Chapter 1

$\pi_0(\mathcal{Q})$ and solitons

A *soliton* is a classical solution of nonlinear field equations which (1) is nonsingular, (2) has finite energy and (3) is localized in space. We will only consider static solitons. In this case the field equations can be obtained by varying a functional that we will call the static energy. In some cases, the solitons are local minima of the static energy, and are separated from the absolute minimum (the vacuum) by a finite energy barrier. Such solitons are called "nontopological solitons". We will only be interested in another class of solitons, which either cannot be deformed continuously into the vacuum, or if they can, are separated from the vacuum by an infinite energy barrier. Such solitons are called "topological solitons".

In order to make this concept mathematically more precise, it is convenient to think of a field theory as a mechanical system with an infinite dimensional configuration space. Let us define the classical configuration space of the theory, $\mathcal{Q}$, to be the space of smooth, finite energy configurations of the field at some instant of time. Note that $\mathcal{Q}$ defines the kinematics of the theory, but also knows about the form of the energy. The theories that we will consider in this chapter will have the common characteristic that their configuration space is not connected. Instead, it will be the disjoint union of several connected components, indexed by a set $\pi_0(\mathcal{Q})$ (the reason for this notation is explained in Appendix A):

$$\mathcal{Q} = \bigcup_{i\in\pi_0(\mathcal{Q})} \mathcal{Q}_i \ ,$$

where $\mathcal{Q}_i$ are connected. Having determined the structure of the configuration space, the natural problem will be to find (if it exists) the absolute minimum of the static energy in each connected component. Such minima will automatically be solutions of the classical equations of motion. The minimum of the energy in some connected components will be the classical vacuum configuration, but in others it will correspond to non trivial solutions; these will be our topological solitons.

The nonconnectedness of the configuration space $\mathcal{Q}$ will manifest itself analytically in the existence of a conserved current known as the topological current. This current is not related to any symmetry of the theory and is identically conserved, i.e. it is conserved without making use of the equations of motion. (By contrast, Noether currents are conserved only upon using the equations of motion). Associated to the topological current is the topological charge, which is a functional on $\mathcal{Q}$ that is locally constant. It is zero in the connected components containing the vacuum, and nonzero in those containing solitons.

The above definition of soliton is tailored to describe a classical extended particle. When the theory is quantized, the solitons behave like a new species of particles, in addition to the perturbative particle states of the field. This can be seen in various ways. In these lectures we will often find it convenient to think of a quantum field theory as the quantum mechanics of a system with configuration space $\mathcal{Q}$. This is a formal definition that would require more technical work to be made precise, but is useful for heuristic considerations. In the Schrödinger picture, the wave functions are complex functionals on $\mathcal{Q}$. If $\mathcal{Q}$ has several connected components, the Hilbert space $\mathcal{H}$ will split into subspaces called the topological sectors:

$$\mathcal{H} = \bigoplus_{i\in\pi_0(\mathcal{Q})} \mathcal{H}_i \ ,$$

where $\mathcal{H}_i$ consists of wave functionals which are nonzero only on $\mathcal{Q}_i$. Each subspace $\mathcal{H}_i$ will be an eigenspace of the topological charge with eigenvalue i. It is clear that with any sensible definition of the measure the spaces $\mathcal{H}_i$ will be orthogonal to each other. The topological charge therefore defines a superselection rule: if the state vector belongs initially to the subspace $\mathcal{H}_i$, it will never leave it in the course of the time evolution. This fact can also be easily understood from the point of view of Feynman's path integral, because there are no paths joining $\mathcal{Q}_i$ to $\mathcal{Q}_j$ when $i \neq j$, so the transition amplitude between states in different sectors must vanish.

1.1 Scalar solitons in $1+1$ dimensions

1.1.1 Classical kinks

We begin by discussing the simplest case, that of a single scalar field in one space dimension, with action:

$$S(\phi) = \int d^2x \left[-\frac{1}{2}\partial_\mu\phi\partial^\mu\phi - V(\phi)\right] \tag{1.1}$$

with $\partial_\mu\phi\partial^\mu\phi = -(\partial_0\phi)^2 + (\partial_1\phi)^2$. We demand that the potential V be bounded from below, and we assume without loss of generality that the minimum value of V be zero. We call y_i, $i \in \mathcal{J}$, the minimum points. For definiteness one can think of the quartic potential

$$V = -\frac{1}{2}m^2\phi^2 + \frac{\lambda}{4}\phi^4 + \frac{m^4}{4\lambda} = \frac{\lambda}{4}\left(\phi^2 - f^2\right)^2, \tag{1.2}$$

with $f = \frac{m}{\sqrt{\lambda}}$ and m real and positive, with minima at points $y_\pm = \pm f$.

With these assumptions, the energy:

$$E = \int\limits_{-\infty}^{\infty} dx \left[\frac{1}{2}(\partial_0\phi)^2 + \frac{1}{2}(\partial_1\phi)^2 + V(\phi)\right] \tag{1.3}$$

is positive semidefinite, and is zero only for the constant field configurations $\phi(x,t) = y_i$. These are the absolute minima of E; they are the classical vacua of the theory. Note that in (1.3) the first term represents the kinetic energy; the rest

$$E_S = \int\limits_{-\infty}^{+\infty} dx \left[\frac{1}{2}(\partial_1\phi)^2 + V\right] \tag{1.4}$$

will be called "static energy". We will reserve the name "potential energy" for the second term in E_S, while the first term could be appropriately called "elastic energy".

The field ϕ belongs to the space $\Gamma(\mathbb{R},\mathbb{R})$ of smooth real functions of one variable. (In general we will use the notation $\Gamma(X,Y)$ for the space of smooth maps from X to Y, where X and Y are manifolds. This space is itself an infinite dimensional smooth manifold. See Appendix E) Finiteness of the energy demands that when $|x|$ tends to infinity ϕ tends to one of the classical vacua, for otherwise the last two terms in E would diverge. We will call $\mathcal{Q}$ the subspace of $\Gamma(\mathbb{R},\mathbb{R})$ for which the static energy E_S is finite.

If V has more that one minimum, $\mathcal{Q}$ will not be connected. In fact, let

$$\mathcal{Q} = \bigcup_{i,j} \mathcal{Q}_{ij} \ , \qquad \mathcal{Q}_{ij} = \{\phi \in \mathcal{Q} \mid \phi \underset{x\to-\infty}{\longrightarrow} y_i, \phi \underset{x\to+\infty}{\longrightarrow} y_j\} \ .$$

Every path in $\Gamma(\mathbb{R},\mathbb{R})$ joining $\mathcal{Q}_{ij}$ to $\mathcal{Q}_{i'j'}$ (with $ij \neq i'j'$) must necessarily pass through the complement of $\mathcal{Q}$. In fact, to change the asymptotic behaviour of ϕ one has to go through fields which do not tend to one of the minima at infinity, and these have infinite energy. So, the spaces $\mathcal{Q}_{ij}$ are separated by infinite energy barriers. For example in the case of the potential (1.2) there are four connected components of $\mathcal{Q}$, labelled $\mathcal{Q}_{++}$, $\mathcal{Q}_{+-}$, $\mathcal{Q}_{-+}$, $\mathcal{Q}_{--}$. In general, the set $\pi_0(\mathcal{Q})$ of connected components of $\mathcal{Q}$ is the cartesian product of two copies of the set indexing the minima: $\pi_0(\mathcal{Q}) = \mathcal{J} \times \mathcal{J}$.

Every $\phi \in \mathcal{Q}_{ij}$ can be written as the sum of an arbitrary given $\phi_0 \in \mathcal{Q}_{ij}$ (which we call the "basepoint" of $\mathcal{Q}_{ij}$) plus a function ψ which tends asymptotically to zero at $\pm\infty$. The function ψ can be regarded as a function $S^1 \to \mathbb{R}$, where $S^1 = \mathbb{R} \cup \{\infty\}$ is the one-point compactification of space. The space of such functions will be denoted $\Gamma_*(S^1, \mathbb{R})$. The subscript $*$ is there to remind us that we are dealing with functions which map a selected "basepoint" of S^1 (namely ∞) to the "basepoint" of $\mathbb{R}$ (namely 0). Therefore all connected components of $\mathcal{Q}$ are vectorspaces isomorphic to $\Gamma_*(S^1, \mathbb{R})$.

The natural problem is then to find the minimum of the energy in each connected component, if it exists. It is clear that in the connected components $\mathcal{Q}_{ii}$ the minima are the constant fields $\phi = y_i$. These are also the absolute minima of E on all $\mathcal{Q}$. In the case of the potential (1.2), one can easily convince oneself by means of the following qualitative argument that with the dynamics considered above there should be absolute minima of the static energy also in the sectors $\mathcal{Q}_{-+}$ and $\mathcal{Q}_{+-}$. Let us denote ℓ the "size of the soliton", *i.e.* the length of the region where the field is significantly different from either vacua. It is clear that the elastic energy is of order f^2/ℓ, and hence decreases with ℓ, while the potential energy is of order $\lambda f^4 \ell$, and hence increases with ℓ. The static energy will have a minimum at some finite value of order $\ell \approx 1/(\sqrt{\lambda} f)$. Inserting in the formula for the energy we also find that both elastic and potential energy of the soliton are of order $\sqrt{\lambda} f^3$. The soliton will therefore be the result of a balance between elastic and potential energy.

In order to find the explicit form of the soliton we have to solve the differential equation

$$\frac{d^2\phi}{dx^2} = \frac{\partial V}{\partial \phi} \tag{1.5}$$

with the appropriate boundary conditions. For the potential (1.2) the solutions of (1.5) in the sectors $\mathcal{Q}_{-+}$ and $\mathcal{Q}_{+-}$ are

$$\phi(x) = \pm \frac{m}{\sqrt{\lambda}} \tanh\left[\frac{m}{\sqrt{2}}(x - x_0)\right] , \tag{1.6}$$

with the upper sign in the first case, the lower sign in the second. These solutions are known as the "kink" and the "antikink" respectively. Note that these solitons are not isolated solutions: they come in one-parameter families, parametrized by the "center of mass" coordinate x_0. This is a reflection of the translational invariance of the action. Figure (1.1) shows a plot of ϕ/f as a function of $x\sqrt{2}/m$ for the kink at $x_0 = 0$. (The horizontal lines correspond to the minima of the potential.)

Inserting (1.6) in (1.4) we obtain

$$E_S = \frac{2\sqrt{2} m^3}{3\lambda} = \frac{2\sqrt{2}}{3} f^3 \sqrt{\lambda} . \tag{1.7}$$

It is useful to note that there is equipartition between elastic and potential energy (*i.e.* each of the two terms in (1.4) contributes exactly $E_S/2$). To see this, multiply both sides of the equation of motion (1.5) by $\frac{d\phi}{dx}$. The resulting equation can be written

$$\frac{d}{dx}\left[\frac{1}{2}\left(\frac{d\phi}{dx}\right)^2 - V\right] = 0 ,$$

implying that the quantity in square brackets is constant. We can evaluate the constant for $x \pm \infty$, and we find it must be zero. Thus, the density of elastic energy and the density of potential energy are equal. In particular, the total elastic and potential energies are equal.

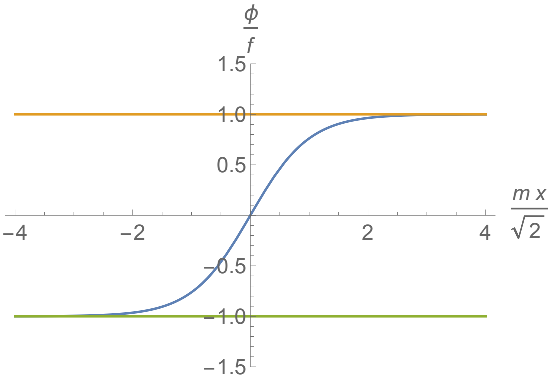

Figure 1.1: The kink of ϕ^4 theory.

In the theory with potential (1.2), consider the current

$$J_T^\mu = \frac{1}{2f}\varepsilon^{\mu\nu}\partial_\nu\phi \; ; \tag{1.8}$$

clearly we have

$$\partial_\mu J_T^\mu = 0 \; . \tag{1.9}$$

This current is conserved without recourse to the equations of motion, and it is not related to any symmetry of the theory. It will be called the topological current. The integral

$$Q_T = \int_{-\infty}^{\infty} dx\, J_T^0 = \frac{1}{2f}\left[\phi(+\infty) - \phi(-\infty)\right] \tag{1.10}$$

is known as the topological charge. It is clear that all fields in $\mathcal{Q}_{-+}$ have $Q_T = 1$, those in $\mathcal{Q}_{+-}$ have $Q_T = -1$ and those in $\mathcal{Q}_{++}$ and $\mathcal{Q}_{--}$ have $Q_T = 0$. Thus Q_T is a measure of the nontriviality of the boundary conditions of the fields.

Another interesting potential is

$$V(\phi) = \frac{m^4}{\lambda}\left[1 - \cos\left(\frac{\sqrt{\lambda}}{m}\phi\right)\right] \; . \tag{1.11}$$

This corresponds to the so called “sine-Gordon” model. The indexing set of minima is the set of the integers $\mathcal{J} = \mathbb{Z}$, so there is a double infinity ($\mathbf{Z} \times \mathbf{Z}$) of connected components. The topological current and the topological charge are given again by (1.8) and (1.10), where f, which is half the distance between two

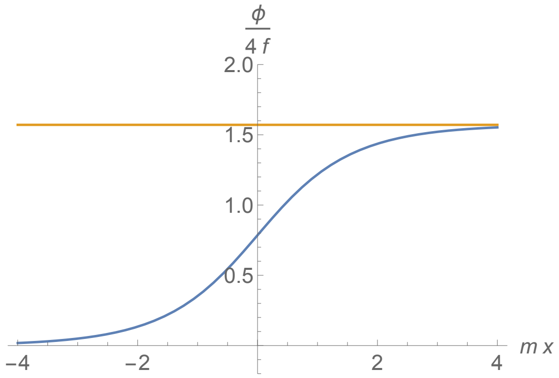

Figure 1.2: The kink of the Sine-Gordon model.

successive minima of the potential, is now equal to $\pi m/\sqrt{\lambda}$. We give the form of the solitons with $Q_T = \pm 1$, which minimize the energy in $\mathcal{Q}_{0\,1}$ and $\mathcal{Q}_{0\,-1}$

$$\phi(x) = \pm \frac{4m}{\sqrt{\lambda}} \arctan\big\{\exp\big[(x-x_0)m\big]\big\} \tag{1.12}$$

This solution is plotted in the Figure (2.7).

Just adding $2nf$ we get the soliton and antisoliton, still with $Q_T = \pm 1$, which minimize the energy in $\mathcal{Q}_{n\,n+1}$ and $\mathcal{Q}_{n\,n-1}$. Note that if in the field equation (1.5) with the potential (1.11) we reinterpret x as time and ϕ as the coordinate of a particle on a line, then we can regard it as Newton's equation of motion of the particle moving in the gravitational potential $-V$. Formula (1.12) represents a motion in which the particle rolls from one maximum of the gravitational potential to the next. Using this analogy it becomes intuitively clear that there cannot be any static soliton of the sine-Gordon model with $|Q_T| > 1$.

Note that this reinterpretation links a field theory in $1+1$ dimensions to mechanics, regarded as a field theory in $0+1$ dimensions. In chapter 2 we shall frequently use this trick of relating theories differing by one in dimension.

1.1.2 Quantum kinks

In this section we consider the quantum version of the ϕ^4 theory with potential (1.2). This simple example already exhibits all the phenomena that characterize quantum solitons of more complicated systems.

We begin from the topologically trivial sectors. We can write a functional integral over fields $\phi \in \mathcal{Q}_{++}$ as an integral over the shifted field $\varphi = \phi - f$, that has trivial boundary conditions $\varphi \to 0$ for $x \to \pm\infty$.

The standard perturbative quantization procedure applied to the small fluctuations around the vacuum state $\varphi = 0$ gives a Fock space of scalar particles, that we shall call "pions" with mass

$$m_\pi = \sqrt{V''(f)} = \sqrt{2}m \ .$$

Note that ϕ is dimensionless and λ has dimensions of mass squared. In this theory weak coupling means $\lambda \ll m_\pi^2$.

The theory is superrenormalizable. The only divergence is logarithmic and renormalizes the pion mass, see figure (1.3). Evaluation of this diagram gives

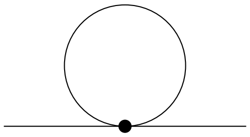

Figure 1.3: Renormalization of the pion mass

for the renormalized mass

$$m_{\pi R}^2 = m_\pi^2 - \frac{3\lambda}{2\pi} \log\left(\frac{\Lambda^2}{m_\pi^2}\right) \ , \tag{1.13}$$

where we employed a simple momentum cutoff Λ.

The theory also contains kinks, that we can view as another type of particles. From (1.7) these particles have mass

$$m_k = \frac{m_\pi^3}{3\lambda} \ . \tag{1.14}$$

Taking the ratio to the pion mass we see that the solitons are much heavier than the pions at weak coupling.

Now one wonders what will become of (1.14) when the pion mass is renormalized. In order to answer this question we will now calculate the quantum corrections to the soliton mass. In the course of this calculation we will learn also several other interesting features of quantum solitons.

The path integral of the theory contains four distinct sectors, corresponding to paths that lie in each of the four connected components of configurations space $\mathcal{Q}_{++}$, $\mathcal{Q}_{+-}$, $\mathcal{Q}_{-+}$, $\mathcal{Q}_{--}$. Standard perturbation theory corresponds to the first or the last of these path integrals. We now consider the other two.

The lowest energy state in $\mathcal{Q}_{-+}$ is given by the kink, so the "vacuum-to-vacuum" amplitude is the sum over fields that are continuous deformations of the kink. We decompose

$$\phi(x) = \bar{\phi}(x) + \eta(x) \ ,$$

where $\bar{\phi}$ is the static solution (1.6), treated as a classical background, and η is the quantum field.

Let us expand the action around the background:

$$\begin{aligned} S(\phi) &= \int dt \left[\frac{1}{2}\int dx \left(\frac{d\phi}{dt}\right)^2 - E_S(\phi)\right] \\ &= S(\bar{\phi}) + \int dt\, dx \left[\frac{1}{2}\left(\frac{d\eta}{dt}\right)^2 - \frac{1}{2}\eta L \eta - \lambda\left(\bar{\phi}\eta^3 + \frac{1}{4}\eta^4\right)\right] \end{aligned} \tag{1.15}$$

where

$$L = -\frac{d^2}{dx^2} + V''(\bar\phi) \tag{1.16}$$

is essentially the second functional derivative of E_S at $\bar\phi$.

Note that the terms on the r.h.s. of (1.15) are ordered in powers of λ: the term $S(\bar\phi) = -m_k \int dt$ is of order λ^{-1} and hence non-perturbative; the first two terms in the square bracket are of order λ^0, the term cubic in η is of order $\sqrt{\lambda}$ ($\bar\phi$ contains a factor $\lambda^{-1/2}$) and the term quartic in η is of order λ. We are going to evaluate quantum corrections at order λ^0, which is equivalent to a standard saddle point (one-loop) evaluation of the path integral.

Most of the complications of this problem derive from the fact that the "mass" term in the operator L is actually a function of x:

$$\begin{aligned} V''(\bar\phi) &= \lambda(3\bar\phi^2 - f^2) \\ &= m^2\left(-1 + 3\tanh^2\left(\frac{mx}{\sqrt{2}}\right)\right) \ . \end{aligned} \tag{1.17}$$

This function is shown in Figure (1.4). Away from the position of the kink it tends quickly to m_π^2, but near the kink it has a dip and becomes even negative.

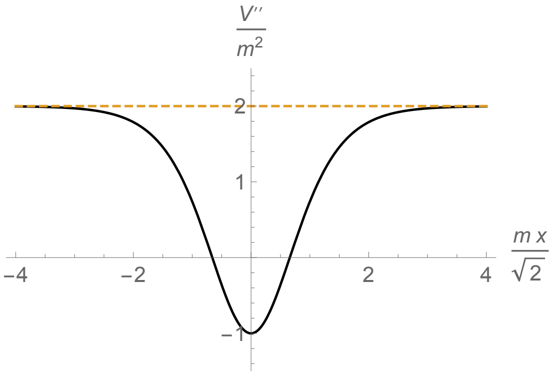

Figure 1.4: The potential in the operator L.

The operator L is a self-adjoint, second order differential operator and therefore its eigenfunctions η_n form a basis for the space of square-integrable functions on the real line:

$$L\eta_n = \omega_n^2 \eta_n \ ; \qquad \int dx\, \eta_n(x)\eta_m(x) = \delta_{nm} \ . \tag{1.18}$$

We have used a notation that is appropriate to a discrete spectrum, as would be obtained if the system was put in a box, but in infinite space the spectrum is actually mixed and consists of the following:

- an isolated eigenvalue $\omega_0^2 = 0$ with eigenfunction $\eta_0 = \frac{1}{\cosh^2\left(\frac{mx}{\sqrt{2}}\right)}$;

- an isolated eigenvalue $\omega_1^2 = \frac{3}{2}m^2$ with eigenfunction $\eta_1 = \frac{\sinh\left(\frac{mx}{\sqrt{2}}\right)}{\cosh^2\left(\frac{mx}{\sqrt{2}}\right)}$ describing an excited state of the kink;
- a continuous spectrum with eigenvalues $\bar{\omega}_p^2 = m_\pi^2 + p^2$, with $-\infty < p < \infty$ describing scattering states of pions in the background of the kink.

We observe that in the absence of the kink there would just be the continuous spectrum with eigenvalues $\omega^2 = m_\pi^2 + p^2$, with $-\infty < p < \infty$. Each eigenvalue corresponds to a normal mode of the field with momentum p, which is e^{ipx}. The presence of the kink deforms the spectrum but in a rather simple way. Mathematically, solving the eigenvalue equation (1.18) is equivalent to solving the Schrödinger equation for a particle moving in the potential (1.17). [1] A "right-moving" mode with momentum $p > 0$ is given for large negative x by $\bar{\eta}_p(x) \approx e^{ipx}$. Near the soliton the solution is more complicated, but it must have again a similar form for large positive x. It turns out that there is no reflected wave, and the transmitted wave, for large positive x is simply

$$\bar{\eta}_p(x) \approx e^{ipx+i\delta_p} \tag{1.19}$$

where the phase shift is given by

$$e^{i\delta_p} = \left(\frac{1+ip/m_\pi}{1-ip/m_\pi}\right)\left(\frac{1+2ip/m_\pi}{1-2ip/m_\pi}\right) \ . \tag{1.20}$$

This is left as an exercise (see Exercise XXX). There is another eigenfunction with the same eigenvalue ω^2, which is given by the "left-mover" $\eta_p(-x)$. The general solution with given p is a linear combination of left- and right-moving waves:

$$A\bar{\eta}_p(x) + B\bar{\eta}_p(-x) \ . \tag{1.21}$$

At this point we put the system in a box of size $L \gg m^{-1}$ and impose boundary conditions on the pions, discretizing the continuous part of the spectrum. Imposing that (1.21) vanishes at $x = \pm L/2$ leads to $A = \pm B$ and $\bar{\eta}_p(L/2) = \pm\bar{\eta}_p(-L/2)$. Then, using the asymptotic behavior of the solutions, one obtains $\exp(ipL - i\delta_p) = \pm 1$, or

$$p = \bar{p}_n \equiv \frac{\pi n}{L} + \frac{\delta_{p_n}}{L} \quad \text{with} \quad n = 0, 1, 2 \ldots$$

We denote $\bar{\omega}_n^2 = m_\pi^2 + \bar{p}_n^2$ the corresponding eigenvalues. We denote

$$p_n = \frac{\pi n}{L} \quad \text{with} \quad n = 0, 1, 2 \ldots$$

the momenta, and $\omega_n^2 = m_\pi^2 + p_n^2$ the eigenvalues, in the absence of the kink.

[1] See Morse and Feschbach, eq (12.3.22) and following.

It is natural to expand the quantum field η on the basis of eigenfunctions of L, instead of ordinary Fourier modes:

$$\eta(t,x) = b_0(t)\eta_0(x) + b_1(t)\eta_1(x) + \sum_{n=0}^{\infty} a_n(t)\bar{\eta}_n(x) \ , \tag{1.22}$$

where the first two terms correspond to the isolated modes and the sum to the "continuous" spectrum. Then, the quadratic part of the Hamiltonian of the fluctuation field becomes a sum of independent oscillators:

$$\begin{aligned} H &= \int dx \left[\frac{1}{2}\dot{\eta}^2 + \frac{1}{2}\eta L \eta\right] \\ &= \frac{1}{2}\dot{b}_0^2 + \frac{1}{2}\left(\dot{b}_1^2 + \omega_1^2 a_n^2\right) + \frac{1}{2}\sum_{n=1}^{\infty}\left(\dot{a}_n^2 + \bar{\omega}_n^2 a_n^2\right) \ . \end{aligned} \tag{1.23}$$

By choosing to work in the basis of eigenfunctions of L we have decomposed the system into infinitely many decoupled oscillators.

There is only one odd degree of freedom, namely the zero mode, which is not an oscillator. Since the potential for this mode is zero, its wave function will not remain localized near the center of the soliton. Recall that the semi-classical approximation rests on the assumption that the quadratic term in the Lagrangian is dominant with respect to the quartic one:

$$\omega^2\langle q^2\rangle \gg \lambda\langle q^4\rangle \ ,$$

where q is to be identified with one of the normal modes. This is true for all the oscillator states, but not for the zero mode.

The physical origin of the zero mode can be understood by noting that η_0 proportional to the derivative of the classical solution (1.6). Among all possible deformations of the kink field, there is one that corresponds simply to an infinitesimal translation of the kink by δx:

$$\delta\phi(x) = \delta x \frac{d\bar{\phi}}{dx} \ .$$

Such a deformation does not change the energy, because a translated kink is a solution of the field equations with the same energy as the original kink. This particular direction in the functional space of the fields corresponds to the bottom of a flat valley for the energy.

This suggests that instead of the zero mode b_0, which amounts to infinitesimal translations of a kink with a fixed center, we take the position of the center of the kink as a dynamical variable. To study the quantization of the center of the kink, let us consider a slowly moving kink, which can be described by the solution (1.6) with x_0 replaced by $x_0(t)$: $\phi(x,t) = \bar{\phi}(x - x_0(t))$. [2] Inserting in

[2] The condition of slow motion is necessary to ensure that the classical field remains at least approximately a solution of the equations of motion. Since the field equations are Lorentz-invariant, a kink in motion will be obtained by operating on the static kink with a boost, and not simply by giving a time dependence to its center. For sufficiently low velocity, however, the two coincide.

the action we find

$$\begin{aligned} S &= \int dt\, dx \left(\frac{1}{2}\dot{\phi}^2 - \frac{1}{2}\phi'^2 - V\right) \\ &= \int dt \left[\dot{x}_0^2 \frac{1}{2}\int dx \phi'^2 - \int dx \left(\frac{1}{2}\phi'^2 + V\right)\right] \ , \end{aligned}$$

where a prime denotes derivative with respect to x. Now we recall that the energy of the kink is equally divided between elastic and potential energy. Thus the coefficient of $\dot{x}_0^2$ is $m_k/2$ and the second integral is m_k:

$$S = \int dt \left[\frac{1}{2} m_k \dot{x}_0^2 - m_k\right] \ . \tag{1.24}$$

The corresponding Hamiltonian is therefore that of a free particle with mass m_k:

$$H = m_k + \frac{p^2}{2m_k} \ . \tag{1.25}$$

This collective degree of freedom can be quantized simply imposing the standard commutation relation $[x_0, p] = i\hbar$. [3] When the motion of the kink is taken into account in this way, we can remove the zero mode from the list of the degrees of freedom.

Then, the energy of the quantum state describing a kink at rest, with the pion field in the Fock vacuum, is given by

$$H = m_k + \frac{\sqrt{3}}{4} m_\pi + \sum_{n=0}^{\infty} \frac{1}{2}\bar{\omega}_n \ , \tag{1.26}$$

where the first term is energy of the classical solution, the second is the vacuum energy of the isolated non-zero mode and the sum extends on the vacuum energy of all the oscillators in the discretized continuous spectrum. For large n, $\bar{\omega}_n \sim n$, so the sum is quadratically divergent. This is the usual divergent contribution to the vacuum energy that one also encounters in any quantum field theory. It is also present in the vacuum sector $\mathcal{Q}_{--}$. We are thus led to define the renormalization of the kink mass as the difference between the sum of the vacuum energies of all the oscillators in the presence of the kink and the sum of the vacuum energies of all the oscillators in the absence of the kink. Both sums are quadratically divergent, and in the difference this divergence is cancelled. The renormalization of the kink mass is therefore

$$\begin{aligned} \delta m_k &= \frac{\sqrt{3}}{4} m_\pi + \frac{1}{2}\sum_{n=1}^{\infty} (\bar{\omega}_n - \omega_n) \\ &= \frac{\sqrt{3}}{4} m_\pi + \frac{1}{2}\sum_{n=1}^{\infty} \frac{p_n \delta_p}{L\omega_n} \ , \end{aligned} \tag{1.27}$$

[3] In the functional integral the transformation of the integration variable from a_0 to x_0 has to be accompanied by a Jacobian. We will not need to compute it here, but it will play a role later in other models.

where, in view of taking the limit $L \to \infty$, in the second step we expanded:

$$\bar{p}_n^2 = p_n^2 + 2\frac{\delta_{p_n}}{L} p_n + O(1/L^2) \; .$$

At this point we can take the limit $L \to \infty$ and we return to continuous momenta:

$$\begin{aligned}
\delta m_k &= \frac{\sqrt{3}}{4} m_\pi + \frac{1}{2\pi} \int dp \frac{p\delta_p}{\sqrt{m_\pi^2 + p^2}} \\
&= \frac{\sqrt{3}}{4} m_\pi + \frac{1}{2\pi} \lim_{\Lambda\to\infty} \delta_p \sqrt{m_\pi^2 + p^2}\Big|_0^\Lambda - \frac{1}{2\pi} \int dp \sqrt{m_\pi^2 + p^2} \frac{d\delta_p}{dp} \; ,
\end{aligned}$$

where in the last line we have performed an integration by parts. Since we are only interested in a logarithmically divergent term, we neglect the first two terms, that are finite ($\delta_\Lambda \sim 1/\Lambda$).

Using the explicit form of the phase shift given in (1.20), we find

$$\frac{d\delta_p}{dp} = \frac{2}{m_\pi} \left(\frac{1}{1 + p^2/m_\pi^2} + \frac{2}{1 + 4p^2/m_\pi^2} \right) \; .$$

A direct calculation (see Exercise XXX) then yields for the renormalized kink mass, up to finite terms,

$$m_{kR} = m_k + \delta m_k = m_k - \frac{3}{4\pi} m_\pi \log\left(\frac{\Lambda^2}{m_\pi^2}\right) \; . \qquad (1.28)$$

For the unrenormalized mass on the r.h.s. we now use equation (1.14), which we can reexpress in terms of the renormalized pion mass, to first order in λ/m_π^2, as

$$m_k = \frac{m_{\pi R}^3}{3\lambda} + \frac{3}{4\pi} m_{\pi R} \log\left(\frac{\Lambda^2}{m_\pi^2}\right) \; .$$

We see that the logarithmic divergence cancels, so that the relation (1.14) is preserved under renormalization:

$$m_{kR} = \frac{m_{\pi R}^3}{3\lambda} \; . \qquad (1.29)$$

1.1.3 Fermions and kinks *

We have considered the quantum properties of the scalar field fluctuating around a kink. Peculiar phenomena happen when fermions propagate in the background of a kink. In this section we consider the scalar theory with potential (1.2) and couple it to a Dirac fermion, a complex two-component field ψ with Lagrangian

$$\mathcal{L}_F = \bar{\psi}(\gamma^\mu \partial_\mu + g\phi)\psi \qquad (1.30)$$

The theory is invariant under global $U(1)$ transformations

$$\psi \to e^{i\alpha}\psi \; ; \qquad \bar{\psi} \to e^{-i\alpha}\bar{\psi}$$

as well as the discrete transformation

$$\phi \to -\phi \ ; \qquad \psi \to \gamma_A \psi \ ; \qquad \bar{\psi} \to -\bar{\psi}\gamma_A$$

where $\gamma_A = \gamma^0\gamma^1$ is the chirality operator. This $\mathbb{Z}_2$ symmetry is broken in the scalar vacuum $\phi = \pm f$, where the fermion acquires a mass $m_F = gf$. [4]

In the sectors $\mathcal{Q}_{--}$ and $\mathcal{Q}_{++}$, *i.e.* in scalar vacuum, the fermion field can be decomposed in plane waves

$$\psi = \int \frac{dp}{2\pi}\frac{1}{\sqrt{2E}}\left[b_p e^{-iEt}u_p(x) + d_p^\dagger e^{iEt}v_p(x)\right] \ .$$

If we choose the representation $\gamma^0 = i\sigma_2$, $\gamma^1 = -\sigma_3$, $\gamma^A \equiv \gamma^0\gamma^1 = \sigma_1$, the elementary spinor solutions are

$$u_p(x) = e^{ipx}\begin{pmatrix} \sqrt{E} \\ \frac{-p-im_F}{\sqrt{E}} \end{pmatrix} \ ; \qquad v_p(x) = e^{-ipx}\begin{pmatrix} \sqrt{E} \\ \frac{-p+im_F}{\sqrt{E}} \end{pmatrix} \ . \qquad (1.31)$$

The field is quantized by imposing the canonical anticommutation relations

$$\{b_p, b_{p'}^\dagger\} = \delta(p-p') \ ; \qquad \{d_p, d_{p'}^\dagger\} = \delta(p-p') \ .$$

which are equivalent to canonical equal-time anticommutation relations for ψ and $\psi^\dagger$.

For the fermion current it is best to use the definition

$$j^\mu = \frac{1}{2}\left(\bar{\psi}\gamma^\mu\psi - \bar{\psi}^c\gamma^\mu\psi^c\right) \ , \qquad (1.32)$$

where $\psi^c = \psi^*$ is the charge conjugate field, obeying the same equation as ψ. This expression has the advantage of avoiding the infinite charge of the Dirac sea that is present in the more familiar expression $j^\mu = \bar{\psi}\gamma^\mu\psi$. Indeed we have

$$Q = \int \frac{dp}{2\pi}\left(b_p^\dagger b_p - d_p^\dagger d_p\right) \qquad (1.33)$$

whereas the Hamiltonian is given by

$$H = \int \frac{dp}{2\pi} E_p \left(b_p^\dagger b_p + d_p^\dagger d_p\right) \ . \qquad (1.34)$$

Let us now see what happens in the presence of a kink. In the chosen representation of the gamma matrices, the Dirac operator has the form

$$\begin{pmatrix} P^\dagger & \partial_t \\ -\partial_t & P \end{pmatrix} \qquad \text{where} \qquad P = \partial_x + g\bar{\phi} \ , \quad P^\dagger = -\partial_x + g\bar{\phi} \ .$$

[4] In general, the sign of the mass term in the fermionic Lagrangian is not physically significant because it can be changed by the field redefinition $\psi \to \gamma_A\psi$, $\bar{\psi} \to -\bar{\psi}\gamma_A$.

Normally squaring the Dirac operator (with a change of sign for the mass term) produces the Klein-Gordon operator times the unit matrix. This calculation requires commuting the mass with derivatives. Now, however, the mass has been replaced by the field $g\bar{\phi}$, which does not commute with the space derivative. We thus find:

$$\begin{pmatrix} -P & \partial_t \\ -\partial_t & -P^\dagger \end{pmatrix} \begin{pmatrix} P^\dagger & \partial_t \\ -\partial_t & P \end{pmatrix} = \begin{pmatrix} -\partial_t^2 - PP^\dagger & 0 \\ 0 & -\partial_t^2 - P^\dagger P \end{pmatrix}$$

where

$$P^\dagger P = -\partial_x^2 + g^2\bar{\phi}^2 - g\partial_x\bar{\phi} \ , \qquad PP^\dagger = -\partial_x^2 + g^2\bar{\phi}^2 + g\partial_x\bar{\phi} \ .$$

The square of the Dirac operator therefore reads $-(\partial_t^2 1 + L)$. where L is the self-adjoint operator

$$L = \begin{pmatrix} PP^\dagger & 0 \\ 0 & P^\dagger P \end{pmatrix}$$

Unlike the normal case, it is not proportional to the unit matrix.

As with the scalar field, it will prove convenient to decompose the spinor on the basis of eigenfunctions of this operator, instead of ordinary Fourier modes. We make the ansatz

$$\psi = e^{-iEt} \begin{pmatrix} \tilde{u}_1(x) \\ \tilde{u}_2(x) \end{pmatrix}$$

and demand that these functions are annihilated by $\partial_t^2 1 + L$. This implies that u_1 must be an eigenfunction of $PP^\dagger$ with eigenvalue E^2 and u_2 must be an eigenfunction of $P^\dagger P$ with the same eigenvalue.

One easily sees that if u is an eigenfunction of $PP^\dagger$ with a given eigenvalue, $P^\dagger u$ is an eigenfunction of $P^\dagger P$ with the same eigenvalue. The converse is also true, so these operators have the same eigenfunctions. If we choose the upper spinor component to be $\tilde{u}_1(x)$, the corresponding lower spinor component must be $\tilde{u}_2(x) = C_2 P^\dagger \tilde{u}_1(x)$, where C_2 is some normalization constant. In the same way we find that if we choose the lower component $\tilde{u}_2(x)$, the upper component must be $\tilde{u}_1(x) = C_1 P \tilde{u}_2(x)$. For these two relations to be compatible we must have $C_1 C_2 = 1/E^2$. [5]

The spectrum of L can be computed analytically, but we shall not need it in the following. Suffice it to say that it consists of a continuum of scattering states and a discrete spectrum with energies $E^2 = 2rg - r^2$, where $r = 0, 1 \ldots$ are integers less than g. The continuum and the discrete states with $r \geq 1$ come in pairs, as described above. The modes $r = 0$, which have zero energy, behave in a drastically different way. The equation $P\tilde{u}_0 = 0$ has solution

$$\tilde{u}_0(x) \sim e^{-g\int^x dy \bar{\phi}(y)} \ .$$

This is a normalizable zero-mode of $P^\dagger P$, due to the asymptotic behavior of the function $\bar{\phi}$. On the other hand the solution of the equation $P^\dagger \tilde{u}_0 = 0$ is

$$\tilde{u}_0(x) \sim e^{g\int^x dy \bar{\phi}(y)} \ .$$

[5] In the case $\bar{\phi} = f$ these relations are satisfied by the solutions in (1.31), with $C_2 = -i/E$.

which is not normalizable, for the same reasons. Therefore $PP^\dagger$ does not have a (normalizable) zero mode.

We can now decompose a spinor in the background of the kink as

$$\psi = b_0 \left(\begin{array}{c} \tilde{0} \\ u_0(x) \end{array} \right) + \int \frac{dp}{2\pi} \frac{1}{\sqrt{2E}} \left[b_p e^{-iEt} \tilde{u}_p(x) + d_p^\dagger e^{iEt} \tilde{v}_p(x) \right] \ . \tag{1.35}$$

where $\tilde{u}_p$ and $\tilde{v}_p$ are the eigenfunctions of L described above.

When this decomposition is used, the Hamiltonian still has the form (1.34), with the integral extending over all the non-zero modes. The zero mode is a discrete fermionic degree of freedom that can be in two quantum states: either free or occupied. The peculiar fact is that the occupied state has zero energy like the empty state. Therefore, the system has two degenerate vacua $|0\rangle$ and $|0'\rangle = b_0^\dagger |0\rangle$.

The surprise comes when we consider the charge of these states. When the decomposition (1.35) is inserted in the fermionic charge

$$Q = \int dx \left(\psi^\dagger \psi - \psi^T \psi^* \right) \ ,$$

due to the fact that they still come in degenerate pairs, the non-zero modes work out as in the absence of the kink and give back (1.33). However, the zero mode does not have a partner and its contribution is different:

$$\frac{1}{2} \left(b_0^\dagger b_0 - b_0 b_0^\dagger \right) = b_0^\dagger b_0 - \frac{1}{2}$$

In the vacuum state where the zero mode is empty

$$Q|0\rangle = -\frac{1}{2}|0\rangle$$

while in the vacuum state where the zero mode is occupied

$$Q|0'\rangle = \frac{1}{2}|0'\rangle$$

So we find that in the presence of the kink the fermionic field does not have a state of zero charge, and the charges are fractional. Creating fermions or antifermions will add integer charges to that of the vacuum, so all the states have a fractional charge. We could say that in the presence of the fermion field the kink itself carries a charge equal to ±1/2.

1.2 Scalar fields in other dimensions

1.2.1 Domain walls *

There is a way to use the preceding solution in higher dimensions. Consider the case of a single scalar field in $d > 1$ space dimensions. The equation of motion for a static solution is

$$\sum_i \partial_i^2 \phi = V' \ , \tag{1.36}$$

We can make an ansatz for the field

$$\phi(x_1,\ldots,x_d)=\phi(x_1)$$

then the equation of motion reduces to that of a scalar in one dimension. We have already discussed solutions for this equation in section 1.1.1. Thus, inserting any of those solutions in the ansatz above gives a solution of the higher dimensional equations.

These kinks in higher dimensions are called *domain walls*. They separate two half-spaces where the scalar is in different vacua. The location of the wall is a linear subspace W of codimension one where the scalar field vanishes. Domain walls are not solitons, because the energy of the solution is infinite:

$$E_S=\int_W d^{d-1}x\,\mathcal{E} \qquad \text{where} \qquad \mathcal{E}=\int dx_1\left[\frac{1}{2}(\partial_1\phi)^2+V(\phi)\right]$$

where $\mathcal{E}$ represents a surface density of energy. For example, for the potential (1.2), one has from (1.7)

$$\mathcal{E}=\frac{2\sqrt{2}}{3}f^3\sqrt{\lambda}\ .$$

(One has to bear in mind that the dimension of f and λ is now different from section 1.1, so that $\mathcal{E}$ has the correct dimension d in mass.)

