^2+a^2}\sin\vartheta\cos\phi\,, \qquad & z &= r\cos\vartheta\,, \\ y &= \sqrt{r^2+a^2}\sin\vartheta\sin\phi\,, \end{aligned} \tag{24.43}$$

Hence the singularity at $r = 0$ and $\vartheta = \pi/2$ corresponds to a ring of radius a in the equatorial plane $z = 0$ of the Kerr black hole.

Horizons We have defined an event horizon as a three-dimensional hypersurface, $f(x^\mu) = 0$, that is null. In a stationary, axisymmetric space-time the general equation of a surface, $f(x^\mu) = 0$, simplifies to $f(r,\vartheta) = 0$. The condition for a null surface becomes

$$0 = g^{\mu\nu}(\partial_\mu f)(\partial_\nu f) = g^{rr}(\partial_r f)^2 + g^{\vartheta\vartheta}(\partial_\vartheta f)^2\,. \tag{24.44}$$

In the case of the surface defined by the coordinate singularity $\Delta = r^2 - 2Mr + a^2 = 0$ that depends only on r,

$$r_\pm = M \pm \sqrt{M^2 - a^2}\,. \tag{24.45}$$

the condition defining a horizons becomes simply $g^{rr}=0$ or $g_{rr}=1/g^{rr}=\infty$. Hence, r_- and r_+ define an inner and outer horizon around a Kerr black hole.

The surface A of the outer horizon follows from inserting r_+ together with $\mathrm{d}r=\mathrm{d}t=0$ into the metric,

$$\mathrm{d}s^2=\rho_+^2\mathrm{d}\vartheta^2+\left(r_+^2+a^2+\frac{2Mr_+a^2\sin^2\vartheta}{\rho_+^2}\right)\sin^2\vartheta\mathrm{d}\phi^2\,, \tag{24.46}$$

Using $r_\pm^2+a^2=2Mr_\pm$, we obtain

$$\mathrm{d}s^2=\rho_+^2\mathrm{d}\vartheta^2+\left(\frac{2Mr_+}{\rho_+}\right)^2\sin^2\vartheta\mathrm{d}\phi^2\,. \tag{24.47}$$

Hence the metric determinant g_2 restricted to the angular variables is given by $\sqrt{g_2}=g_{\vartheta\vartheta}g_{\phi\phi}=2Mr_+\sin\vartheta$ and integration gives the area A of the horizon as

$$A=\int_0^{2\pi}\mathrm{d}\phi\int_0^{\pi}\mathrm{d}\vartheta\sqrt{g_2}=8\pi Mr_+=8\pi M(M+\sqrt{M^2-a^2})\,. \tag{24.48}$$

Note that the area depends on the angular momentum of the black hole that can in turn be manipulated by dropping material into the hole: The horizon area A for fixed mass M becomes maximal for a non-rotating black hole, $A=16\pi M^2$, and decreases to $A=8\pi M^2$ for a maximally rotating one with $a=M$. For $a>M$, the metric component $g^{rr}=\Delta$ has no real zero and thus no event horizon exists.

Ergosphere and dragging of inertial frames The Kerr metric is a special case of a metric with $g_{t\phi}\neq 0$. As result, both massive and massless particles with zero angular momentum falling into a Kerr black hole will acquire a non-zero angular rotation velocity $\omega=\mathrm{d}\phi/\mathrm{d}t$ as seen by an observer from infinity.

We consider a light-ray with $\mathrm{d}\vartheta=\mathrm{d}r=0$. Then the line-element becomes

$$g_{tt}\mathrm{d}t^2+2g_{t\phi}\mathrm{d}t\mathrm{d}\phi+g_{\phi\phi}\mathrm{d}\phi^2=0\,. \tag{24.49}$$

Dividing by $g_{\phi\phi}\mathrm{d}t^2$ we find as two possible solutions for the angular rotation velocity

$$\dot\phi_\pm=-\frac{g_{t\phi}}{g_{\phi\phi}}\pm\sqrt{\left(\frac{g_{t\phi}}{g_{\phi\phi}}\right)^2-\frac{g_{tt}}{g_{\phi\phi}}}\,. \tag{24.50}$$

There are two interesting special cases of this equation. First, on the surface $g_{tt}=0$, the two possible solutions of $\omega=\mathrm{d}\phi/\mathrm{d}t$ for light-rays satisfy

$$\dot\phi_+=-2\,\frac{g_{t\phi}}{g_{\phi\phi}}\qquad and\qquad \dot\phi_-=0\,. \tag{24.51}$$

Hence, the rotating black hole drags space-time at $g_{tt}=0$ so strongly that even a photon can only co-rotate. Similarly, this condition specifies a surface inside which no stationary observers are possible: The normalisation condition $\boldsymbol{u}\cdot\boldsymbol{u}=1$ is inconsistent with $u^a=(1,0,0,0)$ and $g_{tt}<0$: However strong your rocket engines are, your space-ship will not be able to hover at the same point inside the region with $g_{tt} < 0$. Therefore one calls a surface with $g_{tt} = 0$ a stationary limit surface. Solving

$$g_{tt} = 1 - \frac{2Mr}{\rho^2} = 0\,, \tag{24.52}$$

we find the position of the two stationary limit surfaces at

$$r_{1/2} = M \pm \sqrt{M^2 - a\cos\vartheta}\,. \tag{24.53}$$

The *ergosphere* is the space bounded by these two surfaces.

The other interesting special case of (24.50) occurs when the allowed range of values, $\phi_- \le \omega \le \phi_+$, shrinks to a single value, i.e. when

$$\omega^2 = \frac{g_{tt}}{g_{\phi\phi}} = \left(\frac{g_{t\phi}}{g_{\phi\phi}}\right)^2 . \tag{24.54}$$

This happens at the outer horizon r_+ and defines the rotation velocity ω_H of the black hole. In the case of a Kerr black hole, we find

$$\omega_H = \frac{a}{2Mr_+}\,. \tag{24.55}$$

Thus the rotation velocity of the black hole corresponds to the rotation velocity of the light-rays forming its horizon, as seen by an observer at spatial infinity.

Penrose process and the area theorem The total energy of a Kerr BH consists of its rest energy and its rotational energy. These two quantities control the size of the event horizon and therefore it is important to understand how they change dropping matter into the BH.

The energy of any particle moving on a geodesics is conserved, $E = -\boldsymbol{p}\cdot\boldsymbol{\xi}$. Inside the ergosphere, the Killing vector $\boldsymbol{\xi}$ is space-like and the quantity E is thus the component of a spatial momentum which can have both signs. This lead Penrose to entertain the following Gedankenexperiment: Suppose the space craft A starts at infinity and falls into the ergosphere. There it splits into two parts: B is dropped into the BH, while C escapes to infinity. In the splitting process, four-momentum has to be conserved, $\boldsymbol{p}_A = \boldsymbol{p}_B + \boldsymbol{p}_C$. We can now choose a time-like geodesics for B falling into the BH such that $E_B < 0$. Then $E_C > E_A$ and the escaping part C of the space craft has at infinity a higher energy than initially.

The Penrose process decreases both the mass and the angular momentum of the BH by an amount equal to that of the space craft B falling into the BH. We want to show now that the changes are correlated in such a way that the area of the BH increases. Let us first define a new Killing vector,

$$\boldsymbol{K} = \boldsymbol{\xi} + \omega_H \boldsymbol{\eta}\,.$$

This Killing vector is null on the horizon and time-like outside. It corresponds to the four-velocity with the maximal possible rotation velocity. Now we use $E_B = -\boldsymbol{p}_C\cdot\boldsymbol{\xi}$ and $L_B = -\boldsymbol{p}_C\cdot\boldsymbol{\eta}$ and

$$\boldsymbol{p}_B\cdot\boldsymbol{K} = \boldsymbol{p}_B\cdot(\boldsymbol{\xi} + \omega_H\boldsymbol{\eta}) = -(E_B - \omega_H L_B) < 0, \tag{24.56}$$

to obtain the bound $L_B < E_B/\omega_H$. Since $E_B < 0$, the added angular momentum is negative, $L_B < 0$.

The mass and the angular momentum of the BH change by $\delta M = E_B$ and $\delta L = L_B$, when particle B drops into the BH. Thus

$$\delta M > \omega_H \delta L = \frac{a\delta L}{r_+^2 + a^2} \tag{24.57}$$

Now we define the irreducible mass of BH as the mass of that Schwarzschild BH which event horizon has the same area,

$$M_{\mathrm{i}rr}^2 = \frac{1}{2}(M^2 + \sqrt{M^2 - L^2}) \tag{24.58}$$

or

$$M^2 = M_{\mathrm{i}rr}^2 + \left(\frac{L}{2M_{\mathrm{i}rr}}\right)^2 . \tag{24.59}$$

Thus we can interpret the total mass as the Pythagorean sum of the irreducible mass and a contribution related to the rotational energy. Differentiating the relation (24.58) results in

$$\delta M_{\mathrm{i}rr} = \frac{a}{4M_{\mathrm{i}rr}\sqrt{M^2 - a^2}} \left(\omega_H^{-1}\delta M - \delta L\right) \tag{24.60}$$

Our bound implies now $\delta M_{\mathrm{i}rr} > 0$ or $\delta A > 0$. Thus the surface of a Kerr BH can only increase, even when its mass decreases.

24.3. Black hole thermodynamics and Hawking radiation

Bekenstein entropy We have shown that classically the horizon of a black hole can only increase with time. The only other quantity in physics with the same property is the entropy, $\mathrm{d}S \geq 0$. This suggests a connection between the horizon area and its entropy. To derive this relation, we apply the first law of thermodynamics $\mathrm{d}U = T\mathrm{d}S - P\mathrm{d}V + \ldots$ to a Kerr black hole. Its internal energy U is given by $U = M$ and thus

$$\mathrm{d}U = \mathrm{d}M = T\mathrm{d}S - \boldsymbol{\omega}\mathrm{d}\boldsymbol{L} , \tag{24.61}$$

where $\boldsymbol{\omega}\mathrm{d}\boldsymbol{L}$ denotes the mechanical work done on a rotating macroscopic body.