1.2.2 No go theorems

The existence of topological solitons requires that the configuration space has more than one connected components and that the equations of motion admit smooth, localized, finite energy solutions. These are separate conditions. In this section we show that linear scalar theories with the usual two-derivative kinetic term and a potential, do not satisfy either of them.

We begin with a single scalar in higher dimensions. Finiteness of the static energy

$$E_S=\int d^dx\left[\frac{1}{2}\sum_i(\partial_i\phi)^2+V(\phi)\right]$$

demands that when $r=|\vec{x}|\to\infty$, ϕ tends to one of the minima of V. Thus the configuration space $\mathcal{Q}$ will consist again of various connected components:

$$\mathcal{Q}=\bigcup_{i\in\mathcal{J}}\mathcal{Q}_i\ , \qquad \mathcal{Q}_i=\{\phi\in\mathcal{Q}\mid\phi\underset{r\to\infty}{\longrightarrow}y_i\}$$

and $\mathcal{J}$ is the set of the minima of V. The absolute minimum of E_S in each $\mathcal{Q}_i$ is given by the constant $\phi=y_i$. These are just the classical vacua of the model. The essential difference with the case of the previous section is that in $d=1$ the "sphere at infinity" S^0_∞ defined by the limit $r\to\infty$ consists of two disconnected points, and the field can take different values at these two points, whereas in $d\geq 2$ the "sphere at infinity" S^{d-1}_∞ is connected. By continuity the value of the field at infinity must be constant and there cannot be solutions with nontivial boundary conditions.

Let us next consider the case of $N>1$ scalar fields $\phi=\phi^a$ $(a=1,\dots,N)$ in d space dimensions. The space of all such fields is denoted $\Gamma(\mathbb{R}^d,\mathbb{R}^N)$. Assuming symmetry under $SO(N)$, the action is

$$S=\int d^{d+1}x\left[-\frac{1}{2}\partial_\mu\phi^a\partial^\mu\phi^a-V(|\phi|)\right]\ , \qquad (1.37)$$

where $|\phi|=\sqrt{\phi^a\phi^a}$ and repeated indices are summed over. For definiteness we will consider only the case of a quartic potential

$$V=-\frac{1}{2}m^2|\phi|^2+\frac{\lambda}{4}|\phi|^4+\frac{m^4}{4\lambda}=\frac{\lambda}{4}\left(|\phi|^2-f^2\right)^2,$$

where $f=\sqrt{\frac{m^2}{\lambda}}$ and $m^2>0$. The locus of the minima is a sphere S^{N-1}. The static energy is now

$$E_S=\int d^dx\left[\frac{1}{2}\partial_i\phi^a\partial_i\phi^a+V(|\phi|)\right]\ . \qquad (1.38)$$

We are interested in the subspace $\mathcal{Q}\subset\Gamma(\mathbb{R}^d,\mathbb{R}^N)$ for which the static energy is finite. This demands again that as $r\to\infty$, ϕ tends to one of the minima of V.

One can ask whether it is necessary to allow ϕ to go to an *arbitrary* point of S^{N-1} when $r\to\infty$, or it suffices to consider fields that tend to a *specific* point of S^{N-1}. Let ϕ and ϕ' be two field configurations such that $\phi\underset{r\to\infty}{\longrightarrow}y$ and $\phi'\underset{r\to\infty}{\longrightarrow}y'$, where y and y' are two different points on S^{N-1}. Since all maps from $\mathbb{R}^d$ to $\mathbb{R}^N$ are homotopic, there exists a one-parameter family of maps $\phi_\tau(x)$, with $0\le\tau\le 1$, such that $\phi_0=\phi$ and $\phi_1=\phi'$ (for more on homotopy theory see Appendix A). It is convenient to redefine the homotopy parameter to go from $-\infty$ to ∞ instead than from 0 to 1. For example, we can define

$$\tau=\frac{1}{2}+\frac{1}{\pi}\arctan t\ . \qquad (1.39)$$

Writing $\phi_\tau(x)=\hat{\phi}(x,t)$, we can interpret t as time and $\hat{\phi}\in\Gamma(\mathbb{R}^{d+1},\mathbb{R}^N)$ as a *spacetime* field. The energy of this field is $E=E_K+E_S$ where $E_K=\int d^dx\,\frac{1}{2}\left(\frac{d\hat{\phi}}{dt}\right)^2$ is the kinetic energy. Since $\frac{d\hat{\phi}}{dt}$ does not tend to zero as $r\to\infty$, it is clear that for finite t, E_K is divergent. We conclude that to go from ϕ to ϕ' one must go through configurations with infinite kinetic energy, so the boundary value of ϕ cannot change in the course of the time evolution. For this reason, we will always assume that the configuration space consists of field with a fixed boundary condition at infinity. In particular, the only possible constant field is the field that is everywhere equal to the boundary value. It is worth noting that this restriction has a counterpart in homotopy theory, where one usually considers *based* maps, namely maps that have a predetermined value at a predetermined point.

Using the $SO(N)$ invariance of the theory, we can assume without loss of generality that the value of ϕ as $r \to \infty$ be $y_0 = (0, 0, \ldots, 0, f)$. The limit $r \to \infty$ defines a "sphere at infinity" S^{d-1}_∞; since the map ϕ must be constant on S^{d-1}_∞, all its points may be identified to a single point ∞. Then ϕ may be regarded as a map from the one-point compactification $\mathbb{R}^d \cup \{\infty\} = S^d$ into $\mathbb{R}^N$, mapping the "basepoint" ∞ of S^d to the "basepoint" y_0. Therefore $\mathcal{Q} = \Gamma_*(S^d, \mathbb{R}^N)$. All maps with these properties are homotopic to one another, so the space $\mathcal{Q}$ is connected.

These results imply that linear scalar field theories in dimensions $d \geq 2$ cannot have *topological* solitons. There is an independent result, known as Derrick's theorem, saying that linear scalar field theories with action 1.37 do not admit nontrivial static solutions (whether topological or not) when $d \geq 2$. The proof is based on a scaling argument.

Let us rewrite equation (1.38) as $E_S = E_1 + E_2$, where E_1 and E_2 are the "elastic" and "potential" energy, in the terminology introduced in the previous section. Let ϕ_λ be a one-parameter family of configurations defined by $\phi_\lambda(x) = \phi_1(\lambda x)$. We have

$$E_1(\phi_\lambda) = \lambda^{2-d} E_1(\phi_1) \ , \qquad E_2(\phi_\lambda) = \lambda^{-d} E_2(\phi_1) \ .$$

In order for ϕ_1 to be a stationary point of E_S it is necessary that

$$0 = \frac{d}{d\lambda} E_S(\phi_\lambda)\Big|_{\lambda=1} = (2-d) E_1(\phi_1) - d E_2(\phi_1) \ . \tag{1.40}$$

Since E_1 and E_2 are positive semidefinite, for $d \geq 3$ this implies $E_1(\phi_1) = 0$ and $E_2(\phi_1) = 0$, which is only satisfied by the trivial vacuum configuration.

For $d = 2$ we get $E_2(\phi_1) = 0$. This means that the field must be everywhere in the minimum of V, which implies that $\frac{\partial V}{\partial \phi^a} = 0$. Inserting in the equation of motion we obtain $\partial_\mu \partial^\mu \phi^a = 0$, which, together with the given boundary conditions, implies again $\phi = \text{constant}$.

To escape the negative conclusions derived in this section, one has to modify either the kinematics or the dynamics of the theory, or both. One way is to couple the scalars to gauge fields. This will be discussed in section XXX. Another way is to consider nonlinear scalar theories.

1.2.3 Nonlinear sigma models

Let us start from a linear scalar theory with action (1.37). It is invariant under global internal rotations of the fields, forming the group $SO(N)$. In particular, the potential is constant on the orbits of $SO(N)$ in $\mathbb{R}^N$. The minima occur on a particular orbit $S^{N-1} = SO(N)/SO(N-1)$ (see Appendix C). If we take the limit $\lambda \to \infty$ with f kept constant, the potential becomes unbounded everywhere except on the orbit of the minima, where it remains equal to zero. Thus in the strong coupling limit the potential constrains the field to lie on that particular orbit. This is illustrated by the following figure:

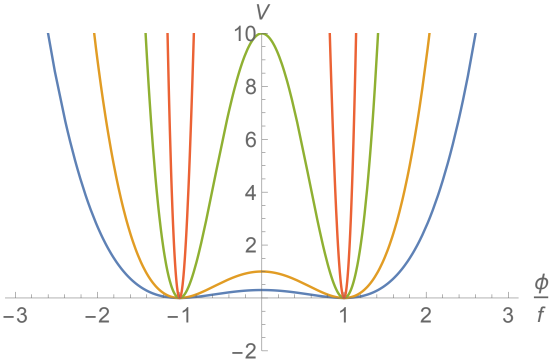

Figure 1.5: The potential with increasing λ.

A mathematically more sensible way of studying the limit is to introduce a Lagrange multiplier field Λ and consider the action

$$S = \int d^{d+1}x \left[-\frac{1}{2}\partial_\mu \phi^a \partial^\mu \phi^a - \frac{2\Lambda}{\sqrt{\lambda}}\sqrt{V} + \frac{\Lambda^2}{\lambda} \right] . \tag{1.41}$$

The equation of motion for Λ is $\Lambda = \sqrt{\lambda V}$ and when this equation is used in (1.41) it gives back (1.37). Thus (1.41) is classically equivalent to (1.37). The advantage of the action (1.41) is that it remains well defined in the limit $\lambda \to \infty$. In fact, it reduces to

$$S = \int d^{d+1}x \left[-\frac{1}{2}\partial_\mu \phi^a \partial^\mu \phi^a - \Lambda(|\phi|^2 - f^2) \right] . \tag{1.42}$$

The second term enforces the constraint $\phi^2 = f^2$. This is called a “nonlinear sigma model with values in S^{N-1}”, or a “$SO(N)$-nonlinear sigma model”.

It is usually quite inconvenient to work with constrained fields. This can be avoided by working directly with the coordinates of the target space. Let us illustrate how this works for the two-dimensional sphere. We can solve the constraint $\phi^a \phi^a = f^2$ expressing the three fields ϕ^a in terms of only two independent fields φ^α. There are infinitely many ways of doing this. For example we could choose φ^α to be the spherical coordinates ($\varphi^1 = \Theta$, $\varphi^2 = \Phi$):

$$\phi^1 = f \sin\Theta \cos\Phi \tag{1.43}$$
$$\phi^2 = f \sin\Theta \sin\Phi \tag{1.44}$$
$$\phi^3 = f \cos\Theta \tag{1.45}$$

Introducing into (1.42), we find the action

$$S = -\frac{f^2}{2} \int d^{d+1}x \, (\partial_\mu \Theta \partial^\mu \Theta + \sin^2\Theta \partial_\mu \Phi \partial^\mu \Phi) \ .$$

Another choice are the stereographic coordinates $\varphi^1 = \omega^1$, $\varphi^2 = \omega^2$:

$$\phi^1 = f\frac{4\omega_1}{\omega_1^2+\omega_2^2+4} \tag{1.46}$$

$$\phi^2 = f\frac{4\omega_2}{\omega_1^2+\omega_2^2+4} \tag{1.47}$$

$$\phi^3 = f\frac{\omega_1^2+\omega_2^2-4}{\omega_1^2+\omega_2^2+4} \tag{1.48}$$

Introducing in (1.42),

$$S = -\frac{f^2}{2}\int d^{d+1}x\,\frac{16}{(\omega_1^2+\omega_2^2+4)^2}(\partial_\mu\omega_1\partial^\mu\omega_1+\partial_\mu\omega_2\partial^\mu\omega_2)\ .$$

In any case the action has the form

$$S = -\frac{f^2}{2}\int d^{d+1}x\,\partial_\mu\varphi^\alpha\partial^\mu\varphi^\beta h_{\alpha\beta}(\varphi)\ , \tag{1.49}$$

where $h_{\alpha\beta}(\varphi)$ is the standard metric on the sphere S^2 of unit radius, written in the chosen coordinate system, and f^2 is a constant.

We recall that the original linear model (1.37) is a standard example of the Goldstone theorem. Picking a vacuum state, for example $\phi = (0,\dots,0,f)$ breaks $SO(N)$ to $SO(N-1)$, and gives rise to $N-1$ massless Goldstone bosons. The model contains an additional "radial" scalar degree of freedom with mass $\sqrt{2\lambda}f$. If we consider phenomena at energies much lower than this mass, the radial mode cannot be excited and we remain just with the Goldstone bosons, whose dynamics is described by the action (1.49).

This discussion can be generalized to scalar fields carrying a representation of any Lie group G. [6] If ϕ_0 is a minimum of the potential, every other point in the orbit of G through ϕ_0 is also a minimum. We assume that all the minima belong to a single orbit. If H is the stabilizer of ϕ_0, the orbit of the minima is diffeomorphic to the coset space G/H. Then, the procedure described above gives a nonlinear sigma model with values in G/H.

In fact, equation (1.49) is valid for any target space G/H, provided we interpret $h_{\alpha\beta}$ as the components of a G-invariant metric. The G-invariance of the action can be proven as follows. Let us first consider a general variation of the field. We have

$$\delta S = -\frac{f^2}{2}\int d^{d+1}x\left[2\partial_\mu\delta\varphi^\alpha\partial^\mu\varphi^\beta h_{\alpha\beta}+\partial_\mu\varphi^\alpha\partial^\mu\varphi^\beta\partial_\gamma h_{\alpha\beta}\delta\varphi^\gamma\right]\ . \tag{1.50}$$

Assume that $\delta\varphi^\gamma = \epsilon^a K_a^\gamma(\varphi)$, where ϵ^a are constant infinitesimal parameters (which can be thought of as an element of the Lie algebra of G) and K_a are vectorfields, satisfying the Killing equation

$$K_a^\gamma\partial_\gamma h_{\alpha\beta}+h_{\alpha\gamma}\partial_\beta K_a^\gamma+h_{\beta\gamma}\partial_\alpha K_a^\gamma = 0\ .$$

[6] See for example L. Michel, "Minima Of Higgs-Landau Polynomials", Contribution to Colloq. on Fundamental Interactions, in honor of Antoine Visconti, Marseille, France, Jul 5-6, 1979. Published in Marseille Collog. 157 (1979) (CERN-TH-2716)

Then is it easy to check that $\delta S = 0$. On the other hand, if we keep the variation arbitrary, but going to zero at infinity so that integrations by parts do not leave any boundary term, then one obtains the field equation

$$\partial_\mu \partial^\mu \varphi^\gamma + \partial_\mu \varphi^\alpha \partial^\mu \varphi^\beta \Gamma^\gamma_{\alpha\beta}(\varphi) = 0 \ . \tag{1.51}$$

where $\Gamma^\gamma_{\alpha\beta}$ are the Christoffel symbols of $h_{\alpha\beta}$.

One can further generalize this discussion by considering nonlinear sigma models with values in a completely arbitrary target manifold, as long as it is endowed with a metric $h_{\alpha\beta}$. This is relevant for example in string theory. However, we will not need to consider such models here. For us a nonlinear sigma model will always be a theory of Goldstone bosons.

1.2.4 Power counting

We conclude this section with some remarks on the quantization of nonlinear sigma models, with the action (1.49). For this discussion it is convenient to define $f = 1/g$. Since the metric is in general a nonpolynomial function, the fields have to be dimensionless. Therefore the constant g^2 must have dimension L^{n-2}, where $n = d + 1$ is the dimension of spacetime. In two spacetime dimensions, and only in two, we can choose $g^2 = 1$. In order to give the scalar fields their canonical dimension, we absorb first the constant g^2 in the fields, defining $\bar\varphi^\alpha = \varphi^\alpha/g$. The dimension of $\bar\varphi$ is then $[\bar\varphi^\alpha] = L^{\frac{2-n}{2}}$. Now the action reads

$$S = -\frac{1}{2} \int d^n x \, \partial_\mu \bar\varphi^\alpha \partial^\mu \bar\varphi^\beta h_{\alpha\beta}(g\bar\varphi) \ . \tag{1.52}$$

The metric $h_{\alpha\beta}(g\bar\varphi)$ is still dimensionless. In order to separate the kinetic term from the interaction terms we have to fix some constant background $\bar\varphi^\alpha_0$, write $\bar\varphi^\alpha = \bar\varphi^\alpha_0 + \eta^\alpha$, and expand the metric in Taylor series in η:

$$h_{\alpha\beta}(g\bar\varphi) = h_{\alpha\beta}(g\bar\varphi_0) + g\partial_\gamma h_{\alpha\beta}(g\bar\varphi_0)\eta^\gamma + g^2 \partial_\gamma \partial_\delta h_{\alpha\beta}(g\bar\varphi_0)\eta^\gamma\eta^\delta + \cdots \tag{1.53}$$

where we write ∂_γ for $\frac{\partial}{\partial\varphi^\gamma}$. The coefficients of this expansion are now field-independent and represent the coupling constant of the theory. Note that there is in general an infinite number of couplings and all couplings involve derivatives of the fields. (In most models of interest, a $\mathbb{Z}_2$ invariance under the transformation $\eta \to -\eta$ forbids terms with an odd number of fields.)

The dimension of the coupling constant in the m-th term, i.e. the coefficient of $\partial\eta\,\partial\eta\,\eta^m$, is $[g^m] = L^{\frac{m}{2}(n-2)}$. In spite of the infinite number of couplings, this theory is renormalizable in a generalized sense for $n = 2$. It is nonrenormalizable for $n > 2$. We note that in some cases, such as the sphere, the metric is entirely determined up to an overall scale by symmetry requirements. In these cases there is really only one independent coupling constant g: All the coefficients of the expansion of the metric are determined by the requirement of G-invariance.

In conclusion, the nonlinear sigma models are not good candidates for fundamental theories in more than two dimensions. Instead, they are widely used in condensed matter physics and also in particle physics, as low energy phenomenological models.

1.3 Nonlinear sigma model in $d=2$

Let us ask whether the nonlinear sigma models could have nontrivial solutions in $d > 1$. All solutions of the nonlinear sigma model equations have $E_2 = 0$, so (1.40) implies that if $d > 2$ the only static solution of the field equations is constant, while in $d = 2$ nontrivial solutions are possible.

For the existence of topological solitons one also needs a suitable target space. The simplest example is S^2, so we now turn to the S^2-nonlinear sigma model in $d = 2$.

1.3.1 Topology

We start by discussing the configuration space. We work with unconstrained fields φ representing a map from $\mathbb{R}^2$ to S^2. Finiteness of the static energy

$$E_S = \frac{f^2}{2}\int d^2x\, \partial_i\varphi^\alpha\partial_i\varphi^\beta h_{\alpha\beta}(\varphi) \tag{1.54}$$

demands that $\partial_i\varphi \to 0$ as $r \to \infty$. Thus φ must tend to a constant at infinity. Without loss of generality we can take this constant value to be the north pole. In spherical coordinates it is given by $\Theta = 0$; in stereographic coordinates it is given by $\sqrt{\omega_1^2+\omega_2^2} \to \infty$. Since from now on we will restrict our attention to this particular class of maps, we can compactify space to a sphere by adding a point at infinity: $S^2 = \mathbb{R}^2 \cup \{\infty\}$. In homotopy theory it is often very convenient to pick a special point in each space, called the "basepoint". In the present context it is natural to choose the basepoint of the spatial S^2 to be the point ∞, and the basepoint of the internal S^2 to be the north pole. There follows that any finite energy configuration can be regarded as a map from S^2 to S^2 preserving basepoints. The space of such maps is denoted $\mathcal{Q}=\Gamma_*(S^2, S^2)$. This space consists of infinitely many connected components: $\pi_0(\mathcal{Q}) = \pi_2(S^2) = \mathbb{Z}$ (see Appendix XXX). So we can write

$$\mathcal{Q} = \bigcup_{n\in\mathbb{Z}} \mathcal{Q}_n \ .$$

The integer n labelling the homotopy classes is known as the winding number. In any coordinate system, it can be written as

$$W(\varphi^\alpha) = \frac{1}{8\pi}\int d^2x\, \varepsilon^{ij}\partial_i\varphi^\alpha\partial_j\varphi^\beta\sqrt{\det h}\,\varepsilon_{\alpha\beta} \ . \tag{1.55}$$

For example, in spherical and stereographic coordinates it has the expression

$$\begin{aligned} W(\Theta,\Phi) &= \frac{1}{4\pi}\int d^2x\,\sin\Theta\varepsilon^{ij}\partial_i\Theta\partial_j\Phi\ , && (1.56)\\ W(\omega^1,\omega^2) &= \frac{1}{4\pi}\int d^2x\,\frac{16}{(\omega_1^2+\omega_2^2+4)^2}\varepsilon^{ij}\partial_i\omega^1\partial_j\omega^2\ , && (1.57)\end{aligned}$$

respectively. It is not obvious from these formulae that it is an integer. However, we can easily prove that W is locally constant. To this effect, let us write

$$\sqrt{\mathrm{det}h(\varphi)}\epsilon_{\alpha\beta}=\omega_{\alpha\beta}(\varphi) \qquad (1.58)$$

for the components of the volume form of S^2. Varying infinitesimally we get

$$\delta W=\frac{1}{8\pi}\int d^2x\,\varepsilon^{ij}\left[2\partial_i\delta\varphi^\alpha\partial_j\varphi^\beta\omega_{\alpha\beta}+\partial_i\varphi^\alpha\partial_j\varphi^\beta\delta\varphi^\gamma\partial_\gamma\omega_{\alpha\beta}\right]\ .$$

Since the variation is supposed to preserve the boundary conditions, it must vanish at infinity. Thus we can integrate the first term by parts. Factoring $\delta\varphi^\gamma$ and antisymmetrizing the first term, we arrive at

$$\delta W=\frac{1}{8\pi}\int d^2x\,\varepsilon^{ij}\partial_i\varphi^\alpha\partial_j\varphi^\beta\delta\varphi^\gamma\left(\partial_\alpha\omega_{\beta\gamma}+\partial_\beta\omega_{\gamma\alpha}+\partial_\gamma\omega_{\alpha\beta}\right)=0\ ,$$

since the exterior derivative of the form ω vanishes. Thus W is a functional on $\mathcal{Q}$ that is constant on each connected component $\mathcal{Q}_n$. We shall encounter in the next section explicit solutions of the field equations for which one can check, by explicit calculation, that $W=n$ is an integer. Then, W is constant and equal to n for all fields belonging to the same connected component $\mathcal{Q}_n$. A more general definition of the winding number and a theorem proving its integrality are discussed in Appendix XXX.

Since the time evolution is a continuous curve in $\mathcal{Q}$, the value of the winding number cannot change: the winding number must be a constant of motion of the theory. This can be confirmed by the following argument. We define a topological current

$$J_T^\lambda=\frac{1}{8\pi}\varepsilon^{\lambda\mu\nu}\partial_\mu\varphi^\alpha\partial_\nu\varphi^\beta\omega_{\alpha\beta}\ ,$$

which is identically conserved:

$$\partial_\lambda J_T^\lambda=\frac{1}{8\pi}\varepsilon^{\lambda\mu\nu}\partial_\lambda\varphi^\gamma\partial_\mu\varphi^\alpha\partial_\nu\varphi^\beta\partial_\gamma\omega_{\alpha\beta}=0\ ,$$

again because the form ω is closed. One sees immediately that the topological charge is equal to the winding number:

$$Q_T=\int d^2x\,J_T^0=\frac{1}{8\pi}\int d^2x\,\varepsilon^{ij}\partial_i\varphi^\alpha\partial_j\varphi^\beta\omega_{\alpha\beta}=W(\varphi)\ . \qquad (1.59)$$

(There is no contribution to the boundary integral coming from spatial infinity, because J^0 is proportional to spatial derivatives, that are required to vanish at infinity.) There follows that $Q_T\!=\!W$ is a constant of motion.

1.3.2 Dynamics

Let us look at the absolute minimum of the static energy (1.54) in each topological sector $\mathcal{Q}_i$. Consider the following inequality [7]

$$0 \leq \int d^2x\, h_{\alpha\beta}\Big(\partial_i\varphi^\alpha \pm \varepsilon_{ik}\partial_k\varphi^\gamma\omega_\gamma{}^\alpha\Big)\Big(\partial_i\varphi^\beta \pm \varepsilon_{ij}\partial_j\varphi^\epsilon\omega_\epsilon{}^\beta\Big) =$$

$$= \int d^2x\Big[2h_{\alpha\beta}\partial_i\varphi^\alpha\partial_i\varphi^\beta \mp 2\varepsilon_{ij}\partial_i\varphi^\alpha\partial_j\varphi^\epsilon\omega_{\alpha\epsilon}\Big] = \frac{4}{f^2}E_S \mp 16\pi W$$

where in the product of the last two terms we used

$$\omega_{\gamma\alpha}\omega_\epsilon{}^\alpha = \epsilon_{\gamma\alpha}\epsilon_{\epsilon\delta}h^{\alpha\delta}\,\mathrm{det}h = h_{\gamma\epsilon}\ .$$

If $W > 0$ (resp. $W < 0$) the inequality with the upper sign (resp. lower sign) is stronger. There follows that

$$E_S \geq 4\pi f^2|W|\ . \tag{1.60}$$

Furthermore, equality holds if and only if

$$\partial_i\varphi^\alpha = \mp\varepsilon_{ik}\partial_k\varphi^\gamma\varepsilon_{\gamma\delta}h^{\delta\alpha}\sqrt{\mathrm{det}h}\ . \tag{1.61}$$

The fields for which this equation is satisfied are the absolute minima of the static energy and are also static solutions of the Euler-Lagrange equations of the theory. Note that (1.61) are first order equations, and therefore simpler than the second order Euler-Lagrange equations.

It is convenient to specialize the discussion to stereographic coordinates ω^1 and ω^2. Equation (1.61) reduces to

$$\partial_i\omega_\alpha = \mp\varepsilon_{ik}\partial_k\omega^\gamma\varepsilon_{\gamma\alpha}\ , \tag{1.62}$$

and spelling these out

$$\begin{aligned}\partial_1\omega^1 &= \pm\partial_2\omega^2\ ,\\ \partial_2\omega^1 &= \mp\partial_1\omega^2\ .\end{aligned} \tag{1.63}$$

If we define $\omega = \omega^1 + i\,\omega^2$ and $z = x^1 + i\,x^2$ we recognize (1.63) as the Cauchy-Riemann equations for the function $\omega = \omega(z)$. The solutions are the functions which are analytic or antianalytic depending on the sign in (1.63). For example $\omega(z) = z^n$ and $\omega(z) = (z^*)^n$, with $n \geq 0$, are solutions of (1.63). Note that for large $|z|$, ω does not tend to an angle-independent limit, but since $|\omega| \to \infty$ it does not matter since all these points represent the north pole of S^2. These functions describe smooth maps $\varphi \in \Gamma_*(S^2, S^2)$ with winding number $W = n$ and $W = -n$ respectively. They are absolute minima of the static energy in the sectors $\mathcal{Q}_n$ and $\mathcal{Q}_{-n}$ respectively ($n \geq 0$).

[7] A.M. Polyakov, A.A. Belavin, "Metastable States of Two-Dimensional Isotropic Ferromagnets", JETP Lett. **22** 245-248 (1975) Pisma Zh. Eksp. Teor. Fiz. **22** 503-506 (1975).

The theory is invariant under rotations, translations and dilatations, so applying these transformations to the solutions we get other solutions. This means that the solitons are not isolated, but rather come in four-parameter families. Applying these transformations to the solutions mentioned above we find

$$\omega(z) = \left(\frac{(z-z_0)e^{i\alpha}}{\lambda}\right)^n \tag{1.64}$$

$$\omega(z) = \left(\frac{(z-z_0)^* e^{-i\alpha}}{\lambda}\right)^n \tag{1.65}$$

where the complex number z_0 gives the position of the center of the soliton, the angle α its "internal orientation" and the positive real number λ its scale. The parameters z, α and λ are the collective coordinates, or moduli, of the soliton. [8]

1.3.3 No ferromagnetic transition in $d=2$

This model can be regarded as the continuum limit of a planar ferromagnetic crystal, with unit spins allowed to point in any direction in a three-dimensional embedding space. Classically, the state of lowest energy of the system is a perfect ferromagnet with all spins aligned in a fixed direction. It has $W=0$. The direction of the spins breaks the rotational invariance of the system and from Goldstone's theorem one expects to find massless excitations in the spectrum. In fact, small perturbations of the field around this state describe massless particles. The field φ is itself the Goldstone boson and its quanta are the fundamental excitations of the system.

However, it is also possible to excite states with $W \neq 0$, namely solitons. Since a soliton with $|W|=1$ has mass $4\pi f^2$, at a fixed temperature T there will be a density of solitons of order $e^{-f^2/kT}$. If the solitons had fixed size (as the kinks of section 1.1), for very small T this would describe an ordered state with a few localized defects. But in this theory solitons can be arbitrarily large without paying any price in energy. Thus in a given box of finite size there will be solitons/antisolitons that occupy much of the (two dimensional) volume and since a soliton has spins pointing in any direction, the ferromagnetic order will be destroyed.

This is a special case of the so-called Mermin-Wagner theorem, stating that in two (or less) space dimensions at temperature $T>0$, there cannot be a phase where a continuous symmetry is spontaneously broken.

[8] Since the theory has internal $SO(3)$ invariance, any rotation of the solution is also a solution. However, as discussed in Section 1.2.2, the boundary condition at infinity breaks $SO(3)$ to $SO(2) \approx U(1)$. A space rotation, which in the chosen coordinates is represented by $z \to e^{i\alpha}z$, can be undone by an internal $U(1)$ transformation, so of these two abelian groups only one remains as a modulus.

1.4 Current algebra and solitons in $d = 3$

Let us consider a nonlinear sigma model with values in some target space N. The scaling argument rules out static solitons for the action (1.49) in dimensions other than two. Nevertheless let us see for what choices of dimension and target space the configuration space would have more than one connected component. Then we shall look for some alternative action functional that could have stationary points in the nontrivial topological sectors.

Following the same reasoning as in the case of the S^2 sigma model, the space of smooth finite energy configurations of the field is $\mathcal{Q}=\Gamma_*(S^d, N)$. Therefore, there is room for the existence of topological solitons whenever $\pi_0(\mathcal{Q}) = \pi_d(N) \neq 0$. One important case is when $N = G$, a Lie group. This is called a principal sigma model. If G is semisimple one has $\pi_3(G)=\mathbb{Z}$, the fundamental class being realized by a homomorphism $SU(2) \equiv S^3 \to G$. These models appear in the description of strong interactions at low energies. To motivate this we will give first a brief review of current algebra.

1.4.1 The chiral models

The strong interactions are described by QCD, a gauge theory with gauge group $SU(3)$. The fields entering the QCD action are the gauge fields A_μ, describing particles called gluons, and spinor fields describing the quarks. There are six known types (or *flavors*) of quarks: u (up), d (down), s (strange), c (charm), b (bottom or beauty) and t (top), in order of increasing mass. Each of them is described by a Dirac spinor. We can collect these quark fields into a column vector q_α, where α is an index that runs over the six flavors. The quark part of the QCD action is

$$S_q = \sum_\alpha \int d^4x\, \bar{\psi}_\alpha \left(i\gamma^\mu D_\mu - m_\alpha\right) \psi_\alpha \ . \tag{1.66}$$

where D_μ denotes the covariant derivative with respect to the gluon fields For arbitrary masses, the only invariance of this action are the constant phase transformations. Infinitesimally, these are given by

$$\delta_{V\alpha}\psi = i\alpha\psi \quad ; \quad \delta_{V\alpha}\bar{\psi} = -i\alpha\bar{\psi} \tag{1.67}$$

The corresponding group is called the vector $U(1)$, or $U(1)_V$. Assuming that N massess are equal, then also the transformations

$$\delta_{V\epsilon}\psi = \epsilon^a T_a \psi \quad ; \quad \delta_{V\epsilon}\bar{\psi} = -\bar{\psi}\epsilon^a T_a \tag{1.68}$$

with T_a a basis in the Lie algebra of $SU(N)$, are symmetries. This group is called $SU(N)_V$.

The masses of the quarks are distributed over a large range, so it is sometimes possible to pretend that some of them are massless. This is a good approximation for the u and d quarks and, to a lesser extent, also for the s quark. Let us suppose that the masses of the N lightest quarks can be neglected (this is usually called the chiral limit of QCD). Then, in addition to the above, the QCD action is invariant also under axial $U(1)$ and $SU(N)$ transformations:

$$\begin{aligned}\delta_{A\alpha}\psi &= \quad i\alpha\gamma^A\psi \;\; ; \quad \delta_{A\alpha}\bar{\psi} = i\alpha\bar{\psi}\gamma^A \quad U(1)_A \; ; \\ \delta_{A\epsilon}\psi &= \epsilon^a T_a \gamma^A \psi \;\; ; \quad \delta_{A\epsilon}\bar{\psi} = \bar{\psi}\epsilon^a T_a \gamma^A \quad SU(N)_A \; .\end{aligned} \tag{1.69}$$

Here $\gamma^A = \gamma^5$ is the chirality operator, which anticommutes with the gamma matrices. We shall now forget about the heavy quarks and have a closer look at the symmetries of massless QCD with N flavors. The generators of the transformations written above are the charges constructed with the following currents:

$$\begin{aligned}j^\mu_V &= \bar{\psi}\gamma^\mu\psi && \text{for } U(1)_V \\ j^\mu_A &= \bar{\psi}\gamma^\mu\gamma^A\psi && \text{for } U(1)_A \\ j^\mu_{V\epsilon} &= \bar{\psi}\gamma^\mu\epsilon^a T_a\psi && \text{for } SU(N)_V \\ j^\mu_{A\epsilon} &= \bar{\psi}\gamma^\mu\gamma^A\epsilon^a T_a\psi && \text{for } SU(N)_A\end{aligned} \tag{1.70}$$

where ϵ is an element of the Lie algebra of $SU(N)$. From the canonical equal-time anticommutation relations

$$\{\psi^{\alpha i}(\vec{x},t), \bar{\psi}_{\beta j}(\vec{y},t)\} = \delta^\alpha_\beta \delta^i_j \delta(\vec{x}-\vec{y}) \; , \tag{1.71}$$

where a, b are Dirac indices and i, j are $SU(N)$ indices, we obtain the following current algebra

$$\begin{aligned}[j^0_V, j^0_V] &= [j^0_V, j^0_A] = [j^0_A, j^0_A] = 0 \; ; \\ [j^0_{V\epsilon_1}, j^0_{V\epsilon_2}] &= j^0_{V[\epsilon_1,\epsilon_2]} \; ; \\ [j^0_{V\epsilon_1}, j^0_{A\epsilon_2}] &= j^0_{A[\epsilon_1,\epsilon_2]} \; ; \\ [j^0_{A\epsilon_1}, j^0_{A\epsilon_2}] &= j^0_{V[\epsilon_1,\epsilon_2]} \; ;\end{aligned} \tag{1.72}$$

One can verify that these are the algebras implied by (1.69).

The vector and axial transformations are entangled; in particular, the axial transformations do not form a subalgebra. It is convenient to reshuffle the $SU(N)_V$ and $SU(N)_A$ transformations in a different way. Since the chirality operator γ^A satisfies $(\gamma^A)^2 = \mathbf{1}$, the operators

$$P_\pm = \frac{1 \pm \gamma^A}{2} \tag{1.73}$$

are projectors and can be used to decompose the Dirac spinors (for each flavor) as the sum of a left handed (negative chirality) and right handed (positive chirality) part: $\psi = \psi_+ + \psi_-$, where $\psi_\pm = P_\pm\psi$. Defining

$$\begin{aligned}j^\mu_{L\epsilon} &= \frac{j^\mu_{V\epsilon} - j^\mu_{A\epsilon}}{2} = \bar{\psi}\gamma^\mu P_- \epsilon^a T_a \psi \; ; \\ j^\mu_{R\epsilon} &= \frac{j^\mu_{V\epsilon} + j^\mu_{A\epsilon}}{2} = \bar{\psi}\gamma^\mu P_+ \epsilon^a T_a \psi \; ;\end{aligned} \tag{1.74}$$

we can rewrite (1.72) as

$$\begin{aligned}
[j^0_{L\epsilon_1}, j^0_{L\epsilon_2}] &= j^0_{L[\epsilon_1,\epsilon_2]} \ ; \\
[j^0_{L\epsilon_1}, j^0_{R\epsilon_2}] &= 0 \ ; \\
[j^0_{R\epsilon_1}, j^0_{R\epsilon_2}] &= j^0_{R[\epsilon_1,\epsilon_2]} \ ;
\end{aligned} \tag{1.75}$$

showing that the global symmetry group is $SU(N)_L \times SU(N)_R$.

The generator of the the group $U(1)_V$ is baryon number, and is therefore an observed symmetry of nature. The group $U(1)_A$ is not realized in nature, because if it was, for every hadron there would be another hadron with the same mass but opposite parity. We shall defer a discussion of the fate of this group to Section 5.3.

In the case $N = 2$ the group $SU(2)_V$ corresponds to isospin (this can be deduced for example by looking at the transformation of the proton and neutron, which are composites of quarks). In the case $N = 3$ the group $SU(3)_V$ corresponds to the $SU(3)$ of the eightfold way (again this follows for example from the action on the octet of baryons). These are not strictly speaking symmetry groups of the real world, because if they were the masses of the proton and neutron (in the case $N = 2$) or of all baryons of the octet (in the case $N = 3$) would be equal. However, to the extent that the mass differences between these particles can be neglected, they are an unbroken symmetry.

The "axial $SU(N)$" transformations cannot be a symmetry of nature, however, not even approximately, for if it was then for each multiplet of baryons and mesons there would exist another multiplet with the same masses but opposite parity. On the other hand, the phenomenology of hadrons shows that the current algebra (1.75) is realized in nature to good approximation for $N = 2$ and to a slightly lesser extent also for $N = 3$. One concludes that $SU(N)_L \times SU(N)_R$ is a symmetry of the Lagrangian but not of the vacuum, or in other words it is a spontaneously broken symmetry. From Goldstone's theorem, then, there should exist $N^2 - 1$ massless scalar particles (Goldstone bosons). There do indeed exist scalar particles whose masses are small compared to those of the other hadrons: these are the pions and, to a lesser extent, all the mesons in the pion/kaon octet. In the case $N = 2$, it is therefore possible to interpret the pions as the Goldstone bosons that come from the spontaneous breaking of $SU(2)_A$. In the case $N = 3$, it is also possible to interpret the pions and kaons as the Goldstone bosons that come from the spontaneous breaking of $SU(3)_A$.

The upshot of this discussion is that in the chiral limit in which N quarks are massless, the vacuum state of QCD breaks $SU(N)_L \times SU(N)_R$, leaving $SU(N)_V$ unbroken, and therefore defines a point U in the coset space $SU(N)_L \times SU(N)_R/SU(N)_V$. This coset space can be geometrically identified with the group $SU(N)$ itself. Suppose now that we want to study low momentum/low energy phenomena. The state of the system is no longer the vacuum state, but in a sufficiently small spacetime region it can still be described as the vacuum. We can describe such a state by giving the vacuum vector a weak dependence on the spacetime point, so at low energy strong interactions can be described by a map from spacetime into $SU(N)$. It is quite convenient to represent this map by a matrix-valued field $U(x) \in SU(N)$ (U is in the fundamental representation).

This is a phenomenological description of low energy QCD, so the action can in principle contain all terms consistent with the symmetries of the theory. However, at low momenta the terms with the lowest number of derivatives will dominate. There cannot be any potential term, and the term with the lowest number of derivatives is

$$S = f^2 \int d^4x \, \mathrm{tr}\big(U^{-1}\partial_\mu U U^{-1}\partial^\mu U\big) \ . \tag{1.76}$$

To relate this to previous formulas, given a coordinate system on $SU(N)$, we call $\varphi^\alpha(x)$ the coordinates of the group element $U(x)$ and we decompose

$$U^{-1}\partial_\mu U = \partial_\mu \varphi^\alpha L_\alpha^a(\varphi) T_a \ , \tag{1.77}$$

where L_α^a are the components of the Maurer-Cartan form on $SU(N)$. In the case of $SU(2)$, these are given explicitly in Appendix XXX. The basis in the Lie algebra, in the fundamental representation, is chosen such that $\mathrm{tr}\, T_a T_b = -\frac{1}{2}\delta_{ab}$. (In the case $N = 2$, $T_a = -\frac{i}{2}\sigma_a$, where σ^a are the Pauli matrices). Then we choose the Ad-invariant inner product in the Lie algebra

$$-2\mathrm{tr}(T_a T_b) = \delta_{ab} \tag{1.78}$$

and we define a Riemannian metric on $SU(N)$ by declaring the Maurer-Cartan forms to be an orthonormal field of co-frames:

$$h_{\alpha\beta} = L_\alpha^a L_\beta^b \delta_{ab} \ . \tag{1.79}$$

In this way we see that the action (1.76) can be is identical to the nonlinear sigma model action (1.49). The advantage of the form (1.76) is that it makes the $SU(N)_L \times SU(N)_R$ invariance of the theory very transparent: if we transform

$$U \to g_L^{-1} U g_R \ , \tag{1.80}$$

the (constant) group elements g_L and g_R cancel (the latter using ciclicity of the trace). [9] On the other hand, choosing a particular U breaks this invariance, leaving a residual unbroken group. For the choice $U = 1$ the unbroken group is the "diagonal" subgroup $SU(N)_V$, defined by $g_L = g_R$.

For phenomenological purposes, the most useful coordinates on $SU(N)$ are the normal coordinates π^α, defined by:

$$U(x) = e^{2\pi^a(x) T_a / f} \tag{1.81}$$

where T_a is a basis in the Lie algebra of $SU(N)$, satisfying $\big[T_a, T_b\big] = f_{ab}{}^c T_c$. Note that the coordinates have been scaled as in (1.52) so as to have the canonical dimension of mass.

[9] When the action is written in the form (1.49), its invariance is less evident. It follows from the fact that the metric $h_{\alpha\beta}$ is both left- and right-invariant, which can be proven by showing that the vectorfields with components L_a^α and R_a^α, that generate right- and left-multiplications, respectively, are Killing vectors for h.

Using (1.81), (1.76) can be expanded as

$$\int d^4x \left[-\frac{1}{2}\partial_\mu\pi^a\partial^\mu\pi^a + \frac{1}{f^2}\varepsilon^{abc}\pi^b\partial_\mu\pi^c\varepsilon^{ade}\pi^d\partial^\mu\pi^e + \dots\right] . \tag{1.82}$$

This corresponds to the expansion (1.53) in normal coordinates in the neighborhood of the identity. One observes that in this model the pions are massless. Furthermore, all interactions contain derivatives of the fields: this is as it should be, since a potential for π would certainly break the global invariance of the theory.

1.4.2 The Skyrmion

We have mentioned in the beginning of this section, that principal models with values in semisimple groups have topological sectors. To describe these sectors in the present formalism let us consider first the case $G = SU(2) = S^3$. The topological sectors in this case are classified by the winding number, which in terms of the fields U can be written (see Exercise XXX):

$$W(U) = -\frac{1}{24\pi^2}\int d^3x\, \varepsilon^{\lambda\mu\nu}\mathrm{tr}\Big(U^{-1}\partial_\lambda UU^{-1}\partial_\mu UU^{-1}\partial_\nu U\Big) . \tag{1.83}$$

For other groups, the generator of $\pi_3(G) = \mathbb{Z}$ can be obtained by embedding $SU(2)$ in G and then considering the composition of this embedding with a map $S^3 \to SU(2)$ of winding number one.

A peculiar feature of principal sigma models is that their configuration space is itself a group. The product of two field configurations is defined by pointwise multiplication: $(U_1U_2)(x) = U_1(x)U_2(x)$. One can then verify directly from (1.83), that

$$W(U_1U_2) = W(U_1) + W(U_2) \quad ; \qquad W(U^{-1}) = -W(U) . \tag{1.84}$$

A field configuration of the form

$$U(\vec{x}) = \exp\big[T_a\hat{x}^a g(r)\big] \tag{1.85}$$

where $\hat{x}^a = \frac{x^a}{r}$ and g is a function which is -2π in the origin and tends to zero as $r \to \infty$, has winding number one. From (1.84), configurations with arbitrary winding numbers can be constructed simply taking powers of (1.85).

Unfortunately, it follows from the discussion in the end of Section 2 that such fields cannot be solutions of the field equations obtained from the action (1.76). In fact, from (1.40) we get

$$\frac{dE(\phi_\lambda)}{d\lambda}\bigg|_{\lambda=1} = -E(\phi_1) < 0 ,$$

so they are unstable against deformations that shrink the size of the soliton to zero. The way of stabilizing the solitons is to add higher order terms to the action. [10] This may seem a bit artificial, but one has to bear in mind that this theory is to be thought of as an effective low energy theory and hence in principle one should consider all terms in the action consistent with the desired symmetry properties. The total action considered by Skyrme was

$$\begin{aligned} S &= \int d^4x \left[\frac{f^2}{4} \mathrm{tr}\big(U^{-1}\partial_\mu U U^{-1}\partial^\mu U\big) \right. \\ &\qquad \left. + \frac{1}{32e^2} \mathrm{tr}\big[U^{-1}\partial_\mu U, U^{-1}\partial_\nu U\big]\big[U^{-1}\partial^\mu U, U^{-1}\partial^\nu U\big] \right] , \end{aligned} \tag{1.86}$$

where e is a new coupling constant. Out of all possible terms containing four derivatives of the fields, only the one with the commutators was chosen, because it contains only two time derivatives of the fields and is therefore better amenable to canonical analysis. This is not essential for what follows, however.

To see that addition of the four-derivative terms circumvents the scaling argument of Derrick's theorem, suppose that the function g goes from $-\pi$ to zero within a distance ℓ of the origin, corresponding to the size of the soliton. Then the static energy is of the order

$$E_S(\ell) \approx \ell^3 \left[\frac{f^2}{\ell^2} + \frac{1}{e^2 \ell^4} \right] .$$

The first term, that comes from the standard two-derivative Lagrangian, is linear in ℓ. It means that one can gain energy by shrinking the profile of the field to zero. The second term has the opposite effect: it favors broad field profiles. In the presence of both terms, the energy has a minimum for some finite value of ℓ, suggesting that solitons can exist. In fact, as in Section 1.1, one can use this argument to derive some qualitative properties of the solutions: the minimum of the energy occurs at $\ell \approx 1/fe$ and inserting this back in the formula for the energy, the mass of the soliton is of the order f/e. For weak coupling ($e \ll 1$), the soliton is much heavier than the pions.

In order to find the soliton with unit winding number, we have to insert the Ansatz (1.85) in the equations of motion that come from (1.86), and solve for the radial function g. Unfortunately the dynamics is sufficiently complicated to prevent an explicit solution. However, solutions can be found numerically.

As in previous examples, the global symmetries of the action imply that the solitons are not isolated solutions but come in families. Restricting our attention to static fields, the energy is invariant under translations, rotations and under $SU(2)_L \times SU(2)_R$, acting on the field as in (1.80). Each of these transformations could in principle give rise to moduli of solutions. However, we must restrict our attention to transformations that preserve the boundary condition $\lim_{r\to\infty} U = \mathbf{1}$ and this means that of $SU(2)_L \times SU(2)_R$ we must only consider the diagonal (isospin) subgroup $SU(2)_V$, acting as

$$U \mapsto g^{-1} U g \ .$$

[10] T.H.R. Skyrme, "A Nonlinear field theory" Proc. Roy. Soc. Lond. **A260** 127-138 (1961); "A Unified Field Theory of Mesons and Baryons" Nucl. Phys. **31** 556-569 (1962).

Regarding rotations, we observe that a rotation matrix $R \in SO(3)$ corresponds (up to a sign) to an $SU(2)$ matrix B by the formula

$$(R\cdot\hat{x})_a T_a = B^{-1}(\hat{x}_a T_a)B\ .$$

When used in the ansatz (1.85), this action exponentiates to

$$U(R\cdot x) = B^{-1}U(x)B\ .$$

We see that if we follow a rotation B by an isospin transformation $g = B^{-1}$ the solution is invariant. This is a symmetry of the skyrmion. [11] Thus the only moduli are given by translations and either rotations or isospin transformations.

These solitons are known as skyrmions. Skyrme suggested that the solitons of the theory (1.86) be interpreted as the baryons. In order to understand this claim, we have to study the quantum numbers of the skyrmions. This we shall do much later, in section 4.3.

1.5 Yang-Mills theories

In this section we will consider the question whether a pure Yang–Mills theory can have static solitons. Before doing this, however, it will be useful to review some generalities about these theories, and to establish the notation. The dynamical variable is a one-form with values in the Lie algebra $\mathfrak{g}$ of a group G: $A = A^a_\mu dx^\mu \otimes T_a$, where $\{T_a\}$ is a basis in $\mathfrak{g}$. With A one can construct the nonabelian field strength

$$F^a_{\mu\nu}{=}\partial_\mu A^a_\nu - \partial_\nu A^a_\mu{+}ef_{abc}A^b_\mu A^c_\nu\ , \qquad (1.87)$$

where $[T_a,T_b]{=}f_{ab}{}^cT_c$ and e is the coupling constant of the theory. The Yang–Mills action is

$$S_{YM} = -\frac{1}{4}\int d^{d+1}x\, F^a_{\mu\nu}F^{a\mu\nu}\ . \qquad (1.88)$$

It is invariant under local gauge transformations

$$A_\mu{\rightarrow}g^{-1}A_\mu g{+}\frac{1}{e}g^{-1}\partial_\mu g\ , \qquad F_{\mu\nu}{\rightarrow}g^{-1}F_{\mu\nu}g\ ,$$

where $g:\mathbb{R}^{d+1}{\rightarrow}G$ and $F_{\mu\nu} = F^a_{\mu\nu}T_a$.

This formulation of the theory is best suited for the perturbative expansion. In many cases it is more convenient to rescale the field A by a factor $1/e$. In this case the Yang–Mills action reads

$$S_{YM} = -\frac{1}{4e^2}\int d^{d+1}x\, F^a_{\mu\nu}F^{a\mu\nu}\ , \qquad (1.89)$$

[11] It is worth noting that this symmetry is a nontrivial mixing of internal and spacetime transformations. It may thus seem to be in contrast with the Coleman-Mandula theorem. It is not, because the Coleman-Mandula theorem only applies in the presence of Poincarè invariance. Here translation invariance is broken by the soliton.

where the curvature is now defined by

$$F^a_{\mu\nu}=\partial_\mu A^a_\nu - \partial_\nu A^a_\mu + f_{abc}A^b_\mu A^c_\nu \ , \tag{1.90}$$

The nonabelian gauge transformations then read

$$A_\mu \to g^{-1}A_\mu g + g^{-1}\partial_\mu g \ , \qquad F_{\mu\nu}\to g^{-1}F_{\mu\nu}g \ . \tag{1.91}$$

This formulation is better suited for the discussion of geometrical properties of the Yang–Mills fields. In this context one often refers to A as a connection and F as its curvature. In the present chapter dealing with solitons we will use the former definition of the theory, with action (1.88). In later chapters we will use the rescaled fields, with action (1.89).

Let us define the Yang–Mills Lagrangian density by $S_{YM} = \int d^{d+1}x\,\mathcal{L}_{YM}$. Separating the space and time components of the curvature we have

$$\mathcal{L}_{YM} = \frac{1}{2}E^a_i E^a_i - \frac{1}{4}F^a_{ij}F^a_{ij} \ ,$$

where $E^a_i = F^a_{0i} = \partial_0 A^a_i - D_i A^a_0$ is the nonabelian "electric" field (we have used the notation $D_i A^a_0 = \partial_i A^a_0 + e f_{abc} A^b_i A^c_0$; this quantity is a covariant derivative with respect to time independent gauge transformations). The space components of the field strength F_{ij} are related to the nonabelian "magnetic" field: in $d = 3$ we define $F_{ij} = \varepsilon_{ijk}B_k$, while in $d = 2$, $F_{ij} = \varepsilon_{ij}B$.

The momenta canonically conjugate to the the fields are

$$P^0_a \equiv \frac{\partial\mathcal{L}_{YM}}{\partial_0 A^a_0} = 0 \ , \qquad P^i_a \equiv \frac{\partial\mathcal{L}_{YM}}{\partial_0 A^a_i} = E^a_i \ .$$

The relation between velocities and momenta is not invertible, so the proper way to formulate the Hamiltonian dynamics is via Dirac's theory of constrained systems. In the present case the equation $P^0_a = 0$ is known as a "primary constraint". The canonical Hamiltonian can be written

$$H_c = \int d^dx \left[\frac{1}{2}P^a_i P^a_i + \frac{1}{4}F^a_{ij}F^a_{ij} - A^a_0 G_a\right] \ , \tag{1.92}$$

where $G_a = D_i P^i_a = D_i E^a_i$. We have to impose that the primary constraint holds for all time. This means that $\{P^0_a(x), H\} = 0$, which results in the "secondary constraint" $G_a = 0$. In the Hamiltonian formalism The fields A^a_0 play the role of Lagrange multipliers enforcing the Gauss law $G_a = 0$.

When studying the canonical formulation of a YM theory it is often very convenient to choose the gauge $A_0 = 0$ (this can be done by performing the gauge transformation $g(x,t) = \mathrm{P}\exp\Big(-e\int^t dt'\,A_0(x,t')\Big)$, where P stands for path ordering). This leaves the freedom of performing time-independent gauge transformations. In this gauge $E^a_i = \dot{A}^a_i$, so the first term in (1.92) is seen as a kinetic term, the second as a potential term. We will mostly use this gauge in later sections.

Let us now come to the question whether a pure Yang–Mills theory can have static solitons. There is here a slight complication: if a gauge field configuration is time-independent, it can acquire a time dependence after a gauge transformation. In a gauge theory one calls a field "static" if there is a gauge in which A_μ is time-independent. This implies that all gauge invariant quantities constructed with the field (such as, for example, the energy density) are time-independent. Note that for a static configuration, the gauge $A_0=0$ may not be the gauge in which $\partial_0 A_\mu=0$, so we do not make this gauge choice here.

We shall now prove that pure YM theory does not admit static solitons if $d\neq 4$ (*i.e.* in five-dimensional spacetime). [12]

For a static field in a gauge in which $\partial_0 A_\mu=0$, the lagrangian is given by $L=E_1-E_2$, where

$$E_1=\frac{1}{2}\int d^dx\,(D_iA_0)^2>0 \qquad \text{and} \qquad E_2=\frac{1}{4}\int d^dx\,(F^a_{ij})^2>0\ .$$

Consider the two-parameter family of configurations $A_{(\sigma,\lambda)}$ defined by

$$\begin{aligned} A_{(\sigma,\lambda)}{}^a_0(x) &= \sigma\lambda A^a_0(\lambda x)\ , && (1.93)\\ A_{(\sigma,\lambda)}{}^a_i(x) &= \lambda A^a_i(\lambda x)\ . && (1.94)\end{aligned}$$

We have $E_1\big(A_{(\sigma,\lambda)}\big)=\sigma^2\lambda^{4-d}E_1\big(A_{(1,1)}\big)$ and $E_2\big(A_{(\sigma,\lambda)}\big)=\lambda^{4-d}E_2\big(A_{(1,1)}\big)$. For $A_{(1,1)}$ to be a solution of the field equations we must have

$$\begin{aligned} 0 &= \frac{d}{d\lambda}L\bigg|_{\lambda=\sigma=1}=(4-d)L\big(A_{(1,1)}\big)\ , && (1.95)\\ 0 &= \frac{d}{d\sigma}L\bigg|_{\lambda=\sigma=1}=2E_1\big(A_{(1,1)}\big)\ , && (1.96)\end{aligned}$$

which implies that for $d\neq 4$, $E_1=E_2=0$, which in turn implies $F^a_{\mu\nu}=0$.

This argument rules out nontrivial static solitons for pure YM theories except in five spacetime dimensions, which are not physically interesting. Static solitons do indeed exist in five spacetime dimensions, but we will discuss them later in a different context, where they have a different physical interpretation and are known as instantons.

1.6 Vortices

1.6.1 The Nielsen-Olesen vortex

We now consider scalar electrodynamics in two space dimensions. [13] The dynamical variables are a $U(1)$ gauge field A_μ coupled to a complex scalar field ϕ, with action

$$S=\int d^3x\left[-\frac{1}{4}F_{\mu\nu}F^{\mu\nu}-\frac{1}{2}|D_\mu\phi|^2-\frac{\lambda}{4}\Big(|\phi|^2-f^2\Big)^2\right]\ , \qquad (1.97)$$

[12] S. Coleman, Comm. Math. Phys. **31** 259 (1973).

[13] H.B. Nielsen and P. Olesen, Nucl. Phys. **B61** 45 (1973).

where $F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu$, $D_\mu\phi = \partial_\mu\phi - ieA_\mu\phi$. (The Lie algebra of $U(1)$ consists of the purely imaginary numbers and one can take as a basis element $T = -i$. The Lie algebra valued gauge potential is therefore an imaginary one-form $A = A^1T = -iA^1$. The field A_μ used in this section is A^1_μ stripped of the index 1. The gauge transformations can then be obtained from (1.5) by putting $g = e^{i\alpha}$.) The theory is invariant under the local gauge transformations

$$A_\mu \to A'_\mu = A_\mu + \frac{i}{e}g^{-1}\partial_\mu g = A_\mu - \frac{1}{e}\partial_\mu\alpha \ , \qquad \phi \to \phi' = g^{-1}\phi = e^{-i\alpha(x)}\phi \ . \tag{1.98}$$

In the gauge $A_0 = 0$, $E_i = F_{0i} = \dot{A}_i$ and $D_0\phi = \dot{\phi}$; in this gauge the energy reads $E = E_K + E_S$, where

$$E_K = \int d^2x \left[\frac{1}{2}\dot{A}_i\dot{A}_i + \frac{1}{2}|\dot{\phi}|^2\right] .$$

and E_S is the static energy

$$E_S = \int d^2x \left[\frac{1}{2}B^2 + \frac{1}{2}|D_i\phi|^2 + \frac{\lambda}{4}\Big(|\phi|^2 - f^2\Big)^2\right] . \tag{1.99}$$

where $B = F_{12}$. The absolute minimum of E_S, the classical vacuum, occurs for

$$B = 0 \ , \qquad D_i\phi = 0 \ , \qquad |\phi| = f \ . \tag{1.100}$$

A particular solution of these conditions is

$$A_i = 0 \ , \qquad \phi = f \ . \tag{1.101}$$

This is the starting point for the usual perturbative discussion of the Higgs phenomenon, showing that the small fluctuations around this vacuum comprise a vector field with mass $m_A = ef$ and a scalar field with mass $m_S = \sqrt{2\lambda}f$. Any gauge transformation of (1.101) $A_i = \frac{i}{e}\, g^{-1}\partial_i g$, $\phi = g^{-1}f$ is obviously still a solution (here $g = e^{i\alpha}$ is a smooth map $\mathbb{R}^2 \to U(1)$). However, there are other interesting states.

We will now look for static solitons, assuming that the gauge in which the field is time-independent is the gauge $A_0 = 0$. The classical configuration space of this theory consists of regular fields with finite static energy. Clearly (A, ϕ) will have finite energy only if the conditions (1.100) are satisfied asymptotically as $r \to \infty$. This requires that

$$\phi(r,\theta) \underset{r\to\infty}{\longrightarrow} \phi_\infty \qquad = f\, e^{-i\alpha_\infty} \ , \tag{1.102}$$

$$A_i(r,\theta) \underset{r\to\infty}{\longrightarrow} \qquad -\frac{1}{e}\partial_i\alpha_\infty \ , \tag{1.103}$$

where α_∞ depends only on the angular variable θ parameterizing the “circle at infinity” S^1_∞. We see that unlike the case of the sigma model, the condition $D_i\phi \to 0$ does not imply that ϕ tends to a constant at infinity: as long as

$|\phi| \to f$, any dependence of ϕ on the angle θ is permitted, because one can always compensate for this dependence by choosing $A_i = \frac{1}{ie}\frac{\partial_i \phi}{\phi}$.

The asymptotic behaviour of the field ϕ as $r \to \infty$ defines a map $\phi_\infty : S^1_\infty \to U(1)$. We have seen that such maps fall into homotopy classes, labelled by the winding number

$$W(\phi_\infty) = \frac{1}{2\pi}\int_0^{2\pi} d\theta \, \frac{d\alpha_\infty}{d\theta} = \frac{i}{2\pi}\int_0^{2\pi} d\theta \, \frac{1}{\phi_\infty}\frac{d\phi_\infty}{d\theta} \ . \tag{1.104}$$

The field ϕ has values in a linear space and therefore any field configuration can be smoothly deformed into any other. The following figure shows a homotopy between a field with $W = 1$ and a constant field $\phi = f$ (having $W = 0$). The circles represent the images in field space of S^1_∞.

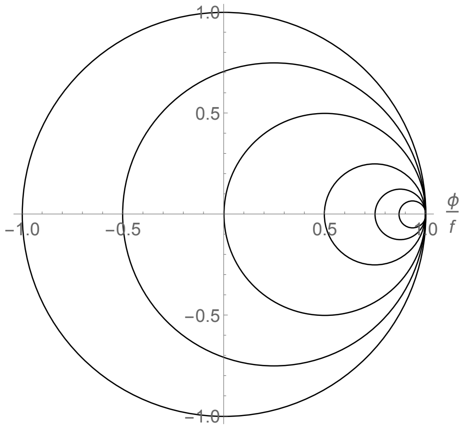

Figure 1.6: A homotopy in fields space. The circle of unit radius is the locus of the minima of the potential.

It is clear that in the intermediate steps of the deformation the field $|\phi|$ does not tend to f as $r \to \infty$. Such fields have infinite static energy, so there is an infinite energy barrier between configurations with different winding numbers of ϕ_∞, or in other words the configuration space consists of infinitely many connected components, labelled by $W(\phi_\infty)$.

The time evolution cannot change the winding number of ϕ_∞, so there must be in the theory a topological conservation law. In fact, consider the topological current

$$J^\lambda_T = \frac{1}{2\pi i}\varepsilon^{\lambda\mu\nu}\partial_\mu \hat{\phi}^* \partial_\nu \hat{\phi} \ ,$$

where $\hat{\phi} = \phi/|\phi|$. This current is identically conserved and the corresponding topological charge is

$$Q_T = \int d^2x \, J^0_T = W(\phi_\infty) \ .$$

The physical meaning of the winding number can be understood by using (1.103) in (1.104) and then applying Stokes' theorem:

$$W(\phi_\infty) = \frac{e}{2\pi} \oint_{S^1_\infty} A_i dx^i = \frac{e}{2\pi} \int_{\mathbb{R}^2} d^2x \, B = \frac{e}{2\pi} \Phi \ ,$$

where Φ is the magnetic flux through $\mathbb{R}^2$ (thinking of B as a magnetic field orthogonal to $\mathbb{R}^2$).

Since W is an integer, we get flux quantization:

$$\Phi = \frac{2\pi}{e} n \ . \tag{1.105}$$

Finally, we would like to find explicit "vortex" solutions in each topological sector. For the solitons with unit flux we make the ansatz

$$\begin{aligned} A_0 &= 0 \ , \\ A_i &= -\varepsilon_{ij} \hat{x}^j A(r) \ , \\ \phi &= F(r) e^{i\varphi} \ , \end{aligned} \tag{1.106}$$

where A and F are functions of the radius such that $A(r) \to \frac{1}{er}$ and $F(r) \to f + O(r^{-1})$ when $r \to \infty$. Clearly the asymptotic conditions are satisfied and $W(\phi_\infty) = 1$. However, it has so far proved impossible to solve explicitly the equations of motions (proofs of existence have been given, though). One has to resort to numerical calculations.

1.6.2 Superconductivity *

Now consider scalar electrodynamics in $d = 3$. If we assume that $A_3 = 0$ and that all the fields are independent of x_3, then the equations of the theory reduce to those of scalar electrodynamics in $d = 2$. Thus, the vortex soliton of $d = 2$ becomes as an infinite vortex line in $d = 3$. It now has infinite energy on account of its infinite length, so it is not a soliton, but it has important physical application that we review briefly here.

What in relativistic quantum fied theory would be called scalar QED, is called in condensed matter physics a Landau-Ginzburg theory. It is an approximation of Bardeen-Cooper-Schrieffer (BCS) theory, which is itself an (approximate) microscopic model. The important properties of superconductors are independent of the approximate nature of these models and follow simply from the assumption that in the bulk of the material, electromagnetic gauge invariance is in the Higgs phase.

In the BCS theory the charge-carriers are weakly-bound pairs of electrons. Such pairs can be described by a field transforming under $U(1)$ as

$$\phi(x) \to \exp(i2e\alpha(x)/\hbar)\phi(x) \ ,$$

where $-e$ is the electron charge and α is identified mod2π. The field is invariant under transformations with $\alpha = \pi\hbar/e$, so a nontrivial VEV for this field would break $U(1)$ to $\mathbb{Z}_2$. In the ungauged case, there would then be a Goldstone boson with values in $U(1)/\mathbb{Z}_2$. It is a real field identified mod$\pi\hbar/e$ and transforming under $U(1)$ by

$$\varphi \to \varphi + \alpha \ . \tag{1.107}$$

Given any other field ψ, transforming linearly under gauge transformation, with charge q, we can construct a gauge invariant field

$$\tilde{\psi}(x) = \psi(x) \exp(iq\varphi(x)/\hbar)., \tag{1.108}$$

The Lagrangian has to have the form

$$\mathcal{L} = -\frac{1}{4} F_{\mu\nu} F^{\mu\nu} + \mathcal{L}_m(\varphi, \tilde{\psi}) \ ,$$

where the matter Lagrangian depends only on the gauge invariant fields and on the Goldstone boson, the latter entering only through the covariant derivative

$$D_\mu \varphi = \partial_\mu \varphi - A_\mu \ .$$

We can define the current

$$J^\mu = \frac{\delta \mathcal{L}_m}{\delta A_\mu} \tag{1.109}$$

The equation of motion of the field φ is then seen to be equivalent to the statement of current conservation:

$$0 = \frac{\delta \mathcal{L}_m}{\delta \varphi} - \partial_\mu \frac{\delta \mathcal{L}_m}{\delta \partial_\mu \varphi} = \partial_\mu \frac{\delta \mathcal{L}_m}{\delta A_\mu} = \partial_\mu J^\mu \ .$$

The main assumption that is needed to describe superconductivity is that the Goldstone boson is covariantly constant:

$$D_\mu \varphi = 0 \ . \tag{1.110}$$

This condition will follow if we make the rather reasonable assumption that in the Lagrangian for φ, the lowest term is quadratic in $D_\mu \varphi$. This is indeed what one has in the Landau-Ginzburg theory: if we consider the action (1.97) and take the strong coupling limit, following the procedure of section 1.2.3, we will arrive at

$$\mathcal{L} = -\frac{f^2}{2}(D\varphi)^2 \ .$$

Conside a static situation with $\partial_0 \varphi = 0$, $A_0 = 0$. The space component of (1.110)

$$A_i = \partial_i \varphi \tag{1.111}$$

implies that the magnetic field must be zero:

$$B_i = 0 \ .$$

This is known as Meissner effect and holds in the bulk of a piece of superconductor. If there is an external magnetic field, the field lines will be deformed so as to avoid going through the semiconductor.

Absence of electrical resistance can be gleaned from the following argument. In any simply connected piece of superconductor, φ can be set to any fixed constant by a transformation (1.107). Now consider a thick torus made of superconductor and let ℓ be a closed loop deep in the material. Integrating (1.111) on this loop and using Stokes' theorem we find that

$$\Delta\varphi = \int_\ell A = \int_S B = \Phi \ ,$$

where Φ is the magnetic flux through a surface S bounded by the loop ℓ. Since the Goldstone field is periodically identified, it can jump by integral multiples of $\pi\hbar/e$. We thus find that flux must be quantized:

$$\Phi = \frac{\pi\hbar}{e} n \ . \tag{1.112}$$

Because of this, the current in the superconductor cannot decay continuously. The fact that the current is not affected by ordinary electrical resistance can be proven more generally by considering the time dependence of the current. We will not discuss this here.

We have seen that the crucial field in the description of the superconducting state is a real Goldstone boson. On the other hand, in the ordinary state of matter, electromagnetic $U(1)$ is not Higgsed, and the fields carry ordinary linear representations of $U(1)$. By continuity, near the transition also the Goldstone boson must be accompanied by a dynamical modulus field ρ that acts as an order parameter: it is zero in the normal state and nonzero in the superconducting state.

If the superconductor is exposed to an external magnetic field, the field lines will penetrate the material but only for a depth of order

$$\lambda = \frac{1}{ef} = \frac{1}{m_V} \ , \tag{1.113}$$

which is called the penetration depth and is the inverse of the mass of the photons in the bulk. In addition to λ, superconductors are characterized by another length scale called the superconducting coherence length

$$\xi = \frac{1}{\sqrt{2} m_S} \ . \tag{1.114}$$

In the Landau-Ginzburg description, these lengths are just the inverse masses of the gauge fields and the scalar. The ratio of these lengths

$$\kappa = \frac{\lambda}{\xi} = \sqrt{2}\frac{m_S}{m_A}$$

characterizes the behavior of the material in a strong magnetic field: a superconductor is said to be of type I if $0 < \kappa < 1/\sqrt{2}$ and of type II if $\kappa > 1/\sqrt{2}$.

In a type I superconductor, when the magnetic field exceeds a critical value, the material undergoes a phase transition to a normal state. In a type II superconductor, when the external magnetic field exceeds a critical value, it is energetically favorable for the magnetic field to penetrates the superconductor in the form of thin tubes, called Abrikosov vortices. In Landau-Ginzburg theory they correspond to the vortex solution discussed in the preceding section. In the core of each tube the modulus field is zero, but elsewhere the material remains superconducting. This is called the vortex phase. The density of vortices increases with the external magnetic field, up to a second, higher, critical field, where supercondutivity is lost.

Since the core of the flux tube is not superconducting, the topology of a piece of superconductor that is pierced by a vortex line is the same as that of the thick torus discussed earlier. Therefore, the flux through the tube must be quantized as in (1.112). Note the similarity between the classical quantization conditions (1.105) of Landau-Ginzburg theory and the quantum condition (1.112). In the former e is a classical parameter in the Lagrangian that could have any value, in the latter it is identified with the electron charge (the factor of two is due to the charge of the Cooper pairs). Yet somehow we see that the topological information is preserved in the approximate phenomenological theory. This is a rather general phenomenon of which we shall encounter other examples later on.

1.7 Monopoles

Maxwell's equations can be written in the form

$$\begin{aligned} \partial_\mu F^{\mu\nu} &= 4\pi J^\nu_{(E)} \ , \\ \partial_\mu {}^*F^{\mu\nu} &= 0 \ , \end{aligned} \qquad (1.115)$$

where ${}^*F_{\mu\nu} = \frac{1}{2} g_{\mu\rho} g_{\nu\sigma} \varepsilon^{\rho\sigma\alpha\beta} F_{\alpha\beta}$ is the dual of the field strength. (Recall that $g_{\mu\nu} = (-\ +\ ++)$ and $\varepsilon^{0123} = 1$. In Minkowski space ${}^{**}F = -F$, whereas in Euclidean space one would have ${}^{**}F = F$). In vacuum ($J^\nu_E = 0$) these equations are invariant under the duality transformation $F \to {}^*F$, ${}^*F \to {}^{**}F = -F$. Writing

$$F_{\mu\nu} = \begin{pmatrix} 0 & E_1 & E_2 & E_3 \\ -E_1 & 0 & +B_3 & -B_2 \\ -E_2 & -B_3 & 0 & B_1 \\ -E_3 & B_2 & -B_1 & 0 \end{pmatrix} \qquad {}^*F_{\mu\nu} = \begin{pmatrix} 0 & -B_1 & -B_2 & -B_3 \\ B_1 & 0 & E_3 & -E_2 \\ B_2 & -E_3 & 0 & E_1 \\ B_3 & E_2 & -E_1 & 0 \end{pmatrix}$$

we see that duality transformations amount to the replacements $E \to -B$, $B \to E$. In fact the vacuum Maxwell equations are invariant under a whole $U(1)$ group of transformations of the form

$$\begin{aligned} F \to &\quad \cos\theta F + \sin\theta\, {}^*F \\ {}^*F \to &\quad -\sin\theta F + \cos\theta\, {}^*F \ . \end{aligned} \qquad (1.116)$$

In the presence of sources an asymmetry is seen to arise, due to the empirical fact that the r.h.s. of the second equation in (1.115) is identically zero. They could be made symmetric under duality transformations by introducing a

$$\partial_\mu {}^*F^{\mu\nu} = 4\pi J_M^\nu \tag{1.117}$$

and postulating the transformation

$$\begin{aligned} J_E &\to \quad \cos\theta J_E + \sin\theta J_M \ , \\ J_M &\to -\sin\theta J_E + \cos\theta J_M \ . \end{aligned} \tag{1.118}$$

That J_M^ν is a magnetic current is seen by observing for example that the time component of (1.117) would read $\mathrm{div}B = 4\pi\rho_M$, and therefore acts as the source of the magnetic potential, i.e. has to be interpreted as the magnetic charge density. Such a modification would introduce essential new features in the theory. Most important, if $J_M \neq 0$ it would become impossible to introduce a magnetic potential A_μ such that $F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu$. This complication does not arise if we limit ourselves to the study of pointlike magnetic sources. The Coulomb–like field

$$B_i = \frac{Q_M}{r^2}\hat{x}^i \ , \tag{1.119}$$

describing a static pointlike magnetic monopole in the origin, solves the equation $\mathrm{div}B = 4\pi Q_M \delta(r)$. Since the field is singular in the origin, one can remove this point from space and regard (1.119) as a smooth field on $\mathbb{R}^3\backslash\{0\}$. Since the field B given in (1.119) is divergence free on $\mathbb{R}^3\backslash\{0\}$, it is possible to introduce the magnetic potential there.

This solution of Maxwell's equations has interesting properties that we shall study in detail in Section 3.1. In paricular we will find that the magnetic monopole can be regarded as a $U(1)$ gauge field only if Q_M is quantized in certain units. For the time being we merely observe that it is a singular field and has infinite energy, so it does not satisfy the requirements for a soliton. The remarkable fact is that certain nonabelian gauge theories with Higgs fields admit solitons whose behaviour at large r approaches that of a Dirac monopole. We will now discuss this type of solutions.

1.7.1 The 't Hooft-Polyakov monopole

We consider the Georgi-Glashow model, consisting of an $SU(2)$ gauge field $A_\mu = A_\mu^a T_a$ coupled to a Higgs field ϕ^a in the adjoint (triplet) representation. We use the unscaled gauge fields, with curvature (1.87) and action (1.88). The total Lagrangian density is

$$\mathcal{L} = -\frac{1}{4}F_{\mu\nu}^a F^{\mu\nu a} - \frac{1}{2}D_\mu\phi^a D^\mu\phi^a - \frac{\lambda}{4}\left(\phi^a\phi^a - f^2\right)^2 \tag{1.120}$$

where $D_\mu\phi^a = \partial_\mu\phi^a + e\varepsilon_{abc}A_\mu^b\phi^c$. The structure constants of the Lie algebra of $SU(2)$ are $f_{abc} = \varepsilon_{abc}$ (in the adjoint representation the generators are $(T_a)_{bc} =$

$-\varepsilon_{abc}$). The action is invariant under the local gauge transformations (1.5), acting on the scalar as $\phi \to g^{-1}\phi$ (here g is in the adjoint representation). It is convenient to choose the gauge so that $A_0^a = 0$. Then $F_{0i}^a = \partial_0 A_i^a$, $D_0\phi^a = \partial_0\phi^a$. The static energy is

$$E_S = \int d^3x \left[\frac{1}{4}\left(F_{ij}^a\right)^2 + \frac{1}{2}\left(D_i\phi^a\right)^2 + \frac{\lambda}{4}\left(\phi^a\phi^a - f^2\right)^2\right] . \tag{1.121}$$

Its absolute minimum is obtained for

$$\begin{aligned} F_{ij} &= 0 \ , \\ D_i\phi^a &= 0 \ , \\ \phi^a\phi^a &= f^2 \ , \end{aligned} \tag{1.122}$$

in which case $E_S = 0$. This is the classical vacuum of the theory. Due to the shape of the potential, the Higgs phenomenon occurs. This can be seen by choosing a gauge in which $A_i^a = 0$, $\phi^a = \bar{\phi}^a = (0, 0, f)$ and expanding the action to second order in A and in the shifted field $\phi - \bar{\phi}$. Invariance under local $SU(2)$ transformations is not broken, however, and any gauge transform of this solution is also a solution.

Finiteness of E_S demands that the conditions (1.123) be satisfied asymptotically when $r \to \infty$. In particular for large r we must have $\phi^2 = f^2 + O(1/r^2)$, so the asymptotic behaviour of ϕ defines a map $\phi_\infty : S^2_\infty \to S^2_{\rm int}$, where S^2_∞ denotes the "sphere at infinity" in $\mathbb{R}^3$ and $S^2_{\rm int}$ is the locus of the minima of the potential in the field space. The covariant derivative and the magnetic field have to go to zero like $1/r^2$. As in the abelian case, discussed in the previous section, the second condition in (1.123) does not restrict the map ϕ itself. The asymptotic field ϕ^a_∞ can depend on the angles in an arbitrary way; the condition $D_i\phi \to 0$ can then be solved by

$$A_i^a = \frac{1}{f^2 e}\varepsilon^{abc}\partial_i\phi^b\phi^c + \alpha_i\phi^a + O(1/r^2) \ , \tag{1.123}$$

for an arbitrary constant α_i.

The scalar fields ϕ fall into classes, labelled by the winding number of the map ϕ_∞. Fields with different winding numbers at infinity are separated by an infinite energy barrier. There follows that the configuration space of smooth finite energy configurations for this model consists of infinitely many connected components, labelled by the winding number of ϕ_∞. The configuration with $W = 0$ is the vacuum, the other one is called a "hedgehog". The winding number cannot be altered in the course of the time evolution, so there will be a topological conservation law. We define the topological current

$$J_T^\mu = \frac{1}{8\pi}\varepsilon^{\mu\nu\rho\sigma}\varepsilon_{abc}\partial_\nu\hat{\phi}^a\partial_\rho\hat{\phi}^b\partial_\sigma\hat{\phi}^c \ , \tag{1.124}$$

where $\hat{\phi}^a = \frac{\phi^a}{\sqrt{\phi^b\phi^b}}$. This current is identically conserved and the corresponding charge is

$$\begin{aligned} Q_T &= \int d^3x J^0_T = \frac{1}{8\pi}\int d^3x \varepsilon^{ijk}\varepsilon_{abc}\partial_i\hat{\phi}^a\partial_j\hat{\phi}^b\partial_k\hat{\phi}^c \\ &= \frac{1}{8\pi}\int\limits_{S^2_\infty} d^2x\varepsilon^{ij}\varepsilon_{abc}\hat{\phi}^a\partial_i\hat{\phi}^b\partial_j\hat{\phi}^c = W(\phi_\infty)\ . \end{aligned} \tag{1.125}$$

The last equality can be proven by choosing a particular coordinate system on S^2, for example the spherical coordinates (1.45), and comparing with (1.57).

We are now in a position to explain why configurations with $W \neq 0$ can be interpreted as monopoles. When the Higgs phenomenon occurs, we can interpret the projection of the gauge field along the Higgs VEV as an abelian gauge field. If $\hat{\phi}^a = (0,0,1)$, the corresponding field strength is $\mathcal{F}_{\mu\nu} = \partial_\mu A^3_\nu - \partial_\nu A^3_\mu$.

Following 't Hooft, we can generalize this to position-dependent Higgs fields. [14] Let $\mathcal{A}_\mu = A^a_\mu\hat{\phi}^a$. We define an abelian electromagnetic field $\mathcal{F}_{\mu\nu}$ by

$$\mathcal{F}_{\mu\nu} = \partial_\mu\mathcal{A}_\nu - \partial_\nu\mathcal{A}_\mu - \frac{1}{e}\varepsilon_{abc}\hat{\phi}^a\partial_\mu\hat{\phi}^b\partial_\nu\hat{\phi}^c\ . \tag{1.126}$$

The last term has been added to compensate the $SU(2)$ non-invariance of $\mathcal{A}$. In fact, this can also be written as

$$\mathcal{F}_{\mu\nu} = \hat{\phi}^a F^a_{\mu\nu} - \frac{1}{e}\varepsilon_{abc}\hat{\phi}^a D_\mu\hat{\phi}^b D_\nu\hat{\phi}^c\ , \tag{1.127}$$

which is manifestly invariant under $SU(2)$ gauge transformations. This tensor does not obey the Bianchi identities. Instead,

$$\partial_\nu\, {}^*\mathcal{F}^{\nu\mu} = \frac{4\pi}{e}J^\mu_T\ . \tag{1.128}$$

as one can check most easily using (1.126). Comparing with (1.117), we see that we can interpret $\frac{1}{e}J^\mu_T$ as a magnetic current. The corresponding magnetic charge is

$$Q_M = \frac{1}{e}Q_T = \frac{1}{e}W\ . \tag{1.129}$$

Since W is an integer, we get a quantization condition for the magnetic charge, analogous to the flux quantization condition (1.105). We shall see in section 3.1 that quantum mechanics requires the magnetic charge to be quantized in units of $\frac{\hbar}{2e}$, where e is the charge of the electron. The relation between these two conditions is the same as that between (1.105) and (1.112).

We would like to get an explicit solution to the Euler-Lagrange equations realizing these nontrivial boundary conditions. Consider the ansatz

$$\begin{cases} \phi^a = \frac{x^a}{r}F(r)\ , \\ A^a_i(x) = \varepsilon_{aij}\frac{x^j}{r}A(r)\ , \\ A^a_0 = 0\ . \end{cases} \tag{1.130}$$

[14] G. 't Hooft, Nucl. Phys. **B79** 276 (1974); A.M. Polyakov, Pisma v. Zh. E.T.F. **20** 430 (1974), JETP Lett. **20** 194 (1974).

In order for the potential energy to be finite, $F(r) - f$ must go to zero faster than $r^{-3/2}$. Then we calculate

$$D_i\phi^a = (\delta_{ia} - \hat{x}_i\hat{x}_a)\left(\frac{1}{r} - eA\right) F + \hat{x}_i\hat{x}_a F' \ , \qquad (1.131)$$

where a prime stands for derivative with respect to r. The contribution to the energy coming from the covariant derivatives will be finite provided $A(r) \to \frac{1}{er}$ for $r \to \infty$. For the non-abelian magnetic field we have

$$B_i^a = -\left(\delta_{ia} - \hat{x}_i\hat{x}_a\right) A' - \frac{1}{r}\delta_{ia}A + \hat{x}_i\hat{x}_a\left(eA^2 - \frac{1}{r}A\right) \ . \qquad (1.132)$$

It behaves at large r like $1/r^2$, so the magnetic field energy will be automatically finite.

Clearly, the conditions for finiteness of the energy are satisfied and this configuration belongs to the sector $W = 1$. Since $D\phi \to 0$ for $r \to \infty$, the abelian magnetic field

$$\mathcal{B}_i = \frac{1}{2}\varepsilon_{ijk}\mathcal{F}_{jk} \to \hat{\phi}^a B_i^a = -\frac{1}{e}\frac{\hat{x}^i}{r^2} \qquad (1.133)$$

while $\mathcal{E}_i = \mathcal{F}_{0i} = 0$. Therefore, for large r, the abelian field strength becomes identical to the one of the Dirac monopole.