Our experience with the thermodynamics of non-gravitating systems suggests that the entropy is an extensive quantity and thus proportional to the volume, $S \propto V$. We give now an argument, that shows that the entropy S of a black hole is proportional to its area A. We introduce the "rationalised area" $\alpha = A/4\pi = 2Mr_+$, cf. (24.48), or

$$\alpha = 2M^2 + 2\sqrt{M^4 - L^2} . \tag{24.62}$$

The parameters describing a Kerr black hole are its mass M and its angular momentum L and thus $\alpha = \alpha(M, L)$. We form the differential $\mathrm{d}\alpha$ and find after some algebra (problem 25.**??**)

$$\frac{\sqrt{M^2 - a^2}}{2\alpha} \mathrm{d}\alpha = \mathrm{d}M + \frac{\boldsymbol{a}}{\alpha}\mathrm{d}\boldsymbol{L} . \tag{24.63}$$

Using now Eqs. (24.48) and (24.55), we can rewrite the RHS as

$$\frac{\sqrt{M^2 - a^2}}{2\alpha} \mathrm{d}\alpha = \mathrm{d}M + \boldsymbol{\omega}_H\mathrm{d}\boldsymbol{L} . \tag{24.64}$$

Thus the first law of black hole thermodynamics predicts the correct angular velocity ω_H of a Kerr black hole. Including the term $\Phi \mathrm{d}q$ representing the work done by adding the charge $\mathrm{d}q$ to a black hole, the area law of a charged black hole together with the first law of BH thermodynamics reproduces the correct surface potential Φ of a charged black hole.

The factor in front of $\mathrm{d}\alpha$ is positive, as its interpretation as temperature requires. We identify

$$T\mathrm{d}S = \frac{\sqrt{M^2 - a^2}}{2\alpha}\,\mathrm{d}\alpha \tag{24.65}$$

and thus $S = f(A)$. The validity of the area theorem requires that f is a linear function, the proportionality coefficient between S and A can be only determined by calculating the temperature of black hole. Hawking could show 1974 that a black hole in vacuum emits black-body radiation ("Hawking radiation") with temperature

$$T = \frac{2\sqrt{M^2 - a^2}}{A} \tag{24.66}$$

and thus

$$S = \frac{kc^3}{4\hbar G}\,A = \frac{A}{4L_{\mathrm{P}l}^2}\,. \tag{24.67}$$

The entropy of a black hole is not extensive, but proportional to its surface. It is large, because its basic unit of entropy, $4L_{\mathrm{P}l}^2$, is so tiny. The presence of $\hbar$ in the first formula, where we have inserted the natural constants, signals that the black hole entropy is a quantum property.

The heat capacity C_V of a Schwarzschild black hole follows with $U = M = 1/(8\pi T)$ from the definition

$$C_V = \frac{\partial U}{\partial T} = -\frac{1}{8\pi T^2} < 0\,. \tag{24.68}$$

As it is typical for self-gravitating systems, its heat capacity is negative: Thus a black hole surrounded by a cooler medium emits radiation, heats up the environment and becomes hotter.

Black hole temperature If the interpretation of the BH area as its entropy is correct, then we should be able to treat a BH as a thermal system. In particular, a BH should emit thermal radiation with temperature T. As the entropy and the temperature of a BH are quantum properties, we should examine the evolution of a quantum field in the background of a gravitational field describing a BH solution. We consider the generating functional $Z[J]$ of a real scalar field in a Schwarzschild background

$$Z[J] = \int \mathcal{D}\phi\, \mathrm{e}^{\mathrm{i}(S[\phi, g^{\mu\nu}] + \langle J\phi\rangle)} \tag{24.69}$$

together with the generating functional $Z[J_{\mu\nu}]$ of the gravitational field coupled to an external stress tensor $J_{\mu\nu}$ as source,

$$Z[J^{\mu\nu}] = \int \mathcal{D}g_{\mu\nu}\, \mathrm{e}^{\mathrm{i}(S[g_{\mu\nu}] + \langle J^{\mu\nu} g_{\mu\nu}\rangle)}\,. \tag{24.70}$$

Here $S[g_{\mu\nu}]$ is the Einstein-Hilbert action of gravity coupled to source $J^{\mu\nu}$. The path integral over $S[g_{\mu\nu}]$ is even less well-behaved than those of other quantum fields. In particular, the gravitational action is not bounded from below and the classical limit for a BH contains singularities. Luckily, we can read the quantity of our interest, the BH temperature, from the path integral without the need to evaluate it.

Using Kruskal coordinates, we have eliminated the coordinate singularity at $r = 2M$ but we are left with the physical curvature singularity at $r = 0$. Recalling the connection between the Kruskal coordinates T and R and the Schwarzschild coordinates t and r given by (24.38) and setting $A = 0$ gives

$$T^2 - R^2 = \exp\left(\frac{r}{2M}\right)\left(\frac{r}{2M} - 1\right) \tag{24.71a}$$

$$\frac{T+R}{T-R} = -\exp\left(\frac{-t}{2M}\right) \tag{24.71b}$$

Thus the singularity at $r = 0$ is mapped onto the surface $R^2 - T^2 = 1$. With the change $T \to \tau = \mathrm{i}T$, the metric becomes (up to an overall sign) Euclidean

$$-\,\mathrm{d}s^2 = \frac{32M^3}{r}\exp\left(-\frac{r}{2M}\right)(\mathrm{d}\tau^2 + \mathrm{d}R^2) + r^2\mathrm{d}\Omega\,. \tag{24.72}$$

Now (24.71a) implies that r is real and larger than $2M$ for real τ and R. Thus we avoid the singularity at $r = 0$ by the change to an Euclidean metric.

If we set $R + \mathrm{i}\tau = \rho \mathrm{e}^{\mathrm{i}\phi}$, then (24.71b) becomes

$$\frac{R+\mathrm{i}\tau}{R-\mathrm{i}\tau} = \mathrm{e}^{2\mathrm{i}\phi} = \exp\left(\mathrm{i}\frac{t_E}{2M}\right), \tag{24.73}$$

where we also introduced Euclidean Schwarzschild time $\mathrm{i}t_E = t$. Our coordinates are single-valued functions only, if $2\phi = t_E/(2M)$ is a periodic function with period $\beta = 2\pi$. Thus the Euclidean time has the period $\beta = 8\pi M$ and consequently the Euclidean path integral has to be restricted to periodic fields, $\phi(t_E, \boldsymbol{x}) = \phi(t_E + \beta, \boldsymbol{x})$ and $g_{\mu\nu}(t_E, \boldsymbol{x}) = g_{\mu\nu}(t_E + \beta, \boldsymbol{x})$. However, this condition describes the partition function of a thermal system with temperature $T = 1/\beta = 1/(8\pi M)$. In this picture, the BH is in thermal equilibrium with its environment filled by the scalar field ϕ, emitting and absorbing the same amount of radiation.

Hawking radiation in 1+1 dimension An alternative, more direct approach to Hawking radiation is to calculate the rate of particle production measured by an observer at $r \to \infty$ using the method of Bogolyubov transformations. As usually, the calculation simplifies if we consider the conformally invariant case of a massless scalar field in 1+1 dimensions. Then we can express the solutions of the action

$$S = \int \mathrm{d}^2x\sqrt{|g|}\,\frac{1}{2}g^{\alpha\beta}\nabla_\alpha\phi\nabla_\beta\phi \tag{24.74}$$

through light-cone coordinates as

$$\phi = f(\tilde{u}) + g(\tilde{v}) = f(u) + g(v)\,. \tag{24.75}$$

Here, f and g are arbitrary smooth functions specifying the wave form, and $(\tilde{u}, \tilde{v})$ and (u, v) are the tortoise and Kruskal light-cone coordinates, respectively.

For an observer at spatial infinity, we can use the tortoise light-cone coordinates $\tilde{u}, \tilde{v}$ and the metric (24.31)

$$\mathrm{d}s^2 = \left(1 - \frac{2M}{r(\tilde{u}, \tilde{v})}\right)\mathrm{d}\tilde{u}\mathrm{d}\tilde{v}\,. \tag{24.76}$$

Unruh effect	Hawking effect
acceleration a	surface gravity $\kappa = 1/(4M)$
$u = -\exp(-a\tilde{u})/a$	$u = -\exp(-\kappa\tilde{u})/\kappa$
$v = \exp(a\tilde{v})/a$	$v = \exp(\kappa\tilde{v})/\kappa$
Minkowski vacuum $\lvert 0_{\mathrm{M}}\rangle$	Kruskal vacuum $\lvert 0_{\mathrm{K}}\rangle$
Rindler vacuum $\lvert 0_{\mathrm{R}}\rangle$	Boulware vacuum $\lvert 0_{\mathrm{B}}\rangle$

Table 24.1.: Comparison of variables used in the Unruh and Hawking effect.

With $r^* \to r$ for $r \to \infty$, the metric becomes at large distances Minkowskian,

$$\mathrm{d}s^2 \to \mathrm{d}\tilde{u}\mathrm{d}\tilde{v} = \mathrm{d}t^2 - \mathrm{d}r^2\,, \tag{24.77}$$

and thus coordinate time t and proper time τ coincide for such an observer. This observer will split the modes using the Schwarzschild coordinate time t,

$$\phi(x) = \sum_i \left[b_i \mathrm{e}^{-\mathrm{i}\Omega\tilde{u}} + b_i^\dagger \mathrm{e}^{\mathrm{i}\Omega\tilde{u}}\right] + \text{left-movers}\,. \tag{24.78}$$

The vacuum defined by $b_i \lvert 0_{\mathrm{B}}\rangle = 0$ is called the Boulware vacuum $\lvert 0_{\mathrm{B}}\rangle$. The tortoise coordinates cover however only the Schwarzschild space-time with $r > 2M$. They, and as a result also the Boulware vacuum $\lvert 0_{\mathrm{B}}\rangle$, are singular on the horizon: the regularised vacuum $\langle 0_{\mathrm{B}}\rvert\, T_{\mu\nu}(r = 2M) \lvert 0_{\mathrm{B}}\rangle$ energy diverges at $r = 2M$ using these coordinates

On the other hand, we know that the Kruskal coordinates u, v cover the whole Schwarzschild space-time; in particular they are not singular on the horizon. Close to the horizon,

$$\mathrm{d}s^2 \to \mathrm{d}u\mathrm{d}v = \mathrm{d}T^2 - \mathrm{d}R^2\,, \tag{24.79}$$

and thus the time T should be used to split the modes into positive and negative frequency modes,

$$\phi(x) = \sum_i \left[a_i \mathrm{e}^{-\mathrm{i}\Omega u} + a_i^\dagger \mathrm{e}^{\mathrm{i}\Omega u}\right] + \text{left-movers}\,. \tag{24.80}$$

Now the Kruskal vacuum is defined by $a_i \lvert 0_{\mathrm{a}}\rangle = 0$ for all i. The energy density of $\lvert 0_{\mathrm{a}}\rangle$ is finite (after subtracting as usually the zero-point energies) in the whole manifold, and increasing towards the singularity. For a sufficiently large BH, we can treat the metric as a classical static background, neglecting any quantum back-reaction.

We find now the spectrum of particle measured by an observer at $r \to \infty$ connecting the two vacua by a Bogolyubov transformation. Comparing the case at hand with the Unruh effect, we find the following analogy shown in Table 24.1. With the identification $a = \kappa = 1/(4M)$, we can translate the result for the Unruh effect to the spectrum emitted by a BH,

$$\left\langle \tilde{N} \right\rangle = \langle 0_K\rvert\, b^\dagger b \lvert 0_K\rangle = \frac{1}{\exp(2\pi\Omega/\kappa) - 1}\delta(0) \tag{24.81}$$

obtaining a thermal spectrum with temperature $T = \kappa/(2\pi) = 1/(8\pi M)$. The quantity κ is the surface gravity of the BH, i.e. it corresponds to the acceleration of a test particle experienced on the horizon, cf. problem 25.**??**.