When the ansatz (1.130) is inserted into the Euler-Lagrange equations, these become coupled second order differential equations for the functions F and A. The exact solution to these equations has not been found; only numerical solutions have been given.

1.7.2 The Prasad-Sommerfield limit

There is one particular limit, known as the Prasad-Sommerfield limit, in which the functions F and A can be solved exactly: it is the limit in which λ and m^2 tend to zero with $f = \sqrt{m^2/\lambda}$ constant. In this limit one can derive a useful bound on the energy. We have

$$\begin{aligned} E &= \int d^3x \left[\frac{1}{4}F_{ij}^aF_{ij}^a + \frac{1}{2}D_i\phi^a D_i\phi^a\right] \\ &= \frac{1}{4}\int d^3x \left(F_{ij}^a \mp \varepsilon_{ijk}D_k\phi^a\right)^2 \pm \frac{1}{2}\int d^3x\, \varepsilon_{ijk}F_{ij}^a D_k\phi^a \ . \end{aligned} \qquad (1.134)$$

In the second term on the r.h.s. the covariant derivative can be integrated by parts, and using the Bianchi identities for F_{ij}^a it becomes

$$\frac{1}{2}\int d^3x\, \partial_k\left(\varepsilon_{ijk}F_{ij}^a\phi^a\right) = f\int_{S^2_\infty} d\sigma^k\mathcal{B}_k = 4\pi f\, Q_M = \frac{4\pi f}{e}W \ , \qquad (1.135)$$

where we have used (1.133). Using this in (1.134) we get the so-called Bogomol'nyi bound [15]

$$E \geq \frac{4\pi f}{e}|W| \ , \qquad (1.136)$$

[15] E.B. Bogomol'nyi, Sov. J. Nucl. Phys. **24** 449 (1976)

with equality holding if and only if

$$F^a_{ij} = \pm\varepsilon_{ijk} D_k \phi^a \ . \tag{1.137}$$

The solutions of these equations are the absolute minima of the static energy and therefore automatically satisfy the Euler-Lagrange equations of the theory. In this way we have been able to replace the second-order Euler-Lagrange equations with the first-order equations (1.137). In the Prasad-Sommerfield limit, [16] the explicit form of the functions appearing in (1.130), for the lower sign in (1.137), is

$$\begin{aligned} F(r) &= \frac{f}{\tanh(efr)} - \frac{1}{er} \ , \\ A(r) &= \frac{1}{er} - \frac{f}{\sinh(efr)} \ . \end{aligned} \tag{1.138}$$

The profiles of these functions is shown in the following figures.

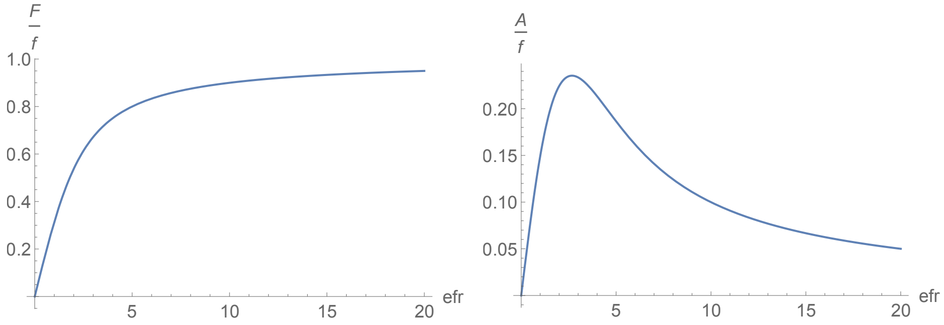

Figure 1.7: Monopole profiles in the Prasad-Sommerfield limit.

1.7.3 Symmetries and moduli

The symmetries of the Georgi-Glashow model are the Poincaré group and internal $SO(3)$ transformations with constant parameters (usually called global gauge transformations). These are transformations that correspond to observable transformations on the fields, and they do not include local gauge transformations, that correspond to unobservable transformations of the fields.

We now ask which of these transformations are also symmetries of the monopole solution. Time translation invariance is preserved, because the solution is static, but space translations are broken, because we can distinguish a monopole from a translated monopole. Boosts are also broken: acting with a boost generates another solution that describes a monopole in motion. There remain to discuss internal rotations and space rotations.

[16] M.K. Prasad, C.H. Sommerfield, Phys. Rev. Lett. **35** 760 (1975).

Let us consider the effect these transformations have on the scalar field $\phi^a = F(r)\hat{x}^a$. An internal transformation with constant parameter ϵ_I^a transforms

$$\delta_I \phi^a = \varepsilon_{abc} \epsilon_I^b \phi^c \ . \tag{1.139}$$

Under the rotation group a scalar transforms by

$$\delta_R \phi^a = \delta x^k \partial_k \phi^a \ ,$$

where

$$\partial_k \phi^a = \frac{1}{r}\left[\delta_{ka} - \frac{x_k x_a}{r^2}(rF' - F)\right]$$

A space rotation corresponds to $\delta_R x^i = \epsilon_R^a \varepsilon_{aij} x^j$, so

$$\delta_R \phi^a = \varepsilon_{abc} \epsilon_R^b \phi^c \ . \tag{1.140}$$

Combining (1.139) and (1.140) we see that ϕ is invariant under the combined transformation

$$(\delta_R - \delta_I)\phi^a = 0$$

where the infinitesimal parameters are the same in the two cases. One can show that also A_i^a is invariant under the same transformations, so the monopole has a symmetry $SO(3)$ consisting of simultaneous internal and space rotations.

This subgroup is unbroken and does not give rise to moduli. There remains one $SO(3)$ subgroup that we can choose to correspond to the internal transformations and could give rise to moduli. However, recall that we work in a functional space with fixed boundary conditions. In this space the field ϕ at infinity is fixed and we do not allow transformations that change it. Since the field ϕ^a at infinity is direction-dependent, the transformations that leave it invariant must also be direction-dependent and are not strictly speaking a subgroup of the rigid internal $SO(3)$ rotations. They can be described in the following way. As is always the case, a field configuration in the Higgs phase can be brought to the unitary gauge, where the Higgs is aligned along the third direction. The transformation that does this for the monopole is

$$T = \begin{bmatrix} \frac{n_2^2 + n_1^2 n_3}{1-n_3^2} & -\frac{n_1 n_2}{1+n_3} & -n_1 \\ -\frac{n_1 n_2}{1+n_3} & \frac{n_1^2 + n_2^2 n_3}{1-n_3^2} & -n_2 \\ n_1 & n_2 & n_3 \end{bmatrix} \tag{1.141}$$

This is clearly a singular transformation, since it changes the winding number of the Higgs field at infinity. but it defines a valid gauge locally. In this gauge the last remaining modulus consists just of the group of internal rotations around the third axis. Alternatively, in the regular gauge the modulus parameterizes the rotations of the form

$$T^{-1} e^{\alpha t_3} T$$

In conclusion, the monopoles come in a four-parameter family, characterized by the coordinates of the center of mass and an internal angle.

1.7.4 Monopoles in GUTs

The preceding discussion of the Georgi-Glashow model can be generalized to arbitrary groups G and H. The condition for the existence of monopoles is that the map ϕ_∞, mapping the sphere at infinity to the minima of the potential, can be topologically nontrivial. Since the orbit where the potential is minimized is diffeomorphic to the coset space G/H, the condition is that $\pi_2(G/H)$ be nontrivial.

The homotopy groups of this space are related to the homotopy groups of G and H by the so-called homotopy exact sequence. This is discussed in general in Appendix XXX. The part of the sequence that is relevant to us is

$$\ldots \xrightarrow{\partial} \pi_2(H) \xrightarrow{\iota_*} \pi_2(G) \xrightarrow{\mu_*} \pi_2(G/H) \xrightarrow{\partial} \pi_1(H) \xrightarrow{\iota_*} \pi_1(G) \rightarrow \ldots \tag{1.142}$$

Here ι_* are the homomorphisms of homotopy groups induced by the embedding $\iota : H \rightarrow G$ and μ_* are the homomorphisms induced by the projection $\mu : G \rightarrow G/H$. The basepoints in the groups are the identity elements e and the basepoint in the coset space is the coset of the identity, eH.

The following facts are known about Lie groups. The homotopy groups of $U(1)$ are given in the table of Appendix XXX. Only the fundamental group is nonzero. If G is a compact, connected, simple Lie group G, $\pi_2(G) = 0$ and $\pi_3(G) = \mathbb{Z}$.

We can use these properties to deduce the second homotopy group of the coset space. One has to use the fact that the maps in the exact sequence are such that the image of each map is the kernel of the next. For example, in the case of GUTs, the group G is compact, simple and simply connected and the subgroup H contains an abelian factor (the unbroken electromagnetic $U(1)_Q$). Thus

$$\ldots \xrightarrow{\partial} 0 \xrightarrow{\iota_*} 0 \xrightarrow{\mu_*} \pi_2(G/H) \xrightarrow{\partial} \mathbb{Z} \xrightarrow{\iota_*} 0 \rightarrow \ldots \tag{1.143}$$

The fact that $\pi_2(G) = 0$ implies that the map μ_* is injective and the fact that $\pi_1(G) = 0$ implies that μ_* is surjective. Thus $\pi_2(G/H)$ is isomorphic to $\mathbb{Z}$, and the theory will have monopoles.

This argument does not apply to the Standard Model, because the group G contains an abelian factor $U(1)_Y$. Even though this subgroup is not the same as the electromagnetic, unbroken group $U(1)_Q$, one can continuously deform one into the other by

$$Q_t = tT_3 + Y \ , \qquad 0 \leq t \leq 1$$

and therefore $\iota_* : \pi_1(U(1)) \rightarrow \pi_1(G)$ is still an isomorphism. This implies that the image of ∂ is zero. On the other hand, since $\pi_2(SU(2) \times U(1)) = 0$, the map ∂ is still injective. Therefore we must have $\pi_2(SU(2) \times U(1)/U(1)) = 0$, and we conclude that the Standard Model does not admit monopole solutions. In this case one could have come to the same conclusion more easily by noting that the orbit of the minima is the locus where the norm of the complex Higgs doublet vanishes:

$$|\phi_1|^2 + |\phi_2|^2 = v^2 \ ,$$

which is a three-sphere.

Instead of appealing to the existence of the homotopy exact sequence, one can give an *ad hoc* proof of the above results, that goes as follows. [17] We start from the map $\phi_\infty : S^2_\infty \to G/H$. As usual in homotopy, we think of it as a map $I \times I \to G/H$, such that $\phi_\infty(t_1, t_2) = eH$ whenever t_1 or t_2 is equal to 0 or 1. Using the gauge field A we construct a map $g_\infty : S^2_\infty \to G$ as follows:

$$g_\infty(t_1, t_2) = P \exp \int_0^{t_1} dt A(t, t_2) \ .$$

The integral is along the line (t, t_2) with constant t_2. Since $D\phi_\infty = 0$,

$$g_\infty(t_1, t_2) eH = \phi_\infty \ ,$$

or in other words $\phi_\infty = \mu \circ g_\infty$. Clearly $g(t_1, 0) = g(t_1, 1) = g(0, t_2) = e$. Since $\mu(g_\infty(1, t_2)) = \phi_\infty(1, t_2) = eH$, $g_\infty(1, t_2) \in H$. We define $h(t) = g_\infty(1, t_2)$. In this way we have constructed a map $\partial : \pi_2(G/H) \to \pi_1(H)$ that maps $[\phi_\infty]$ to $[h]$.

Next we observe that the map g_∞ defines a homotopy of $\iota \circ h$ (for $t_1 = 1$) to a constant (for $t_1 = 0$). Thus $\mathrm{im}\,\partial \subset \ker \iota_*$. Running the above argument backwards, given $h : S^1 \to H$ such that $\iota \circ h$ is homotopic to a constant, we construct a map ϕ_∞ such that $\partial([\phi_\infty]) = [h]$. Thus ∂ is surjective. [18]

To complete the proof, we must show that ∂ is also injective, *i.e.* that if h is homotopic to a constant, also $[\phi_\infty]$ is homotopic to a constant. To this end, let us define $\gamma : I \times I \to G$ by

$$\gamma(t_1, t_2) = \begin{cases} g_\infty(2t_1, t_2) & \text{for } \ 0 \le t_1 \le \frac{1}{2} \\ g_\infty(1, t_2) & \text{for } \ \frac{1}{2} \le t_1 \le 1 \ . \end{cases}$$

and $\varphi : S^2 \to G/H$ by

$$\varphi(t_1, t_2) = \begin{cases} \phi_\infty(2t_1, t_2) & \text{for } \ 0 \le t_1 \le \frac{1}{2} \\ eH & \text{for } \ \frac{1}{2} \le t_1 \le 1 \ . \end{cases}$$

These maps are such that $\varphi = \mu \circ \gamma$. Then, let h_t be a homotopy between $h_0 = h$ and $h_1 = eH$. If we replace γ by the map

$$\gamma'(t_1, t_2) = \begin{cases} g_\infty(2t_1, t_2) & \text{for } \ 0 \le t_1 \le \frac{1}{2} \\ h_{2t_1 - 1}(t_2) & \text{for } \ \frac{1}{2} \le t_1 \le 1 \ . \end{cases}$$

we have again $\varphi = \mu \circ \gamma'$, but now the map γ' is equal to e on the boundary of $I \times I$, and therefore can be viewed as a map $S^2 \to G$. Since $\pi_2(G) = 0$, γ' is homotopic to a constant, and therefore also φ is homotopic to a constant, which is equivalent to saying that ϕ_∞ is homotopic to a constant, QED.

[17] See Coleman's lectures on "Classical lumps and their quantum descendants".

[18] This proves the exactness of the homotopy sequence at $\pi_1(H)$.

Chapter 2

$\pi_1(\mathcal{Q})$, θ-sectors and instantons

We have seen in the previous chapter that when the configuration space of a theory is not connected, there is a conserved topological charge and, if the dynamics is properly chosen, topological solitons. In this chapter we will assume that the configuration space is connected but not simply connected. This leads again to a splitting of the Hilbert space in superselection sectors, but the physical interpretation is different. The paradigm of this phenomenon is the Aharonov–Bohm effect. In the first part of the chapter we give several examples of theories, both in quantum mechanics and quantum field theory, with multiply connected configuration spaces. In the second part we introduce the instantons, which are solutions of the field equations representing the motion of the system in Euclidean time through a non-contractible loop in configuration space.

2.1 Theta sectors

2.1.1 The Aharonov Bohm effect

Consider an experiment where electrons emerge from a source, graze a solenoid carrying a magnetic flux Φ and thereafter form an interference pattern on a screen. As the magnetic flux is varied, the interference pattern is observed to vary. It is found that the interference pattern repeats itself when the flux is changed by $\frac{2\pi\hbar}{e}$, where e is the charge of the electron. [1]

We will now give a theoretical interpretation of this phenomenon. Consider the idealized situation of an infinite perfect solenoid lying along the z axis. The core of the solenoid is assumed to be totally impenetrable to the electrons (the core of the solenoid is made e.g. of iron, and we neglect the probability of

[1] Y.Aharonov and D. Bohm "Significance of electromagnetic potentials in quantum theory", Phys. Rev. **115** 485–491 (1959). "Further Considerations on Electromagnetic Potentials in the Quantum Theory", Phys. Rev. **123** 1511–1524 (1961).

an electron tunnelling through the solenoid). When the current flows, there is a constant magnetic field inside the solenoid but the magnetic field is zero outside (a real solenoid is not infinitely long and the distance between the coils is not zero, so the magnetic field has a weak "tail" outside the solenoid; this we also neglect). As a result of these approximations, the electrons move in a configuration space $\mathcal{Q}$ which is all of $\mathbb{R}^3$ with the solenoid removed and the magnetic field vanishes on $\mathcal{Q}$. The space $\mathcal{Q}$ is multiply connected, with $\pi_1(\mathcal{Q})=\mathbb{Z}$. Consider the magnetic potential

$$\mathcal{A}=\theta\frac{\hbar}{2\pi e}d\varphi\ , \tag{2.1}$$

where θ is an arbitrary real parameter, and φ is the azimuthal cylindrical coordinate around the z axis. The magnetic field corresponding to $\mathcal{A}$ is zero, so $\mathcal{A}$ is a good gauge potential on $\mathcal{Q}$. To find the meaning of the parameter θ, consider the line integral of $\mathcal{A}$ along a loop encircling once the z axis: $\oint\mathcal{A}=\theta\frac{\hbar}{e}$. On the other hand, using Stokes' theorem, the line integral is equal to the integral of $\mathcal{F}=d\mathcal{A}$ on a surface bounded by the loop; such a surface cuts through the solenoid, so the integral is equal to the magnetic flux through the solenoid, Φ. So we find $\theta=\frac{e}{\hbar}\Phi$. We conclude that $\mathcal{A}$ is the potential seen by an electron travelling outside the solenoid when the flux in the solenoid is $\frac{\hbar}{e}\theta$.

The interference pattern on the screen arises from the phase difference between waves that travel above and below the solenoid. Consider first the case when there is no flux, $\theta=0$. The wave function satisfies the free Schrödinger equation $H_0\psi_0=E\psi_0$, with the free hamiltonian $H_0=-\frac{\hbar^2}{2m}\partial_i\partial_i$. Let us now turn the flux on. The Hamiltonian becomes

$$H=-\frac{\hbar^2}{2m}\mathcal{D}_i\mathcal{D}_i \tag{2.2}$$

where $\mathcal{D}_i=\partial_i-\frac{ie}{\hbar}\mathcal{A}_i$ is the covariant derivative with respect to $\mathcal{A}$. It is immediate to check that

$$\psi(q)=\psi_0(q)e^{\frac{ie}{\hbar}\int^q\mathcal{A}}\ , \tag{2.3}$$

obeys the Schrödinger equation with Hamiltonian (2.2) and the same energy eigenvalue E. The phase difference between waves that travel above and below the solenoid in the presence of the magnetic flux is equal to the phase difference in the absence of magnetic flux, plus $\frac{e}{\hbar}\oint\mathcal{A}=\theta$. This phase, and hence the interference pattern, varies linearly with flux. When θ changes by 2π, the phase repeats itself. So the interference pattern has to be periodic in Φ with period $\frac{2\pi\hbar}{e}$, as observed. This concludes the theoretical explanation of the Aharonov-Bohm effect.

Let us see a bit more closely what this phenomenon means. The effect of a gauge transformation on the wave function and on the gauge potential is

$$\psi'=g^{-1}\psi\ ,\qquad \mathcal{A}'=\mathcal{A}-\frac{\hbar}{ie}g^{-1}dg=\mathcal{A}-\frac{\hbar}{e}d\alpha\ , \tag{2.4}$$

where $g(x) = e^{i\alpha(x)}$ is a function from $\mathcal{Q}$ into $U(1)$. We assume that the wavefunctions ψ are periodic both before and after the gauge transformation, and therefore g has to be a well-defined, single valued function into $U(1)$. [2] Two gauge potentials $\mathcal{A}$ and $\mathcal{A}'$ are $U(1)$-gauge related only if the function g in (2.4) is single valued.

Now consider two gauge potentials $\mathcal{A} = \theta\frac{\hbar}{2\pi e}d\varphi$ and $\mathcal{A}' = \theta'\frac{\hbar}{2\pi e}d\varphi$ corresponding to different values of the flux. Are they gauge related in the strict sense defined above? We have

$$\mathcal{A}' - \mathcal{A} = (\theta' - \theta)\frac{\hbar}{2\pi e}d\varphi \ ,$$

and comparing with (2.4) we see that $\alpha(\varphi) = \frac{\theta - \theta'}{2\pi}\varphi$. The gauge potentials $\mathcal{A}$ and $\mathcal{A}'$ are $U(1)$-gauge related if $e^{i\alpha}$ is single valued, which is equivalent to $\theta - \theta' = 2\pi n$, with n integer. Thus we learn that the interference patterns are the same whenever the gauge potentials are $U(1)$ gauge-related, and differ otherwise.

There is here a significant difference between classical and quantum mechanics. In classical mechanics the electron moves according to the Lorentz force. Any two gauge potentials giving the same field strength will follow the same trajectories. In particular, the value of θ is physically irrelevant. In quantum mechanics there is not a single trajectory, rather a superposition of all possible trajectories, and the phase accrued by the wave function along different trajectories depends in a nontrivial way on the gauge potential. The classical electron "sees" only the field strength, but the quantum electron is sensitive to the gauge equivalence class of the potential. One can thus say that there is an ambiguity in the quantization: to a single classical theory there correspond infinitely many inequivalent quantum theories parametrized by the angle θ.

2.1.2 Generalization

In mathematics, a gauge potential is interpreted as a connection and its field strength as the corresponding curvature. The connections with zero curvature are called flat connections. It emerged in the discussion of the Aharonov-Bohm effect that *the inequivalent quantum theories are in one-to-one correspondence with $U(1)$ gauge equivalence classes of flat connections.* We can now take this as the paradigm for a new class of phenomena and look for generalizations in other theories.

For this, we need the answer to the following mathematical question: given a manifold $\mathcal{Q}$, what is the set of gauge equivalence classes of $U(1)$ flat connections on $\mathcal{Q}$? To this end, recall that all the gauge invariant information about a connection is contained in its holonomies ("Wilson loops")

$$\chi(\ell) = e^{\frac{ie}{\hbar}\oint_\ell \mathcal{A}} \ . \qquad (2.5)$$

[2] It is not strictly necessary to assume that ψ is periodic. See the discussion below. However we will see there that one does not lose any generality by making this assumption.

In the case of a flat connection, these holonomies are invariant under continuous deformations of the loop (homotopies). Thus they only depend on the homotopy class of the loop: $\chi(\ell) = \chi([\ell])$. It is easy to see, using the definition of product of homotopy classes given in Appendix A, that

$$\chi([\ell_1]\cdot[\ell_2]) = e^{\frac{ie}{\hbar}\oint_{\ell_1}\mathcal{A}+\frac{ie}{\hbar}\oint_{\ell_2}\mathcal{A}} = \chi([\ell_1])\chi([\ell_2]) \ ,$$

so χ defines a homomorphism from $\pi_1(\mathcal{Q})$ into $U(1)$. Conversely, given any character χ, it can be shown that there exists a flat connection $\mathcal{A}$ such that (2.5) holds.

Thus, the set of flat $U(1)$ connections modulo gauge transformations is in bijective correspondence with the set of characters of the fundamental group:

$$\mathrm{Hom}\big(\pi_1(\mathcal{Q}), U(1)\big) \ .$$

Note that if $\mathcal{Q}$ is simply connected, there is no quantization ambiguity of this type.

In the following we shall encounter only two cases: $\pi_i(\mathcal{Q}) = \mathbb{Z}$ and $\pi_i(\mathcal{Q}) = \mathbb{Z}_2$. In the former case the characters are given by $\chi_\theta(n) = e^{i\theta n}$. Since θ and $\theta + 2\pi m$, with $\in \mathbb{Z}$ define the same character, we have

$$\mathrm{Hom}\big(\mathbb{Z}, U(1)\big) = U(1) \ ,$$

where $U(1)$ is parameterized by $0 \le \theta < 2\pi$. In the other case the characters are $\chi_+(1) = 1$, $\chi_+(-1) = 1$ and $\chi_-(1) = 1$, $\chi_-(-1) = -1$, so

$$\mathrm{Hom}\big(\mathbb{Z}_2, U(1)\big) = \mathbb{Z}_2 \ .$$

2.1.3 The topological term

The lesson of the Aharonov-Bohm effect can now be carried over to an arbitrary configuration space. Consider a particle with mass m, electric charge e, moving on a manifold $\mathcal{Q}$ with metric $g_{ij}(q)$, potential $V(q)$, magnetic field $\mathcal{F}_{ij}(q)$. Also, let $\mathcal{A}_i(q)$ be a gauge potential such that $\mathcal{F}_{ij} = \partial_i\mathcal{A}_j - \partial_j\mathcal{A}_i$. Everything that follows is true also in the case when $\mathcal{Q}$ is infinite dimensional. The most general Lagrangian quadratic in time derivatives of q is

$$L = \frac{1}{2}m g_{ij}(q)\dot{q}^i\dot{q}^j + e\mathcal{A}_i(q)\dot{q}^i - V(q) \ . \qquad (2.6)$$

The momentum conjugate to q^i is

$$p_i = m g_{ij}(q)\dot{q}^j + e\mathcal{A}_i(q) \ , \qquad (2.7)$$

and the canonical hamiltonian is

$$H = \frac{1}{2m}g^{ij}(q)\big(p_i - e\mathcal{A}_i\big)\big(p_j - e\mathcal{A}_j\big) + V(q) \ ,$$

where $g^{ij}g_{jk}=\delta^i_k$. In the Schrödinger picture, coordinate representation, quantization is achieved by replacing q^i with the multiplicative operator $\hat{q}_i=q^i$ and p_i with the derivative operator $\hat{p}_i=-i\hbar\frac{\partial}{\partial q^i}$. Then we have

$$\widehat{p_i - e\mathcal{A}_i} = -i\hbar\left(\frac{\partial}{\partial q^i} - \frac{ie}{\hbar}\mathcal{A}_i\right) = -i\hbar\mathcal{D}_i \ , \tag{2.8}$$

where $\mathcal{D}_i$ is the covariant derivative with respect to $\mathcal{A}_i$, acting now on wavefunctions $\psi(q)$. Under local phase transformations (2.4) we have $\mathcal{D}'_i\psi'=e^{-i\alpha}(\mathcal{D}_i\psi)$. The hamiltonian becomes the operator

$$\hat{H} = -\frac{\hbar^2}{2m}\frac{1}{\sqrt{g}}\mathcal{D}_i\sqrt{g}g^{ij}\mathcal{D}_j + V \tag{2.9}$$

where $g=\det(g_{ij})$. [3] Now let us consider the special case when the connection is flat: $\mathcal{F}=0$. There exists at least locally a function Λ such that

$$\mathcal{A}_i = \frac{\hbar}{e}\partial_i\Lambda \tag{2.10}$$

so the second term in (2.6) is a total derivative:

$$L_T = e\dot{q}^i\mathcal{A}_i(q) = \hbar\frac{d\Lambda}{dt} \ . \tag{2.11}$$

This term does not affect the equations of motion and therefore can be neglected in the classical theory. It is called a "topological term". We shall understand better the reason for this terminology when we consider concrete examples.

2.1.4 Multivalued wave functions

There is an alternative description of the θ sectors that does not rely on the existence of a topological term in the Lagrangian. To arrive at it, we observe that (2.10) can be interpreted by saying that $\mathcal{A}$ is locally a gauge transform of $\mathcal{A}=0$, with gauge function $\alpha=-\Lambda$. The corresponding function in $U(1)$ is

$$\mathcal{U}=e^{-i\Lambda(q)} \ .$$

We can view this as a unitary transformation leading to an alternative form of the quantum theory, with operators $\mathcal{O}'=\mathcal{U}\,\mathcal{O}\,\mathcal{U}^{-1}$ and states $\psi'=\mathcal{U}\psi$.

We have

$$\mathcal{U}\,\mathcal{D}_i\,\mathcal{U}^{-1} = \partial_i \ . \tag{2.12}$$

Therefore, acting on the Hamiltonian (2.9) it gives

$$\hat{H}_\theta \mapsto \mathcal{U}\hat{H}_\theta\mathcal{U}^{-1} = -\frac{\hbar^2}{2m}\frac{1}{\sqrt{g}}\partial_i\sqrt{g}g^{ij}\partial_j + V(\varphi) = \hat{H}_0 \ . \tag{2.13}$$

[3] We have chosen a certain factor ordering in the first term which makes it equal to the covariant laplacian in the metric g_{ij}. This will be of no relevance in what follows.

We see that by this transformation we can remove the dependence on θ from the Hamiltonian.

In this way, however, the dependence on θ appears in the states. In fact, let $\mathcal{H}_0$ be the Hilbert space of single-valued wave functions, that we have considered so far. If $\psi \in \mathcal{H}_0$, the transform $\psi' = \mathcal{U}\psi$ does not belong to $\mathcal{H}_0$ anymore. To see this, let us consider the case when $\pi_1(\mathcal{Q}) = \mathbb{Z}$ and let $0 \leq t < 1$ be a coordinate along the fundamental non-contractible loop generating this homotopy group. If $\Lambda(1) = \Lambda(0) + 2\pi n$, then $e^{i\Lambda}$ is a proper $U(1)$ gauge transformation, $\mathcal{A}$ is a $U(1)$ pure gauge and we are in the trivial θ-sector. Let us consider instead the general case when

$$\Lambda(1) = \Lambda(0) + \theta \ ,$$

with $\theta \neq 2\pi n$. Then

$$\psi \mapsto \psi' = \mathcal{U}\psi$$

and we have

$$\psi'(1) = \mathcal{U}(1)\psi(1) = e^{i\theta}\psi'(0) \ . \tag{2.14}$$

We see that the transformed wave function is only periodic up to a phase. Wave functions satisfying these conditions form a Hilbert space $\mathcal{H}_\theta$, and clearly $\mathcal{H}_{\theta+2\pi} = \mathcal{H}_\theta$, so the set of different Hilbert spaces is parametrized by $0 \leq \theta < 2\pi$. Thus, in this alternative description, the information about the θ angle is contained in the wave functions rather than the Hamiltonian.

We thus see that the theta sectors always admit two descriptions: either with a topological term in the lagrangian and single-valued wave functions or without topological term and with multiple-valued wave functions. In the first description the θ dependence is in the Hamiltonian, in the second in the states, so the first is sometimes called the "θ-Heisenberg" picture while the second is called "θ-Schrödinger" picture. The transformation between the two descriptions has the form of a gauge transformation with multiple-valued gauge function. Thus it is not a $U(1)$ gauge transformation in the strict sense. We will stick mostly to the first description, but the second is more familiar in certain examples.

In the next four sections we shall consider increasingly complicated systems with multiply connected configuration spaces. In most cases they will have $\pi_1(\mathcal{Q}) = \mathbb{Z}$. These theories can be quantized in inequivalent ways parametrized by an angle $0 \leq \theta < 2\pi$. These inequivalent quantum theories are called "theta sectors".

2.2 Quantum mechanical examples

2.2.1 Spin and statistics

Before coming to the field theoretic examples let us see how quantum spin and statistics can be seen as manifestation of the same type of ambiguity that leads to the existence of theta sectors. In these cases the comparison with standard quantum mechanical formalism is easier if we use the "θ-Schrödinger" picture.

A classical model for a particle with spin is the rigid rotator. The configuration space of this system is $\mathcal{Q} = \mathbb{R}^d \times SO(d)$, where d is the dimension of space. The group $SO(d)$ has fundamental group $\mathbb{Z}$ for $d = 2$ and $\mathbb{Z}_2$ for $d > 2$. Thus one would expect inequivalent quantizations labelled by an angle in two dimensions and by $Hom(\mathbb{Z}_2, U(1)) = \mathbb{Z}_2$ in higher dimensions. This is indeed what happens. We have seen that the inequivalent quantizations can be described by choosing the periodicity conditions on the wave function:

$$\psi(\omega + 2\pi) = e^{i\theta}\psi(\omega) \ , \tag{2.15}$$

where ω is a parameter along the loop. In the case of the rotation group, fixing the axis of rotation, the parameter ω is the angle of rotation and the fundamental noncontractible loop consists of rotating the body by 2π. Therefore (2.15) describes the behaviour of the wave function under a 2π rotation. It can be compared with the definition of spin in quantum mechanics. The wave function of a system with spin s acquires a phase $e^{2\pi i s}$ when the system is rotated by 2π. So we learn that θ is equal to $2\pi s \bmod 2\pi$. In $d > 2$, s can be either integer or half-integer. Integer spin corresponds to $\theta = 0 \bmod 2\pi$, giving single-valued wave functions, whereas half integer spin corresponds to $\theta = \pi \bmod 2\pi$ giving wave functions that change sign under 2π rotations.

In two dimensions the spin can take any real value and the corresponding particles are called *anyons*. Their wave functions change by a phase under a 2π rotation.

For a multiparticle system, the statistical parameter σ is defined by

$$\psi(\ldots, \vec{x}_i, \ldots, \vec{x}_j, \ldots) = e^{2\pi i\sigma}\psi(\ldots, \vec{x}_j, \ldots, \vec{x}_i, \ldots) \ . \tag{2.16}$$

The usual Bose–Einstein and Fermi–Dirac statistics correspond to σ integer and half-integer respectively. To see the connection between statistics and inequivalent quantizations, consider the classical configuration space of two identical particles in d dimensions. Let us also assume that the particles cannot be at the same point in space. [4] The configuration space is then $\mathcal{Q} = (\mathbb{R}^{2d} \setminus \Delta)/S_2$, where Δ is the subset of $\mathbb{R}^{2d}$ for which the particle positions $\vec{x}_1$ and $\vec{x}_2$ coincide, and $S_2 = \mathbb{Z}_2$ is the permutation group of two objects. Passing from the coordinates $(\vec{x}_1, \vec{x}_2)$ to the center-of-mass coordinates $(x_{\rm CM}, \Delta\vec{x}) = (\frac{\vec{x}_1+\vec{x}_2}{2}, \frac{\vec{x}_2-\vec{x}_1}{2})$ shows that the topology of the space $\mathbb{R}^{2d} \setminus \Delta$ is $\mathbb{R}^d \times \mathbb{R}^+ \times S^{d-1}$ (here $\mathbb{R}^d$ is parametrized by $\vec{x}_{\rm CM}$, $\mathbb{R}^+$ is parametrized by $|\Delta\vec{x}|$ and S^{d-1} is parametrized by the angular variables of $\Delta\vec{x}$). For $d > 2$ this space is simply connected. The group S_2 acts on it by $(x_{\rm CM}, \Delta\vec{x}) \to (x_{\rm CM}, -\Delta\vec{x})$ and therefore acts antipodally on S^{d-1}; the quotient has topology $\mathbb{R}^d \times \mathbb{R}^+ \times RP^{d-1}$ where $RP^{d-1} = S^{d-1}/\mathbb{Z}_2$ is a real projective space, whose fundamental group is $\mathbb{Z}_2$. The system of two particles can therefore be quantized in two inequivalent ways, corresponding to bosonic and fermionic statistics.

[4] This is necessary for $\mathcal{Q}$ to be a smooth manifold. Furthermore, equation (2.16) is compatible with $\vec{x}_1 = \vec{x}_2$ only for integer σ, so if we allowed this case, the statistics could only be bosonic.

For $d = 2$, $\mathbb{R}^{2d} \setminus \Delta$ already has a nontrivial fundamental group $\mathbb{Z}$, and $\pi_1(\mathcal{Q}) = \mathbb{Z}$ too. In this case the inequivalent quantizations are labelled by an angle σ; one then speaks of fractional statistics. These considerations can be generalized to the case of N indistinguishable particles.

2.2.2 The pendulum

The simplest system admitting theta vacua is the pendulum. Its configuration space is $\mathcal{Q} = S^1$, and since $\pi_1(S^1) = \mathbb{Z}$, we expect to find inequivalent quantizations labelled by an angle θ. The usual lagrangian for the pendulum is, in suitable units,

$$L_0 = \frac{1}{2}\dot{\varphi}^2 - V(\varphi) \ , \tag{2.17}$$

where $0 \leq \varphi < 2\pi$ is the coordinate on S^1 and $V(\varphi) = 1 - \cos\varphi$ is the gravitational potential. The explicit form of the kinetic and potential terms will not enter in the considerations of this section, but will become relevant later. In particular, the presence of the gravitational potential will be necessary in Section 3.7 for the application of the WKB method.

In the "θ-Heisenberg" picture the Lagrangian contains in addition a total time derivative term

$$L_T = \theta \frac{\hbar}{2\pi} \frac{d\varphi}{dt} \tag{2.18}$$

where θ is an arbitrary real parameter. This does not change the equations of motion, so the classical theory is independent of the value of θ. Assuming that for $|t| \to \infty$, $\varphi(t) \to 0$, this corresponds to adding to the action the term

$$S_T(\varphi) = \theta \frac{\hbar}{2\pi} \int dt \frac{d\varphi}{dt} = \theta\hbar W(\varphi) \ ,$$

where $W(\varphi)$ is the winding number of the history $\varphi(t)$, counting the total number of times the pendulum rotates about its center in the course of the time evolution. Because of this topological significance, the term S_T is known as a "topological term".

From a physical point of view, the term L_T can be seen as the interaction of the particle (with charge $e = 1$) with the magnetic potential $\mathcal{A}_{(\theta)} = \theta \frac{\hbar}{2\pi} d\varphi$. This is the same as the Aharonov-Bohm potential (2.1). In fact, apart from giving the explicit form of L_0, we have simply restricted the motion of the particle by fixing the value of z (the axial coordinate along the solenoid) and r (the distance from the center of the solenoid). The potential is physically unimportant in the case of the actual Aharonov-Bohm experiment, but it will be important later on.

The Hamiltonian operator in this "Heisenberg" picture is

$$\hat{H}_{(H)} = -\frac{\hbar^2}{2} \mathcal{D}_\varphi \mathcal{D}_\varphi + V(\varphi) \ , \tag{2.19}$$

where $\mathcal{D}_\varphi = \frac{\partial}{\partial\varphi} - i\frac{\theta}{2\pi}$. [5] The Hilbert space is always $\mathcal{H} = L^2(S^1)$, the space of complex functions $\psi_{(H)}(\varphi)$ such that

$$\psi_{(H)}(\varphi + 2\pi) = \psi_{(H)}(\varphi) \tag{2.20}$$

and $\int_0^{2\pi} d\varphi\, \psi^*_{(H)}\psi_{(H)} < \infty$.

Alternatively, in the "θ-Schrödinger" picture there is no topological term, so the Hamiltonian is always

$$\hat{H}_{(S)} = -\frac{\hbar^2}{2}\partial^2_\varphi + V(\varphi) \ .$$

Instead, the wave functions are periodic up to a phase:

$$\psi_{(S)}(\varphi + 2\pi) = e^{i\theta}\psi_{(S)}(\varphi).$$

These wave functions form a Hilbert space $\mathcal{H}_\theta$.

The transformation from the "θ-Schrödinger" to the "θ-Heisenberg" picture is given by the unitary operator

$$\mathcal{U} = \exp\left(-i\,\frac{\theta\hbar}{2\pi}\,\hat{\varphi}\right) \ .$$

In fact

$$\mathcal{U}\,\hat{H}_{(H)}\,\mathcal{U}^{-1} = \hat{H}_{(S)} \ , \qquad \mathcal{U}\,\psi_{(H)} = \psi_{(S)} \ .$$

2.3 Spherical sigma models

Let us now consider the S^2 nonlinear sigma model in 1+1 dimensions. This is perhaps the simplest field theoretic example showing the existence of theta sectors. It is easier to discuss than gauge theories, because one can work directly with the true, unconstrained degrees of freedom of the theory and there are no complications due to gauge invariance. We work in the "intrinsic" formulation, in terms of two fields φ^α which have the meaning of coordinates on S^2. We choose a metric $h_{\alpha\beta}(\varphi)$ on S^2 and write the action as

$$S_0 = -\frac{f^2}{2}\int d^2x\, \partial_\mu\varphi^\alpha\partial^\mu\varphi^\beta h_{\alpha\beta}(\varphi)$$

The canonical configuration space of this model is $\mathcal{Q} = \Gamma_*(S^1, S^2)$, where the constant time spacelike surfaces $\mathbb{R}$ have been compactified to S^1 due to the requirement of finiteness of the energy This space is called the loop space of S^2. Its fundamental group is $\pi_1(\mathcal{Q}) = \pi_2(S^2) = \mathbb{Z}$ (see Appendix XXX). So this theory will admit theta sectors, labelled by an angle $0 \leq \theta < 2\pi$. The fundamental non-contractible loop in $\mathcal{Q}$ (i.e. the loop whose homotopy class generates $\pi_1(\mathcal{Q})$) can be described as follows. Points on $\mathcal{Q}$ are loops in S^2 beginning and ending at some basepoint y_0, i.e. maps $c : [0, 1] \to S^2$ such that $c(0) = c(1) = y_0$.

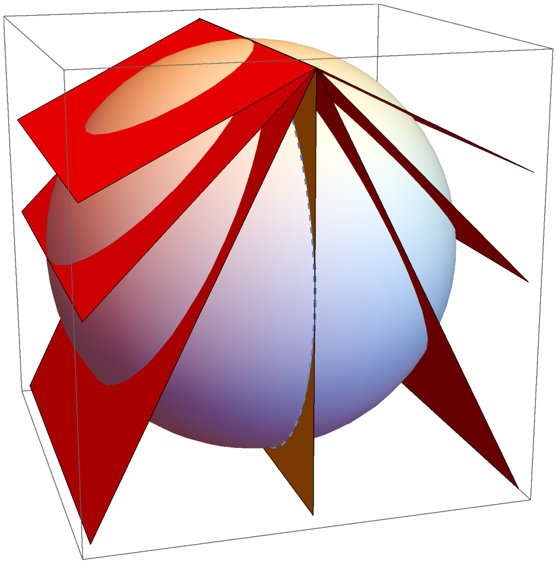

Figure 2.1: The intersections of the planes $z = ax + 1$ with the unit sphere, are the images of loops in the sphere based at the north pole. For $a = tan(\pi t)$ and $0 \leq t \leq 1$, this defines a noncontractible loop in the loop space of the sphere.

The basepoint of $\mathcal{Q}$ itself is the constant loop which maps all of $[0, 1]$ into y_0. Consider the one-parameter family of loops c_t depicted in Fig. 2.1.

When $t = 0$ we have the constant loop. For growing t, the loops sweep out the whole sphere, and for $t \to 1$ it shrinks back to the constant loop. Clearly c_t is a non-contractible loop of loops. More formally, the isomorphism between $\pi_0(\mathcal{Q})$ and $\pi_2(S^2)$ can be described as follows: if $c : I \to \mathcal{Q}$ is a loop in $\mathcal{Q}$ we define $\hat{c} : I \times I \to S^2$ by $\hat{c}(t, s) = \big(c(t)\big)(s)$, where $c(t)$, for fixed t, is regarded as a map $I \to S^2$. We have $\hat{c}(t, s) = y_0$ whenever t or s are equal to 0 or 1, so $\hat{c}$ defines a map $S^2 \to S^2$. Clearly homotopies of c correspond to homotopies of $\hat{c}$. So the desired isomorphism correspond to mapping $[c]$ to $[\hat{c}]$.

In order to make the theta sectors manifest, we add to the action a topological term $S_T = \theta W(\varphi)$, where

$$W(\varphi) = \frac{1}{4\pi} \int d^2x\, \varepsilon^{\mu\nu} \partial_\mu \varphi^\alpha \partial_\nu \varphi^\beta \frac{1}{2!} \sqrt{h}\, \varepsilon_{\alpha\beta}$$

is the winding number of the map φ (see Appendix A). The addition of W does not change the equations of motion, nor the form of the energy because it is a total derivative. In fact, we have locally $\sqrt{h}\varepsilon_{\alpha\beta} = \partial_\alpha \tau_\beta - \partial_\beta \tau_\alpha$ for some one-form τ. Then $W(\varphi) = \int d^2x\, \partial_\mu \omega^\mu$, where

$$\omega^\mu = \frac{1}{4\pi} \varepsilon^{\mu\nu} \partial_\nu \varphi^\alpha \tau_\alpha(\varphi) \ .$$

[5] Note that since the metric on S^1 is independent of φ in this case there are no ordering ambiguities.

However, the addition of the topological term affects the relation between velocities and momenta:

$$\pi_\alpha = f^2 h_{\alpha\beta}\partial_0\varphi^\beta + \mathcal{A}_\alpha \ ,$$

where

$$\mathcal{A}_\alpha(x) = \frac{\theta}{4\pi}\partial_1\varphi^\beta\sqrt{h}\,\varepsilon_{\alpha\beta} \ . \tag{2.21}$$

Comparing with equation (2.7) we see that $\mathcal{A}_\alpha$ can be regarded as a "functional magnetic potential" on $\mathcal{Q}$. In fact we can write the action $S = S_0 + S_T = \int dt(L_0 + L_T)$, with

$$L_0 = \frac{f^2}{2}h(\dot\varphi,\dot\varphi) - V(\varphi) \quad ; \quad L_T = \mathcal{A}_\varphi(\dot\varphi)$$

This is an infinite dimensional version of the form (2.6) where we replaced the index i with the infinite indexing set (α, x). The potential is $V(\varphi) = \frac{f^2}{2}g_\varphi(\partial_1\varphi,\partial_1\varphi)$, the magnetic potential is the one-form $\mathcal{A} = \int dx\mathcal{A}_\alpha(x)\delta\varphi^\alpha(x)$ and the riemannian metric is $g = \int dx h_{\alpha\beta}\big[\varphi(x)\big]\delta\varphi^\alpha(x)\delta\varphi^\beta(x)$. In these formulae $\delta\varphi^\alpha(x)$ play the role of the differentials dq^i in the finite dimensional case. This terminology is further explained in appendix E.

Since the topological term (i.e. the magnetic field) does not appear in the equation of motion, we expect that $\mathcal{F} = d\mathcal{A} = 0$. This is what one gets from a direct calculation based on formula (E.16)

$$d\mathcal{A}(v,w) = v\big(\mathcal{A}(w)\big) - w\big(\mathcal{A}(v)\big) - \mathcal{A}\big([v,w]\big) \ .$$

There follows that, at least locally on $\mathcal{Q}$,

$$\mathcal{A} = d\Lambda.$$

In fact we have

$$\Lambda = \theta\int dx\,\omega^0 = \frac{\theta}{4\pi}\int dx\,\partial_1\varphi^\alpha\tau_\alpha \ .$$

The function Λ is not single valued. The polidromy of Λ on the fundamental loop in $\mathcal{Q}$ is

$$\begin{aligned}\oint d\Lambda = \oint \mathcal{A} \quad &= \quad \int d\tau\left[\frac{\theta}{4\pi}\int dx\partial_1\varphi^\alpha\frac{d\varphi^\beta}{d\tau}\sqrt{h}\,\varepsilon_{\alpha\beta}\right] \\ &= \quad \frac{\theta}{4\pi}\int d^2x\varepsilon^{\lambda\mu}\partial_\lambda\hat\varphi^\alpha\partial_\mu\hat\varphi^\beta\frac{1}{2}\sqrt{h}\,\varepsilon_{\alpha\beta} = \theta W(\hat\varphi) = \theta \ .\end{aligned} \tag{2.22}$$

Therefore, Λ is single-valued only if $\theta = 0$. However, if $\theta = 2\pi n$, with $n \in \mathbb{Z}$, $e^{i\Lambda}$ is a single-valued function $\Gamma_*(S^1, S^2) \to U(1)$ and so the gauge potentials $\mathcal{A}_{\theta+2\pi n}$ and $\mathcal{A}_\theta$ are gauge-related in the strict sense. The gauge inequivalent magnetic potentials, and hence the inequivalent quantizations, are labelled by $0 \le \theta < 2\pi$.

The pendulum and the sigma model discussed in this section are the $d = 0$ and $d = 1$ cases of an infinite sequence of theories that behave all in the same way. The S^{d+1}-valued sigma model in d space dimensions has configuration space $\mathcal{Q} = \Gamma_*(S^d, S^{d+1})$ and $\pi_1(\mathcal{Q}) = \pi_d(S^d) = \mathbb{Z}$. The topological term is given again by the winding number.

2.4 QED in 1+1 dimensions

Next consider a $U(1)$ gauge field A_μ in one space dimension. A pure gauge theory would not have physical degrees of freedom, so in order to have a non-empty theory it is necessary to include also some matter fields, either fermionic (QED proper) of bosonic (scalar QED) or both. For the purposes of this section it does not matter what matter field one chooses, as long as it carries a linear representation of $U(1)$. The action is

$$S = S_{YM} + S_T + S_m$$

where

$$S_{YM} = -\frac{1}{4}\int d^2x\, F_{\mu\nu}F^{\mu\nu} \tag{2.23}$$

is the usual Maxwell action, S_m is the matter action and $S_T = \theta c_1$, with

$$c_1 = \frac{1}{4\pi}\int d^2x\, \varepsilon^{\mu\nu}F_{\mu\nu} \tag{2.24}$$

is a "topological term". The topological significance of this term will be understood better in section 2.8. For the time being we merely observe that $\frac{1}{4\pi}\varepsilon^{\mu\nu}F_{\mu\nu} = \partial_\mu C^\mu$, where

$$C^\mu = \frac{1}{2\pi}\varepsilon^{\mu\nu}A_\nu \tag{2.25}$$

is known as the (dual of the) one-dimensional Chern-Simons form. There follows that c_1 is invariant under infinitesimal variations of the field A_μ that vanish at infinity, and therefore does not contribute to the classical equations of motion. However, it does enter the canonical definition of momentum and hamiltonian

$$P^1(x) = \frac{\partial\mathcal{L}}{\partial\partial_0 A_1(x)} = E_1(x) + \frac{\theta}{2\pi} \tag{2.26}$$

$$H = \int dx \left[\frac{1}{2}\left(P^1 - \frac{\theta}{2\pi}\right)^2 - A_0 G\right] \tag{2.27}$$

where $E_1 = F_{01} = \partial_0 A_1 - \partial_1 A_0$. The field A_0 enters as a Lagrange multiplier enforcing the Gauss law constraint $0 = G \equiv \partial_1 E_1 - \rho$, where ρ is the charge density of matter.

Our discussion will be simplified by choosing the gauge $A_0 = 0$. This leaves a residual gauge freedom consisting of time-independent gauge transformations. With this choice of gauge $E_1 = \dot{A}_1$, so the energy of the gauge field $E = \int dx \frac{1}{2}E_1^2$ is seen to be of purely kinetic character: the static energy is zero.

The configuration space $\mathcal{Q}$ of this theory consists of gauge and matter fields modulo gauge transformations. We denote $\mathcal{C} = \{(A_1, \Phi)\}$ the space of gauge and matter fields. and $\mathcal{G} = \Gamma_*(S^1, U(1))$ the gauge group, consisting of maps $g : \mathbb{R} \to U(1)$ such that $g \to 1$ for $|x| \to \infty$ (hence the possibility of compactifying

$\mathbb{R}$ to S^1). So $\mathcal{Q}=\mathcal{C}/\mathcal{G}$. The action of $\mathcal{G}$ on $\mathcal{C}$ is free, because a gauge field A_1 that is a fixed point for a gauge transformation $g = e^{i\alpha}$ must satisfy

$$A_1 + \partial_1\alpha = A_1$$

and since $\alpha = 0$ at infinity, $\alpha = 0$ everywhere. Since the action is free, $\mathcal{Q}$ is an infinite dimensional manifold and $\mathcal{C}$ is a principal bundle over $\mathcal{Q}$ with fibers $\mathcal{G}$. Since the topological term depends only on the gauge field, the matter fields do not play a role in what follows, so they will not be indicated explicitly, but one should bear in mind that when we talk of a connection A_1 we really mean a pair of a connection and a matter field (A_1, ϕ).

The space $\mathcal{C}$ has trivial topology, but $\mathcal{Q}$ is multiply connected. In fact,

$$\pi_1(\mathcal{Q}) = \pi_0(\mathcal{G}) = \pi_1(S^1) = \mathbb{Z} \ .$$

The fact that $\pi_1(\mathcal{Q})$ and $\pi_0(\mathcal{G})$ are isomorphic can be proven using the homotopy exact sequence discussed in Appendix D. Here we describe the isomorphism. The gauge group $\mathcal{G}$ consists of infinitely many connected components $\mathcal{G}_n=\{g : S^1 \to U(1) \mid W(g)=n\}$. Now choose a basepoint $A_{(0)}=0\in\mathcal{C}$ (for definiteness we will take $A_{(0)} = 0$, but this is by no means necessary) and consider the orbit through $A_{(0)}$, i.e. the set of all connections of the form $A^g_{(0)} = g^{-1}dg$ for $g\in\mathcal{G}$. Since the action of $\mathcal{G}$ is free, there is a one-to-one correspondence between points of $\mathcal{G}$ and points of the orbit through $A_{(0)}$. (See Appendix C). So the topology of the orbit is the same as the topology of $\mathcal{G}$. There is a natural projection $p : \mathcal{C}\to\mathcal{Q}$ which associates to A its gauge equivalence class $[A]$. Under this projection all points in the orbit through $A_{(0)}$ are mapped to the same point $[A_{(0)}]$ in $\mathcal{Q}$. It is natural to take $A_{(0)}$ as the basepoint in $\mathcal{C}$, $[A_{(0)}]$ as a basepoint in $\mathcal{Q}$. Now consider a gauge transformation g with $W(g)=1$. There is no continuous path in $\mathcal{G}$ joining g to the identity, and therefore there is also no path in the orbit through $A_{(0)}$ joining $A^g_{(0)}=g^{-1}dg$ to $A_{(0)}$. However, the space $\mathcal{C}$ is connected and so there is some path $\tilde{\ell}_t$ in $\mathcal{C}$, with $t\in[0,1]$ such that $\tilde{\ell}_0=A_{(0)}$ and $\tilde{\ell}_1=A^g_{(0)}$. For instance one can take $c_t = t\,g^{-1}dg = t\,d\alpha$. The natural projection p maps this path in $\mathcal{C}$ to a path $\ell_t = [\tilde{\ell}_t]$ in $\mathcal{Q}$ beginning and ending at $[A_{(0)}]$. The desired isomorphism $\pi_1(\mathcal{Q})\to\pi_0(\mathcal{G})$ is obtained by mapping the homotopy class of the loop ℓ_t in $\mathcal{Q}$ to the homotopy class of g. See fig. XXX.

Returning to equations (2.26) and (2.27) we see that the topological term θc_1 in the action can be written, in the gauge $A_0 = 0$, as $\int dt \int dx \dot{A}_1 \frac{\theta}{2\pi}$ and hence can be regarded as the interaction of a particle with unit charge and coordinate $A_1(x)$ with a magnetic potential (a one-form on $\mathcal{C}$)

$$\tilde{\mathcal{A}} = \int dx\, \frac{\theta}{2\pi} \delta A_1(x) \ .$$

Since the components of the vector potential are constant, it is easy to verify that the corresponding magnetic field $\tilde{\mathcal{F}}=d\tilde{\mathcal{A}}=0$. This is in accordance with the fact that the topological term does not contribute to the equation of motion: if it did, one could interpret the corresponding term in the equation of motion as a Lorentz force due to a nonzero $\tilde{\mathcal{F}}$. Since $d\tilde{\mathcal{A}}=0$, we can write at least locally $\tilde{\mathcal{A}}=d\tilde{\Lambda}$. The functional $\tilde{\Lambda}$ on $\mathcal{C}$ that has this property is

$$\tilde{\Lambda}=\frac{\theta}{2\pi}\int dx A_1(x) \ .$$

All this is on the contractible space $\mathcal{C}$.

We would like now to see the corresponding steps being carried out on $\mathcal{Q}$. It is convenient to write a time-independent gauge transformation in the form $g(x)=e^{i\alpha(x)}$, where $\alpha\rightarrow 2\pi n_-$, for $x\rightarrow-\infty$ and $\alpha\rightarrow 2\pi n_+$, for $x\rightarrow\infty$. The winding number of g is just n_+-n_-. Infinitesimal gauge transformations are real-valued functions $\epsilon(x)$ such that $\epsilon\rightarrow 0$ for $|x|\rightarrow\infty$.

We now consider again the function $\tilde{\Lambda}$ and ask whether it is the pullback of a function on $\mathcal{Q}$. This will be the case provided $\tilde{\Lambda}$ is constant on the orbits, *i.e.* if it is gauge invariant. Under a gauge transformation g,

$$\tilde{\Lambda}(A^g)-\tilde{\Lambda}(A)=\frac{\theta}{2\pi i}\int dx\, g^{-1}dg=\frac{\theta}{2\pi}\int dx\,\frac{d\alpha}{dx}=\theta W(g) \ . \qquad (2.28)$$

Therefore, $\tilde{\Lambda}$ is invariant under gauge transformations which are connected to the identity, but not under "large" gauge transformations, *i.e.* transformations that have winding number different from zero. Under these circumstances, $\tilde{\Lambda}$ does not define a function Λ on $\mathcal{C}/\mathcal{G}$, but only a function which is defined up to integer multiples of θ.

Similarly, we can ask if $\tilde{\mathcal{A}}=p^*\mathcal{A}$ for some one-form $\mathcal{A}$ on $\mathcal{C}/\mathcal{G}$. This is true provided: [6]

1) $\tilde{\mathcal{A}}$ is gauge invariant;

2) $\tilde{\mathcal{A}}(v)=0$ when v is a vertical vector (i.e. v is tangent to the orbit).

The first condition is obviously satisfied, and for the second we observe that a vertical vector has the form $v_\epsilon=\int dx\,\partial_1\epsilon\frac{\delta}{\delta A_1}$, where ϵ is an infinitesimal gauge parameter; then

$$\tilde{\mathcal{A}}(v_\epsilon)=\frac{\theta}{2\pi}\int dx\partial_1\epsilon=\frac{\theta}{2\pi}\big(\epsilon(\infty)-\epsilon(-\infty)\big)=0 \ .$$

So there is a one-form $\mathcal{A}$ on $\mathcal{Q}$ such that $\tilde{\mathcal{A}}=p^*\mathcal{A}$. Since p is surjective, $\mathcal{A}$ is entirely determined by $\tilde{\mathcal{A}}$, and since $p^*d=dp^*$, $d\mathcal{A}=0$ and, locally, $\mathcal{A}=d\Lambda$.

According to the general discussion in section 3.1, inequivalent quantizations correspond to the gauge inequivalent magnetic potentials $\mathcal{A}$. The magnetic potential $\mathcal{A}(\theta)$ will be gauge equivalent to $\mathcal{A}(\theta=0)$ if the function $e^{i\Lambda}$ is single-valued, i.e. if the polydromy of Λ is an integral multiple of 2π. From the construction of the fundamental loop ℓ in $\mathcal{Q}$ we see that the polydromy of Λ on ℓ is equal to $\oint_\ell\mathcal{A}=\int_{\tilde{\ell}}\tilde{\mathcal{A}}$, where $\tilde{\ell}$ is a lift of ℓ, i.e. a path joining $A_{(0)}$ to $A^g_{(0)}$, with $W(g)=1$. But then $\int_{\tilde{\ell}}\tilde{\mathcal{A}}=\tilde{\Lambda}(A^g)-\tilde{\Lambda}(A)=\theta$, by equation (2.28). So, whenever $\theta=2\pi n$, $\mathcal{A}(\theta)$ is a pure gauge. The classes of gauge inequivalent $\mathcal{A}$'s are parameterized again by $0\leq\theta<2\pi$.

[6] S.Kobayashi and K.Nomizu "Foundations of differential geometry", vol. II, p. 294, lemma 1

2.5 Nonabelian Yang–Mills theory in 3+1 dimensions

Except for algebraic complications, the discussion of a nonabelian Yang–Mills theory in 3+1 dimensions follows step by step that of the abelian theory in 1+1 dimensions. It is convenient to use the rescaled, geometrical gauge fields, so that the curvature is given by (1.90) and the gauge transformations act as in (1.91). The total action is $S = S_{YM} + S_T$ where S_{YM} is given by (1.89) (with $d{=}3$) and $S_T = \theta c_2$, where

$$c_2 = \frac{1}{64\pi^2}\int d^4x\, \varepsilon^{\mu\nu\rho\sigma} F^a_{\mu\nu} F^a_{\rho\sigma}$$

is a topological term, known as the second Chern class. This term does not modify the classical equations of motion since

$$\frac{1}{64\pi^2}\varepsilon^{\mu\nu\rho\sigma} F^a_{\mu\nu} F^a_{\rho\sigma} = \partial_\mu C^\mu \ , \tag{2.29}$$

where

$$C^\mu = \frac{1}{16\pi^2}\varepsilon^{\mu\nu\rho\sigma}\left(A^a_\nu \partial_\rho A^a_\sigma + \frac{1}{3} f_{abc} A^a_\nu A^b_\rho A^c_\sigma\right) \tag{2.30}$$

is known as the (dual of the) three dimensional Chern-Simons form. Thus c_2 is invariant under infinitesimal variations of A^a_μ. However, it changes the relation between velocities and momenta. We have

$$P^0_a = \frac{\partial L}{\partial \partial_0 A^a_0} = 0 \tag{2.31}$$

$$P_a = \frac{\partial L}{\partial \partial_0 A^a_i} = \frac{1}{e^2} E^a_i + \frac{\theta}{8\pi^2} B^a_i \tag{2.32}$$

where $E^a_i = F^a_{0i} = \partial_0 A^a_i - D_i A^a_0$ and $B^a_i = \frac{1}{2}\varepsilon_{ijk} F^a_{jk}$. The hamiltonian is

$$H = \int d^3x \left[\frac{e^2}{2}\left(P^a_i - \theta \frac{e^2}{8\pi^2} B^a_i\right)^2 + \frac{1}{2e^2}\left(B^a_i\right)^2 - A^a_0 D_i E^i_a\right] \ . \tag{2.33}$$

We now choose the gauge $A_0 = 0$. In this case the last term in H drops out, while the first and the second are recognized as kinetic and static energy respectively (in this gauge $E^a_i {=} \partial_0 A^a_i$). Let $\mathcal{C}$ be the space of all gauge potentials A^a_i with finite static energy, i.e. such that $\int d^3x (B^a_i)^2$ is finite. Let $\mathcal{G}$ be the residual gauge group, consisting of time-independent gauge transformations such that $g(x) {\to} \mathbf{1}$ for $|\vec{x}| {\to} \infty$. With these boundary conditions, $\mathbb{R}^3$ can be compactified to S^3 and $\mathcal{G} = \Gamma_*(S^3, G)$. As in the previous section, $\mathcal{G}$ acts freely on $\mathcal{C}$. To see this note that if A is a fixed point for a gauge transformation g, we have $g^{-1}Ag + g^{-1}dg = A$. Thus g satisfies the equation $dg + [A, g] = 0$, which means that g is covariantly constant. If g is covariantly constant, its value at any point can be obtained from its value at another point by parallel transport. Since

$g(\infty)=1$ this implies $g=1$ everywhere. Thus, the physical configuration space of the theory is the orbit space $\mathcal{Q}=\mathcal{C}/\mathcal{G}$, and the projection $p:\mathcal{C}\to\mathcal{Q}$ is a smooth infinite dimensional bundle. [7]

Since $\mathcal{C}$ is topologically trivial we have, following again the arguments of Appendix D, $\pi_1(\mathcal{Q})=\pi_0(\mathcal{G})=\pi_3(G)=\mathbb{Z}$. The isomorphism between $\pi_1(\mathcal{Q})$ and $\pi_0(\mathcal{G})$ is described again by fig. 10. The homotopy class $[g]$ of a gauge transformation corresponds to the homotopy class of the loop ℓ which is obtained by projecting to $\mathcal{Q}$ a curve $\tilde{\ell}$ joining $A=0$ to $A^g=g^{-1}dg$. Comparing equations (2.32) and (2.33) with (2.9) and (2.8) we see that the topological term has given rise to a magnetic potential $\tilde{\mathcal{A}}$ on $\mathcal{C}$ defined by

$$\tilde{\mathcal{A}}(A)=\frac{\theta}{8\pi^2}\int d^3x\,B_i^a\delta A_i^a\ .$$

A direct calculation shows that $d\tilde{\mathcal{A}}=0$. In fact, we have $\tilde{\mathcal{A}}=d\tilde{\Lambda}$, with

$$\tilde{\Lambda}=\theta\int d^3x\,C^0=\frac{\theta}{16\pi^2}\int d^3x\,\varepsilon^{ijk}\left(A_i^a\partial_jA_k^a+\frac{1}{3}f_{abc}A_i^aA_j^bA_k^c\right)\ . \tag{2.34}$$

See Exercise 2.5.1. As in the previous section, one would like to describe the theory as a particle moving in $\mathcal{Q}$, rather than $\mathcal{C}$, so the question arises again whether the function $\tilde{\Lambda}$ and the form $\tilde{\mathcal{A}}$ can be projected onto a function Λ and a form $\mathcal{A}$ on $\mathcal{Q}$. Under a gauge transformation g, one finds

$$\tilde{\Lambda}(A^g)-\tilde{\Lambda}(A)=\theta W(g)\ . \tag{2.35}$$

So $\tilde{\Lambda}$ is invariant under gauge transformations connected to the identity, but not under "large" transformations: it projects to a function Λ on $\mathcal{Q}$ which is only defined modulo integral multiples of θ.

To see if $\tilde{\mathcal{A}}$ projects, we have to verify whether the conditions given in the preceding section are satisfied. Given an infinitesimal gauge transformation parameter ϵ, a map from $\mathbb{R}^3$ to the Lie algebra of $SU(2)$ which goes to zero at infinity, we construct the corresponding vertical vectorfield in $\mathcal{C}$

$$v_\epsilon=\int d^3x\,D_i\epsilon^a\frac{\delta}{\delta A_i^a}\ .$$

Then we have:

1) $\tilde{\mathcal{A}}$ is gauge invariant (B_i^a and δA_i^a both transform homogeneously);

2) $\tilde{\mathcal{A}}(v_\epsilon)=\frac{\theta}{8\pi^2}\int d^3xB_i^aD_i\epsilon^a=0$ upon integrating by parts, using Bianchi's identity and the fact that $\epsilon\to0$ for $|\vec{x}|\to\infty$.

So $\tilde{\mathcal{A}}$ satisfy the two conditions which are needed for it to be the pullback of a one-form $\mathcal{A}$ on $\mathcal{Q}$. The relation between $\mathcal{A}$ and Λ is again, locally, $\mathcal{A}=d\Lambda=\frac{1}{i}e^{-i\Lambda}de^{i\Lambda}$. The polydromy of Λ on the loop ℓ which generates $\pi_1(\mathcal{Q})$ is

[7] P.K. Mitter and C.M. Viallet, Comm. Math. Phys. **79** 457-472 (1981).

$\oint_\ell \mathcal{A} = \int_{\tilde\ell} \tilde{\mathcal{A}} = \tilde\Lambda(A^g) - \tilde\Lambda(A) = \theta$. So we come again to the conclusion that there is a $U(1)$'s worth of quantum Yang-Mills theories, parameterized by the angle $0 \le \theta < 2\pi$.

Before closing this section we note for future reference the following interpretation of the Gauss law of the theory. Let $G_\epsilon = \int d^3x\, \epsilon^a G_a$. If the theory is quantized before eliminating all unphysical degrees of freedom, the wave functions are complex functionals on $\mathcal{C}$ and Gauss' law has to be imposed as a constraint on the physical states: $G_\epsilon \psi_{\rm phys} = 0$ for all ϵ. Upon using the quantization rule $P^a_i = -i\frac{\delta}{\delta A^a_i}$, we find

$$\begin{aligned} G_\epsilon \psi &= e^2 \int d^3x\, \epsilon^a D_i \left(P^a_i - \frac{\theta}{8\pi^2} B^a_i \right) \psi \\ &= ie^2 \int d^3x\, D_i \epsilon^a \left(\frac{\delta\psi}{\delta A^a_i} - i\frac{\theta}{8\pi^2} B^a_i \psi \right) \\ &= ie^2 \left(v_\epsilon \psi + i\tilde{\mathcal{A}}(v_\epsilon)\psi \right) = ie^2 v_\epsilon \psi \ . \end{aligned} \tag{2.36}$$

Therefore, Gauss' law states that the physical wave functions are precisely those wave functions that are locally constant along the gauge orbits. Since the orbits are not connected, they need not be globally constant, as the preceding discussion shows.

2.6 Instantons

In the preceding sections we have established that certain quantum field theories have a multiply connected configuration space and that this gives rise to a superselection rule in the quantum theory. The discussion has been essentially kinematical and very abstract. Non-contractible paths in configuration space have been shown to exists, but no explicit formulas were given. We would like now to make these notions more concrete. For example, we ask whether there are solutions of the field equations that correspond to non-contractible paths in configuration space. It turns out that there are no such solutions in real Minkowski space, but they do exist in Euclidean signature.

Generically, we call *instanton* a classical solution of nonlinear Euclidean field equations which (1) is nonsingular, (2) has finite Euclidean action and (3) is localized in Euclidean spacetime.

It is not an accident that this definition closely resembles the definition of soliton given in the beginning of chapter 1. Indeed, in all the cases that we shall consider, *the Euclidean action of a theory, regarded as a functional of fields defined on d-dimensional Euclidean space, is identical to the static energy of the same theory in one more dimension, regarded as a functional of fields defined on a constant-time slice of $d+1$-dimensional Minkowski space.* Therefore, the soliton solutions discussed in chapter 1, can be recycled as instantons for the same theories in one less dimension. In the following we shall discuss several examples in detail.

As time evolves, the system traces out a path in configuration space. We will sometimes refer to such paths also as "histories". In chapter 1 we have found it useful to to visualize alternatively a field configuration as a function on space or as a point in the infinite-dimensional configuration space $\mathcal{Q}$. It will now be useful to visualize field histories alternatively as functions on spacetime or as paths in $\mathcal{Q}$. In particular, the histories beginning and ending at the vacuum are loops in configuration space. A third point of view is to think of these fields as points in the loop space of $\mathcal{Q}$.

As with solitons, not all instantons have a topological meaning. Restricting ourselves to histories with finite action. we must demand that the system be in the vacuum in the far past, in the far future and at spacial infinity. Some of these histories can be continuously deformed into the vacuum. We are mostly interested in those that cannot. Let us now consider some examples.

2.6.1 The instanton of the pendulum and of the sigma model

The pendulum can be viewed as a field theory in zero space dimensions. The analog of the pendulum in one higher dimension is a scalar theory in 1+1 dimensions with a potential of the form $1-\cos\phi$. We have encountered this theory in section 1.1.1: it was called the Sine-Gordon model. The Euclidean action of the pendulum is

$$S_E(\varphi) = \int dt \left[\frac{1}{2}\left(\frac{d\varphi}{dt}\right)^2 + \beta(1-\cos\varphi)\right] \tag{2.37}$$

Apart from trivial changes of name of the variables, this is the same functional as the static energy (1.4) with potential (1.11). The soliton of the Sine-Gordon model, given in (1.12), is the instanton for the pendulum:

$$\varphi(t) = \pm 4 \arctan\left\{\exp\left[\sqrt{\beta}(t-t_0)\right]\right\} \tag{2.38}$$

This solution of the field equations describes a history of the system that starts in the vacuum $\varphi = 0$ in the far past, swings once around its center and settles again in the vacuum in the far future. If we compactify Euclidean time to S^1, as is allowed by the boundary conditions, it has winding number one.

Next consider the S^2 nonlinear sigma model in 1+1 dimension. Its Euclidean action is exactly the static energy of the S^2 nonlinear sigma model in 2+1 dimension, equation (1.54). The instantons of this model in 1+1 Euclidean dimensions are the functions given in (1.65). Viewed as spacetime fields, the solutions with $n=1$ start out in the past in the vacuum, then sweep once the target space and finally settle again in the vacuum. Compactifying spacetime to S^2, they have winding number one. Viewed as paths in configuration space, they trace out the noncontractible path in $\mathcal{Q}$.

2.6.2 The instanton of scalar QED

The case of scalar QED requires a little more discussion. Its Euclidean action is

$$S_{0E} = \int dx d\tau \left[\frac{1}{4}F_{\mu\nu}F_{\mu\nu} + \frac{1}{2}\left(D_\mu\phi\right)^*\left(D_\mu\phi\right) + \frac{\lambda}{4}\left(|\phi|^2 - f^2\right)^2\right] \tag{2.39}$$

and coincides with the static energy of the same theory in 2+1 dimensions, in the gauge $A_0 = 0$, given in (1.99).

The one-instanton of this theory is going to be a solution of the euclidean field equations describing the tunnelling of the system through the fundamental noncontractible loop in $\mathcal{Q}$. It follows from the discussion of section 3.4 that this loop in $\mathcal{Q}$ is the projection of a path in $\mathcal{C}$ joining the classical vacuum $(A_{(0)}, \phi_{(0)}) = (0, f)$ to $(A^{g_1}_{(0)}, \phi^{g_1}_{(0)}) = (\frac{i}{e}g_1^{-1}dg_1, g_1^{-1}f)$, where $g_1 = e^{i\alpha}$ is a time-independent gauge transformation with winding number one, i.e. $\alpha(x \to -\infty) = 0$, $\alpha(x \to +\infty) = 2\pi$. We have found in section 2.7 a stationary point of this functional with the boundary condition that when $r = \sqrt{x^2+\tau^2} \to \infty$, $A_i \to \frac{i}{e}g_1^{-1}dg_1$ and $\phi \to fg_1$, where $g_1(\theta)$ is a map from S^1_∞ to $U(1)$ with winding number one: it was called the vortex. These are exactly the boundary conditions that we need (fig. 13). The explicit form of the solution was given in (1.107); it can be rewritten in the gauge $A_0 = 0$ (in the sense of the 1+1-dimensional theory; remember that one of the spatial coordinates of section 1.7 should be reinterpreted as Euclidean time). See exercise XXXXX. Therefore, the vortex solution of the Abelian Higgs model in 2+1 dimensions with unit flux is the desired instanton solution of the Abelian Higgs model in 1+1 dimensions.

We can now understand better in what sense the integral c_1 in equation (2.24) is a "topological number". We restrict our attention to spacetime fields with finite euclidean action. This demands that when $r = \sqrt{x^2+\tau^2} \to \infty$, $A_i \to \frac{i}{e}g_\infty^{-1}dg_\infty$ and $\phi \to g_\infty^{-1}f$, where $g_\infty(\theta)$ is a map from S^1_∞ to $U(1)$. Such maps are classified by their winding number, so the fields with finite action fall into disjoint classes, characterized by different asymptotic behaviour. These classes are usually called the topological sectors. We can now evaluate the quantity c_1 on such a field. Using (2.25) and the asymptotic form of A we get

$$c_1 = \frac{1}{4\pi}\int d^2x\, \varepsilon^{\mu\nu}F_{\mu\nu} = \frac{1}{2\pi}\int_{S^1_\infty} A = W(g_\infty)\ .$$

Thus c_1 is an integer measuring the nontriviality of the asymptotic behaviour of the gauge field.

2.6.3 The BPST instanton

In our search for solitons in chapter 1 we left out the case of pure Yang-Mills theory, because we showed that the Yang-Mills action only has static solitons in 4+1 dimensions. It is now time to describe this case in detail, because these solutions have the interpretation of instantons of 4-dimensional Yang-Mills theory.

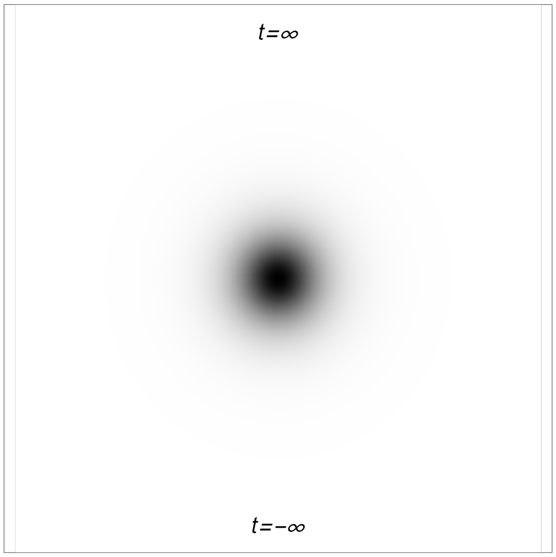

Figure 2.2: Density of action of the instanton. The field becomes pure gauge far away from the central core.

We begin by giving a topological classification of four dimensional Yang-Mills fields. From the fact that the time evolution traces a continuous curve in $\mathcal{Q}$ and from the multiple connectedness of $\mathcal{Q}$, there follows that four-dimensional Yang-Mills fields must fall into disjoint classes labelled by the integers. These classes can be described more explicitely as follows. We impose that $A^a_\mu(\vec{x},\tau)$ has finite Euclidean action. This requires that at spacetime infinity, i.e. for $|x|=\sqrt{|\vec{x}|^2+\tau^2}\to\infty$, $F^a_{\mu\nu}\to 0$. This in turn implies

$$A_\mu \to g_\infty^{-1}\partial_\mu g_\infty \ , \tag{2.40}$$

where g_∞ is a function of the angles or equivalently a function from the sphere at infinity S^3_∞ to the gauge group $SU(2)$. Since $\pi_3\big(SU(2)\big)=\mathbb{Z}$, we find that the finite action gauge potentials A^a_μ fall into topologically distinct classes distinguished by their asymptotic behaviour. The topological invariant c_2 precisely measures these classes. In fact using (2.29) we can write

$$c_2 = \int\limits_{\mathbb{R}^4} d^4x\,\partial_\mu C^\mu = \frac{1}{16\pi^2}\int_{S^3_\infty} d^3x\,\varepsilon^{ijk}\left(A^a_i\partial_j A^a_k + \frac{1}{3}f_{abc}A^a_iA^b_jA^c_k\right) = W(g_\infty)\ .$$

The last equality is obtained by noting that on S^3_∞ we can replace A^a_i by its asymptotic form (2.40); the result then follows from Exercise 2.5.2. This calculation shows that c_2 is an integer for YM fields with finite action. [8]

[8] A mathematically more sophisticated understanding of the topology of Yang–Mills fields requires the language of fiber bundles and characteristic classes.

The instanton has to represent the motion of the system through the fundamental noncontractible loop in $\mathcal{Q}$. We recall from section 2.5 that in terms of the spatial gauge potential A_i this means a path joining, for example $A_i = 0$ to $A_i = g^{-1}\partial_i g$, where g is a time-independent gauge transformation with winding number one. We take the parameter on this path to be euclidean time τ. In this way the instanton will be represented by a gauge potential $A_\mu(x, \tau)$ with $A_\tau = 0$ and A_i a pure gauge both at spatial and temporal infinity. In fact the gauge function g_∞ for this field will be one everywhere except for $\tau \to \infty$, where it coincides with g_1. Such a field will therefore have $c_2 = 1$.

To find the explicit form of the instanton we consider the inequality [9]

$$\begin{aligned} 0 &\leq \int d^4x\,\left(F^a_{\mu\nu} \pm {}^*F^a_{\mu\nu}\right)\left(F^{\mu\nu a} \pm {}^*F^{\mu\nu a}\right) \\ &= 2\int d^4x\,F^a_{\mu\nu}F^{\mu\nu a} \pm 2\int d^4x\,F^a_{\mu\nu}\,{}^*F^{\mu\nu a}\ , \end{aligned}$$

which implies

$$S_E \geq \frac{8\pi^2}{e^2}|c_2|\ . \tag{2.41}$$

The absolute minima of the action in each sector are the gauge fields for which F is either self-dual or anti-self-dual

$$F^a_{\mu\nu} = \pm^* F^a_{\mu\nu}\ . \tag{2.42}$$

These fields are automatically solutions of the Yang-Mills equations. So we have succeeded in replacing the second order Yang-Mills equation by the simpler first order equations (2.42).

For the instanton with $c_2 = 1$, motivated by (2.40), we make an ansatz of the form

$$A_\mu(x) = f(r^2)g_1^{-1}\partial_\mu g_1 \tag{2.43}$$

Here g_1, a function of the angles only, has the following explicit representation:

$$g_1(x) = \begin{bmatrix} \hat{x}^4 + i\hat{x}^3 & \hat{x}^2 + i\hat{x}^1 \\ -\hat{x}^2 + i\hat{x}^1 & \hat{x}_4 - i\hat{x}^3 \end{bmatrix} = \hat{x}^\mu\tau_\mu\ , \tag{2.44}$$

where $\tau_k = i\sigma_k$ for $k = 1, 2, 3$ and $\tau_4 = \mathbf{1}$. This function clearly has $W(g_1) = 1$. From here one finds

$$g_1^{-1}\partial_\mu g_1 = -2i\bar{\Sigma}_{\mu\rho}\frac{\hat{x}^\rho}{r}\ ; \qquad g_1\partial_\mu g_1^{-1} = -\partial_\mu g_1 g_1^{-1} = -2i\Sigma_{\mu\rho}\frac{\hat{x}^\rho}{r}\ , \tag{2.45}$$

where

$$\Sigma_{\mu\nu} = \frac{1}{2}\begin{bmatrix} 0 & \sigma_3 & -\sigma_2 & \sigma_1 \\ -\sigma_3 & 0 & \sigma_1 & \sigma_2 \\ \sigma_2 & -\sigma_1 & 0 & \sigma_3 \\ -\sigma_1 & -\sigma_2 & -\sigma_3 & 0 \end{bmatrix}\ ; \qquad \bar{\Sigma}_{\mu\nu} = \frac{1}{2}\begin{bmatrix} 0 & \sigma_3 & -\sigma_2 & -\sigma_1 \\ -\sigma_3 & 0 & \sigma_1 & -\sigma_2 \\ \sigma_2 & -\sigma_1 & 0 & -\sigma_3 \\ \sigma_1 & \sigma_2 & \sigma_3 & 0 \end{bmatrix}\ .$$

[9] A.A. Belavin, A.M. Polyakov, A.S. Schwartz, Yu.S. Tyupkin, "Pseudoparticle Solutions of the Yang-Mills Equations", Phys. Lett.B 59 (1975) 85-87.

These matrix-valued tensors are self-dual and anti-self-dual respectively.

The function f in (2.43) must satisfy $f(r^2)\to 1$ for $r^2\to\infty$ and $f(0)=0$ to avoid singularities in A (the form $g^{-1}dg$ is ill-defined in the origin). In order to determine the function f we compute the curvature of (2.43):

$$F_{\mu\nu} = 4i\left(\bar{\Sigma}_{\mu\rho}\hat{x}^\rho\hat{x}^\nu - \bar{\Sigma}_{\nu\rho}\hat{x}^\rho\hat{x}^\mu\right)\left(f' + \frac{1}{r^2}f(f-1)\right) - 4i\bar{\Sigma}_{\mu\nu}\frac{1}{r^2}f(f-1)\ .$$

Here f' denotes the derivative of f with respect to r^2. In order to compute the dual we use

$$\varepsilon_{\mu\nu\alpha\beta}\bar{\Sigma}_{\rho\beta} = -\delta_{\mu\rho}\bar{\Sigma}_{\nu\alpha} + \delta_{\nu\rho}\bar{\Sigma}_{\mu\alpha} - \delta_{\alpha\rho}\bar{\Sigma}_{\mu\nu}$$

and find

$$^*F_{\mu\nu} = 4i\left(\bar{\Sigma}_{\mu\rho}\hat{x}^\rho\hat{x}^\nu - \bar{\Sigma}_{\nu\rho}\hat{x}^\rho\hat{x}^\mu\right)\left(f' + \frac{1}{r^2}f(f-1)\right) - 4i\bar{\Sigma}_{\mu\nu}f'\ .$$

The anti-self-duality equation $0 = F_{\mu\nu} +^* F_{\mu\nu}$ implies

$$f' + \frac{1}{r^2}f(f-1) = 0\ , \tag{2.46}$$

which is solved by

$$f(r^2) = \frac{r^2}{\lambda^2 + r^2} \tag{2.47}$$

where λ is an arbitrary constant. This function has indeed the desired behavior in the origin and at infinity.