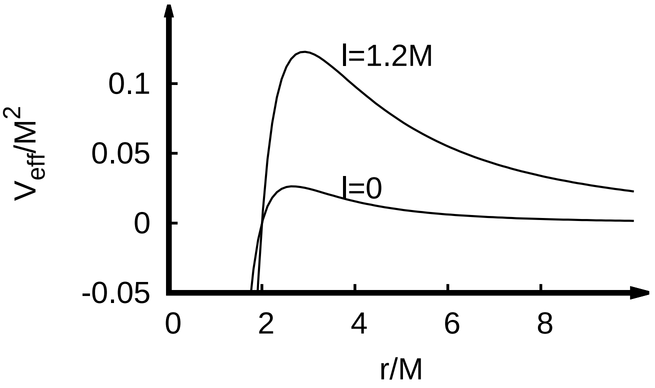


Figure 24.3.: Effective potential d V_{eff} describing the centrifugal barrier in $d = 4$ for two different values of the angular momentum l/M.

The state $|0_K\rangle$ contains not only outgoing right-movers, but also incoming left-movers with the same temperature. Since we are assuming a static system in thermal equilibrium, the amount of energy emitted and absorbed by the BH should be equal. This assumes that the BH is static and eternal, i.e. does not originate from the collapse of star. In the latter case, the past horizon at $v = 0$ does not exists and the space-time approaches the flat Minkowski metric. Therefore it is more appropriate to choose a_i as annihilation operators only for the out-going right-movers, and b_i for the left-movers. Hence an observer at spatial infinity sees only thermal radiation emitted by the BH.

Comment on 1+3 dimension How are our results modified in $d > 2$ dimensions? We know that the classical Newtonian potential $V(r) = M/r$ acquires an additional, effective term due to the centrifugal barrier. The same happens in the relativistic case: We can separate the Klein-Gordon equation $\Box\phi = 0$, if we insert the ansatz

$$\phi(r_*, t, \vartheta, \phi) = \sum_{l,m} f_l^m(r_*, t) Y_l^m(\vartheta, \phi) \,. \tag{24.82}$$

The resulting equation for $f_l^m(r_*, t)$ is

$$\left[\frac{\partial^2}{\partial t^2} - \frac{\partial^2}{\partial r^2} + \left(1 - \frac{2M}{r}\right)\left(\frac{2M}{r^3} + \frac{l(l+1)}{r^2}\right)\right] f_l^m(t, r) = 0 \,, \tag{24.83}$$

where the two terms in the round brackets define the effective potential $V_{eff} = \mathrm{d}V_l/\mathrm{d}r$. A plot of the effective potential V_{eff} describing the centrifugal barrier is shown in Fig. 24.3. Note that even for $l = 0$ a barrier exists and therefore a particle has to tunnel through it. Since the tunnelling probability is energy dependent, the spectrum emitted by a BH is modified by a grey-factor $\Omega_l(E)$. The condition for thermal radiation, namely an equal emission and absorption rate, is however satisfied also for $\Omega_l(E) < 1$.

Information paradox The discovery of Hawking that black holes evaporate led to the so called information paradox: The event horizon formed during the collapse of a massive star hides all information about its initial state in a set of conserved quantum numbers, M, J, and Q. If the black hole is stable, the information about the infalling material can be stored in its microphysical states. If the black hole evaporates however, then the question arises where the information goes—or if it is lost.

In order to see why a loss of information during black hole evaporation seems to contradict basic properties of quantum theory, let us recall first the familiar case of scattering in Minkowski space. The S-matrix maps initial states $|\psi_{\rm in}\rangle$ at $t = -\infty$ on final states states $|\psi_{out}\rangle = S\,|\psi_{\rm in}\rangle$ at $t = +\infty$. Since the S-matrix is unitary, we can recover the initial state measuring the final state, $|\psi_{\rm in}\rangle = S^\dagger\,|\psi_{out}\rangle$. In other words, the unitary time-evolution of quantum theory guaranties the preservation of information.

Next we consider the case that a BH is formed at finite time t. We neglect first Hawking radiation so that the BH is stable. At $t \to -\infty$, we are in Minkowski space and we can choose the initial state $|\psi_{\rm in}\rangle$ as a pure state from the elements of the usual Minkowski Hilbert space, $|\psi_{\rm in}\rangle \in \mathscr{H}_{\rm in}(M)$. Some of the scattered particles will end in the BH and hit finally the singularity; others will stay outside the horizon at $t \to \infty$. We denote the Hilbert space of states ending in the singularity as $\mathscr{H}_{out}(BH)$, while the Hilbert space of states escaping is $\mathscr{H}_{out}(M)$. Now the question arises what the complete Hilbert space $\mathscr{H}_{out}(M, BH)$ of final states is. Hawking argued that all operators defined on the Hilbert spaces $\mathscr{H}_{out}(M)$ and $\mathscr{H}_{out}(BH)$ commute, because the BH singularity is at space-like distances to the outside future. Therefore the complete Hilbert space of final states, $\mathscr{H}_{out}(M, BH)$, is the tensor product of the individual Hilbert spaces, $\mathscr{H}_{out}(M, BH) = \mathscr{H}_{out}(M) \otimes \mathscr{H}_{out}(BH)$. Consequently, the final states $|\psi_{\rm out}\rangle$ are product states, $|\psi_{\rm out}(M, BH)\rangle = |\psi_{\rm out}(M)\rangle \otimes |\psi_{\rm out}(BH)\rangle$. An observer outside the BH horizon can perform only measurements on $|\psi_{out}(M)\rangle$. Thus the outcome of its measurements is described by a density matrix,

$$\rho = {\rm t}r_{BH}\{|\psi_{\rm out}(M)\rangle\,\langle\psi_{\rm out}(M)|\}\,, \tag{24.84}$$

where the trace is over a complete set of states in $\mathscr{H}_{out}(BH)$. Now we add the effect of Hawking radiation. If the black hole evaporates completely, then a pure state has been transformed into a mixed state described by the density matrix (24.84). Since the time evolution in quantum theory is unitary, this cannot happen without assuming that quantum gravity violates the basic principles of quantum theory. Various proposals how this paradox can be avoided have been made, but no clear picture has emerged yet.

Summary

Black holes are solutions of Einstein's equations containing a physical singularity which—according to the cosmic censorship hypothesis—is always covered by an event horizon. This horizon is a global property of the space-time, being independent of the observer and influenced by the complete space-time. Within classical physics, the event horizon can only increase and has been therefore associated by Bekenstein with the entropy of a black holes. However, the event horizon is a infinite redshift surface and emits in the semi-classical picture thermal radiation. This Hawking radiation leads in turn to the information paradox.

Further reading

The Kerr solution is derived using symmetry arguments by citeLudvigsen199905. Additional material on classical BHs can be found e.g. in [HEL06]. [BD82] treat Hawking radiation in four dimensions. The entropy of a BH is calculated by [BL04]. For a description of the information paradox see e.g. [Mat09], the more recent discussion can be traced from [AMPS13] and its descents.

25. Cosmological constant

We have encountered three contributions to the cosmological constant: A non-zero Λ term in the Einstein-Hilbert Lagrangian would mean that an empty space-time has classically a non-zero energy density. Another classical contribution to the cosmological constant arises naturally in theories with SSB, because the minimum of the potential is non-zero either before or after the symmetry breaking. We expect a sequence of broken symmetries during the cosmological evolution, and it looks therefore mysterious why we end up today with $V \simeq 0$. Finally, the vacuum fluctuations of quantum fields contribute to the cosmological constant and it is often only this aspect which is called the cosmological constant problem. We start this chapter reconsidering these quantum fluctuations, studying the influence of the used regularisation scheme. Then we will introduce alternative explanations for the observed accelerated expansion of the universe in the present epoch which either modify gravity or add a new component of matter, dubbed dark energy. This approach assumes that the cosmological constant problem is solved, i.e. that $\rho_\Lambda + \langle\rho\rangle = 0$. One may argue that we need a theory of quantum gravity to understand how this works, and we close with some comments on this issue.

25.1. Vacuum energy density

Let us recall that we obtained for the regularised vacuum energy density $\langle\rho\rangle$ two expressions with a very different dependence on the regularisation parameter. Using in the Wilsonian approach an effective action including only modes up to the scale M, we found that $\rho \propto M^4$. In contrast, in DR power-like divergences are absent and thus ρ depends only logarithmically on the renormalisation scale μ. We will now reconsider these calculations, evaluating additionally the pressure P exerted by vacuum fluctuations. This will allow us to check which one of the two results reproduce correctly the equation of state $w = -1$ required for a Lorentz invariant vacuum.

As usually, we look at the simplest case, a real scalar field, where we only have to recollect our previous results from section 3.4 and example 5.1. The contribution of a free scalar field to the expectation value of the vacuum energy density $\langle\rho\rangle$ measured by an observer at rest is given by

$$\langle\rho\rangle = \langle 0|T^{00}|0\rangle = \frac{1}{2}\langle 0|\dot\phi^2 + (\boldsymbol{\nabla}\phi)^2 + m^2\phi^2|0\rangle = \int \frac{\mathrm{d}^3k}{(2\pi)^3}\,\frac{1}{2}\omega_k \tag{25.1}$$

with $\omega_k = \sqrt{m^2 + \boldsymbol{k}^2}$. The same observer measures an isotropic contribution of the scalar zero-point fluctuations to the vacuum pressure, $P^{ij} = P\delta^{ij}$, given by

$$\langle P\rangle = \frac{1}{3}\langle 0|T^{ii}|0\rangle = \frac{1}{6}\langle 0|(\dot\phi)^2 + (\boldsymbol{\nabla}\phi)^2 - m^2\phi^2|0\rangle = \frac{1}{3}\int \frac{\mathrm{d}^3k}{(2\pi)^3}\,\frac{\boldsymbol{k}^2}{2\omega_k}\,. \tag{25.2}$$

The scalar Feynman propagator $\Delta_F(0)$ at coincident points can be written as the limit of two fields at nearby points,

$$\mathrm{i}\Delta_F(0) = \langle 0|\phi(x')\phi(x)|0\rangle_{x'\searrow x} = \int \frac{\mathrm{d}^3k}{(2\pi)^3 2\omega_k}\mathrm{e}^{-\mathrm{i}k(x'-x)}\bigg|_{x'\searrow x} = \int \frac{\mathrm{d}^3k}{(2\pi)^3 2\omega_k}\,. \tag{25.3}$$

Thus we can connect the Feynman propagator $\mathrm{i}\Delta_F(0)$ with the stress tensor at the one-loop level forming the trace of the stress tensor of an ideal fluid and inserting Eqs. (25.1) and (25.2) for the energy density and pressure, respectively,

$$\langle T_\mu^\mu\rangle = \langle\rho\rangle - 3\langle P\rangle = \frac{m^2}{4\pi^2}\int_0^M \mathrm{d}k\,\frac{k^2}{\sqrt{m^2+\boldsymbol{k}^2}} = m^2\,\mathrm{i}\Delta_F(0)\,. \tag{25.4}$$

Using again the equation of state $w = \langle P\rangle/\langle\rho\rangle = -1$ valid for a contribution to the cosmological constant, we can rewrite this relation as

$$\langle\rho\rangle = -\langle P\rangle = \frac{m^2}{4}\,\mathrm{i}\Delta_F(0)\,. \tag{25.5}$$

Thus we should impose two physical conditions on the vacuum fluctuations: First, Lorentz invariance requires that the energy density and the pressure of the vacuum satisfy the EoS $w=-1$. Second, the trace $\langle T_\mu{}^\mu\rangle$ of the stress tensor and the Feynman propagator $\Delta_F(0)$ are connected by (25.5) at the one-loop level (or at $\mathcal{O}(\lambda^0)$). Since at $\mathcal{O}(\lambda^0)$ the mass m is cutoff independent, this implies that a M^4 term is absent at one-loop in the vacuum energy density.

As a side remark, we recall that classically $T_\mu{}^\mu \to 0$ for $m\to 0$. The trace anomaly will generate an additional contribution proportional to the beta function $\beta(\lambda) = \mu\partial\lambda/\partial\mu$ of the $\lambda\phi^4$ interaction,

$$\langle T_\mu^\mu\rangle = \frac{1}{2}m^2\phi^2 + \frac{\beta}{2\lambda}\phi^4\,. \tag{25.6}$$

The last term is only logarithmically sensitive to the UV cutoff M. This suggests that a term $\propto M^4$ is absent in general. Note also that while the trace anomaly gives a contribution to the vacuum energy, its EoS does not qualify it as dark energy.

Cutoff in spatial momenta We start investigating a sharp cutoff in the spatial momenta, as used in many discussions of the cosmological constant problem. Integrating the energy density (25.7a) up to the maximal scale $|\boldsymbol{k}|\le M$, we obtain

$$\langle\rho\rangle = \frac{1}{4\pi^2}\int_0^M \mathrm{d}|\boldsymbol{k}|\,\boldsymbol{k}^2\sqrt{m^2+\boldsymbol{k}^2} = \frac{1}{4\pi^2}M^4\int_0^1 \mathrm{d}z\,z^2\sqrt{x^2+z^2} \tag{25.7a}$$

$$= \frac{1}{16\pi^2}\left[M^4\sqrt{1+x^2}\left(1+\frac{1}{2}x^2\right) - \frac{1}{2}m^4 arcsinh\left(\frac{1}{x}\right)\right], \tag{25.7b}$$

with $z\equiv|\boldsymbol{k}|/M$ and $x\equiv m/M \ll 1$. Similarly, we find for the pressure

$$\langle P\rangle = = \frac{1}{3}\frac{1}{4\pi^2}\int_0^M \mathrm{d}|\boldsymbol{k}|\frac{\boldsymbol{k}^4}{\sqrt{m^2+\boldsymbol{k}^2}} = \frac{1}{3}\frac{1}{4\pi^2}M^4\int_0^1 \mathrm{d}z\frac{z^4}{\sqrt{x^2+z^2}} \tag{25.8a}$$

$$= \frac{1}{3}\frac{1}{16\pi^2}\left[M^4\sqrt{1+x^2}\left(1+\frac{1}{2}x^2\right) + \frac{3}{2}m^4 arcsinh\left(\frac{1}{x}\right)\right] \tag{25.8b}$$

and for the Feynman propagator at coincident points

$$\mathrm{i}\Delta_F(0) = \frac{1}{4\pi^2}\int_0^M \mathrm{d}|\boldsymbol{k}|\,\frac{\boldsymbol{k}^2}{\sqrt{m^2+\boldsymbol{k}^2}} = \frac{1}{4\pi^2}\,M^2\int_0^1 \mathrm{d}z\,\frac{z^2}{\sqrt{x^2+z^2}} \tag{25.9a}$$

$$= \frac{1}{8\pi^2}\left[M^2\sqrt{1+x^2} - m^2 \mathrm{a}rcsinh\left(\frac{1}{x}\right)\right]. \tag{25.9b}$$

All three results show the behaviour expected from naive power-counting, $\rho \propto P \propto M^4$ and $\Delta_F(0) \propto M^2$ in the limit $M \gg m$. We can check now, if these results fulfil the two constraints. We first neglect the mass of the scalar particle, i.e. consider the limit $x \to 0$: Then vacuum fluctuations are predicted to have the EoS of radiation, $w = \langle P\rangle/\langle\rho\rangle = 1/3 + \mathcal{O}(x^4)$ and to break scale invariance, $\langle T_\mu^\mu\rangle \neq 0$ for $m^2 \to 0$. Moreover, the relation (25.5) between the stress tensor and the Feynman propagator $\Delta_F(0)$ is violated.

The natural explanation for this behaviour is that regularisation schemes that break symmetries lead to spurious terms reflecting this violation. In our case, a momentum cutoff breaks Lorentz invariance and thus we cannot expect that a relation like $w = -1$ for the vacuum fluctuations is correctly reproduced. Using a regularisation scheme which breaks symmetries requires therefore to add additional, symmetry breaking counter-terms and to subtract the offending terms by hand. This suggests that the subleading m^4 terms do satisfy the two conditions, what is indeed the case.

Dimensional regularisation We replace the three-dimensional integrals in (25.1)–(25.3) by $d-1$ dimensional ones. Using then DR and the definition of Euler's Beta function, we obtain

$$\langle\rho\rangle = \frac{\mu^{4-d}}{(2\pi)^{(d-1)}}\frac{1}{2}\int \mathrm{d}^{d-1}k\,\sqrt{m^2+\boldsymbol{k}^2} = \frac{\mu^4}{2\,(4\pi)^{(d-1)/2}}\frac{\Gamma(-d/2)}{\Gamma(-1/2)}\left(\frac{m}{\mu}\right)^d. \tag{25.10}$$

Similarly, it follows

$$\langle P\rangle = \frac{\mu^{4-d}}{(2\pi)^{(d-1)}}\frac{1}{2(d-1)}\int \mathrm{d}^{d-1}k\,\frac{k^2}{\omega_k} = \frac{\mu^4}{4\,(4\pi)^{(d-1)/2}}\frac{\Gamma(-d/2)}{\Gamma(1/2)}\left(\frac{m}{\mu}\right)^d. \tag{25.11}$$

and

$$\mathrm{i}\Delta_F(0) = \mu^{4-d}\frac{(m^2)^{\frac{d}{2}-1}}{(4\pi)^{d/2}}\,\Gamma\left(1-\frac{d}{2}\right). \tag{25.12}$$

Expanding next in $\varepsilon = 4-d$ results in

$$\langle\rho\rangle = -\langle P\rangle = -\frac{m^4}{64\pi^2}\left[\frac{2}{\varepsilon}+\frac{3}{2}-\gamma+\ln\left(\frac{4\pi\mu^2}{m^2}\right)\right]+\mathcal{O}(\varepsilon). \tag{25.13}$$

It follows that DR reproduces the correct EoS as well as the relation (25.5). The latter ensures also that $T_\mu^\mu \to 0$ for $m \to 0$, as expected at $O(\lambda^0)$. Note also that the integrands in both Eqs. (25.10) and (25.11) are positive definite. Therefore, any cutoff scheme has to fail reproducing the correct EoS. In contrast, we know that the integration measure of DR is not positive definite and here we see an example where this property is required to reproduce the correct physics.

In order to obtain the observed value of ρ and P at a certain scale μ, we have to add a corresponding counter-term. After subtraction of the M^4 term and identifying $4\pi\mu$ with the cutoff scale M, we find in both schemes the same logarithmic dependence

$$\langle\rho\rangle = -\frac{m^4}{32\pi^2} arcsinh\ (M/m) \approx -\frac{m^4}{32\pi^2} \ln(M/m)\ , \tag{25.14}$$

in the limit $M = 4\pi\mu \gg m$. We thus conclude that the natural value of the vacuum energy density is $\langle\rho\rangle \sim m^4$. While this behaviour reduces the "numerical size" of the cosmological constant problem, it does not solve it, since the natural value of $\langle\rho\rangle$ in the SM would be determined by the top quark, $m_t^4 \gg \rho_\Lambda$.

25.2. Dark energy

Since the theoretically expected value of the cosmological constant is much larger than the observed one, one may hope that the cosmological constant is set to zero by some principle yet to be discovered. In this case, we require an alternative explanation for the accelerated expansion of the Universe in the present epoch. Such explanations can be divided in modifications of gravity and dark energy, depending on if the LHS or RHS of the Einstein equations are changed. Since we can reshuffle matter and gravity by field redefinitions, such a distinction is somewhat ambiguous. As a practical criterion, one can use the presence of additional long-range forces and thus the violation of the strong equivalence principle as the characteristic feature of modified gravity models.

Quintessence These models introduce a scalar field ϕ with a canonically normalised kinetic term that evolves in the slow-roll regime. Its potential $V(\phi)$ has to be chosen such that the slow-roll regime sets in only recently, i.e. at redshift $z \simeq 0.7$. In general, quintessence models predict time-dependent deviations from $w = -1$, which can be observationally tested. While these models cannot solve the cosmological constant problem on the fundamental level, they may address the coincidence problem. The observed value of ρ_Λ implies that the potential energy of the new scalar field is tiny: For instance, the potential $V(\phi) = \lambda\phi^4$ requires as coupling $\lambda \sim 10^{-122}$, while $V(\phi) = m^2\phi^2/2$ leads to $m \sim 10^{-33}$ eV. Thus quintessence models are extremely fine-tuned. They require therefore some stabilising mechanism, if they are embedded in a more complete theory. We neglect this problem and discuss here only how these models attempt to solve the coincidence problem.