For a solution with $c_2 = -1$ (an anti-instanton) it is enough to replace g_1 by its inverse:

$$g_{-1} = g_1^{-1} = g_1^\dagger = \hat{x}^\mu\bar{\tau}_\mu\ ,$$

where $\bar{\tau}_k = -i\sigma_k$ for $k=1,2,3$ and $\tau_4 = \mathbf{1}$. Following the preceding steps and solving the self-duality equation $0 = F_{\mu\nu} -^* F_{\mu\nu}$ leads to the same function f. Altogether the regular (antiself-dual) instanton and (self-dual) anti-instanton solutions can be written in the form

$$A_\mu = -2i\frac{\bar{\Sigma}_{\mu\nu}x^\nu}{\lambda^2 + r^2}\ ; \qquad A_\mu = -2i\frac{\Sigma_{\mu\nu}x^\nu}{\lambda^2 + r^2}\ . \tag{2.48}$$

The respective field strengths are

$$F_{\mu\nu} = -4i\frac{\bar{\Sigma}_{\mu\nu}\lambda^2}{(\lambda^2 + r^2)^2}\ ; \qquad F_{\mu\nu} = -4i\frac{\Sigma_{\mu\nu}\lambda^2}{(\lambda^2 + r^2)^2}\ . \tag{2.49}$$

What happens if we try to impose self-duality on the configuration $\hat{x}^\mu\tau_\mu$ or antiself-duality on the configuration $\hat{x}^\mu\bar{\tau}_\mu$? In this case we arrive at the equation

$$f' - \frac{1}{r^2}f(f-1) = 0\ , \tag{2.50}$$

which is solved by

$$f(r^2) = \frac{\lambda^2}{\lambda^2 + r^2} \ . \tag{2.51}$$

This solution does not satisfy the desired conditions in the origin and infinity. Nevertheless, we can write the corresponding self-dual and anti self-dual gauge fields:

$$A_\mu = -2i\lambda^2 \frac{\Sigma_{\mu\nu} x^\nu}{r^2(\lambda^2 + r^2)} \ ; \qquad A_\mu = -2i\lambda^2 \frac{\bar{\Sigma}_{\mu\nu} x^\nu}{r^2(\lambda^2 + r^2)} \ . \tag{2.52}$$

These fields are singular in the origin. However, they are mere gauge transforms of the regular instanton and anti-instanton with a gauge transformation that is singular in the origin. In fact, let A be the antiself-dual instanton based on the ansatz (2.43) and consider the gauge transformation

$$\begin{aligned} A'_\mu &= g_1 A_\mu g_1^{-1} + g_1 \partial_\mu g_1^{-1} && (2.53)\\ &= (f-1)\partial_\mu g_1 g_1^{-1} && (2.54)\\ &= -2i\Sigma_{\mu\rho} \frac{x^\rho}{r^2} \frac{\lambda^2}{\lambda^2 + r^2} \ , && (2.55) \end{aligned}$$

which coincides with the first field in (2.52). We observe that using (2.45) the same field can also be written

$$\frac{\lambda^2}{\lambda^2 + r^2} (g_{-1})^{-1} \partial_\mu (g_{-1}) \ ,$$

so it becomes "pure gauge" at the origin, but with a gauge function that is the inverse of the one that describes its behavior at infinity in the regular gauge.

As usual these instantons and anti-instantons are not isolated solutions but come in families parametrized by collective coordinates, or moduli. In order to discover these moduli we have to act on a solution with all the global symmetries of the theory and find when this gives rise to physically distinct solutions. The YM action is invariant under global $SU(2)$ gauge transformations and under the 15-dimensional conformal group, that consists of Poincaré transformations (10 parameters) the so-called special conformal transformations (4 parameters) and dilatations (1 parameter).

The free parameter λ of the solution is clearly a modulus associated to the latter transformations. It can be shown that the special conformal transformations lead to gauge fields that are gauge equivalent to the original ones. Translations also generate four moduli: one just has to replace x^μ by $x^\mu - x_0^\mu$ in the solutions. There remains gauge transformations (3 parameters) and Euclidean rotations (6 parameters). We recall that the group $SO(4)$ is locally isomorphic to $SU(2)_L \times SU(2)_R$. The correspondence is as follows: if a rotation transforms x to x', then in (2.44)

$$g(x') = \hat{x}'^\mu \tau_\mu = g_L g(x) g_R^{-1} \ .$$

Thus we see that the rotation transforms the gauge field (2.43) to

$$A'_\mu = g_R A_\mu g_R^{-1} \ .$$

The ansatz (2.43) is invariant under $SU(2)_L$ and is also invariant under the simultaneous action of $SU(2)_R$ and of the global gauge group $SU(2)$, with the same transformation parameters. Thus, of these nine transformation parameters, only three give rise to moduli, for example the global gauge transformations taken by themselves. Altogether the moduli space of the simple instantons and anti-instantons are eigth-dimensional.

Much work has been done to find all self-dual and anti-self-dual solutions with $|c_2|>1$. This line of research has led to important developments in mathematics, such as Donaldson theory. While mathematically of the greatest interest, this work has not had much impact in the present context since these exact solutions would give a negligible contribution to the path integral compared to the approximate multi-instanton solutions which we use in the dilute instanton gas approximation.

2.7 Path integrals

2.7.1 Path integrals on multiply connected spaces

Recall that for a system with configuration space $\mathcal{Q}$ and action $S_0(q)$, the transition amplitude to go from position q_1 at the time t_1 to position q_2 at the time t_2 can be written as

$$K(q_2,t_2\,|\,q_1,t_1)=\int_{q_1,t_1}^{q_2,t_2}(dq)e^{\frac{i}{\hbar}S_0(q)}\ , \tag{2.56}$$

where the integral is performed over all paths joining q_1 to q_2. We assume that the action has the form

$$S_0=\int dt\,\left[\frac{1}{2}mg_{ij}(q)\dot{q}^i\dot{q}^j-V(q)\right]\ . \tag{2.57}$$

We want to discuss the effects that arise when $\mathcal{Q}$ is multiply connected. We observe that the paths from q_1 to q_2 fall into homotopy classes. Clearly there are as many homotopy classes of paths from q_1 to q_2 as there are homotopy classes of loops beginning and ending at the basepoint q_0, i.e. elements of $\pi_1(\mathcal{Q})$. However, the correspondence between homotopy classes of paths and elements of $\pi_1(\mathcal{Q})$ is not unique. To construct one such correspondence, choose two paths c_1 and c_2 joining q_0 to q_1 and q_2 respectively (fig. 11). Then we associate the homotopy class of the path $q(t)$ to the homotopy class of the loop $c_2^{-1}\cdot q\cdot c_1$. Having chosen this correspondence, we can consider the partial amplitude

$$K_\alpha(q_2,t_2\,|\,q_1,t_1)=\int_{q_1,t_1}^{q_2,t_2}(dq)_\alpha e^{\frac{i}{\hbar}S_0(q)}\ ,$$

where the subscript α in the measure means that the integral is performed over all paths such that $c_2^{-1}\cdot q\cdot c_1$ is in the class $\alpha\in\pi_1(\mathcal{Q})$. Since paths in different homotopy classes form disjoint sets, we can weigh differently the contribution of each homotopy class and write the total amplitude as

$$K(q_2, t_2 \,|\, q_1, t_1) = \sum_{\alpha \in \pi_1(\mathcal{Q})} \chi(\alpha) K_\alpha(q_2, t_2 \,|\, q_1, t_1) \ . \tag{2.58}$$

The complex weights $\chi(\alpha)$ have to be chosen so that the following requirements are satisfied:
1) the total amplitude must be independent of the choice of the paths c_1 and c_2
2) the total amplitude must satisfy the factorization property

$$K(q_2, t_2 \,|\, q_1, t_1) = \int dq\, K(q_2, t_2 \,|\, q, t) K(q, t \,|\, q_1, t_1)$$

for $t_1 < t < t_2$.

It can be shown [10] that these conditions imply that $\chi \in U(1)$ and $\chi(\alpha \cdot \beta) = \chi(\alpha)\chi(\beta)$, where $\alpha \cdot \beta$ is the product in the fundamental group of $\mathcal{Q}$. Thus χ has to be a character of $\pi_1(\mathcal{Q})$. Each choice of χ defines an inequivalent quantum theory, so we have reached again the conclusion that inequivalent quantizations are labelled by $\mathrm{Hom}\big(\pi_1(\mathcal{Q}), U(1)\big)$.

This can be related to the discussion of the preceding section as follows. As mentioned in Section 3.1 there is a one-to-one correspondence between the characters of $\pi_1(\mathcal{Q})$ and the gauge equivalence classes of flat $U(1)$ connections on $\mathcal{Q}$. If $\mathcal{A}$ is a flat connection, the corresponding character is given by

$$\chi(\alpha) = e^{\frac{ie}{\hbar} \oint\limits_{\ell} \mathcal{A}} \ ,$$

where ℓ is a loop in the homotopy class α (χ depends only on the homotopy class of ℓ since $\mathcal{A}$ is flat). There follows that if we define a “topological term”

$$S_T = e \int\limits_{t_1}^{t_2} dt\, \dot{q}^i \mathcal{A}_i \ ,$$

this term has the same value for all curves joining the point q_1 at the time t_1 to the point q_2 at the time t_2 and such that $c_2^{-1} \cdot q \cdot c_1$ is in a fixed homotopy class α. Thus we can absorb this term in the functional integral and write:

$$\sum_\alpha \chi(\alpha) K_\alpha(q_2, t_2; q_1, t_1) = \sum_\alpha \int_{q_1, t_1}^{q_2, t_2} (dq)_\alpha e^{\frac{i}{\hbar}(S_0 + S_T)} = \int_{q_1, t_1}^{q_2, t_2} (dq) e^{\frac{i}{\hbar}(S_0 + S_T)} \ .$$

So the effect of performing the functional integral with the action S_0 and weighting the partial amplitudes with characters of $\pi_1(\mathcal{Q})$ is exactly the same as performing the functional integral with the action $S_0 + S_T$.

[10] M.G.G. Laidlaw, C. Morette DeWitt “Feynman functional integrals for systems of indistinguishable particles”, Phys. Rev.D 3 (1971) 1375-1378

2.7.2 Euclidean path integrals

To make the integral convergent we perform a Wick rotation to imaginary time $\tau{=}it$. The euclidean action is given by

$$S_{0E} = -iS_0(t = -i\tau) = \int d\tau \left[\frac{1}{2} m g_{ij} \left(\frac{dq^i}{d\tau}\right)\left(\frac{dq^j}{d\tau}\right) + V(q)\right] \, . \tag{2.59}$$

We will omit the subscript E from now on for notational simplicity. Putting aside the issues due to multiple connectedness for a moment, the euclidean amplitude is

$$K_E(q_2, \tau_2 \,|\, q_1, \tau_1) = \int_{q_1,\tau_1}^{q_2,\tau_2} (dq) e^{-\frac{1}{\hbar} S_{0E}(q)} \, .$$

Let q_0 denote the vacuum (the state of lowest energy). To start with, we assume that it is unique. The vacuum-to-vacuum amplitude is also called the partition function:

$$Z_E(T) = K_E(q_0, T/2; q_0, -T/2) \, . \tag{2.60}$$

We can extract the ground state energy of the system from the vacuum-to-vacuum amplitude, using the following trick.

Denoting $\hat{H}$ the hamiltonian, the (Euclidean) evolution operator is $e^{-\frac{1}{\hbar}\hat{H}t}$, and we have

$$Z_E(T) = \langle q_0 | e^{-\frac{1}{\hbar}\hat{H}T} | q_0 \rangle = \sum_n |\langle q_0 | E_n \rangle|^2 e^{-\frac{1}{\hbar}E_n T} \, ,$$

where $\{|E_n\rangle\}$ is a complete set of eigenstates of the hamiltonian with eigenvalues E_n. For $T \to \infty$ the lowest energy eigenstate dominates the sum, so

$$\lim_{T\to\infty} Z_E(T) = \lim_{T\to\infty} |\langle q_0 | E_0 \rangle|^2 e^{-\frac{1}{\hbar}E_0 T} \, . \tag{2.61}$$

In this way, if we are able to compute the l.h.s. of the equation, we can read off the lowest energy eigenvalue E_0.

For example in the case of a harmonic oscillator, with $m = 1$ and $V(q) = \frac{1}{2}\omega^2 q^2$, the vacuum to vacuum amplitude turns out to be equal to

$$\left(\frac{\omega}{\pi\hbar}\right)^{1/2} e^{-\frac{\omega T}{2}} \, . \tag{2.62}$$

Comparing with (2.61) one finds the ground state energy $E_0 = \frac{1}{2}\hbar\omega$.

Finally consider the case when $\mathcal{Q}$ is multiply connected and a topological term is present in the action. In general, it can be written

$$S_T = \int dt \, \mathcal{A}_i(q) \dot{q}^i \, ,$$

where we have absorbed the charge in the definition of the "magnetic potential" $\mathcal{A}$. Since the topological Lagrangian only contains one time derivative, the Euclidean version of this action becomes imaginary:

$$S_{TE} = -i \int d\tau \, \mathcal{A}_i(q) \dot{q}^i = -iS_T \, , \tag{2.63}$$

Thus in the path integral its contribution remains oscillatory:

$$K_E(q_2,\tau_2\,|\,q_1,\tau_1) = \int_{q_1,\tau_1}^{q_2,\tau_2} (dq) e^{-\frac{1}{\hbar}(S_{0E}(q) - iS_T(q)} \ .$$

2.8 Path integral of the pendulum

The most important complications arising in the approximate calculations of path integrals on multiply connected configuration spaces appear already in the simplest case of the pendulum. We will therefore begin by discussing in some detail this example. As a preliminary, we observe that this problem is very similar to that of a particle in a periodic potential. This problem is well-known in solid-state physics. In a "zeroth-order" approximation one would expand the potential around a minimum $2\pi n$ and the lowest energy eigenfunction, with energy $E_0 = \frac{1}{2}\hbar\omega$, would be the one of the harmonic oscillator centered around $2\pi n$. There would be one such eigenfunction for each minimum, so the ground state would consist of infinitely degenerate states with energy $\frac{1}{2}\hbar\omega$. However, this approximation neglects tunnelling between neighbouring minima. When taken into account, this breaks the degeneracy and one gets a continuous band of states whose energy depends on the parameter θ (see Exercise 2.7.1). However, there is an important physical difference. Although the pendulum and the particle in the periodic potential have the same classical Lagrangian, they are different because in the former case all points on the line are identified $\mathrm{mod}\, 2\pi$, whereas in the latter they are not. This leads to a different physical interpretation of the results.

2.8.1 The $n = \pm 1$ contributions

We are going to study the vacuum energy of the pendulum as a function of θ using the trick of Section 2.7.2. We begin by observing that the classical "vacuum state" of the pendulum is $\varphi = 0 \bmod 2\pi$ (independent of time). The vacuum-to-vacuum transition amplitude is

$$K(0\, mod\, 2\pi, T/2\,|\,0, -T/2) = \sum_{n=-\infty}^{\infty} e^{in\theta} K_n(2\pi n, T/2\,|\,0, -T/2) \ ,$$

where $n \in \mathbb{Z} = \pi_1(S^1)$ labels the homotopy classes of $\varphi(t)$ and we assume without loss of generality that $\varphi = 0$ for $T \to -\infty$. The partial amplitudes are computed here with the action S_0 corresponding to the Lagrangian (2.17) and we have introduced the characters of $\mathbb{Z}$ in the sum over the partial amplitudes, as discussed earlier. Since K_n is a path integral over loops in a fixed homotopy class, we can bring the character inside the path integral and write

$$e^{in\theta} K_n(2\pi n, T/2\,|\,0, -T/2) = \int_{0,-T/2}^{2\pi n, T/2} (d\varphi)_n e^{\frac{i}{\hbar} S_0 + i\theta n}$$

$$= \int_{0,-T/2}^{2\pi n, T/2} (d\varphi)_n e^{\frac{i}{\hbar}(S_0 + \theta\hbar W)} = \tilde{K}_n(2\pi n, T/2\,|\,0, -T/2) \ .$$

We have defined $\tilde{K}$ the amplitude in the presence of the topological term.

Next we perform the Wick rotation. As already mentioned, the euclidean topological term is imaginary: $S_{T,E} = -iS_T = -i\theta\hbar W$. Thus, the euclidean amplitude is

$$\tilde{K}_E(0 \bmod 2\pi, T/2\,|\,0, -T/2) = \sum_n \tilde{K}_{E,n}(2\pi n, T/2\,|\,0, -T/2) \ ,$$

$$\tilde{K}_{E,n}(2\pi n, T/2\,|\,0, -T/2) = \int_{0,-T/2}^{2\pi n, T/2} (d\varphi)_n\, e^{-\frac{1}{\hbar}S_{0E} + i\theta n} \ ,$$

The partial amplitudes can be evaluated using the WKB, or saddle point approximation: we will now compute the contribution of fields which are near a stationary point of the euclidean action.

Let us begin by evaluating certain contributions to $\tilde{K}_{E1}(2\pi, T/2\,|\,0, -T/2)$, i.e. the sum over paths with winding number one. The action is minimized by the instanton solutions (**??**), which are parametrized by the coordinate of the 'center" τ_0, and the path integral will be dominated by configurations that are near one of these solutions. Thus we expand the action around $\varphi_{\rm cl}(\tau)$. We get

$$S_E(\varphi) = S_E(\varphi_{\rm cl}) + \frac{1}{2}\int d\tau d\tau' \eta(\tau)\mathcal{O}(\tau,\tau')\eta(\tau') \ ,$$

where $\eta = \varphi - \varphi_{\rm cl}$ and

$$\mathcal{O}(\tau,\tau') = \left.\frac{\delta^2 S_E}{\delta\varphi(\tau)\delta\varphi(\tau')}\right|_{\varphi_{\rm cl}} = \delta(\tau-\tau')\left(-\frac{d^2}{d\tau^2} + V''(\varphi_{\rm cl})\right) \ . \qquad (2.64)$$

In the WKB approximation

$$\begin{aligned} K_{E1}(2\pi, T/2\,|\,0, -T/2) &= e^{-\frac{1}{\hbar}S_E(\varphi_{\rm cl})}\int (d\eta)\, e^{-\frac{1}{2}\int \eta\mathcal{O}\eta} \\ &= e^{-\frac{1}{\hbar}S_E(\varphi_{\rm cl})} B(T)\,[{\rm Det}\mathcal{O}]^{-1/2} \qquad (2.65) \end{aligned}$$

where $B(T)$ is a measure factor .

The operator $-\frac{d^2}{dt^2} + V''(\varphi_{\rm cl})$ has a translational zero mode, corresponding to the fact that the position of the instanton is arbitrary (see Exercise 2.7.2). The integration on the zero mode is replaced by an integration on the "collective coordinate" τ_0, which yields a factor T. The change in the integration variable produces a Jacobian J, whose evaluation we postpone to section 2.8.3. The main result that is needed here is that J is independent of T for $T \to \infty$. So (2.65) can be rewritten

$$e^{-\frac{1}{\hbar}S_E(\varphi_{\rm cl})} B(T) J T \left[{\rm Det}'\mathcal{O}\right]^{-1/2} \ ,$$

where ${\rm Det}'$ is the product of the nonzero eigenvalues. The evaluation of the determinant is difficult because $\varphi_{\rm cl}$, which appears in the operator (2.64) depends explicitly on time. However, the size of the instanton was fixed by the form of the potential and is independent of T, so if we are only interested in the limit of large T, we see that "most of the time" $\varphi_{\rm cl} = 0 \bmod 2\pi$ and therefore $V''(\varphi_{\rm cl}) = \omega^2$. We can then write

$$\left[\mathrm{Det}'\mathcal{O}\right]^{-1/2} = K\left[\mathrm{Det}\left(-\frac{d^2}{dt^2}+\omega^2\right)\right]^{-1/2} \tag{2.66}$$

where K, the ratio of the determinants, becomes a constant independent of T for large T. The determinant on the r.h.s. together with the factor $B(T)$ is the partition function of a harmonic oscillator, which is given by (2.62). We thus find

$$K_{E,1}(2\pi, T/2\,|\,0, -T/2) = e^{-\frac{1}{\hbar}S_{0E}+i\theta} KJTe^{-\frac{\omega T}{2}}\left(\frac{\omega}{\pi\hbar}\right)^{1/2}$$

where we have written $S_E(\varphi_{\rm cl}) = S_{0E}(\varphi_{\rm cl}) - i\theta\hbar W(\varphi_{\rm cl}) = S_{0E} - i\theta\hbar$. This is the contribution of the one-instanton sector to the total amplitude. The one anti-instanton sector gives

$$K_{E,-1}(-2\pi, T/2\,|\,0, -T/2) = e^{-\frac{1}{\hbar}S_{0E}-i\theta} KJTe^{-\frac{\omega T}{2}}\left(\frac{\omega}{\pi\hbar}\right)^{1/2}$$

2.8.2 The dilute instanton gas

In principle we should now evaluate the contributions of paths with higher winding numbers and then sum over the winding numbers. However, we have already observed in Exercise 1.1.1 that there are no classical solutions to the equation $-\frac{d^2\varphi}{dt^2}+\frac{dV}{d\varphi}=0$ in the sectors $\mathcal{Q}_{0i}$ with $|i|>1$, i.e. solutions interpolating between nonadjacent minima. This means that there are no exact multi-instanton solutions around which to expand the action. Thus, we cannot directly apply the WKB method to compute the contribution of paths with winding number greater than one. In practice the calculation can still be done, but in a different way.

We observe that a configuration consisting of m_1 instantons and m_2 anti-instantons, all widely separated, will provide an approximate solution to the classical equation of motion with $W = m_1 - m_2$. Such a configuration will contribute to the partial amplitude $\tilde{K}_E\big((m_1-m_2)2\pi, T/2\,|\,0, -T/2\big)$. The evaluation of the functional integral for this case proceeds much as in the one-instanton case, with the following changes: every instanton gives a contribution to $S_E(\varphi_{\rm cl})$ equal to $S_{0E} - i\theta\hbar$ and each anti-instanton gives a contribution $S_{0E} + i\theta\hbar$; every instanton and anti-instanton has a translational zero mode contributing a factor TJ; as long as they are widely separated, every instanton and anti-instanton contributes a factor K when $V''(\varphi_{\rm cl})$ is replaced by ω^2 in the determinant. Altogether the contribution to the total amplitude due to configurations containing m_1 instantons and m_2 anti-instantons is

$$\frac{1}{m_1!m_2!}\exp\left[-\frac{1}{\hbar}(m_1+m_2)S_{0E}+i(m_1-m_2)\theta\right](KJT)^{m_1+m_2}\left(\frac{\omega}{\pi\hbar}\right)^{1/2}e^{-\frac{\omega T}{2}}$$

The factor $\frac{1}{m_1!m_2!}$ is due to the indistinguishability of the instantons and anti-instantons (in the integral over the collective coordinates, the situation when instanton 1 is in position τ_1 and instanton 2 is in position τ_2 is physically the same as when instanton 1 is in position τ_2 and instanton 2 is in position τ_1). The total amplitude is obtained by summing over m_1 and m_2. This automatically includes a sum over winding numbers. The sums can be performed explicitly and we get

$$\begin{aligned} Z_\theta(T) &= \exp\left(KJTe^{-\frac{1}{\hbar}S_{0E}+i\theta}\right)\exp\left(KJTe^{-\frac{1}{\hbar}S_{0E}-i\theta}\right)\left(\frac{\omega}{\pi\hbar}\right)^{1/2}e^{-\frac{\omega T}{2}} \\ &= \left(\frac{\omega}{\pi\hbar}\right)^{1/2}\exp\left[-\frac{1}{\hbar}T\left(\frac{1}{2}\hbar\omega-2\hbar KJe^{-\frac{1}{\hbar}S_{0E}}\cos\theta\right)\right] . \qquad (2.67) \end{aligned}$$

Comparing with (2.61) we find that the energy of the vacuum in the presence of the θ-term in the action is

$$E_\theta = \frac{1}{2}\hbar\omega - 2\hbar KJe^{-\frac{1}{\hbar}S_{0E}}\cos\theta \ . \qquad (2.68)$$

This is much the same result that one obtains for a particle in a periodic potential but with an important difference: there all states in the band belong to the same Hilbert space and therefore transitions between states with different values of θ are permitted. Here every value of θ defines a different theory and no transition between different θ-states can occur.

This way of computing the functional integral for a theory with multiply connected $\mathcal{Q}$ is known as the dilute instanton gas approximation. We will see that it can be easily generalized to the case of fields theories.

The preceding discussion shows that instantons are related to tunnelling. The static energy E_S, as a function on $\mathcal{Q}$, has its minimum at some point φ_0 that we call the classical vacuum. We take φ_0 as "basepoint" in $\mathcal{Q}$. Without loss of generality we can assume the vacuum energy to be zero; elsewhere it is positive. Therefore if we consider a non-contractible loop in $\mathcal{Q}_0$ parameterized by $-\infty \le t \le \infty$, we see that the energy as a function of t has the shape shown in fig. 12. If the system is in a low energy state, it cannot classically follow such a trajectory, but it can do it in the quantum theory by tunnelling. In the WKB approximation the tunnelling amplitude is evaluated as a sum over trajectories that are near a classical solution of the equations of motion. No classical solutions exists in the real time, minkowskian section, but as we have seen, solutions exist in the imaginary time, euclidean section. Thus it is the WKB approximation that requires performing the Wick rotation. In the end the amplitude is analytically continued back to real time.

2.8.3 Evaluation of the Jacobian

The substitution of the integral over the zero mode by the integral over the collective coordinate can be justified by a procedure resembling the Faddeev-Popov method. We are going to constrain the projection of the quantum field

φ onto the zero mode η_0 to be equal to the projection of the instanton on the zero mode. This requires a compensating factor that we shall evaluate.

We begin by recalling (from section 1.1.1) that the action of the instanton comes in equal amounts from the kinetic and potential term. Thus the action of the instanton is

$$S_{cl} = \int dt \left[\frac{1}{2}\dot{\varphi}_{cl}^2 + V(\varphi_{cl})\right] = \int dt\, \dot{\varphi}_{cl}^2 \ . \tag{2.69}$$

There follows that the normalized zero mode is

$$\eta_0(t-t_0) = \frac{1}{\sqrt{S_{cl}}}\frac{d\varphi_{cl}(t-t_0)}{dt} \ . \tag{2.70}$$

In what follows both the instanton φ_{cl} and the zero mode η_0 are located at a particular time t_0. Thus using (2.70) we find that the projection of the instanton on the zero mode is

$$\begin{aligned}(\varphi_{cl},\eta_0) &= \int_{-\infty}^{\infty} dt\, \eta_0(t-t_0)\varphi_{cl}(t-t_0) \\ &= \frac{1}{2\sqrt{S_{cl}}}\int_{-\infty}^{\infty} dt \frac{d\varphi_{cl}(t-t_0)^2}{dt} \\ &= \frac{1}{2\sqrt{S_{cl}}}\varphi_{cl}^2\Big|_{-\infty}^{\infty} = \frac{2\pi^2}{\sqrt{S_{cl}}}\end{aligned} \tag{2.71}$$

and is independent of t_0.

Then, the following relation holds

$$\Delta[\varphi]\int dt_0\, \delta\left[(\varphi,\eta_0) - (\varphi_{cl},\eta_0)\right] = 1 \tag{2.72}$$

where

$$(\varphi,\eta_0) = \int_{-\infty}^{\infty} dt\, \eta_0(t-t_0)\varphi(t)$$

will in general depend on t_0. The quantity Δ is

$$\Delta[\varphi] = \frac{d}{dt_0}\left[(\varphi,\eta_0) - (\varphi_{cl},\eta_0)\right]$$

evaluated at a point (here assumed unique) where the argument of the delta function is zero. We have

$$\begin{aligned}\Delta[\varphi] &= \int_{-\infty}^{\infty} dt\, \frac{d\eta_0(t-t_0)}{dt_0}\varphi(t) \\ &= \int_{-\infty}^{\infty} dt\, \eta_0(t-t_0)\frac{d\varphi(t)}{dt} \ .\end{aligned} \tag{2.73}$$

Now we insert the identity (2.72) in the functional integral

$$Z = \int dt_0 \int (d\varphi) e^{-\frac{1}{\hbar}(S_0+i\theta)}\Delta[\varphi]\delta\left[(\varphi,\eta_0) - (\varphi_{cl},\eta_0)\right]$$

When we expand the action around the instanton located at t_0 the argument of the delta function becomes (η, η_0), with η_0 centered again at t_0. The compensating factor, evaluated at the classical solution, is

$$\Delta[\varphi_{cl}] = \int_{-\infty}^{\infty} dt\, \eta_0(t-t_0) \frac{d\varphi_{cl}(t)}{dt} = \sqrt{S}_{cl} \int_{-\infty}^{\infty} dt\, \eta_0(t-t_0)^2 = \sqrt{S_{cl}}$$

Thus we obtain

$$Z = e^{-\frac{1}{\hbar}(S_{cl}+i\theta)} \sqrt{S_{cl}} \int dt_0 \int (d\eta) e^{-\frac{1}{2\hbar}(\eta, L\eta)} \delta[(\eta, \eta_0)]$$

The path integral over η is now performed on fields that have no projection on the zero mode. It is therefore given by the primed determinant $(\det L)^{-1/2}$. The integral over the collective coordinate t_0 gives a factor T and the Jacobian for the change of variable is

$$J = \sqrt{S}_{cl} \ , \tag{2.74}$$

which in the limit $T \to \infty$ is manifestly independent of t_0.

2.9 The abelian Higgs model

The perturbative spectrum of scalar QED depends on the sign of the mass term. For $f^2 < 0$, in dimension $d > 2$, it consists of a charged massive scalar, its antiparticle and a massless photon. In two dimensions there is no photon. Furthermore, the Coulomb potential grows linearly with distance and the force between two oppositely charged scalars is independent of distance. This means that the charged particles are confined in neutral bound states. When $f^2 > 0$, in any dimension including two, the theory is in the Higgs phase: there is only a neutral scalar (the radial mode) and a massive photon. Because of this, the force between charges falls off exponentially with distance. We shall see now that the effect of instantons changes the picture quite drastically: when $\theta \neq 0 \mathrm{mod} 2\pi$, there is no Higgs phase.

As already discussed in section XXX, the instanton of scalar QED$_2$ is the vortex of QED$_3$ with unit flux. We put the system in a large spacetime box of spacial extent L and time duration T. In the limit $T \to \infty$, the partition

$$Z_\theta(L,T) = \int (dA\, d\phi\, d\phi^*)\, e^{-S_{0E}+i\theta c_1}$$

equals e^{-TE_θ}, and by evaluating Z_θ we shall obtain E_θ and other observables. Since the instantons have a fixed finite size that is negligible in the limit of large T and L, we can evaluate Z_θ with a dilute instanton gas. The functional integral can be evaluated following the steps of the previous section. The main novelty is that now there are two translation zero modes for each instanton and anti-instanton, so the integration over the corresponding collective coordinates yields a factor LT for each instanton and anti-instanton. Thus we find

$$\lim_{T\to\infty} Z_\theta(T) = Ae^{-LT\left(C - e^{-S_{0E}} 2B\cos\theta\right)} \tag{2.75}$$

for some constants A,B,C, where S_{0E} denotes the action for the single instanton solution. From here one reads off the energy density

$$\frac{E_\theta}{L} = C - e^{-S_{0E}} 2B\cos\theta$$

analogous to the result (2.68).

The physical meaning of the parameter θ can be further clarified by considering the vacuum expectation value of the electric field $\langle E_1\rangle_\theta = i\langle F_{01}\rangle_\theta$. Due to translational invariance

$$\langle F_{01}(x,\tau)\rangle_\theta = \frac{1}{LT}\left\langle \int dx d\tau\, F_{01}\right\rangle_\theta = \frac{1}{2LT}\left\langle \int dx d\tau\, \varepsilon^{\mu\nu} F_{\mu\nu}\right\rangle_\theta = \frac{2\pi}{LT}\langle c_1\rangle_\theta$$

We have

$$\langle c_1\rangle_\theta = i\frac{d}{d\theta}\ln Z_\theta = -i\frac{d}{d\theta}(E_\theta T) = -iLTe^{-S_0}2B\sin\theta\ .$$

Therefore

$$\langle E_1(x,\tau)\rangle_\theta = 4\pi e^{-S_0} B\sin\theta\ .$$

Therefore, in the theta vacuum, there is a uniform background electric field. This fact leads us to suspect the existence of long range forces, in spite of the fact that at tree level, due to the occurrence of the Higgs phenomenon, we would expect only short range forces. We will now prove that instantons do indeed give rise to long range forces and confinement, even for $f^2 > 0$.

Consider two (nondynamical, external) charges q and $-q$ at a fixed distance $\tilde{L}$. The potential energy between these charges is given by the difference of the energy of the system in the presence and in the absence of the charges. If the system is quasi static, these energies in turn can be evaluated as the effective actions divided by the time.

More precisely, suppose that the pair of charge and anticharge is created at some instant, brought to distance $\tilde{L}$, then left there for a large time $\tilde{T}$ and finally annihilated again. The classical contribution to the action due to the presence of the charges is

$$\int d^2x\, j^\mu A_\mu = q\oint A$$

where $j^\mu(x) = q\delta^{(2)}(x - x(t))\frac{dx^\mu}{dt}$ is the current generated by the charges. The quantity $W = e^{iq\oint A}$ is called the Wilson loop. As before, we enclose the system in a spacetime volume of sides $L \gg \tilde{L}$ and $T \gg \tilde{T}$. In the limit $T \to \infty$ the Euclidean functional integral gives the exponential of the energy in the presence of the charges:

$$\int (dAd\phi d\phi^*) e^{-S_{0E}+i\theta c_1} W = \exp(-TE_\theta - \tilde{T}\Delta E_\theta(\tilde{L}))\ . \qquad (2.76)$$

We have then

$$\lim_{\tilde{T}\to\infty}\langle W\rangle = \frac{1}{Z_\theta(T)}\int (dAd\phi d\phi^*) e^{-S_{0E}+i\theta c_1 - iq\oint A} = e^{-\tilde{T}\Delta E_\theta(\tilde{L})}$$

Therefore we can compute the interaction between the charges from the Wilson loop:

$$\Delta E_\theta(\tilde{L}) = -\lim_{\tilde{T}\to\infty}\frac{1}{\tilde{T}}\ln\langle W\rangle_\theta \ . \tag{2.77}$$

We use again the dilute instanton gas approximation. We divide $n_\pm = n_\pm^{(\text{in})} + n_\pm^{(\text{out})}$, counting separately instantons and anti-instantons that lie inside or outside the spacetime loop traced by the charges. The reason for this is that the Wilson loop can be rewritten:

$$W = e^{iq\oint A} = e^{\frac{iq}{2}\int_U d^2x\epsilon^{\mu\nu}F_{\mu\nu}} = e^{2\pi iq(n_+^{(\text{in})}-n_-^{(\text{in})})}$$

where U is the region enclosed by the loop. Then, the functional integral (2.76) can be evaluated as follows:

$$\begin{aligned} &Ae^{-LTC}\sum_{n_+^{(\text{in})},n_-^{(\text{in})},n_+^{(\text{out})},n_-^{(\text{out})}}\frac{1}{n_+^{(\text{in})}!n_-^{(\text{in})}!n_+^{(\text{out})}!n_-^{(\text{out})}!} \\ &\quad\times\exp\big(\big[-(n_+^{(\text{in})}+n_-^{(\text{in})}+n_+^{(\text{out})}+n_-^{(\text{out})})S_{0E} \\ &\qquad\qquad +i\theta(n_+^{(\text{in})}-n_-^{(\text{in})}+n_+^{(\text{out})}-n_-^{(\text{out})})\big] \\ &\quad\times\left[B(LT-\tilde{L}\tilde{T})\right]^{n_+^{(\text{out})}+n_-^{(\text{out})}}(B\tilde{L}\tilde{T})^{n_+^{(\text{in})}+n_-^{(\text{in})}}\exp\left[2\pi iq(n_+^{(\text{in})}-n_-^{(\text{in})})\right] \\ &\quad = A\exp\left\{-LTC+2Be^{-S_{0E}}\left[\tilde{L}\tilde{T}\cos(\theta+2\pi q)+(LT-\tilde{L}\tilde{T})\cos\theta\right]\right\} \ . \end{aligned} \tag{2.78}$$

Using (2.75) and (2.78) in (2.77) we get

$$\Delta E_\theta(\tilde{L}) = 2Be^{-S_{0E}}\tilde{L}\left[\cos\theta-\cos(\theta+2\pi q)\right] \ . \tag{2.79}$$

From this formula we see that the potential grows with distance, leading again to confinement of the charges. Thus, the physical picture is the same for $f^2 > 0$ as for $f^2 < 0$. From the factor $e^{-S_{0E}}$ we see, however, that the force is strictly nonperturbative (the numerator contains a hidden factor $1/\hbar$) and that it vanishes exponentially in the classical limit.

To get a physical intuition for the θ- and q-dependence, we can expand for small θ and q and find

$$\begin{aligned} \langle E_1\rangle_\theta &= 4\pi Be^{-S_{0E}}\theta \\ E_\theta &= BLe^{-S_{0E}}\theta^2 \\ \Delta E_\theta(\tilde{L}) &= B\tilde{L}e^{-S_{0E}}\left[(\theta+2\pi q)^2-\theta^2\right] \ . \end{aligned}$$

In the θ vacuum there is a constant electric field and an energy density proportional to the square of this electric field. External charges in one dimension act as the plates of a capacitor and produce an additional constant electric field in the space between them. The shift in energy due to the charges is the distance between the charges, times the difference in the energy density in the presence and in the absence of the charges.

Returning to the full result (2.79), we see that if the charges are integer, the force vanishes. This can be understood if we assume that the spectrum contains also particles of unit charge and their antiparticles. In this case we can think that partticle-antiparticle pairs will be created between the two test charges and will move towards them until the electric field is completely screened. If q is not an integer, the creening cannot be complete, leaving a residual force which is independent of distance.

2.10 False vacuum decay

We discuss here an application of instantons that is not related to a multiply connected configuration space, namely the decay of a metastable vacuum. This issue arises in first order phase transitions. [11] To be specific we will assume that the system is described by a scalar theory with quartic potential, tilted by the addition of an infinitesimal linear term:

$$V(\phi) = \frac{\lambda}{4}\left(\phi^2 - f^2\right)^2 + \frac{\epsilon}{2f}\phi \ , \tag{2.80}$$

where $\epsilon > 0$. To first order in ϵ, the two minima and the maximum are located at

$$\phi_\pm = \pm f - \frac{\epsilon}{4\lambda f^3} \ , \qquad \phi_{\rm max} = \frac{\epsilon}{2\lambda f^3} > 0$$

and the respective potentials are

$$V(\phi_+) = \epsilon/2 \ , \qquad V(\phi_-) = -\epsilon/2 \ , \qquad V(\phi_{\rm max}) = \lambda f^4/4 + O(\epsilon^2) \ .$$

The difference in energy density of the two minima is $\mathcal{E} = \epsilon$. Recall from Section 1.2.1 that in $d > 1$ the kink can be reinterpreted as the profile of the walls separating domains of different vacua. The thickness of the wall is $\ell \sim 1/(\sqrt{\lambda} f)$ and the "surface" energy density of the wall is $\mathcal{T} \sim \sqrt{\lambda} f^3$. These formulas were derived for planar domain walls, but they will still be approximately correct if the radius of curvature of the surface is $R \gg \ell$. This is called the *thin wall approximation.*

The other simplifying assumption is spherical symmetry. Thus consider a d-dimensional system in the homogeneous metastable state ϕ_+ and suppose that a spherical bubble of true vacuum forms inside the metastable vacuum. In the thin wall approximation, this is an easily understandable one-dimensional problem, whose only variable is the radius of the bubble. If we shift the potential

[11] A typical example in statistical physics would be a ferromagnet that is in thermodynamic equilibrium in an external magnetic field, and the direction of the magnetic field is reversed. In our treatment we do not consider thermal fluctuations and the transition is driven by quantum fluctuations.

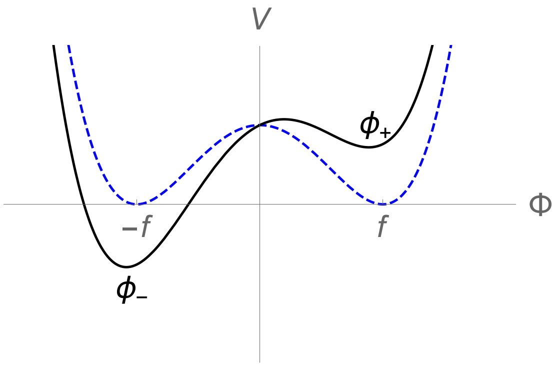

Figure 2.3: Quartic potential with linear tilt (black curve) vs. the original potential (dashed). The two minima have moved to the left and the maximum has moved to the right.

in such a way that the energy of the metastable vacuum is zero, the energy of the bubble has a negative term proportional the bulk volume and a positive one proportional to the surface:

$$E(r) = -\mathcal{E} V_{B^d}(r) + \mathcal{T} V_{S^{d-1}}(r) \ . \tag{2.81}$$

The volume of the $(d-1)$-dimensional sphere of radius r is

$$V_{S^{d-1}}(r) = (4\pi)^{(d-1)/2} r^{d-1} \Gamma((d-1)/2)/\Gamma(d-1)$$

and the volume of the d-dimensional ball of radius r is

$$V_{B^d}(r) = \pi^{d/2} r^d / \Gamma((d+2)/2) \ .$$

The energy difference is plotted for various dimensions in Figure 2.4.

The bubble will be in equilibrium if

$$\frac{dE(r)}{dr} = 0$$

and this happens at

$$r_e = (d-1)\frac{\mathcal{T}}{\mathcal{E}} \ . \tag{2.82}$$

(Note that there can be no equilibrium in $d = 1$.) If the bubble has radius $r < r_e$ the surface tension will dominate and make the bubble shrink and disappear. If it has radius $r > r_e$ the bulk energy will dominate causing the bubble to expand to infinity.

It is also of interest to consider bubbles that involve neither a gain nor a loss of energy. Such bubbles can form spontaneously by quantum tunnelling, as we shall discuss below. Imposing that $E = 0$ we see that the radius of such a bubble is

$$r_0 = d\frac{\mathcal{T}}{\mathcal{E}} > r_e \tag{2.83}$$

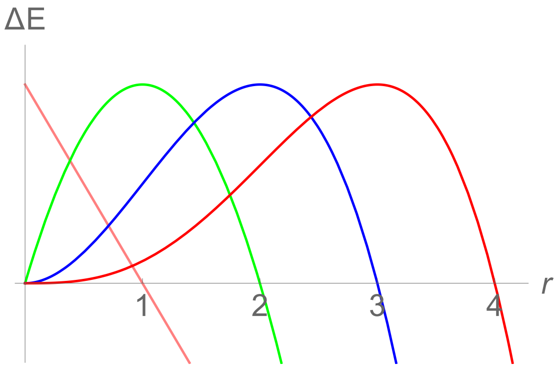

Figure 2.4: Energy of a vacuum bubble with $\mathcal{T} = 1$ and $\mathcal{E} = 1$ in $d = 1$ (straight pink line), $d = 2$ (green), $d = 3$ (blue) and $d = 4$ (red), all rescaled so as to have the same maximum. Note that the zero-energy radius in d dimension is the same as the equilibrium radius in $d + 1$ dimensions.

and therefore such zero-energy bubbles will expand.

Let us see under what circumstances the results of this simplified one-dimensional picture are reliable. The thin wall approximation is justified if $r_e \gg \ell$. Neglecting numerical factors of order one, this means

$$1 \ll \frac{r_e}{\ell} \approx \frac{\mathcal{T}}{\mathcal{E}\ell} \approx \lambda \frac{f^4}{\epsilon} ,$$

which can always be satisfied if ϵ is sufficiently small. In the following we will always assume that this condition is satisfied and that the radius of the bubble is not much smaller than r_e.

So far we have limited ourselves to discussing equilibrium conditions. Let us now study the bubble dynamics in the physically most interesting case $d = 3$. As above, the assumption of spherical symmetry reduces this to a one-dimensional problem. The bubble acts like a particle at position r with a position-dependent mass $M = 4\pi r^2 \mathcal{T}$, so the Lagrangian is given by a kinetic term, minus the “potential” energy (2.81):

$$L = \frac{1}{2} M \dot{r}^2 + \frac{4}{3}\pi r^3 \mathcal{E} - M \ . \tag{2.84}$$

The first and third terms can be seen as coming from the expansion of the square root of the following relativistic Lagrangian:

$$L = -M\sqrt{1 - \dot{r}^2} + \frac{4}{3}\pi r^3 \mathcal{E} \ . \tag{2.85}$$

From here we now derive the relativistic equations for the system. The momentum conjugate to r is

$$p = \frac{M\dot{r}}{\sqrt{1 - \dot{r}^2}} , \tag{2.86}$$

which can be inverted to

$$\dot{r} = \frac{p}{\sqrt{p^2 + M^2}} \ . \tag{2.87}$$

The Hamiltonian, written as a function of the velocity, is

$$H = \frac{M}{\sqrt{1 - \dot{r}^2}} - \frac{4}{3}\pi r^3 \mathcal{E} \ . \tag{2.88}$$

From here, using (2.86,2.87), we obtain the relation

$$\left(H + \frac{4}{3}\pi r^3 \mathcal{E}\right)^2 = \left(\frac{M}{\sqrt{1 - \dot{r}^2}}\right)^2 = \left(\frac{p}{\dot{r}}\right)^2 = p^2 + M^2 \tag{2.89}$$

Thus for a bubble nucleated from vacuum, with $H = 0$, we have

$$p^2 = \left(\frac{4}{3}\pi r^3 \mathcal{E}\right)^2 - M^2 = M^2 \left(\frac{r^2}{r_0^2} - 1\right) \tag{2.90}$$

and inserting this in (2.87) we obtain

$$\dot{r} = \sqrt{1 - \frac{r_0^2}{r^2}} \tag{2.91}$$

This is real only if $r \geq r_0$. Let us begin with a bubble of radius $r = r_0$ and $\dot{r} = 0$. Solving the equation of motion (2.91) one finds

$$r(t) = \sqrt{r_0^2 + t^2} \tag{2.92}$$

This is a relativistic uniformly accelerated motion, starting at rest and approaching asymptotically a light cone.

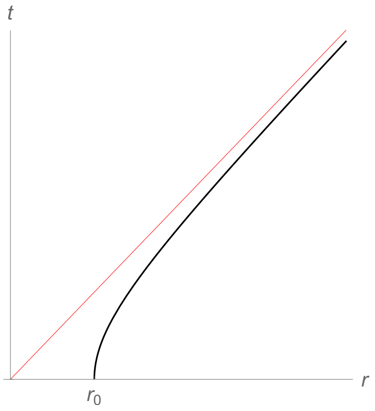

Figure 2.5: Expansion of a vacuum bubble starting at rest.

We now wish to discuss the process of bubble nucleation. This will be done using instanton methods. If the system is at rest in the state ϕ_+, there is no classical solution of the equations of motion that will generate a spherical bubble of true vacuum; it does not have the energy that is necessary to overcome the potential barrier. However, there are solutions of the Euclidean equations that can do this. The simplest way to see this is to observe that, as in all cases considered before, the Euclidean action of the theory in d dimensions is the same as the static energy of the same theory in $d+1$ dimensions. Thus, for a spherically symmetric Euclidean bubble in the thin wall approximation, the action is equal to (2.81), with $d=4$. We know already that this action has a stationary point: it is the maximum of the red curve in Figure (2.4). Therefore this represents the instanton in the simplified model with one degree of freedom.

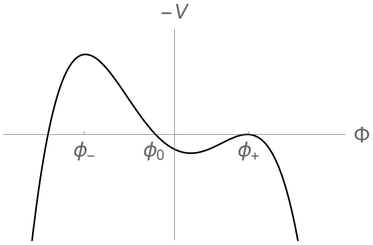

Figure 2.6: For the solution of (2.82).

For a proper evaluation of the nucleation rate we need to know the instanton of the original scalar theory from which (2.81) was derived, and in this case it is a bit more complicated to see that an instantons exists. The instanton will be spherically symmetric in the four-dimensional sense [12] and thus will be given by a function $\phi(r)$ (where $r=\sqrt{x_1^1+x_2^2+x_3^2+\tau^2}$) with the boundary conditions that $\phi \to \phi_+$ for $r \to \infty$, and $\phi'(0)=0$ (as required by spherical symmetry). The Euclidean field equation for this function is

$$\frac{d^2\phi}{dr^2}+\frac{3}{r}\frac{d\phi}{dr}=V'(\phi)\ . \tag{2.93}$$

By the usual device of reinterpreting ϕ as position and x as time, this can be seen as Newton's equation for a particle in the potential $-V$, in the presence of a time-dependent friction term. Following Coleman, the existence of a solution with the desired boundary conditions can be argued as follows. Let us shift the potential such that $V(\phi_+)=0$ and let ϕ_0 be the zero of V that lies nearest to ϕ_+, see Figure (2.6). The boundary condition $\phi'(0)=0$ means that the particle is released at time zero at rest. If $\phi(0)$ is on the right of ϕ_0, the perticle does not have enough energy to reach ϕ_+. If $\phi(0)$ is on the left of ϕ_0 and very close to ϕ_-, the particle will spend a long time near ϕ_-. During this time the friction term can become arbitrarily small. When the particle finally rolls down the potential, the friction term can be neglected and the particle overshoots.

[12] It can be shown that an $SO(4)$-symmetric instanton has lower action than any non-symmetric instanton, see S. Coleman, V. Glaser and A. Martin, Comm. Math. Phys. **58**, 211 (1978).

Thus there must exist an initial position ϕ_* between ϕ_- and ϕ_0 such that if $\phi(0) = \phi_*$, the particle will just reach ϕ_+ and come to rest there. When we replace the fictitious time by radius, the resulting function $\phi(r)$ is a solution of (2.82) with the desired boundary conditions. The profile of the solution is not known analytically, but in the thin wall limit $\epsilon \to 0$, $\phi_* \to \phi_-$ and the profile becomes that of a kink at radius r_e. In this limit the action of the instanton is

$$S_{cl} = -\mathcal{E}V_{B^4}(r_e) + \mathcal{T}V_{S^3}(r_e) = \frac{27}{2}\pi^2\frac{\mathcal{T}^4}{\mathcal{E}^3} \tag{2.94}$$

and the tunnelling amplitude in the classical approximation is proportional to $e^{-S_{cl}}$. We note that this conclusion is in agreement with a standard one-dimensional quantum mechanical treatment. Indeed, in the WKB approximation, the tunnelling amplitude is proportional to

$$\exp\left(-2\int_0^{r_0} dr|p(r)|\right) = \exp\left(-2\int_0^{r_0} dr\, 4\pi r^2\mathcal{T}\sqrt{1-\frac{r^2}{r_0^2}}\right) = \exp\left(-\frac{\pi^2}{2}r_0^3\mathcal{T}\right) ,$$

and we see that the exponent agrees with (2.94) for $d = 3$.

Returning to the field-theoretic picture, we could proceed as with the other instanton calculations, and try to interpret the partition function as e^{-ET}, where E is the energy of the ground state. Unlike the other calculations, however, the system does not have a ground state because the energy is unbounded from below. The instability of the state manifests itself in the energy having an imaginary part, as we shall see next.

Let ϕ_{cl} be the instanton solution described above and let $\phi_\lambda(x) = \phi_{cl}(\lambda x)$. We have

$$S(\phi_\lambda) = \frac{1}{2}\lambda^{-2}\int d^4x(\partial\phi_{cl})^2 + \lambda^{-4}\int d^4xV(\phi_{cl}) .$$

Deriving with respect to λ and putting $\lambda = 1$ we find that

$$0 = \frac{dS(\phi_\lambda)}{d\lambda}\Big|_{\lambda=1} = -\int d^4x(\partial\phi_{cl})^2 - 4\int d^4xV(\phi_{cl}) .$$

This implies that $S(\phi_{cl}) = \frac{1}{4}\int d^4x(\partial\phi_{cl})^2 > 0$. Deriving a second time

$$0 = \frac{d^2S(\phi_\lambda)}{d\lambda^2}\Big|_{\lambda=1} = 3\int d^4x(\partial\phi_{cl})^2 + 20\int d^4xV(\phi_{cl}) = -2\int d^4x(\partial\phi_{cl})^2 < 0 .$$

This shows that the kinetic operator has one negative mode. It corresponds to the radius in the simplified model with one degree of freedom, and we know already that the solution is a maximum for the radius. It can be shown that there are no other negative modes. Therefore the determinant $\left(\det\left(\frac{\delta^2 S}{\delta\phi\delta\phi}\right)\right)^{-1/2}$, is purely imaginary. This fact is important for the following reason. We can evaluate the functional integral in the dilute gas approximation leading to the expression

$$Ae^{-VT(C-J^4Ke^{-S_{cl}})} ,$$

where A is the determinant of the harmonic oscillator states, J is the Jacobian associated to each zero mode, calculated in (2.74), and K is the ratio of determinants defined as in (2.66), adapted to the present problem. J appears to the fourth power because the instanton has four translational zero modes that get converted to the spacetime volume VT. It follows from the previous remark that K is purely imaginary. When we read off the energy from the previous expression we find that it has an imaginary part. However, this was to be expected, because the system does not have a stable ground state. Instead, the imaginary part of the energy is just the decay probability. The decay probability per unit time and unit volume is equal to

$$\Gamma/V = J^4 |K| e^{-S_{cl}} \ . \tag{2.95}$$

Finally we see from equation (2.82) and (2.83) that the radius r_e of the instanton (in $d = 4$) is equal to the radius r_0 of a bubble of zero energy (in $d = 3$). Thus, the instanton can be seen as interpolating continuously between the initial false vacuum state and a state where there is a bubble of true vacuum at rest and zero total energy (this state being represented by the evaluation of the instanton at the time $\tau = 0$). Actually, since the Minkowskian field equations are the analytic continuation of the Euclidean ones, and since at time $\tau = 0$ the four-dimensional radius is the same as the three-dimensional radius, the same profile $\phi(r)$ that gives the $SO(4)$-invariant Euclidean solutions, when evaluated at $\tau = 0$, also solves the static field equations for an $SO(3)$-invariant bubble of zero energy. Thus at the midpoint of its time evolution, the instanton is exactly a bubble of radius r_0. One can then smoothly match the Euclidean solution to the Minkowskian solution representing the expanding bubble.

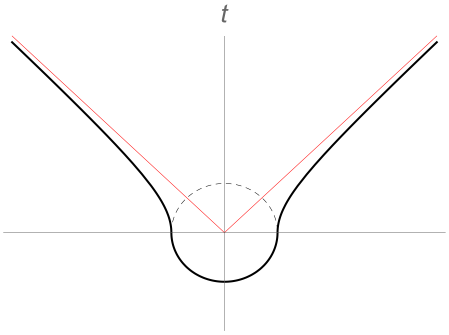

Figure 2.7: Euclidean instanton ($t < 0$) matching the expanding bubble in Minkowski space ($t > 0$).

Chapter 3

$H^2(\mathcal{Q})$ and the quantization of parameters

Let us consider again the motion of a charged particle on a manifold $\mathcal{Q}$, in a background magnetic field $\mathcal{F}$, with a potential $\mathcal{A}$, and lagrangian (2.6). In chapter 2 we have discussed at length the case in which $\mathcal{Q}$ is multiply connected and $\mathcal{F}=0$. We will now consider situations in which $\mathcal{Q}$ is simply connected and $\mathcal{F}\neq 0$. As discussed in section 2.1, quantization demands that $\mathcal{A}$ be a $U(1)$-connection, rather than an $\mathbb{R}$-connection. How can we tell whether this is the case?

A necessary and sufficient condition for $\mathcal{F}$ to be the field strength of a $U(1)$ connection is that

$$\int_m \mathcal{F} = \frac{2\pi\hbar}{e} n \qquad\qquad n \in \mathbf{Z}\ , \tag{3.1}$$

for any two-dimensional submanifold m of $\mathcal{Q}$ without boundary. In mathematical terms, $\mathcal{F}$ has to define an integral cohomology class in $H^2(\mathcal{Q})$ (see Appendix **???**). A general proof of this result is given in Appendix **???**. We will give a proof only in the simplest case, namely when $\mathcal{Q} = S^2$. This leads to the famous quantization condition of the monopole charge given by Dirac. We then move on to discuss some field theoretic analogues of this phenomenon: nonlinear sigma models with Wess–Zumino–Witten terms, and odd dimensional gauge theories with Chern–Simons terms. These are all terms in the action that give a nontrivial contribution to the equations of motion, just like the monopole field enters in the equations of motion of a charged particle. Nevertheless, because of their topological origin, we will still call them ”topological terms”.

The common denominator of all these theories is given by the integrality condition (3.1). In the case of a charged particle moving in the monopole field, it leads to the quantization of the monopole charge; in the field theoretic examples it leads to a certain parameter in the lagrangian taking quantized values.

As is clear from the preceding discussion, the proper topological setting for these phenomena is cohomology rather than homotopy. However if $\mathcal{Q}$ is simply connected, Hurewicz' theorem (Appendix B) states that $H^2(\mathcal{Q},\mathbb{Z}) = \pi_2(\mathcal{Q})$, so one could loosely say that these phenomena are related to a nontrivial second homotopy group.

3.1 The Dirac quantization condition

Let us consider the magnetic field $B_i = \frac{Q_M}{r^2}\hat{x}^i$. We will regard it as a fixed background, and seek consistency conditions for the quantization of a charged particle moving in this background. As discussed in section 1.7, this field is singular in the origin and therefore should be regarded as a vacuum solution of Maxwell's equations on $\mathcal{Q} = \mathbb{R}^3 \setminus \{0\}$. On this manifold, $d\mathcal{F} = 0$ and so one expects that there exists a magnetic potential $\mathcal{A}$. It turns out that there is no magnetic potential for the monopole which is regular everywhere on $\mathbb{R}^3 \setminus \{0\}$. To see this, suppose a magnetic potential $\mathcal{A}$ is given and consider the line integral $\phi(\theta) = \oint_{\ell(\theta)} \mathcal{A}$, where $\ell(\theta)$ is a parallel at colatitude θ on a sphere of radius r, see Fig. 3.1.

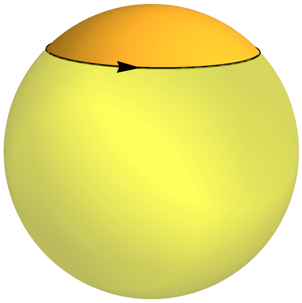

Figure 3.1: A parallel at colatitude θ and the cap it bounds.

Clearly $\phi(0) = \phi(\pi) = 0$. On the other hand using Stokes' theorem $\phi(\theta) = \oint_{\ell(\theta)} \mathcal{A} = \int_{U(\theta)} \mathcal{F}$, where $U(\theta)$ is the cap bounded by $\ell(\theta)$. Thus $\phi(\theta)$ is the flux through the cap. This can be easily computed to be $\phi(\theta) = 2\pi Q_M(1 - \cos\theta)$. For $\theta = \pi$ it is equal to $4\pi Q_M$. Thus we get a contradiction.

In order to understand more clearly what happens we can try to look for explicit forms of the magnetic potential. Using a natural basis in spherical coordinates the field strength reads

$$\mathcal{F} = Q_M \sin\theta \, d\theta \wedge d\varphi \ . \tag{3.2}$$

A solution of the equation $\mathcal{F} = d\mathcal{A}$ is given by

$$\mathcal{A} = \mathcal{A}^{(+)} = Q_M(1 - \cos\theta)d\varphi \ . \tag{3.3}$$

This potential is singular on the negative z-axis ($\theta = \pi$). In fact, the form $d\varphi$ is singular on the whole z-axis but its coefficient $(1-\cos\theta)$ vanishes along the positive z-axis ($\theta = 0$). This singularity of the magnetic potential is known as the Dirac string. Its does not correspond to any singularity of the field, however, since it can be moved by gauge transformations. For example, another choice of magnetic potential is

$$\mathcal{A}^{(-)} = -Q_M(1+\cos\theta)d\varphi \ , \tag{3.4}$$

which is singular on the positive axis. Let U^+ and U^- be the subsets of $\mathcal{Q}$ with $\theta \neq \pi$ and $\theta \neq 0$ respectively. Even though one cannot introduce a magnetic potential everywhere on $\mathcal{Q}$, it is still possible to give a satisfactory description of the monopole field by giving the potential $\mathcal{A}^+$, which is regular on U^+ and the potential $\mathcal{A}^-$, which is regular on U^-. Together, these two open sets cover all of $\mathcal{Q}$. On the intersection $U^+ \cap U^- = \mathbb{R}^3 \setminus \{z\text{-axis}\}$, the two potentials are related by a gauge transformation:

$$\mathcal{A}^+ - \mathcal{A}^- = 2Q_M d\varphi \ . \tag{3.5}$$

As emphasized in section 2.1, quantization of the particle with lagrangian (2.6) demands that $\mathcal{A}$ be a $U(1)$ gauge field.

In the present case, to quantize the particle one would introduce wavefunctions $\psi^\pm$ which need only be well-defined on $U^\pm$ respectively, and are related by a gauge transformation $g = e^{i\alpha}$ in the intersection:

$$\psi^+(\theta,\varphi) = g(\theta,\varphi)^{-1}\psi^-(\theta,\varphi) \qquad \text{on}\, U^+ \cap U^- \ . \tag{3.6}$$

Note that $U^+ \cap U^-$ is multiply connected and g is required to be a single-valued function from $U^+ \cap U^-$ to $U(1)$. The corresponding transformation of the gauge potential is

$$\mathcal{A}^+ = \mathcal{A}^- - \frac{\hbar}{ie}g^{-1}dg = \mathcal{A}^- - \frac{\hbar}{e}d\alpha \ , \tag{3.7}$$

so comparing with (3.5) we see that the appropriate gauge transformation is

$$g(\theta,\varphi) = e^{-i\frac{e}{\hbar}2Q_M\varphi} \ . \tag{3.8}$$

This will be single-valued if the magnetic charge satisfies the following Dirac quantization condition:

$$Q_M = \frac{\hbar}{2e}n \ . \tag{3.9}$$

Noting that $Q_M = \frac{1}{4\pi}\int_{S_2}\mathcal{F}$, this is precisely the same as (3.1).

One can give a path integral argument leading to (3.9). The action of a particle moving in a background magnetic field is

$$I = \int\limits_{-\infty}^{\infty} dt \left[\frac{m}{2}\left(\frac{dq^i}{dt}\right)^2 + e\frac{dq^i}{dt}\mathcal{A}_i \right] \ . \tag{3.10}$$

This action suffers from two related problems. First, in the case of the monopole, $\mathcal{A}$ has singularities, as we have seen. This form of the action is therefore only appropriate for those histories of the particle that do not cross the Dirac string. On the other hand we know that the Dirac string is not a physical singularity, so this must be a shortcoming of our description of the system, not of the system itself. Second, the action is not gauge invariant. Under the gauge transformation (3.7), $I' = I - \hbar\left(\alpha(\infty) - \alpha(-\infty)\right)$.

There is a way of rewriting the action that avoids both problems. Consider a closed orbit c, with $\vec{x}(\infty) = \vec{x}(-\infty)$ (this is analogous to the choice $\varphi(\infty) = \varphi(-\infty)$ in the discussion of the pendulum in Section 2.2). Then we can apply Stokes' theorem and write

$$e\int_c \mathcal{A} = e\int_U \mathcal{F}\ , \tag{3.11}$$

where U is a two dimensional surface having c as boundary. This way of writing the action is gauge invariant and insensitive to the Dirac string, but it makes reference to the surface U, which is not uniquely defined by the trajectory of the particle. Since $d\mathcal{F} = 0$, the integral (3.11) is invariant under infinitesimal deformations of the surface that keep the boundary fixed, but it may change for large deformations. In fact, consider two surfaces U_1 and U_2 both having c as boundary, but one passing "above", the other "below" the origin (see Fig. ???). The difference ΔI in the actions $e\int_{U_1}\mathcal{F}$ and $e\int_{U_2}\mathcal{F}$ is equal to the integral of $\mathcal{F}$ on the closed surface formed by joining U_1 and U_2 along the boundary. Since this surface contains the origin, the integral is equal to $4\pi e Q_M$. This arbitrariness in the action will not affect the functional integral if $e^{\frac{i}{\hbar}\Delta I} = 1$, which directly implies (3.1).

Finally we observe that the quantization condition can also be seen as an application of the old Bohr–Sommerfeld quantization conditions $\oint pdq = 2\pi\hbar n$. From (2.7) we have $\oint pdq = \int\left(mg_{ij}\dot{q}^i\dot{q}^j + e\dot{q}^i\mathcal{A}_i\right)dt$. Now consider a very small loop encircling the Dirac string. When the radius of the loop goes to zero, the first term goes to zero but the second becomes $e\oint\mathcal{A} = e\int_{S^2}\mathcal{F}$, having applied Stokes' theorem to a surface bounded by the loop and not containing the string. The Bohr–Sommerfeld rule then gives (3.1).

3.2 Wess–Zumino–Witten terms

Consider a non-linear sigma model with values in $SU(2) \equiv S^3$, in $d=1$ space dimensions. The configuration space is $\mathcal{Q} = \Gamma_*(S^1, S^3)$. Using the, by now familiar, technique of Appendix **???** we find that $\pi_0(\mathcal{Q}) = \pi_1\big(SU(2)\big) = 0$, $\pi_1(\mathcal{Q}) = \pi_2\big(SU(2)\big) = 0$ and $\pi_2(\mathcal{Q}) = \pi_3\big(SU(2)\big) = \mathbf{Z}$. The generator of $\pi_2(\mathcal{Q})$ is a map $m : S^2 \to \mathcal{Q}$ which is defined by $\big(m(t_1,t_2)\big)(t_3) = \hat{m}(t_1,t_2,t_3)$, where $\hat{m}$ is a map of S^3 (a cube $I\times I\times I$ with the boundary identified to a point) to $SU(2)$, sending $\partial(I\times I\times I)$ into the identity element, and with winding number $W(\hat{m}) = 1$. If the map c in section 2.3 could be referred to as a "loop of loops", the map m defined here could be called a "sphere of loops". By Hurewicz' theorem, together with (B.), one concludes that $H^0(\mathcal{Q},\mathbb{Z}) = 0$, $H^1(\mathcal{Q},\mathbb{Z}) = 0$

and $H^2(\mathcal{Q},\mathbb{Z}) = \mathbb{Z}$. The low homotopy and cohomology groups of $\mathcal{Q}$ are the same as in the previous section, so one may expect to find some analogue of the Dirac quantization condition in this theory. This is indeed the case. As with theta sectors, in order to reveal the occurrence of topological phenomena, it is necessary to add an appropriate term to the action.

To guess the right term we may look for inspiration in Section 2.3, where we discussed the same theory in one more dimension. We saw that the integrand of the topological term $\theta W(\varphi)$ was a total derivative and therefore W could be written as a surface integral (an integral on a two dimensional space). Suppose now that the boundary conditions on the fields are such that they go to a constant at spacetime infinity (this is the case if we demand that the action be finite), so that spacetime can be compactified to a sphere S^2. We can think of this sphere as the boundary of some (fictitious) three dimensional ball B^3 and regard the fields φ as boundary values of some field $\tilde{\varphi}$ defined on B^3. This is always possible because $\pi_2(SU(2)) = 0$, so all fields φ are homotopically trivial and therefore have a continuation in the interior of B^3. The topological term we are after is just the topological term of $\tilde{\varphi}$, which depends only on φ and not on the value of the fields in the interior of the ball. We therefore have

$$S_{WZW} = c\int \varphi^*\tau = \frac{c}{2}\int d^2x\,\varepsilon^{\mu\nu}\partial_\mu\varphi^\alpha\partial_\nu\varphi^\beta\tau_{\alpha\beta} \tag{3.12}$$

where $\omega = d\tau$ or, in components,

$$\omega_{\alpha\beta\gamma} = 3(\partial_\alpha\tau_{\beta\gamma} + \partial_\beta\tau_{\gamma\alpha} + \partial_\gamma\tau_{\alpha\beta})\ . \tag{3.13}$$

and ω is the volume form on $SU(2)$, normalized so that $\int_{SU(2)}\omega = 1$. We have renamed c the constant that previously was called θ. For example, suppose we choose on $SU(2)$ a coordinate system given by the Euler angles (Θ,Φ,Ψ) (see Appendix G). The volume form is given by

$$\omega = \frac{1}{8\pi}\sin\Theta\, d\Theta\wedge d\Phi\wedge d\Psi\ . \tag{3.14}$$

and a choice of τ is

$$\tau = -\frac{1}{8\pi}\cos\Theta\, d\Phi\wedge d\Psi\ . \tag{3.15}$$

Since $d\tau \neq 0$, the Wess–Zumino–Witten term (3.12) is not a total derivative term and, as we shall see in a moment, it does contribute to the equations of motion of the theory.

An important consequence of (3.13) is that τ is not uniquely defined: if τ satisfy (3.13), also $\tau' = \tau + d\beta$ does. This amounts to adding a total derivative to the action 3.12). Another fact of the greatest importance is that τ is not globally defined. If it was, ω would define a trivial cohomology class. But we know from Appendix B that the volume-form on a compact manifold always defines a non-trivial cohomology class. This means that the form τ is singular somewhere on $SU(2)$. For example the form τ defined in (3.15) is singular for $\Theta = 0$ or π.

Now consider a field $\varphi(x,t)$; we regard it as a map from S^2 (the one-point compactification of spacetime) into $SU(2)$. Generically, φ will not meet the singular points of τ. Thus there will be an open subset $\mathcal{U}$ of $\Gamma_*\big(S^2, SU(2)\big)$ where $\mathrm{Im}\varphi \cap \{\text{singular set}\} = \emptyset$, and the WZW action (3.12) will be well defined on $\mathcal{U}$. However, there are also maps φ whose image intersects the singular set of τ. For such maps (3.12) is not well defined. We can however use the freedom of adding a total derivative term to the action. The form $\tau' = \tau + d\beta$ will have a different singular set from τ, and the action $c\int \varphi^*\tau' = c\int \varphi^*\tau + c\int d(\varphi^*\beta)$ will be well defined on another open subset $\mathcal{U}'$ of $\Gamma_*\big(S^2, SU(2)\big)$. Since the image of φ has dimensions 2 and the singular set of any form τ has dimension zero, it is clear that for every $\varphi \in \Gamma_*\big(S^2, SU(2)\big)$ one can find a one-form β such that τ' does not have any singularity on the image of φ. In this way one can cover $\Gamma_*\big(S^2, SU(2)\big)$ with open sets, such that on each set there is a well defined function $c\int \varphi^*\tau$, and on the intersection of two sets these functions differ by a total derivative term $c\int d(\varphi^*\beta)$. The collection of these locally defined functions is the WZW term. It is not a functional of φ in the ordinary sense. Instead, it is a section of a certain line bundle over $\Gamma_*\big(S^2, SU(2)\big)$.

Let us consider the non-linear sigma model with the action given by $S = S_0 + S_{WZW}$, where

$$S_0 = -\frac{1}{2}\int d^2x\, \partial_\mu\varphi^\alpha \partial^\mu\varphi^\beta h_{\alpha\beta}\ . \tag{3.16}$$

The equation of motion reads

$$h_{\alpha\beta}\partial_\mu\partial^\mu\varphi^\beta + \Gamma_{\alpha,\beta\gamma}\partial_\mu\varphi^\beta\partial^\mu\varphi^\gamma + \frac{c}{2!}\varepsilon^{\mu\nu}\partial_\mu\varphi^\beta\partial_\nu\varphi^\gamma\omega_{\alpha\beta\gamma} = 0 \tag{3.17}$$

where $\Gamma_{\alpha,\beta\gamma} = \frac{1}{2}\left(\partial_\beta h_{\alpha\gamma} + \partial_\gamma h_{\alpha\beta} - \partial_\alpha h_{\beta\gamma}\right)$ are the Christoffel symbols of the metric $h_{\alpha\beta}$ on $SU(2)$. The last term is the contribution of the WZW term. It can be interpreted as follows. Note that the WZW term is linear in the time derivative and therefore can be written as $\int dt\, \dot\varphi^\alpha \mathcal{A}_\alpha(\varphi)$, where

$$\mathcal{A} = -c\int dx\, \partial_1\varphi^\alpha \tau_{\alpha\beta}\delta\varphi^\beta \tag{3.18}$$

is a one-form on $\mathcal{Q}$. When we think of the sigma model as a particle moving on $\mathcal{Q}$, $\mathcal{A}$ can be interpreted as a “functional vector potential”. Unlike the cases discussed in chapter 2, the corresponding “functional magnetic field” is now non-vanishing. A direct calculation using the methods of Appendix E yields

$$\mathcal{F} = d\mathcal{A} = \frac{c}{2}\int dx\, \partial_1\varphi^\alpha \omega_{\alpha\beta\gamma}\delta\varphi^\beta\delta\varphi^\gamma\ . \tag{3.19}$$

To confirm our interpretation of $\mathcal{A}$ and $\mathcal{F}$ note that the last term in (3.17) can be written $c\dot\varphi^\beta\mathcal{F}_{\alpha\beta}$ and therefore can be interpreted as the Lorentz force due to $\mathcal{F}$. The one- form $\mathcal{A}$ is only well defined on a subset $\mathcal{V}$ of $\mathcal{Q}$ such that the image of $\varphi(x)$ $\big($a loop in $SU(2)\big)$ does not intersect the singular set of τ. By contrast, $\mathcal{F}$ is well defined everywhere on $\mathcal{Q}$. We are therefore in a situation which resembles very closely that of the previous section, with a magnetic field $\mathcal{F}$ that cannot be derived from a globally defined vector potential $\mathcal{A}$. Rather than repeating the discussion of the Dirac quantization condition in the present context, we will apply directly the general result (3.1). Let us therefore compute the integral of $\mathcal{F}$ on the fundamental two-cycle m. We have

$$\begin{aligned}\int\limits_m \mathcal{F} &= \int\limits_0^1 dt_1 \int\limits_0^1 dt_2 \left\{ c \int dx\, \partial_1 \varphi^\alpha \omega_{\alpha\beta\gamma} \frac{\partial \varphi^\beta}{\partial t_1} \frac{\partial \varphi^\gamma}{\partial t_2} \right\} \\ &= \frac{c}{3!} \int d^3x\, \varepsilon^{\lambda\mu\nu} \frac{\partial \hat{\varphi}^\alpha}{\partial x^\lambda} \frac{\partial \hat{\varphi}^\beta}{\partial x^\mu} \frac{\partial \hat{\varphi}^\gamma}{\partial x^\nu} \omega_{\alpha\beta\gamma} \\ &= cW(\hat{\varphi}) = c\ . \end{aligned} \tag{3.20}$$

Using (3.1) we find that the theory can be quantized only for

$$c = 2\pi n\ . \tag{3.21}$$

This is the analogue of the Dirac quantization condition.

The quantization of the parameter c can also be proven in the path integral formalism by means of the following argument. The extension $\bar{\varphi}$ is not unique. Consider two extensions $\bar{\varphi}_1 : B_1^3 \to SU(2)$ and $\bar{\varphi}_2 : B_2^3 \to SU(2)$, with $\bar{\varphi}_1\big|_{S^2} = \bar{\varphi}_2\big|_{S^2} = \varphi$. Since they coincide on $\partial B_1 = \partial B_2 = S^2$, we can think of them as a single map $\bar{\varphi} : S^3 \to SU(2)$, where S^3 is obtained by glueing the two balls S^3 along their boundaries (in this picture the two balls are the hemispheres of S^3 and S^2 is the equator of S^3). The maps φ_1 and φ_2 together define a map $\tilde{\varphi} : S^3 \to SU(2)$. The difference of the WZW actions is therefore equal to $\Delta S_{WZW} = S_{WZW}(\varphi_2) - S_{WZW}(\varphi_1) = cW(\tilde{\varphi})$. This arbitrariness will not affect the functional integral if $e^{i\Delta S_{WZW}} = 1$, which again implies (3.21).

Can one write a Wess–Zumino–Witten term for a sigma model in 3+1 dimensions? To answer this question, let us review what we have done in this section. We have started from a closed three-form ω representing a nontrivial cohomology class of the target space. This form could be written locally as the exterior differential of a two-form τ. The Wess–Zumino–Witten action was the integral of the pullback of τ. In 3+1 dimensions ω would have to be a closed five form and τ a four-form. There are no five-forms on $SU(2)$, but there are nontrivial five-forms on $SU(N)$ for $N \geq 3$. In fact, $H^5(SU(N), \mathbb{R}) = \mathbb{R}$, and the generator of this group is (the cohomology class of) $-\frac{i}{240\pi^2}\mathrm{tr}R^5$. Therefore, the Wess–Zumino–Witten term can be written in either one of the following two forms:

$$\begin{aligned} S_{WZW} &= c \int d^4x\, \varepsilon^{\mu\nu\rho\sigma} \partial_\mu \varphi^\alpha \partial_\nu \varphi^\beta \partial_\rho \varphi^\gamma \partial_\sigma \varphi^\delta \tau_{\alpha\beta\gamma\delta} \\ &= -\frac{ic}{240\pi^2} \int d^5x\, \varepsilon^{\lambda\mu\nu\rho\sigma} \mathrm{tr}(R_\lambda R_\mu R_\nu R_\rho R_\sigma)\ , \end{aligned} \tag{3.22}$$

where spacetime has been compactified to a four-sphere and in the last line the integral is over a ball having spacetime as a boundary. This gives rise to the magnetic potential

$$\mathcal{A} = -c \int d^3x\, \varepsilon^{ijk} \partial_i \varphi^\alpha \partial_j \varphi^\beta \partial_k \varphi^\gamma \tau_{\alpha\beta\gamma\delta} \delta\varphi^\delta \tag{3.23}$$

on $\mathcal{Q}$. The corresponding field strength is

$$\mathcal{F} = d\mathcal{A} = \frac{c}{2} \int d^3x\, \varepsilon^{ijk} \partial_i \varphi^\alpha \partial_j \varphi^\beta \partial_k \varphi^\gamma \omega_{\alpha\beta\gamma\delta\eta} \delta\varphi^\delta \delta\varphi^\eta \ . \tag{3.24}$$

One can now repeat the arguments given above leading to the quantization of the parameter c.

Finally we observe that the relation between ω and $\mathcal{F}$ is a special example of a general construction which relates cohomology classes of N to cohomology classes of $\Gamma(M, N)$. This is discussed in Appendix F.

3.3 Chern–Simons terms

Next we consider an $SU(2)$ gauge theory in $2+1$ dimensions. As in the previous chapter, we will use the rescaled, geometrical gauge fields, with curvature defined by (1.90), gauge transformations (1.91) and action (1.89). Instead of writing explicitly the Lie algebra indices, we will use a matrix notation and write $A_\mu = A^a_\mu T_a$, where T_a are matrices satisfying $[T_a, T_b] = \varepsilon_{abc} T_c$ and $\mathrm{tr}(T_a T_b) = -\frac{1}{2}\delta_{ab}$ (for example in the spinor representation, $T_a = -\frac{i}{2}\sigma_a$, where σ_a are the Pauli matrices). In this notation the Yang–Mills action (1.89) reads

$$S_{YM} = \frac{1}{2e^2} \int d^3x\, \mathrm{tr}\, F_{\mu\nu} F^{\mu\nu} \ . \tag{3.25}$$

As in Sections 3.4 and 3.5, we choose the gauge $A_0 = 0$; then the static energy reads

$$E_S = -\frac{1}{e^2} \int d^2x\, \mathrm{tr}\, B^2 \ , \tag{3.26}$$

where $B = F_{12}$ is the nonabelian magnetic field. The configuration space is then $\mathcal{Q} = \mathcal{C}/\mathcal{G}$, where $\mathcal{C}$ is the space of connections $A_i(\vec{x})$, $i = 1, 2$, such that E_S is finite, and $\mathcal{G} = \Gamma_*(S^2, SU(2))$ is the residual gauge group consisting of time independent gauge transformations.

This configuration space is connected and furthermore has $\pi_1(\mathcal{Q}) = \pi_0(\mathcal{G}) = \pi_2(SU(2)) = 0$ and $\pi_2(\mathcal{Q}) = \pi_1(\mathcal{G}) = \pi_3(SU(2)) = \mathbf{Z}$. The generator of the group $\pi_2(\mathcal{Q})$ can be described as follows. The gauge group $\mathcal{G}$ is connected but not simply connected. Let $\ell(t)$ be a loop whose homotopy class generates $\pi_1(\mathcal{G})$. Fix a reference point $A_{(0)}$ in $\mathcal{C}$ and consider the loop in the orbit through $A_{(0)}$ given by $A^{\ell(t)}_{(0)}$. This loop cannot be shrunk to a point within the orbit but it can be shrunk to a point in $\mathcal{C}$. Thus there is a map $\tilde{m}$ from a two dimensional ball B^2 to $\mathcal{C}$ which is equal to $A^{\ell(t)}_{(0)}$ on the boundary. Now compose this map with the projection $\mathcal{C} \to \mathcal{Q}$. Since all points on the boundary of the disk are mapped to the same orbit, we get a map m from S^2 to $\mathcal{Q}$ which is not homotopic to a constant (see Fig. ???). The isomorphism between $\pi_1(\mathcal{G}) = \mathbf{Z}$ and $\pi_2(\mathcal{Q})$ is the map that sends (the homotopy class of) ℓ to (the homotopy class of) m.

Once again we have exactly the same homotopy groups as in Section 3.1, so we expect that some parameter will have to be quantized. But what parameter? As in the previous Section, we impose boundary conditions such that spacetime can be compactified to a three dimensional sphere S^3, and regard this sphere as the boundary of a four dimensional ball B^4. Gauge fields A_μ on a three sphere are topologically trivial ($\pi_2(SU(2)) = 0$) and therefore can be thought of as boundary values of gauge fields $\tilde{A}_\mu$ defined on B^4. The $SU(2)$ gauge theory in 3+1 dimensions was discussed in Section 2.5, where, in order to reveal the existence of theta sectors, we added to the action a topological term $S_T = \theta c_2$. With the boundary conditions of Section 2.5, c_2 was an integer; with the boundary conditions used here, the integral $c_2(\tilde{A})$ becomes a function of the boundary values A. Using that the integrand of c_2 is the exterior differential of the Chern–Simons three form (2.30), the appropriate topological term to be added to S_{YM} in three dimensions is the Chern–Simons term

$$S_{CS} = \mu\frac{8\pi^2}{e^2}\int d^3x\Omega\ , \tag{3.27}$$

where

$$\Omega = -\frac{1}{8\pi^2}\varepsilon^{\lambda\mu\nu}\mathrm{tr}\left(A_\lambda\partial_\mu A_\nu + \frac{2}{3}A_\lambda A_\mu A_\nu\right)\ . \tag{3.28}$$

(Note that apart from replacing the coefficient θ by the coefficient $\mu\frac{8\pi^2}{e^2}$, S_{CS} is identical to the functional $\tilde{\Lambda}$ defined in (2.34).) The constant μ has dimension of mass. In fact simple manipulations on the equations of motion (Exercise 3.2.1) show that this theory describes spin one particles with mass $|\mu|$. For this reason it was called a "topologically massive gauge theory".

From our previous discussion of the WZW action, we are led to believe that the coefficient of the CS action, the mass μ, has to be quantized in certain units. This is indeed what happens. The proof of this fact turns out to be rather involved at the canonical level, so we will depart from our standard procedure and give first a proof at the level of path integrals.