We considered already in Eqs. (23.26a) and (23.27) the inflaton ϕ coupled to the Friedmann equations as a dynamical system. Re-interpreting ϕ as a quintessence field and adding normal matter with density ρ and EoS $w_{\rm m} = P/\rho$, we obtain as equations of motion

$$H^2 = \frac{8\pi G}{3}\left(\frac{1}{2}\dot{\phi}^2 + V + \rho\right)\ , \tag{25.15a}$$

$$\dot{H} = -4\pi G\left[\dot{\phi}^2 + (1 + w_{\rm m})\rho\right]\ . \tag{25.15b}$$

Additionally, we have the usual Klein-Gordon equation for ϕ. Next we introduce the dimensionless variables

$$x = \frac{\dot{\phi}}{\sqrt{48\pi G}H} \quad and \quad y = \frac{\sqrt{V}}{\sqrt{24\pi G}H}\ . \tag{25.16}$$

The energy fraction of the quintessence field, $\Omega_\phi = \rho_\phi/\rho_{\rm cr}$, becomes then $\Omega_\phi = x^2 + y^2$, while its EoS is given by

$$w_\phi = \frac{x^2 - y^2}{x^2 + y^2}\,. \tag{25.17}$$

We obtain dimensionless equations of motion taking derivatives of x and y with respect to $N = \ln(a)$. Setting also $\lambda \equiv -(8\pi G)^{1/2} V_{,\phi}/V$, they are given by

$$\frac{{\rm d}x}{{\rm d}N} = 3x + \frac{\sqrt{6}}{2}\lambda y^2 + \frac{3}{2}x[(1-w_{\rm m})x^2 + (1+w_{\rm m})(1-y^2)]\,, \tag{25.18a}$$

$$\frac{{\rm d}y}{{\rm d}N} = -\frac{\sqrt{6}}{2}\lambda xy + \frac{3}{2}y[(1-w_{\rm m})x^2 + (1+w_{\rm m})(1-y^2)]\,. \tag{25.18b}$$

For comparisons with observations, it is useful to define the effective equation of state $w_{eff} \equiv P_{tot}/\rho_{tot}$, since this is the EoS deduced from measuring H and $\dot{H}$. Using $\dot{H} = \ddot{a}/a - H^2$, we can express the Einstein equations (19.47) and (19.48) for a flat FLRW metric as

$$3H^2 = \kappa\rho \qquad and \qquad 3H^2 + 2\dot{H} = -\kappa P\,. \tag{25.19}$$

Thus the effective equation of state becomes

$$w_{eff} = \frac{P_{\rm tot}}{\rho_{\rm tot}} = -1 - \frac{2\dot{H}}{3H^2}\,. \tag{25.20}$$

Applied to quintessence, Eq. (25.15), we find

$$\frac{\dot{H}}{H^2} = -3x^2 - \frac{3}{2}(1+w_{\rm m})(1-x^2-y^2) \tag{25.21}$$

and thus

$$w_{eff} = w_{\rm m} + (1-w_{\rm m})x^2 - (1+w_{\rm m})y^2\,. \tag{25.22}$$

An interesting simplification appears, if λ is constant. Then the two equations (25.18) are closed and can be analysed as a two-dimensional dynamical system. The derivation of the fixed points of this dynamical system, which can be obtained setting ${\rm d}x/{\rm d}N = {\rm d}y/{\rm d}N = 0$, is the topic of problem 26.??. In particular, there exist trajectories connecting the saddle point $(x, y) = (0, 0)$ with $\Omega_\phi = 0$ and $w = w_m$ to the stable fixed point $(x, y) = (\lambda/\sqrt{6}, \sqrt{1-\lambda^2/6})$ with $\Omega_\phi = 1$ and $w_{eff} = w_\phi = -1 + \lambda^2/3$. Thus this case corresponds to a transition from a matter-dominated universe to an accelerated expansion with $w_{eff} = -1 + \lambda^2/3 > -1$. Integrating the definition of λ, we see that the case $\lambda = const.$ corresponds to the special case of exponential potentials, $V(\phi) = V_0 \exp(-\lambda\phi/\widetilde{M}_{\rm Pl})$. For other potentials, Eqs. (25.18) have to analysed combined with the Klein-Gordon equation for the quintessence field.

Tracker solutions Another subclass of quintessence models uses the potentials

$$V(\phi) = \frac{M^{4+n}}{n\phi^n} \tag{25.23}$$

with $n > 2$. Since $V(\phi)$ is unbounded from below, it violates our standard stability requirement and should be therefore modified by quantum correction at small ϕ. The Klein-Gordon equation in a FLRW background with $a(t) \propto t^\alpha$ becomes for this potential

$$\ddot{\phi} + \frac{3\alpha}{t}\dot{\phi} - \frac{M^{4+n}}{\phi^{n+1}} = 0\,. \tag{25.24}$$

A special solution called tracker solution is provided by

$$\phi_*(t) = CM^{1+\alpha}t^\alpha \tag{25.25}$$

with $\alpha = 2/(2+n)$, cf. problem 26.??. This solutions is an attractor, as we can check considering the driving force

$$\mathcal{F}(\phi) = -V'(\phi) = \frac{M^{4+n}}{\phi^{n+1}} \tag{25.26}$$

acting on ϕ: For $\phi(t) < \phi_*(t)$, the force is larger, $\mathcal{F}(\phi) > \mathcal{F}(\phi_*)$, and the solution ϕ makes up leeway. Similarly, a field with $\phi(t) > \phi_*(t)$ has a smaller driving force and falls back to the tracking solution. At late times, the evolution of the scalar field $\phi(t)$ is thus given by the tracking solution for a wide range of initial conditions.

We consider in the following always the tracking regime and set for simplicity $\phi \simeq \phi_*$. Using (25.25), it follows that the kinetic field energy scales as $\dot{\phi}^2 \propto t^{2\alpha-2}$ and that $\dot{\phi}^2 \simeq V(\phi)$ holds. Thus the tracker solution violates the slow-roll condition. The energy density ρ_ϕ decreases with $a \propto t^\beta$ as

$$\rho_\phi \propto \frac{1}{t^{2-2\alpha}} \propto \frac{1}{a^{(2-2\alpha)/\beta}} \,. \tag{25.27}$$

Using $\rho \propto a^{-3(1+w)}$, we obtain as EoS for the tracker field

$$w_\phi = -1 + \frac{2}{3}\frac{1-\alpha}{\beta} = \frac{n}{2+n}w_{\rm m} - \frac{2}{2+n} \,. \tag{25.28}$$

Thus for large n, it tracks the EoS of the dominant form of usual matter in the universe. In particular, w_ϕ is able to follow the change of $w_{\rm m}$ during the radiation-matter transition.

Since the field does not evolve in the slow-roll regime, it seems that it cannot lead to the desired accelerated expansion. However, the relative contribution Ω_ϕ of the field to the total energy density increases, $\Omega_\phi \propto t^{2\alpha}$, and at some point Eq. (25.24) assuming an evolution in a fixed background is no longer valid. This happens when $\rho_\phi \sim (\phi/t)^2$. The corresponding matter density follows with $H \sim 1/t$ as $\rho_m \sim 3H^2\widetilde{M}_{\rm Pl}^2 \sim (\widetilde{M}_{\rm Pl}/t)^2$. Thus the contributions of ordinary matter and the tracker field are equal for $\phi \sim M_{\rm Pl}$. This coincides with the generic slow-roll condition[1] for power-law potentials: Thus the tracker potential leads to an accelerated expansion, as soon as it dominates the energy density of the universe. Since Ω_ϕ in this class of models was never extremely small, the coincidence problem is alleviated relative to the ΛCDM model. However, the question why $\phi \sim M_{\rm Pl}$ happens at the present epoch has still to be addressed in a specific dark energy model.

Additionally to reproduce correctly the evolution of $a(t)$, dark energy models have to reproduce the observed large-scale structure. Deviations to the standard ΛCDM model can become important only at late times, and thus these models can be constrained mainly using the evolution of large-scale structures at redshifts $z \lesssim 5$.

25.3. Modified gravity

An important class of modified gravity models are the so-called $f(R)$ gravity models, which generalise the Einstein-Hilbert form replacing R by a general function $f(R)$. Thus the action

[1] The model belongs to the class of large-field models, with its associated problems.

of $f(R)$ gravity coupled to matter has the form

$$S = \int \mathrm{d}^4x\sqrt{|g|}\left\{\frac{1}{2\tilde{\kappa}}f(R) + \mathscr{L}_\mathrm{m}\right\}, \tag{25.29}$$

where $\mathscr{L}_\mathrm{m}$ may contain both non-relativistic matter and radiation. Note that for $f(R) \neq R$, the gravitational constant $\tilde{\kappa} = 8\pi\tilde{G}$ deviates from Newton's constant G measured in a Cavendish experiment. The field equations can be derived from the action (25.29) either by a variation w.r.t. the metric or the connection. The dynamics and the number of the resulting degrees of freedom differ in the two treatments. Following the first approach, we obtain generalising the derivation in section 18.2

$$F(R)R_{\mu\nu} - \frac{1}{2}f(R)g_{\mu\nu} - \nabla_\mu\nabla_\nu F(R) + g_{\mu\nu}\Box F(R) = -\kappa T_{\mu\nu} \tag{25.30}$$

with $F \equiv \mathrm{d}f/\mathrm{d}R$. Taking the trace of this expression, we find

$$F(R)R - 2f(R)g_{\mu\nu} + 3\Box F(R) = -\kappa T\,. \tag{25.31}$$

The term $\Box F(R)$ acts as a kinetic term, so that these models contain an additional propagating scalar degree of freedom, $\phi = F(R)$. Applied to a flat FLRW metric, one obtains from the 00 and ii part of the field equation (25.30) the modified Friedmann equations as

$$3FH^2 = \tilde{\kappa}\left(\rho_\mathrm{m} + \rho_{\mathrm{r}ad}\right) + \frac{1}{2}(FR - f) - 3H\dot{F}\,, \tag{25.32a}$$

$$-2F\dot{H} = \tilde{\kappa}\left[\rho_\mathrm{m} + \frac{4}{3}\rho_{\mathrm{r}ad}\right] + \ddot{F} - H\dot{F}\,. \tag{25.32b}$$

Dividing (25.32a) by $3FH^2$, we can express the RHS through dimensionless variables x_i which correspond at the present epoch to the density parameters $\Omega_i = \tilde{\kappa}\rho_i/(3F_0H_0^2)$,

$$x_1 = -\frac{\dot{F}}{HF}\,, \qquad x_2 = -\frac{f}{6FH^2}\,, \qquad x_3 = \frac{R}{6H^2} \quad and \quad x_4 = \frac{\tilde{\kappa}^2\rho_{\mathrm{r}ad}}{3FH^2}\,. \tag{25.33}$$

For a flat universe, the matter component has then to satisfy

$$\frac{\tilde{\kappa}\rho_\mathrm{m}}{3FH^2} = 1 - x_1 - x_2 - x_3 - x_4\,. \tag{25.34}$$

We are interested in the "late" universe, when the effect of radiation is negligible, and set therefore $x_4 = 0$. Following the example of quintessence, we obtain dimensionless equation of motions taking derivatives of x_i with respect to $N = \ln(a)$,

$$\frac{\mathrm{d}x_1}{\mathrm{d}N} = -1 - x_3 - 3x_2 + x_1^2 - x_1x_3\,, \tag{25.35a}$$

$$\frac{\mathrm{d}x_2}{\mathrm{d}N} = \frac{x_1x_3}{m} - x_2(2x_3 - x_1 - 4)\,, \tag{25.35b}$$

$$\frac{\mathrm{d}x_3}{\mathrm{d}N} = -\frac{x_1x_3}{m} - 2x_3(x_3 - 2)\,. \tag{25.35c}$$

Here, $m \equiv \mathrm{d}\ln F/\mathrm{d}\ln R$ characterises the deviation from a ΛCDM model which has $m = 0$, while the second parameter r is defined by

$$r \equiv -\frac{\mathrm{d}\ln f}{\mathrm{d}\ln R} = \frac{x_3}{x_2}\,. \tag{25.36}$$

In order to simplify the comparison with observations which assume typically standard Einstein gravity, we keep the definition (25.20) of the effective equation of state w_{eff}. Then we recall from Eq. (19.46) the curvature of a flat FLRW metric,

$$R = g^{\mu\nu} R_{\mu\nu} = 6\,\frac{a\ddot{a} + \dot{a}^2}{a^2} = 6(\dot{H} + 2H^2)\,. \tag{25.37}$$

Combined with the definition of x_3, we can express the effective EoS as a function of only x_3,

$$w_{eff} = -\frac{2}{3}\left(x_3 - \frac{1}{2}\right)\,. \tag{25.38}$$

The cosmological evolution can now be studied for a chosen $f(R)$ solving numerically the system (25.35). Alternatively, one can determine analytically the fixed points of this dynamical system, study their stability and the behaviour of w_{eff}: As result, one finds that models with an accelerated expansion as attractor solution either contradict observations (behaving as $a(t) \propto t^{1/2}$ before acceleration), violate stability constraints (having $w_{eff} < -1$) or are hardly distinguishable from quintessence models in standard gravity [AGPT07]. In addition to reproduce the observed large-scale structure, modified gravity models have to pass solar system tests: Since they contain typically a light scalar (as $\phi = F(R)$ in $f(R)$ gravity) which mediates a 5th force, they require a screening mechanism to evade constraints of local gravity tests.

25.4. Comments on quantising gravity

Gravity as an effective theory The Einstein-Hilbert action contains with the Planck scale $\widetilde{M}_{\mathrm{Pl}}$ a dimensionfull coupling, and gravity is therefore a non-renormalisable theory. As any other non renormalisable theory, we can treat Einstein gravity as an effective field theory, at least in the limit that horizons play no role. Including higher-order operators, we should be able to calculate quantum corrections to observables as e.g. the Newtonian potential following the scheme outlined in section 12.5. If we recall that $[R] = m^2$ and thus also $[R_{\mu\nu}] = [R_{\mu\nu\rho\sigma}] = m^2$, we can order possible higher-order terms according their dimension $d = 6, 8, \dots$ as

$$S_{eff} = \kappa \int \mathrm{d}^4x \sqrt{|g|}\, \big\{(R + 2\Lambda) + L^2 \left[A_1 R^2 + A_2 R_{\mu\nu} R^{\mu\nu} + A_3 R_{\mu\nu\rho\sigma} R^{\mu\nu\rho\sigma}\right] + \mathcal{O}(L^4)\big\}. \tag{25.39}$$

Here, the length scale $L \sim 1/M$ indicates when the higher-order operators become important. Suppressing Lorentz indices, we can write schematically the gravitational wave equation (with $A_i \sim 1$) as

$$\Box h + L^2 \Box\Box h \sim \kappa T\,. \tag{25.40}$$

Solving for the propagator, we obtain in momentum space

$$D(k^2) \sim \frac{1}{k^2 + L^2 k^4} = \frac{1}{k^2} - \frac{1}{k^2 + M^2}\,. \tag{25.41}$$

In the static limit, the first term corresponds to the usual Newtonian $1/r$ potential, while the second one is of the Yukawa type,

$$\Phi \sim -GM \left[\frac{1}{r} - \frac{\exp(-r/L)}{r}\right]\,. \tag{25.42}$$

Thus the higher-order operators lead in tree-level amplitudes to short-range interactions, which are exponentially suppressed on scales $r \gg L$. Newton's law is not tested on sub-mm scales and thus L is only bounded as $L \lesssim 1\,\mathrm{mm}$. If L would be close to this limit, as e.g. in scenarios with large extra dimensions, new gravitational effects may be detectable both in Cavendish-like experiments and at accelerators. However, one usually associates L with the Planck length $L_{\mathrm{P}l}$, and then the action becomes

$$S_{\mathrm{e}ff} = \int \mathrm{d}^4x\sqrt{|g|}\left\{\kappa(R-2\Lambda) - a\kappa^2R^2 + b\kappa^2 R_{\mu\nu}R^{\mu\nu} + \ldots\right\}. \tag{25.43}$$

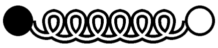
Performing perturbation theory, we expand the metric $g_{\mu\nu}$ around the classical Minkowski background. Applied to the inverse $g^{\mu\nu}$ and $\sqrt{|g|} = \sqrt{\det(-g_{\mu\nu})}$, this will generate an infinite tower of gravitational self-interactions and couplings to matter. Suppressing again Lorentz indices, we can write schematically

$$S_{\mathrm{e}ff} = \frac{\widetilde{M}^2_{\mathrm{P}l}}{4}\int \mathrm{d}^4x\left[\left(\frac{1}{2}h\Box h + h\Box h^2 + \ldots\right) + \left(-ah\Box^2 h + bh\Box^2 h^2 + \ldots\right) + \right]. \tag{25.44}$$

Since we are interested in quantum effects, it is convenient to rescale the gravitational field as $\mathrm{h}_{\mu\nu} = \frac{1}{2}\widetilde{M}_{\mathrm{P}l}h_{\mu\nu}$ such that it is canonically normalised and has mass dimension one,

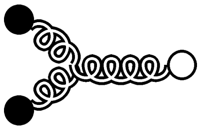
$$S_{\mathrm{e}ff} = \int \mathrm{d}^4x\left[\left(\frac{1}{2}\mathrm{h}\Box\mathrm{h} + \frac{1}{\widetilde{M}_{\mathrm{P}l}}\mathrm{h}\Box\mathrm{h}^2 + \ldots\right) + \left(\frac{a}{\widetilde{M}^2_{\mathrm{P}l}}\mathrm{h}\Box^2\mathrm{h} + \frac{b}{\widetilde{M}^3_{\mathrm{P}l}}\mathrm{h}\Box^2\mathrm{h}^2 + \ldots\right) + \right]. \tag{25.45}$$

The terms in the first round bracket correspond to an expansion of the Einstein-Hilbert action. Their tree-level contributions to the Newtonian potential of a static source with mass M have long-range. The leading term corresponding to single graviton exchange, cf. with Eq. (7.47), is given by



$$= \Phi_2 \sim \frac{M}{\widetilde{M}_{\mathrm{P}l}}\frac{1}{r}\frac{1}{\widetilde{M}_{\mathrm{P}l}} \sim \frac{M}{\widetilde{M}^2_{\mathrm{P}l}}\frac{1}{r}. \tag{25.46}$$

Here, the heavy source contributes the factor $\sqrt{G}M$, while the coupling to the test particle adds only $\sqrt{G}$. We can estimate the contribution of the next, trilinear term as follows



$$= \Phi_3 \sim \left(\frac{M}{\widetilde{M}_{\mathrm{P}l}}\right)^2\frac{1}{\widetilde{M}_{\mathrm{P}l}}\frac{1}{r^2}\frac{1}{\widetilde{M}_{\mathrm{P}l}} \sim \frac{1}{\widetilde{M}^2_{\mathrm{P}l}}\left(\frac{M}{\widetilde{M}_{\mathrm{P}l}}\right)^2\frac{1}{r^2}. \tag{25.47}$$

The two heavy sources add each a factor $M/\widetilde{M}_{\mathrm{P}l}$, while the trilinear coupling contributes the factor $1/\widetilde{M}_{\mathrm{P}l}$. The dependence $1/r^2$ follows then by dimensional analysis. At first sight, one might guess that the effect of the higher-order term Φ_3 is unobservable: However, classical sources of gravitational fields can be strong, $M/\widetilde{M}_{\mathrm{P}l} \gg 1$, and compact, $r \sim few \times R_s \sim M/\widetilde{M}^2_{\mathrm{P}l}$. If the latter condition is satisfied, then $\Phi_2 \sim \Phi_3 \sim \Phi_n \sim 1$ and all higher-order terms in the post-Newtonian expansion become important. Ideal objects to test this expansion are therefore close binary systems of neutron stars or black holes. Using the virial theorem, we can express Φ_3 as $\Phi_3 \sim GM/r \times (v/c)^2$, what shows that Φ_3 is classical, $\Phi_3 \propto \hbar^0$.

Next we consider how loop correction to the Einstein-Hilbert action affect observables like the Newtonian potential. As we know from our general discussion in section 11.4.1, the divergences have the structure of local operators. Since the one-loop corrections come with a factor κ, the one-loop divergences of the classical Einstein-Hilbert Lagrangian are connected to κ^2 terms in the effective Lagrangian. Thus the sole effect of these terms is to renormalise the coefficients a and b. Moreover, we know that the non-analytic terms induced by the loop corrections are finite and computable. In particular, the one-loop correction to the graviton propagator in the harmonic gauge is given by

$$\frac{k^4}{M_{\mathrm{Pl}}^2}\left[\frac{21}{120}(\eta_{\mu\rho}\eta_{\nu\sigma}+\eta_{\mu\sigma}\eta_{\nu\rho})+\frac{1}{120}\eta_{\mu\nu}\eta_{\rho\sigma}\right]\left[\frac{1}{\varepsilon}-\ln(k^2/\mu^2)+const.\right]. \tag{25.48}$$

As in the case of QCD, one has to add the effect of Faddeev-Popov ghosts to obtain a consistent results using a covariant gauge. Performing then e.g. on-shell renormalisation, we are left with the logarithm which turns in the static limit after Fourier transformation into a $1/r^3$ term. Combining these results, the Newtonian potential is

$$\Phi=-\frac{GM}{r}\left[1-\frac{3M}{c^2r^2}+\frac{41}{10\pi}\frac{\hbar G}{c^2r^3}-\frac{2}{3}\exp\left(-\frac{1}{a}\sqrt{\frac{\hbar G}{96\pi c^3}}r\right)\right], \tag{25.49}$$

where we set for simplicity $b=0$, and added c, $\hbar$ and the numerical prefactors [Ste78, BBDH03]. This example shows that loop corrections to observables in classical gravity can be calculated following the usual effective theory approach. These corrections are—as expected—tiny for $E \ll M_{\mathrm{Pl}}$. Additional loop corrections due to higher-order operators will be even more suppressed. Therefore any consistent theory of gravity will lead to experimentally indistinguishable predictions for the loop corrections in the limit $E \ll M_{\mathrm{Pl}}$. However, deviations from Einstein gravity are testable via gravitational wave emission of compact objects and observations on astronomical and cosmological scales.

Approaches beyond Einstein gravity In addition to being a non-renormalisable theory, gravity poses specific technical problems: For instance, canonical quantisation relies on the light-cone structure of space-time as input, while the metric should be the output of a quantum theory of gravity. Using the path integral avoid this problem and allows even the summation over space-times with different topologies. However, also the path integral formulation of quantum gravity is plagued by problems: The justification for a Wick rotation from a metric with Lorentzian to Euclidean signature is obscure—but even if we formally perform the Wick rotation, the Euclidean gravitational action remains unbounded from below.

In the traditional view, one neglects these principal problems and stays within the framework of QFT in $d=4$. The question arises then what UV completion the Einstein-Hilbert action has. One possibility is that the graviton is a composite particle, similar to the pion. However, the constraint that the agent of gravity couples in $d=4$ QFT to the conserved stress tensor seems to forbid this option. Another possibility is that the graviton is a fundamental particle, and additional particles and symmetries lead to an unitarisation of gravity. In particular, adding supersymmetries improves the UV behaviour of gravity and makes it renormalisable up to the two-loop level, since all possible one and two-loop divergences are forbidden by supersymmetric Ward identities. While no explicit three-loop calculations have been performed to date, power counting arguments suggest however that divergences exist at the three-loop level. An even more minimalistic ansatz is the idea of asymptotic safety which assumes that Einstein gravity, defined non-perturbatively e.g. on a lattice, is a consistent theory. This requires that the non-perturbatively calculated running gravitational coupling constant $G_N(q^2)$ has a UV fixed point. In this case, a consistent continuum limit of gravity could be defined. However, even the concept of a universal running coupling constant is, at least in perturbative calculations, ill-defined [Don12]. While it is not excluded that this is an artifact of perturbation theory, little evidence is known that supports asymptotic safety.

We give next a simple argument suggesting that quantum gravity does not admit local observables. As starter, let us consider the case of gauge theories. Here, measurable quantities are associated to n-point functions $\langle O(x_1)\cdots O(x_n)\rangle$ of gauge-invariant operators $O(x)$, or to S-matrix elements. While the one-particle states used in the initial and final states of the S-matrix are not gauge-invariant, the possible gauge transformations are reduced in the limit $t \to \pm\infty$ to global transformations which map one physical state onto a different physical state: Tying the states to $t = \pm\infty$ eliminates thus the redundancy of local gauge transformations.

Now we consider the case of gravity where the gauge group is the group of all invertible coordinate transformations. This means that no local operators $O(x)$ exist which are gauge-invariant, leaving the S-matrix as observable. More explicitly, we can show this as follows: Let us consider the expectation value of observables which are scalar functions of the metric and a scalar field, $O(x) \equiv O(g_{\mu\nu}(x), \phi(x))$. We write $\langle O(x_1)\cdots O(x_n)\rangle$ as a path integral average, and use that a scalar transforms as $O(x) = \tilde{O}(\tilde{x})$ under a general coordinate transformation,

$$\langle O(x_1)\cdots O(x_n)\rangle = \int \mathcal{D}\phi\mathcal{D}g_{\mu\nu}\, O(x_1)\cdots O(x_n)\, \mathrm{e}^{\mathrm{i}S[\phi,g_{\mu\nu}]} \tag{25.50a}$$

$$= \int \mathcal{D}\phi\mathcal{D}g_{\mu\nu}\, \tilde{O}(\tilde{x}_1)\cdots \tilde{O}(\tilde{x}_n)\, \mathrm{e}^{\mathrm{i}S[\phi,g_{\mu\nu}]}\,. \tag{25.50b}$$

Next we apply the invariance of the action, $S[\tilde{\phi}, \tilde{g}_{\mu\nu}] = S[\phi, g_{\mu\nu}]$, and the measure $\mathcal{D}\tilde{\phi}\mathcal{D}\tilde{g}_{\mu\nu} = \mathcal{D}\phi\mathcal{D}g_{\mu\nu}$, and relabel then the integration variables as ϕ and $g_{\mu\nu}$,

$$\langle O(x_1)\cdots O(x_n)\rangle = \int \mathcal{D}\tilde{\phi}\mathcal{D}\tilde{g}_{\mu\nu}\, \tilde{O}(\tilde{x}_1)\cdots \tilde{O}(\tilde{x}_n)\, \mathrm{e}^{\mathrm{i}S[\tilde{\phi},\tilde{g}_{\mu\nu}]} \tag{25.51a}$$

$$= \int \mathcal{D}\phi\mathcal{D}g_{\mu\nu}\, O(\tilde{x}_1)\cdots O(\tilde{x}_n)\, \mathrm{e}^{\mathrm{i}S[\phi,g_{\mu\nu}]} = \langle O(\tilde{x}_1)\cdots O(\tilde{x}_n)\rangle\,. \tag{25.51b}$$

Thus $\langle O(x_1)\cdots O(x_n)\rangle$ cannot depend on the space-time points x_i and has to be constant.

We can compare this behaviour to the case of space-time symmetries in Minkowski space: Applying as a symmetry transformation e.g. translations results in the constraint that an observable can depend only on the differences $\tilde{x}_i - x_i$, i.e. it should be translation invariant. Increasing the symmetry group to general coordinate transformations, the restriction on observables becomes so severe that no non-trivial solution is possible. This argument suggests that a quantum theory of gravity is not simply a version of Einstein-Hilbert gravity with improved UV properties, but should include some fundamentally new features compared to the local quantum field theories of point particles we have considered. Examples for such non-local theories may contain a minimal length scale like loop gravity and string theory, or may be based on the non-commutativity of space-time.

Finally, we comment on the idea of gravity as an emergent phenomenon. We have usually assumed that the Universe at higher energies becomes more and more symmetric. However, several condensed matter systems show the opposite behaviour, where new symmetries emerge at low energies. Applying this picture to gravity, the gravitational field and thus space-time would be no fundamental degrees of freedom, but would appear at $E \ll M_{\mathrm{Pl}}$ as an approximate symmetry of the low-energy world. Support for this idea comes from few directions:

First, the cosmological constant problem is solved, if the coupling of the gravitational field to the stress tensor is shift-invariant, $T_{\mu\nu} \to T_{\mu\nu} + cg_{\mu\nu}$. However, this simple solution excludes that the gravitational action is a functional of the metric $g_{\mu\nu}$. Second, one can derive the Einstein equations using a thermodynamical language, an attempt which is suggested by the thermodynamical description of black holes and horizons in general. More importantly, it is possible to re-interpret the gravitational field equations as equations describing the heating and cooling of null surfaces within this thermodynamical picture. This thermodynamic description poses the question, if the gravitational field is not merely a macroscopic description for the unknown fundamental degrees of freedom. A concrete example for a theory where space-time is emergent is string theory, since there the metric is not a fundamental degree of freedom. Last but not least, black hole formation provides a fundamental limitation to the measurement of space-time: Probing space-time distances below the Planck scale is impossible, because otherwise the measuring device would collapse to a black hole (problem 26.**??**). This implies that the uncertainty relations derived from the usual commutation relations of quantum fields break down on scales $\lesssim 1/M_{\mathrm{Pl}}$, or in other words that local quantum field theory should be replaced by a new theoretical setting. Matvei Bronstein [Bro36] used first this argument to conclude:

> “The elimination of the logical inconsistencies connected with this requires a radical reconstruction of the theory, and . . . perhaps also the rejection of our ordinary concepts of space and time, modifying them by some much deeper and nonevident concepts. Wer’s nicht glaubt, bezahlt einen Thaler.”

Summary

Alternatives to the standard ΛCDM model involve either new scalar fields (dark energy) or modify gravity. While dark energy models are generically plagued by fine-tuning problems, modifications of gravity contain new additional degrees of freedom which pose both theoretical and observational challenges. The cosmological constant problem raises the question, if our (effective) field theory approach has a restricted validity. Few hints suggest that locality has to be abandoned in a quantum theory of gravity and that space-time may be an emergent phenomenon.

Further reading

[Mar12] discusses exhaustively the cosmological constant problem. Dark energy and modified gravity models and their observational consequences are reviewed by [JLS16]. The effective field theory approach to gravity is discussed by [Don12], while [Pad16] reviews links between gravity and thermodynamics.

Bibliography

[A$^+$12] S. Ask et al. From Lagrangians to Events: Computer Tutorial at the MC4BSM-2012 Workshop. 2012.

[A$^+$16] B.P. Abbott et al. Observation of Gravitational Waves from a Binary Black Hole Merger. *Phys. Rev. Lett.*, 116:061102, 2016.

[ABCRM10] R. Allahverdi, R. Brandenberger, F.-Y. Cyr-Racine, and A. Mazumdar. Reheating in Inflationary Cosmology: Theory and Applications. *Ann. Rev. Nucl. Part. Sci.*, 60:27–51, 2010.

[ABDM11] A. Anisimov, W. Buchmüller, M. Drewes, and S. Mendizabal. Quantum Leptogenesis I. *Ann. Phys.*, 326:1998–2038, 2011. [Erratum: Ann. Phys. 338, 376 (2011)].

[AdBF91] U. Amaldi, W. de Boer, and H. Fürstenau. Comparison of Grand Unified Theories with Electroweak and Strong Coupling Constants Measured at LEP. *Phys. Lett.*, B260:447–455, 1991.

[AGPT07] L. Amendola, R. Gannouji, D. Polarski, and S. Tsujikawa. Conditions for the Cosmological Viability of $f(R)$ Dark Energy Models. *Phys. Rev.*, D75:083504, 2007.

[AMPS13] A. Almheiri, D. Marolf, J. Polchinski, and J. Sully. Black Holes: Complementarity or Firewalls? *JHEP*, 02:062, 2013.

[BBDH03] N.E.J Bjerrum-Bohr, J.F. Donoghue, and B.R. Holstein. Quantum Gravitational Corrections to the Nonrelativistic Scattering Potential of two Masses. *Phys. Rev.*, D67:084033, 2003. [Erratum: Phys. Rev. D71, 069903 (2005)].

[BD82] N.D. Birrell and P.C.W. Davies. *Quantum Fields in Curved Space.* Cambridge University Press, Cambridge, 1982.

[BDBP05] W. Buchmüller, P. Di Bari, and M. Plümacher. Leptogenesis for Pedestrians. *Annals Phys.*, 315:305–351, 2005.

[BL93] D. Bailin and A. Love. *Introduction to Gauge Field Theory (Graduate Student Series in Physics).* Institute of Physics Publishing, Bristol, revised edition, 1993.

[BL04] D. Bailin and A. Love. *Cosmology in Gauge Field Theory and String Theory (Graduate Student Series in Physics).* Institute of Physics Publishing, Bristol, 2004.

[Bla11] J.-P. Blaizot. Quantum Fields at Finite Temperature from Tera to nano Kelvin. 2011.

[Bro36] M. Bronstein. Kvantovanie Gravitatsionnykh Voln. *Zh. Eksp. Tear. Fiz.*, 6:195–236, 1936.

[Car03] S. Carroll. *Spacetime and Geometry: An Introduction to General Relativity.* Addison-Wesley, 2003. A draft version is available as arXiv:gr-qc/9712019.

[CC77] Jr. Callan, C.G. and S.R. Coleman. The Fate of the False Vacuum. 2. First Quantum Corrections. *Phys.Rev.*, D16:1762–1768, 1977.

[Cha06] A. Challinor. Constraining fundamental physics with the cosmic microwave background. In *International Scientific Workshop on Cosmology and Gravitational Physics Thessaloniki, Greece, December 15-16, 2005*, 2006.

[CL88] T.-P. Cheng and L.-F. Li. *Gauge Theory of Elementary Particle Physics.* Oxford University Press, New York, 1 edition, 1 1988.

[CL02] P. Coles and F. Lucchin. *Cosmology.* Wiley, Chichester, 2 edition, 2002.

[CM16] M. Coleman Miller. Implications of the Gravitational Wave Event GW150914. *Gen. Rel. Grav.*, 48:95, 2016.

[Das06] A. Das. *Field Theory: A Path Integral Approach (World Scientific Lecture Notes in Physics).* World Scientific, Singapore, 2 edition, 2006.

[DGH14] J.F. Donoghue, E. Golowich, and Barry R. Holstein. *Dynamics of the Standard Model.* Cambridge University Press, Cambridge, 2 edition, 2014.

[DKS09] G. Dissertori, I.G. Knowles, and M. Schmelling. *Quantum Chromodynamics: High Energy Experiments and Theory.* Oxford University Press, Oxford, 2009.

[Dod03] S. Dodelson. *Modern Cosmology.* Academic Press, San Diego, 2003.

[Don12] J.F. Donoghue. The Effective Field Theory Treatment of Quantum Gravity. *AIP Conf. Proc.*, 1483:73–94, 2012.

[Esp14] J.R. Espinosa. Vacuum Stability and the Higgs Boson. *PoS*, LATTICE2013:010, 2014.

[ESW03] R.K. Ellis, W.J. Stirling, and B.R. Webber. *QCD and Collider Physics.* Cambridge University Press, Cambridge, 2003.

[FHS12] M. Flory, R.C. Helling, and C. Sluka. How I Learned to Stop Worrying and Love QFT. 2012.

[GK90] K. Griest and M. Kamionkowski. Unitarity Limits on the Mass and Radius of Dark Matter Particles. *Phys. Rev. Lett.*, 64:615, 1990.

[GR08] W. Greiner and J. Reinhardt. *Field Quantization.* Springer, Berlin, 2008.

[GR11a] D.S. Gorbunov and V.A. Rubakov. *Introduction to the Theory of the Early Universe: Cosmological Perturbations and Inflationary Theory.* World Scientific, Singapore, 2011.

[GR11b] D.S. Gorbunov and V.A. Rubakov. *Introduction to the Theory of the Early Universe: Hot Big Bang Theory.* World Scientific, Singapore, 2011.

[Gre00] W. Greiner. *Relativistic Quantum Mechanics. Wave Equations.* Springer, Berlin, 3rd edition, 2000.

[Hab94] H.E. Haber. Spin Formalism and Applications to New Physics Searches. 1994.

[HEL06] M.P. Hobson, G.P. Efstathiou, and A.N. Lasenby. *General Relativity: An Introduction for Physicists.* Cambridge University Press, Cambridge, 2006.

[Hil51] E.L. Hill. Hamilton's Principle and the Conservation Theorems of Mathematical Physics. *Rev. Mod. Phys.*, 23:253–260, 1951.

[HNN15] J. Hisano, R. Nagai, and N. Nagata. Effective Theories for Dark Matter Nucleon Scattering. *JHEP*, 05:037, 2015.

[Hui01] L. Hui. Unitarity Bounds and the Cuspy Halo Problem. *Phys. Rev. Lett.*, 86:3467–3470, 2001.

[Jeg07] F. Jegerlehner. Essentials of the Muon g-2. *Acta Phys. Polon.*, B38:3021, 2007.

[JLS16] A. Joyce, L. Lombriser, and F. Schmidt. Dark Energy vs. Modified Gravity. *Ann. Rev. Nucl. Part. Sci.*, 66:95–122, 2016.

[JO01] J.D. Jackson and L.B. Okun. Historical Roots of Gauge Invariance. *Rev. Mod. Phys.*, 73:663–680, 2001.

[Kaz00] D.I. Kazakov. Beyond the Standard Model: In Search of Supersymmetry. In *2000 European School of high-energy physics, Caramulo, Portugal, 20 Aug-2 Sep 2000: Proceedings*, pages 125–199, 2000.

[KG11] J.I. Kapusta and Ch. Gale. *Finite-Temperature Field Theory: Principles and Applications.* Cambridge University Press, Cambridge, 2 edition, 2011.

[KN13] M. Kawasaki and K. Nakayama. Axions: Theory and Cosmological Role. *Ann. Rev. Nucl. Part. Sci.*, 63:69–95, 2013.

[KT94] E.W. Kolb and M.S. Turner. *The Early Universe (Frontiers in Physics 69).* Westview Press, 1994.

[LB92] M. Le Bellac. *Quantum and Statistical Field Theory (Oxford Science Publications).* Clarendon Press, Oxford, 1992.

[Lis16] M. Lisanti. Lectures on Dark Matter Physics. In *Theoretical Advanced Study Institute in Elementary Particle Physics: New Frontiers in Fields and Strings (TASI 2015) Boulder, CO, USA, June 1-26, 2015*, 2016.

[LL76] L.D. Landau and E.M. Lifshitz. *Mechanics (Course of Theoretical Physics, Volume 1).* Butterworth-Heinemann, Oxford, 3 edition, 1976.

[LL80] L.D. Landau and E.M. Lifshitz. *The Classical Theory of Fields (Course of Theoretical Physics Series, Volume 2).* Butterworth-Heinemann, Oxford, 4 edition, 1980.

[LL81] L.D. Landau and L.M. Lifshitz. *Quantum Mechanics (Course of Theoretical Physics Series, Volume 3).* Butterworth-Heinemann, Oxford, 3 edition, 1981.

[LL09] D.H. Lyth and R. Liddle. *The Primordial Density Perturbation: Cosmology, Inflation and the Origin of Structure.* Cambridge University Press, Cambridge, 2009.

[LP13] E. Leader and E. Predazzi. A Note on the Implications of Gauge Invariance in QCD. *J. Phys.*, G40:075001, 2013.

[LV16] M. Laine and A. Vuorinen. *Basics of Thermal Field Theory (Lecture Notes in Physics 925).* Springer, Heidelberg, 2016. arXiv:1701.01554.

[Mac00] R. MacKenzie. Path Integral Methods and Applications – Lectures given at Rencontres du Vietnam: VIth Vietnam School of Physics. 2000.

[Mag07] M. Maggiore. *Gravitational Waves: Volume 1: Theory and Experiments.* Oxford University Press, USA, 2007.

[Mar12] J. Martin. Everything You always Wanted to Know about the Cosmological Constant Problem (But Were Afraid to Ask). *Comptes Rendus Physique*, 13:566–665, 2012.

[Mat09] S.D. Mathur. What Exactly is the Information Paradox? *Lect. Notes Phys.*, 769:3–48, 2009.

[Muk05] V. Mukhanov. *Physical Foundations of Cosmology.* Cambridge University Press, Cambridge, 2005.

[MW07] Viatcheslav Mukhanov and Sergei Winitzki. *Introduction to Quantum Effects in Gravity.* Cambridge University Press, Cambridge, 1 edition, 6 2007.

[Nac90] O. Nachtmann. *Elementary Particle Physics: Concepts and Phenomena.* Springer, 1990.

[O$^+$14] K.A. Olive et al. Review of Particle Physics. *Chin. Phys.*, C38:090001, 2014. For up-dates see the on-line version at `http://pdg.lbl.gov`.

[Pad16] T. Padmanabhan. Exploring the Nature of Gravity. In *The Planck Scale II Wroclaw, Poland, September 7-12, 2015*, 2016.

[Pes11] M.E. Peskin. Simplifying Multi-Jet QCD Computation. 2011.

[Poi07] E. Poisson. *A Relativist's Toolkit: The Mathematics of Black-Hole Mechanics.* Cambridge University Press, Cambridge, 2007.

[Pok87] S. Pokorski. *Gauge Field Theories.* Cambridge University Press, Cambridge, 1987.

[PT09] L. Parker and D. Toms. *Quantum Field Theory in Curved Spacetime: Quantized Fields and Gravity.* Cambridge University Press, Cambridge, 2009.

[Rac16] J. Racker. Mini-Review on Baryogenesis at the TeV Scale and Possible Connections with Dark Matter. In *Proceedings, 37th International Conference on High Energy Physics (ICHEP 2014): Valencia, Spain, July 2-9, 2014*, 2016.

[Ram94] P. Ramond. *Field Theory: A Modern Primer (Frontiers in Physics).* Perseus Books (Sd), 2 revised edition, 1994.

[Rav05] F. Ravndal. Lecture Notes for Cosmological Physics. 2005.

[RS12] J.C. Romao and J.P. Silva. A Resource for Signs and Feynman Diagrams of the Standard Model. *Int. J. Mod. Phys.*, A27:1230025, 2012.

[Rub02] V.A. Rubakov. *Classical Theory of Gauge Fields.* Princeton University Press, 2002.

[Sch05] S.S. Schweber. The Sources of Schwinger's Green's Functions. *Proc. Nat. Acad. Sci.*, 102:7783–7788, 2005.

[Sch13] M.D. Schwartz. *Quantum Field Theory and the Standard Model.* Cambridge University Press, Cambridge, 2013.

[Sre07] M. Srednicki. *Quantum Field Theory.* Cambridge University Press, Cambridge, 2007. A draft version is available at `http://web.physics.ucsb.edu/~mark/qft.html`.

[Ste78] K. S. Stelle. Classical Gravity with Higher Derivatives. *Gen. Rel. Grav.*, 9:353–371, 1978.

[Ste93] G. Sterman. *An Introduction to Quantum Field Theory.* Cambridge University Press, Cambridge, 1993.

[Wei65] S. Weinberg. Infrared photons and gravitons. *Phys.Rev.*, B140:516–524, 1965.

[Wei05] S. Weinberg. *The Quantum Theory of Fields, Volume 1: Foundations.* Cambridge University Press, Cambridge, 2005.

[Wei16] S. Weinzierl. Tales of 1001 Gluons. 2016.

[Whi15] C.D. White. An Introduction to Webs. 2015.

[Zee13] A. Zee. *Einstein Gravity in a Nutshell.* Princeton University Press, Princeton, 2013.