Let us restrict our attention to field configurations with S_{YM} finite. This implies that spacetime can be compactified to a sphere S^3. Therefore the group of gauge transformations is $\Gamma_*(S^3, SU(2))$, and it consists of infinitely many connected components, labelled by their winding number. The (dual of the) Chern–Simons form transforms as follows

$$\Omega(A^g) = \Omega(A) - \frac{1}{8\pi^2}\varepsilon^{\lambda\mu\nu}\mathrm{tr}\,\partial_\lambda\left(\partial_\mu g g^{-1}A_\nu\right) + \frac{1}{24\pi^2}\varepsilon^{\lambda\mu\nu}\mathrm{tr}\left(g^{-1}\partial_\lambda g^{-1}\partial_\mu g g^{-1}\partial_\nu g\right)\ , \tag{3.29}$$

and since we assume g to tend to the identity at infinity, upon integration we find

$$S_{CS}(A^g) = S_{CS}(A) + \mu\frac{8\pi^2}{e^2}W(g)\ . \tag{3.30}$$

This is essentially the same calculation that led to (???). Thus, the Chern–Simons action is gauge invariant under gauge transformations which are homotopic to the identity (in particular, it is invariant under infinitesimal gauge transformations), but not under "large" gauge transformations. We demand that the functional integral be insensitive to this ambiguity. This requires that $e^{i\Delta S_{CS}} = 1$, or

$$\mu = \frac{e^2}{4\pi} n \ , \qquad n \in \mathbf{Z} \ . \tag{3.31}$$

Note that in the Euclidean path integral one would demand $e^{-\Delta S_{CS,E}} = 1$, where $S_{CS,E}$ is the Euclidean Chern–Simons action. Since S_{CS} is linear in the time derivative, $S_{CS,E} = iS_{CS}$, so we are led again to (3.31).

To see what would go wrong if we did not impose the quantization condition (3.31), consider the formal procedure for eliminating the volume of the gauge group from the functional integral. Having chosen a gauge condition $f(A) = 0$, one inserts in the functional integral $Z = \int (dA) e^{iS(A)}$ the identity $1 = \Delta_{FP}(A) \int (dg) \delta(f(A^g))$, where $\Delta_{FP}(A)$ is the Faddeev–Popov determinant, a gauge invariant functional of the gauge potential. In the present case, since the gauge group has infinitely many connected components, it is convenient to write the integral over the gauge group as a sum of integrals over the connected components: $\int (dg) = \sum_n \int (dg)_n$. Now we have

$$Z = \sum_n \int (dg)_n \int (dA) \Delta_{FP}(A) \delta(f(A^g)) e^{iS(A)} \ .$$

At this point one usually invokes invariance of the measure, of the Faddeev–Popov determinant and of the action, to rewrite the argument of all functionals on the r.h.s. as A^g (and then A, since it is an integration variable). In the present case the action is not invariant, so taking into account (3.30) we find

$$Z = V_0 \sum_n e^{-i\mu \frac{8\pi^2}{e^2} n} \int (dA) \Delta_{FP}(A) \delta(f(A)) e^{iS(A)} \ ,$$

where V_0 is the volume of one connected component of the gauge group. The sum in front of the integral gives zero unless μ satisfies the quantization condition (3.31). Thus if (3.31) is not satisfied, the functional integral, and similarly the expectation value of any gauge invariant observable, is ill-defined.

Chapter 4

The spin of solitons

In previous chapters we have considered examples of field theories that have either solitons or theta sectors or quantized parameters. For clarity we have considered these phenomena in isolation, but there are interesting cases when they occur together. In this chapter we shall discuss three examples where solitons, theta vacua and/or quantized parameters are simultaneously present. We will see that the topological terms (whether total derivatives or not) have an effect on the physical properties of the solitons. In particular, there are various cases where they determine the spin of the soliton. There are various ways of seeing this effect, but one way that works in all cases is the following.

Let $|i\rangle$ be some quantum state describing a soliton, whose spin we want to compute, Imagine a process whereby the soliton is rotated by 2π, adiabatically slowly. The final state is $|f\rangle = e^{2\pi i R}|i\rangle$, where R is a suitable generator of the rotation. As discussed in Section 2.2.1, the final state is identical to $|i\rangle$, except possibly for a phase that is related to the spin:

$$e^{2\pi i s} = \langle f|i\rangle \ .$$

On the other hand, the amplitude can be expressed as a path integral over all paths with the appropriate boundary conditions and in the appropriate homotopy class:

$$\langle f|i\rangle = \int (d\bar{\varphi}) e^{iS(\bar{\varphi})} \ .$$

In the leading order of a semiclassical approximation, the functional integral is given by $e^{iS(\bar{\varphi}_{\rm cl})}$, where $\bar{\varphi}_{\rm cl}$ is a field configuration describing the rotation of the soliton by 2π in a time T. An important fact about the topological terms S_T, whether total derivatives or not, is that they are linear in time derivatives. In the adiabatic limit $T \to \infty$ such terms dominate over other terms in the Lagrangian that contain two or more derivatives. In fact, they are the only terms that give a finite contribution in the limit. We conclude that the spin is just given by the topological term, evaluated on the classical field:

$$2\pi s = S_T(\varphi_{cl}) \ . \tag{4.1}$$

The identification of the appropriate topological term, its parameter and the classical field have to be made case by case.

4.1 Sigma model anyons

In section 2.4 we discussed the S^2 nonlinear sigma model in 2+1 dimensions and showed that its configuration space consists of infinitely many connected components, labelled by the winding number:

$$\mathcal{Q} = \bigcup_{n=-\infty}^{\infty} \mathcal{Q}_n \ .$$

In each component we were able to find the absolute minimum of the static energy, and these minima were the Belavin–Polyakov solitons. We will now further examine the topology of the configuration space, more precisely we want to compute its fundamental group.

In general, when a space has many connected components, one has to choose a basepoint in each of them and compute the fundamental group separately for each component. In principle, these groups could be all different. Let us therefore consider first the component consisting of homotopically trivial maps, $\Gamma_*(S^2, S^2)_0$. Using the familiar rule we have

$$\pi_1(\Gamma_*(S^2, S^2)_0) = \pi_3(S^2) \ .$$

The generator of this group is the (homotopy class of the) Hopf map $h : S^3 \to S^2$, which is just the projection of the Hopf bundle. [1] The geometry of the Hopf bundle is discussed in detail in Appendix C. The homotopy groups of S^2 are related to the homotopy groups of S^3 by the homotopy exact sequence. In particular, we have

$$\ldots \to \pi_3(S^1) \to \pi_3(S^3) \xrightarrow{h_*} \pi_3(S^2) \to \pi_2(S^1) \to \ldots$$

Since the first and the last group in this sequence are trivial, the map h_* is an isomorphism. Since $\pi_3(S^3) = \mathbb{Z}$, we have proven that also $\pi_3(S^2) = \mathbb{Z}$. The integer that labels the homotopy classes of maps from S^3 to S^2 is called the Hopf invariant. The conclusion we draw from this discussion is that $\pi_1(\mathcal{Q}_0) = \mathbb{Z}$. It can be shown that also the other connected components have the same homotopy group. According to the general discussion in section 2.1, the quantum theory must depend on an angle $0 \le \theta < 2\pi$.

[1] Identifying S^3 with the group $SU(2)$, S^1 with the subgroup $U(1)$ of rotations around the third axis, and S^2 with the coset space $SU(2)/U(1)$, the Hopf map is the natural coset projection.

As with the other examples in Chapter 2, in order to make the theta sectors manifest we have to add a suitable topological term to the action. We consider only fields for which the action is finite. Such fields have to go to a constant at spacetime infinity, so we can compactify spacetime to S^3. The functional integral is now over maps from S^3 to S^2. As we have seen above, such maps fall into homotopy classes characterized by an integer, the Hopf invariant H, which is equal to one for the Hopf map. When the sigma model is formulated in terms of two coordinates (as in Section 2.4) or three fields subject to a constraint (as in Equation (1.42)), the Hopf invariant turns out to be given by a nonlocal expression. We will therefore not write it down explicitly. Rather, we assume that the action contains the topological term θH. Then, using the general argument of (4.1), the spin of the $H = 1$ sigma model solitons is

$$s = \frac{\theta}{2\pi} \ . \tag{4.2}$$

4.2 Dyons

The Georgi-Glashow model is a non-abelian gauge theory with gauge group $SO(3)$. Even though this group is topologically different from $SU(2)$, it is again true that $\pi_3(SO(3)) = \mathbb{Z}$, so also this theory has θ-sectors. We shall now discuss the effect of the θ-parameter on the monopoles.

As we have discussed in Section 1.7.4, magnetic monopoles come in four-parameter families, depending on the coordinates of the center of mass, $\vec{x}_0$ and an internal angular parameter α When we try to quantize the system semi-classically, along the lines of the quantization of the kink in Section 1.1.2, the zero modes of the small fluctuation kinetic operator have to be traded for these moduli, that can then be treated as ordinary quantum mechanical degrees of freedom.

In the case of the translational zero mode, this corresponds to the quantization of a free particle with mass M, equal to the energy of the classical solution. (If we consider the Prasad-Sommerfield limit, $M = 4\pi f/e$.) Its Lagrangian is simply

$$\frac{1}{2}M\dot{\vec{x}}_0^2 \ .$$

The quantization of the fourth modulus is more interesting. By definition, moduli are flat direction of the static energy, so their Lagrangian cannot contain a potential term. In order to write the Lagrangian for $\alpha(t)$ we have to first understand how the classical fields depend on $\dot{\alpha}$. We recall that α parametrizes a “global” $SO(2)$ rotation around the direction of the Higgs field. Here “global” means that the parameter of the tranformation is constant, but since the direction of the Higgs field depends on position, this looks formally like a gauge transformation. Under an infinitesimal variation of the parameter α we must therefore have, in the gauge $A_0 = 0$, [2]

$$\delta\phi^a = 0 \ , \qquad \delta A_i^a = \frac{1}{ef}\delta\alpha D_i\phi^a \ .$$

Thus, the Higgs part of the classical Lagrangian does not contribute anything. In the Yang-Mills Lagrangian there will be a contribution coming from the electric components of the field strength.

$$E_i^a = \dot{A}_i^a = \frac{1}{ef}\dot{\alpha}D_i\phi^a \tag{4.3}$$

Inserting in the Lagrangian we find

$$L = \frac{1}{2}\int d^3x E_i^a E_i^a = \frac{\dot{\alpha}^2}{2f^2e^2}\int d^3x(D_i\phi^a)^2$$

Using the BPS condition (1.137), $\int d^3x(D_i\phi^a)^2 = M$, so we can rewrite the Lagrangian as

$$L = \frac{1}{2}\frac{M}{m_A^2}\dot{\alpha}^2 \ ,$$

where $m_a = fe$. Since $0 \leq \alpha \leq 2\pi$ is an angular variable, this can be interpreted as the Lagrangian of a rotator with moment of inertia $I = M/m_A^2$.

Now we observe that motion in the α-direction corresponds to the presence of an electric field:

$$\mathcal{E}_i = E_i^a\hat{\phi}^a = \dot{A}_i^a\hat{\phi}^a = \frac{\dot{\alpha}}{fe}\hat{\phi}^a D_i\phi^a = \frac{\dot{\alpha}}{fe}\hat{\phi}^a B_i^a = \frac{\dot{\alpha}}{fe}\mathcal{B}_i \ ,$$

where in the second last step we used the BPS condition. Integrating over a two-dimensional sphere at $r \to \infty$ we obtain

$$Q_E = \frac{\dot{\alpha}}{fe}Q_M \ . \tag{4.4}$$

We see that a monopole moving the direction of the angular modulus has an electric charge proportional to the velocity. Electrically charged monopoles are called *dyons*.

The quantum mechanical wave functions must be periodic in α, hence they must be of the form $\psi = \exp(ik\alpha)$ with $k \in \mathbb{Z}$. The momentum conjugate to α therefore has integer eigenvalues

$$\pi_\alpha\psi \equiv -i\frac{\partial\psi}{\partial\alpha} = k\psi$$

and the velocity has eigenvalues

$$\dot{\alpha} = \frac{m_A^2}{M}k \ .$$

[2] Note that the angle of rotation depends on r, being α at infinity and zero in the origin. This is necessary to have a smooth gauge transformation, but ultimately only the transformation at infinity matters.

Using this in (4.4) we see that the electric charge of the dyon is quantized:

$$Q_E = \frac{e}{4\pi} k \ , \qquad k \in \mathbb{Z} \tag{4.5}$$

Let us now add to the Lagrangian of the theory the topological term

$$S_T = \frac{\theta}{8\pi^2} \int d^4x E_i^a B_i^a \ .$$

This contributes to the Lagrangian of the modulus a term

$$L_\theta = \frac{\theta}{8\pi^2} \frac{\dot{\alpha}}{ef} \int d^3x B_i^a D_i \phi^a \ .$$

Integrating by parts, only the surface term survives, and it gives

$$L_\theta = \frac{\theta}{8\pi^2} \frac{\dot{\alpha}}{e} \int_{S^2_\infty} \mathcal{B} = \frac{\theta}{2\pi} \frac{\dot{\alpha}}{e} Q_M = \frac{\theta}{2\pi e^2} \dot{\alpha} \ .$$

Now the Lagrangian of the modulus reads

$$L = \frac{1}{2} \frac{M}{m_A^2} \dot{\alpha}^2 + \frac{\theta}{2\pi e^2} \dot{\alpha} \ .$$

The presence of the term linear in the time derivative changes the relation between velocity and momentum. The total Hamiltonian of the moduli is then

$$H = M + \frac{\vec{p}^2}{2M} + \frac{1}{2} \frac{m_A^2}{M} \left(\pi_\alpha - \frac{\theta}{2\pi e^2} \right) \ . \tag{4.6}$$

Repeating the preceding steps, we find

$$\dot{\alpha} = \frac{m_A^2}{M} \left(k - \frac{\theta}{2\pi e^2} \right)$$

and therefore

$$Q_E = \frac{e}{4\pi} \left(k - \frac{\theta}{2\pi e^2} \right)$$

In the presence of the θ angle, the quantized charge is shifted by a constant amount. This is known as the Witten effect.

4.3 Skyrmions as baryons

As discussed in Section 1.4, the nonlinear sigma model is quite successful in describing the low energy scattering of mesons. It is natural to ask, how should baryons be described in this theory? In a work that was way ahead of its time, Skyrme suggested that the nucleons be described by the solitons that were discussed in section 1.4. This may seem impossible at first, since the nonlinear sigma model is a purely bosonic theory and the baryons are spin 1/2 fermions.

However, it was also observed that in principle the skyrmions could be quantized as spin 1/2 fermions. To see this one has to look at the configuration space of the $SU(N)$ sigma model in three space dimensions, which is $\mathcal{Q} = \Gamma_*(S^3, SU(N))$. We have $\pi_0(\mathcal{Q}) = \pi_3(SU(N)) = \mathbb{Z}$, so $\mathcal{Q} = \cup_{n\in\mathbb{Z}}\mathcal{Q}_n$ is the disjoint union of infinitely many connected components characterized by the winding number. $\mathcal{Q}$ is a group under pointwise multiplication of maps; since the group action maps any connected component into any other, and the group action is given by diffeomorphisms, there follows that all connected components $\mathcal{Q}_n$ are diffeomorphic, and in particular have the same fundamental group. The fundamental group of the $n = 0$ component can be computed using the familiar rule $\pi_1(\Gamma_*(S^3, SU(N))_0) = \pi_4(SU(N))$. The discussion now follows different paths in the cases $N = 2$ and $N > 2$.

For $N = 2$, $\pi_4(SU(2)) = \mathbb{Z}_2$. As with the sigma model anyons of Section 4.1, the fundamental noncontractible loop in $\mathcal{Q}_1$ can be realized as a loop in the moduli space $\mathcal{M}$, which, according to the discussion in Section 1.4.2, contains a factor $SO(3)$. The nontrivial loop in $\mathcal{Q}_1$, whose homotopy class generates the first homotopy group, therefore consist of a rotation of the soliton by 2π. Skyrme's observation is that the configuration space of the fields in the soliton sectors is doubly connected and therefore, following the reasoning of section 2.2, the solitons could be quantized either as bosons or as fermions.

Let us also see how this result is obtained in the functional integral. In this case one works with space-time dependent fields. Imposing finiteness of the action, Euclidean spacetime can be compactified to S^4, so that the space that one is formally integrating on is $\Gamma_*(S^4, SU(2))$. This function space consists of exactly two connected components, corresponding to the two homotopy classes in $\pi_4(SU(2))$. One of them consists of homotopically trivial maps. The other contains a map that describes the following process: a skyrmion-antiskyrmion pair is created in the far past, the skyrmion is rotated by 2π and finally the pair annihilates again. To understand that this is the right map, note the following: This map tends to the identity at infinity in all directions, as is required for elements of $\Gamma_*(S^4, SU(2))$, and it describes the non-contractible loop in $\mathcal{Q}_0$.

As discussed in Section 2.7.1, in the functional integral, homotopically distinct classes of paths can be summed with arbitrary weights given by characters of $\pi_1(\mathcal{Q})$. In this case we have

$$Z = \int (dU)_0 e^{-S} \pm \int (dU)_1 e^{-S} \ , \tag{4.7}$$

where the two terms correspond to the two homotopy classes of paths and the sign of the second term is a character of $\mathbb{Z}_2$. Choosing the lower sign corresponds to quantizing the skyrmion as a fermion.

For $N > 2$, $\pi_4(SU(N)) = 0$ and the previous arguments do not apply. However, in this case $\pi_2(\Gamma_*(S^3, SU(N)))$=$\pi_5(SU(N)) = \mathbb{Z}$ and one can add to the action a new topological term, the Wess–Zumino–Witten term defined in section 3.2. In fact, this term is necessary for phenomenological reasons. The expansion of the usual sigma model action only contains terms with even numbers of mesons. On the other hand, there exist processes such as

$$K^+K^- \to \pi^+\pi^0\pi^-$$

with an odd number of external legs, which are allowed in QCD. If such a process is to be accounted for in the sigma model, then there must exist other terms in the effective theory Lagrangian whose expansion contains an odd number of mesons. This is precisely the WZW term, as one can see by inserting (1.81) in (3.22):

$$c\frac{2}{15\pi^2 f^2}\int d^4x\, \varepsilon^{\mu\nu\rho\sigma}\pi^a\partial_\mu\pi^b\partial_\nu\pi^c\partial_\rho\pi^d\partial_\sigma\pi^e B_{abcde} + \text{higher order terms}$$

where $B_{abcde} = \mathrm{tr}(T_aT_bT_cT_dT_e)$. This term can be used to describe the process of two kaons going to three pions.

Now we can address the question of the spin of the skyrmion. Using the general argument (4.1), the spin of the soliton is given by the WZW action evaluated for a slowly rotating skyrmion. A direct calculation shows that $\Gamma(U_{cl}) = c\pi$, so the have integer spin when c is even and half-integer spin when c is odd. The question that still remains is then: what is the value of c in the real world? This question can be answered by comparing the axial anomaly of QCD with that of the chiral model.

Chapter 5

Anomalies

WARNING:
IN THIS CHAPTER THERE ARE STILL INCONSISTENCIES IN THE NOTATION

It is sometimes impossible to quantize a system preserving all its classical symmetries. One then says that there is an anomaly. There are various types of anomalies, both from a mathematical and from a physical point of view. One can distinguish between anomalies for discrete groups of tranformations, for finite dimensional continuous groups (Lie groups) and for infinite dimensional groups. Another distinction of a more physical nature is whether the invariance that cannot be preserved is a genuine symmetry of the system (meaning that it consists of transformations that can be physically observed) or a gauge invariance (in which case the transformed object is physically indistinguishable from the original one).

In the case of a (finite- or infinite-dimensional) continuous group, the anomaly manifests itself in the failure of a conservation law. The physical implications of the anomaly are then very different depending on the nature of the classical conservation law. In the case of a genuine continuous symmetry (typically a symmetry with constant transformation parameters) the current of interest is the Noether current. The anomaly appears as a nonzero divergence of this current and does not have harmful consequences. The prototype is the Adler-Bell-Jackiw (ABJ) anomaly in the axial current. This case will be discussed in sections 5.1 and 5.2. On the other hand in the case of a current coupled to gauge fields (when the transformation parameters are functions on spacetime), failure of current conservation jeopardizes the consistency of the theory. We will generally refer to such anomalies as “gauge anomalies”. [1] In these cases the anomaly is a pathology of the theory and requiring the absence of anomalies becomes a powerful tool for selecting physically viable theories.

One of the most important examples of anomalies affects scale transformations. The running of couplings breaks scale invariance even in theories that are classically scale invariant. This anomaly manifests itself in a nonvanishing trace of the energy-momentum tensor and is therefore usually called the trace anomaly. We shall not consider the trace anomaly, nor other anomalies that affect space-time transformations. We shall restrict ourselves to anomalies for internal transformations, which are closely related to the topics developed in the preceding chapters.

From the mathematical point of view the existence of anomalies is related to a very rich vein of algebraic and geometrical results. In particular, global (axial) anomalies are intimately related to the index theorem for the Dirac operator and the existence of gauge anomalies can be proven using a generalization of the index theorem involving two-parameter families of Dirac operators. The whole subject can also be recast in cohomological language. Altogether, there are few other fields where the progress of physics and mathematics have been so close.

5.1 The axial anomaly

Here we consider the ABJ anomaly, which was historically one of the earliest examples of anomaly. It appears in the case of a single complex massless Dirac field coupled to electromagnetism. One finds that it is impossible to satisfy simultaneously the conservation of the vector and of the axial symmetry. Therefore this theory is anomalous. Depending on the regularization we choose, we can decide which symmetry is actually realized in the quantum theory: since the vector symmetry is in some sense more important than the axial symmetry, one usually preferes to give up the latter. Once this is understood, one then says that the axial symmetry is anomalous. In the next section we shall generalize the results to the case of a multiplet of fermions fields, carrying a representation of some global symmetry group.

We begin by setting up some notation. The action for a fermion in n spacetime dimensions, coupled to an external electromagnetic potential A_μ is

$$S_F(\psi,\bar\psi,A) = -\int d^n x\,\bar\psi\,(\gamma^\mu D_\mu + m)\,\psi\ . \tag{5.1}$$

[1] The term “local anomaly” is also used, referring to the fact that such an anomaly affects local gauge transformations. By the same token, then, the anomalies for finite dimensional symmetry groups could be called “global anomalies”. Unfortunately, the same terms is also used for certain anomalies that have to do with transformations that are not homotopic to the identity.

Our conventions for the gamma matrices are given in Appendix XXX. In particular, we recall the chirality operator γ^A (often called γ^5), that anticommutes with the gamma matrices and squares to one.

The action (5.1) is invariant under the (global) vector transformations

$$\psi' = e^{-i\alpha}\psi \;\; ; \qquad\qquad \bar{\psi}' = \bar{\psi}e^{i\alpha} \tag{5.2}$$

with associated vector current

$$j^\mu_V = \bar{\psi}\gamma^\mu\psi \; . \tag{5.3}$$

One is also interested in the axial transformations

$$\psi' = e^{i\beta\gamma^A}\psi \;\; ; \qquad\qquad \bar{\psi}' = \bar{\psi}e^{i\beta\gamma^A} \; . \tag{5.4}$$

Under an infinitesimal axial transformation the variation of the action (5.1) is

$$\delta S = -2i\beta m \int d^n x\, \bar{\psi}\gamma^A\psi \; , \tag{5.5}$$

showing that (5.4) is a symmetry of the Dirac action only if the mass vanishes. In this case the corresponding Noether current is

$$j^\mu_A = \bar{\psi}\gamma^A\gamma^\mu\psi \; . \tag{5.6}$$

In general, the divergence of this current is

$$\partial_\mu j^\mu_a = -2m\bar{\psi}\gamma^A\psi \; , \tag{5.7}$$

so the axial current is conserved only if the mass is zero: in the following we will consider the massless situation without losing any generality.

5.1.1 Point splitting

Let us now quantize the theory and ask whether $\partial_\mu\langle j^\mu_A\rangle = 0$ in the massless case (it is understood that j^μ_A now denotes the quantum operator corresponding to the axial current (5.6) and the brackets its vacuum expectation value in the A_μ background). The formal manipulations leading to the result (5.7) cannot be trusted because the operator j^μ_A is the product of two fields at the same spacetime point and is therefore singular: in other words the naive definition of composite operator in quantum field theory leads to divergent result. One has to resort to some kind of regularization. Physically, the most transparent regularization for problems of this type is point splitting: the axial current operator is defined to be the $\epsilon \to 0$ limit of the following expression:

$$j^\mu_A(x,\epsilon) = \bar{\psi}\left(x + \frac{\epsilon}{2}\right)\gamma^A\gamma^\mu \exp\left(iea\int\limits_{x-\frac{\epsilon}{2}}^{x+\frac{\epsilon}{2}} A\right)\psi\left(x - \frac{\epsilon}{2}\right) \; . \tag{5.8}$$

The regulator ϵ is a vector representing an infinitesimal displacement in space-time; in order not to break Lorentz invariance it will be necessary, in taking the limit $\epsilon \to 0$, to average over all directions.

The exponential contains an arbitrary parameter a. Under the local gauge transformation

$$\psi'(x) = e^{-i\alpha(x)}\psi(x) \quad ; \qquad \bar{\psi}'(x) = e^{i\alpha(x)}\bar{\psi}(x) \quad ; \qquad A'_\mu = A_\mu - \frac{1}{e}\partial_\mu \alpha$$

it becomes

$$\exp\left(iea \int\limits_{x-\frac{\epsilon}{2}}^{x+\frac{\epsilon}{2}} A' \right) = \exp\left(-ia\alpha\left(x+\frac{\epsilon}{2}\right)\right) \exp\left(iea \int\limits_{x-\frac{\epsilon}{2}}^{x+\frac{\epsilon}{2}} A \right) \exp\left(ia\alpha\left(x-\frac{\epsilon}{2}\right)\right) .$$

For $a = 1$ the two outer exponentials cancel the transformation of the fermions and the regulated current (5.8) is gauge invariant. To compute the divergence of the current in the quantum theory, the prescription is to take first the divergence and then the limit.

Using the equations of motion

$$\begin{aligned} \gamma^\mu \partial_\mu \psi &= ie\gamma^\mu A_\mu \psi - m\psi \ , \\ \partial_\mu \bar{\psi}\gamma^\mu &= -ie\bar{\psi}\gamma^\mu A_\mu + m\bar{\psi} \ , \end{aligned} \tag{5.9}$$

one finds that

$$\partial_\mu j^\mu_A(x,\epsilon) = iej^\mu_A(x,\epsilon) \left[A_\mu\left(x-\frac{\epsilon}{2}\right) - A_\mu\left(x+\frac{\epsilon}{2}\right) + a\partial_\mu \int\limits_{x+\frac{\epsilon}{2}}^{x+\frac{\epsilon}{2}} A \right] , \tag{5.10}$$

plus the classical term given in (5.7), that we shall disregard from now on. For small ϵ the square bracket can be expanded as

$$\epsilon^\alpha \left(\partial_\alpha A_\mu - a\partial_\mu A_\alpha\right) + O(\epsilon^2) = \epsilon^\alpha (F_{\alpha\mu} + (1-a)\partial_\mu A_\alpha) + O(\epsilon^2).$$

Note that the classical result is recovered if one takes the limit $\epsilon \to 0$ naively. We have already said that this incorrect, being the limit in $j^\mu_A(x,\epsilon)$ singular. To see this concretely, let us take the vacuum expectation value of both sides of (5.10). We find

$$\langle \partial_\mu j^\mu_A(x,\epsilon)\rangle = -ie\langle j^\mu_A(x,\epsilon)\rangle \epsilon^\alpha (F_{\alpha\mu} + (1-a)\partial_\mu A_\alpha) + O(\epsilon^2) \ . \tag{5.11}$$

5.1.2 Calculation of the anomaly

We will now show that the second term does not vanish in the limit, because the coefficient of ϵ is divergent. The v.e.v. on the r.h.s. can be rewritten, for $\epsilon^0 > 0$,

$$\begin{aligned} \langle j^\mu_A \rangle &= -\mathrm{Tr}\gamma^A\gamma^\mu \left\langle T\psi\left(x-\frac{\epsilon}{2}\right)\bar{\psi}\left(x+\frac{\epsilon}{2}\right)\right\rangle e^{iea\int A} \\ &= -\mathrm{Tr}\gamma^A\gamma^\mu G\left(x-\frac{\epsilon}{2}, x+\frac{\epsilon}{2}\right) e^{iea\int A}, \end{aligned} \tag{5.12}$$

where the trace is over Dirac indices, T denotes time ordering and $G(x,y)$ denotes the Dirac propagator in an external electromagnetic field, defined by

$$\gamma^\mu \left(\partial_\mu - ieA_\mu(x)\right) G(x,y) = \delta(x-y) \ . \tag{5.13}$$

Note that due to the presence of the external field, G is not simply a function of the difference $x-y$. Let $S(x-y)$ denote the free Dirac propagator, which is defined by equation (5.13) with A_μ set equal to zero. Multiplying (5.13) by S on the left (in the sense of kernel composition) and using

$$-\frac{\partial}{\partial y^\mu} S(x-y)\gamma^\mu = \delta(x-y)$$

one finds the equation

$$G(x,y) = S(x-y) + ie\int dz S(x-z)\gamma^\mu A_\mu(z) G(z,y) \ . \tag{5.14}$$

This equation can be solved by iteration:

$$\begin{aligned} & G\left(x-\frac{\epsilon}{2}, x+\frac{\epsilon}{2}\right) \\ & = S(-\epsilon) + ie\int d^{2n}y S\left(x-\frac{\epsilon}{2}-y\right)\gamma^\mu A_\mu(y) S\left(y-x-\frac{\epsilon}{2}\right) \\ & -e^2\int dydz S\left(x-\frac{\epsilon}{2}-y\right)\gamma^\mu A_\mu(y) S(y-z)\gamma^\nu A_\nu(z) S\left(z-x-\frac{\epsilon}{2}\right) + \dots \end{aligned} \tag{5.15}$$

This is represented graphically in Fig. XXX, where the single lines represent free propagators, and the crosses represent insertions of the external field.

At this point the analysis begins to depend upon the dimension of spacetime. The free propagator can be written in Fourier space

$$S(x) = \int \frac{d^{2n}p}{(2\pi)^{2n}} e^{-ip\cdot x} \frac{\gamma^\rho p_\rho}{p^2} = \gamma^\rho S_\rho(x)$$

and inserting in (5.15) we see that the first term diverges for $\epsilon \to 0$ like $\epsilon^{-(n-1)}$, the second like $\epsilon^{-(n-2)}$ and so on, until the n–th term, which diverges logarithmically. All subsequent terms are convergent.

The expression (5.15) contains an odd number of gamma matrices. When inserted in (5.12) we have a trace of γ^A times an even number of gamma matrices. The first nonzero term occurs when the number of gamma matrices is equal to the spacetime dimensions. This leading term is proportional to the totally antisymmetric Levi-Civita symbol. It is always linearly divergent and inserted in (5.12) it gives a finite contribution. The subsequent logarithmically divergent and finite terms of G are irrelevant in the limit $\epsilon \to 0$.

Let us discuss first the case $n=2$. We have to take into account only the first term of the expansion (5.15) so the right hand side of (5.11) becomes:

$$ie\mathrm{Tr}[\gamma^A\gamma^\mu\gamma^\nu] S_\nu(-\epsilon)\epsilon^\alpha F_{\alpha\mu} + O(\epsilon).$$

The trace gives

$$\mathrm{Tr}\left[\gamma^A\gamma^\mu\gamma^\nu\right] = 2i\varepsilon^{\mu\nu}$$

while for the fermionic propagator we obtain:

$$S_\nu(-\epsilon) = -\frac{1}{2\pi}\frac{\epsilon_\nu}{\epsilon^2}. \tag{5.16}$$

Taking the limit in ϵ and averaging over all the directions gives

$$\lim_{\epsilon\to 0}\frac{\epsilon_\alpha\epsilon_\nu}{\epsilon^2} = \frac{1}{2}\eta_{\alpha\nu}\ . \tag{5.17}$$

The factor 1/2 comes from imposing that both sides of the equation have the same trace. In this way we arrive at the following expression for the anomaly:

$$\langle\partial_\mu j_A^\mu\rangle = \frac{e}{2\pi}\varepsilon^{\mu\nu}F_{\mu\nu}. \tag{5.18}$$

We note that this is twice the integrand of the topological invariant c_1 defined in (2.24).

A slightly more complicated calculation leads to the following result in four dimensions:

$$\partial_\mu j_A^\mu = \frac{e^2}{16\pi^2}\varepsilon^{\mu\nu\rho\lambda}F_{\mu\nu}F_{\rho\lambda}. \tag{5.19}$$

Some remarks are now in order: first of all we notice that the anomaly appears to be a finite object and this is not a coincidence. Looking at the previous computations we learn that a *classical* zero with a *quantum* infinity give rise to a finite term in the conservation laws: this is always the way in which anomalies arise in quantum field theory. Moreover if we restore the Planck constant h we find that the coefficient of the anomaly depends linearly on it, manifesting the pure quantum mechanical nature of the phenomenum.

At this point one may ask what is the destiny of the vector current (5.3) and of its conservation law: it is easy to understand that no anomaly arises using this regularization procedure. The trace over Dirac matrices produces a tensor that, after averaging over the ϵ directions, saturates two symmetric indices with the strenght field generated by the exponantial string: a simple computation in $d = 2, 4$ may convince the reader.

5.1.3 Other axial anomalies

In the preceding calculation we have considered a fermion coupled to the electromagnetic field. Let us now generalize the previous results to the case of a multiplet of fermions ψ^A, carrying a representation of a global symmetry group G (in realistic applications, G is $SU(N)$, and is called the flavor group). The matrices representing the generators in the Lie algebra of G will be denoted T_a. They are assumed to be antihermitian and to satisfy

$$T_a^\dagger = -T_a\ ,\quad [T_a, T_b] = f_{ab}{}^cT_c\ ,\quad f_{abc}^* = f_{abc}\ ,\quad \mathrm{tr}T_aT_b = -\frac{1}{2}\delta_{ab}\ . \tag{5.20}$$

For example for $SU(2)$, $T_a = -\frac{i}{2}\sigma_a$ whereas for $SU(3)$, $T_a = -\frac{i}{2}\lambda_a$, where λ_a are the Gell-Mann matrices. We will not write explicitly the indices of the fermions, neither the spinor indices nor the indices pertaining to the representation of G. Thus ψ will now denote a column vector on which the group acts by left multiplication and the Dirac conjugate $\bar{\psi}$ is a row vector on which the group acts by right multiplication. The action can be written again as in (5.1), with our new interpretation of symbols. The field

$$A_\mu = A_\mu^a T_a \tag{5.21}$$

is an external (non-dynamical) non-abelian gauge field that couples to the fermions. [2] This action has a global symmetry $U(1)_V \times G_V$, where $U(1)_V$ is defined by (5.2), with all components of ψ transforming by the same phase, and G_V is defined by

$$\psi' = g^{-1}\psi \ ; \qquad \bar{\psi}' = \bar{\psi} g \ , \tag{5.22}$$

where $g = e^{-i\alpha^a T_a}$. The Noether current associated to $U(1)_V$ is (5.3), and the Noether current associated to G_V is

$$j_{V}{}_a^\mu = \bar{\psi} T_a \gamma^\mu \psi \ . \tag{5.23}$$

It transforms according to the adjoint representation. In the case $m = 0$ (5.1) is also invariant under a group $U(1)_A \times G_A$, where $U(1)_A$ is defined by (5.4), and G_A is given by the transformations

$$\psi' = e^{i\alpha^a T_a \gamma^A}\psi \ ; \qquad \bar{\psi}' = \bar{\psi} e^{i\alpha^a T_a \gamma^A} \ . \tag{5.24}$$

The Noether current corresponding to $U(1)_A$ is (5.6), and the current associated to G_A is

$$j_{A}{}_a^\mu = \bar{\psi} T_a \gamma^A \gamma^\mu \psi \ . \tag{5.25}$$

As in the abelian case, the vector and the axial current cannot be simultaneously conserved. The anomaly can be computed using the method described above, the only difference consists in replacing the exponential in (5.8) by a path ordered exponential, and one finds

$$\partial_\mu j_A^\mu = \frac{ie}{2\pi}\epsilon^{\mu\nu}\mathrm{tr}F_{\mu\nu} \qquad \text{for} \quad n = 2 \tag{5.26}$$

$$\partial_\mu j_A^\mu = \frac{e^2}{16\pi^2}\varepsilon^{\mu\nu\rho\sigma}\mathrm{tr}F_{\mu\nu}F_{\rho\sigma} \qquad \text{for} \quad n = 4 \tag{5.27}$$

$F_{\mu\nu}$ now being the non-abelian field strenght. Note that the r.h.s. of (5.26) is zero for the group $SU(N)$, and the r.h.s. of (5.27) is twice the topological invariant c_2. The argument given above can be generalized straightforwardly

[2] In the Abelian case there is a single anti-Hermitian generator $T_1 = i/\sqrt{2}$ and we should write $A_\mu = A_\mu^1 i/\sqrt{2}$. We will avoid this awkward notation, by treating this case separately.

also to the calculation of the (covariant) divergence of the non-singlet current (5.25). The result is

$$
\begin{aligned}
(D_\mu j_A^\mu)_a &= \frac{e}{2\pi}\varepsilon^{\mu\nu}\mathrm{tr}\,T_a F_{\mu\nu}\ , &\text{for } \ n=2 \qquad (5.28)\\
(D_\mu j_A^\mu)_a &= \frac{ie^2}{16\pi^2}\varepsilon^{\mu\nu\rho\sigma}\mathrm{tr}\,T_a F_{\mu\nu}F_{\rho\sigma}\ , &\text{for } \ n=4 \qquad (5.29)
\end{aligned}
$$

We note that the factors of i in the formulas (5.26) and (5.29) are needed for these traces to be real, since they contain an odd number of antihermitian matrices. We end the section stressing again that the axial symmetry we have discussed is a *global* symmetry: no local invariance has been allowed in the classical action for the axial transformation. The anomalies we have studied are therefore global anomalies.

5.2 The index theorem

The careful reader will have noticed that the anomaly of the G_A-current is twice the integrand of the topological invariants c_1 and c_2 introduced in sections XXX. This is no coincidence, and in fact the axial anomaly can be shown to affect the phenomenon of theta vacua in gauge theories in the presence of fermion fields. We begin by observing that since the fermionic configuration space is linear, the general topological arguments for the existence of theta sectors given in sections 2.4 and 2.5 continue to hold (the same remark had been made in section 2.8 concerning the coupling of the gauge theory to scalars). However, *massless* fermions have a dramatic influence on the dynamics of such theories: it turns out that the v.e.v.s of gauge invariant operators become independent of θ. The proof of this statement relies on a profound mathematical result, known as the Atyiah-Singer index theorem, that encodes the topological meaning of the axial anomaly.

In order to state this theorem in a simple form, we pass to the Euclidean signature and assume that spacetime is even dimensional, compact and without boundary. As usual, this can be achieved by imposing suitable boundary conditions on all fields: for istance we can take S^n as the compact space-time. In this case the Dirac operator:

$$D = \gamma^\mu\left(i\partial_\mu + A_\mu\right)$$

is self-adjoint and its spectrum is discrete in the space V of fermionic fields such that $\int d\mu\bar{\psi}\psi$ is finite, $d\mu$ being the natural measure of S^n. Since $(\gamma^A)^2 = 1$, we can split

$$V = V_+ \oplus V_-$$

where $V_\pm$ are eigenspaces of γ^A with eigenvalues ± 1 respectively. Now let $\{\psi_n\}$ be a complete set orthonormal eigenfunctions of D:

$$D\psi_n = \lambda_n\psi_n \ \ ; \qquad\qquad \int d\mu\,\bar{\psi}_m\psi_n = \delta_{mn}\ .$$

Since γ^A anticommutes with γ^μ, if ψ is in V_+, $D\psi$ is in V_-, and vice-versa. So, if $\lambda_n \neq 0$, ψ_n cannot be an eigenfunction of γ^A. However, the eigenfunctions with zero eigenvalue (the zero modes) can be chosen to belong either to V_+ or V_-. We will call n_+ and n_- the numbers of linearly independent zero modes of D with positive and negative chirality respectively. The index theorem states that

$$n_+ - n_- = c_n \ , \tag{5.30}$$

where c_n is the topological invariant defined by (2.4.3) and (2.5.1) for $n = 2$ and 4 respectively. The proof of this theorem is beyond the scope of these notes. Here we shall instead give a heuristic derivation that highlights its connection to the axial anomaly.

5.2.1 Derivation from the anomaly

Let us start with a massive fermion interacting with an external gauge field via the vector current (5.23) (or (5.3) if $G = U(1)$) in four dimension. Here the mass plays the role of a regulator and will be sent to zero in the end. From the anomaly, the divergence of the axial current is

$$\partial_\mu \langle j^\mu_A \rangle = -2m \langle \bar\psi \gamma^A \psi \rangle + \frac{i}{8\pi} \mathrm{Tr} F_{\mu\nu}{}^* F^{\mu\nu}$$

(The factor i appears because we are now in Euclidean signature.) We integrate both sides and take the expectation value in the fermionic vacuum. The l.h.s. gives $\int d^4x \partial_\mu j^\mu_A = \int d\Sigma_\mu j^\mu_A = 0$ because the fermion field is massive (in this case the current has a smooth behaviour in the limit or if were in the compact situation no singularity occurs). From the r.h.s. we obtain

$$2m \int d^4x \, \langle \bar\psi \gamma^A \psi \rangle = -2ic_2 \ . \tag{5.31}$$

Now we want to evaluate the v.e.v. on the left:

$$\langle \bar\psi \gamma^A \psi \rangle = \frac{\int (d\psi d\bar\psi) e^{-S_F} \left(\int d^4x \, \bar\psi \gamma^A \psi \right)}{\int (d\psi d\bar\psi) e^{-S_F}} \ . \tag{5.32}$$

The eigenfunctions of D defined in () are also eigenfunctions of $D - im$ with eigenvalues $\lambda_n - im$. Thus we can decompose

$$\psi(x) = \sum_n a_n \psi_n(x) \ ; \qquad \bar\psi(x) = \sum_n \bar a_n \bar\psi_n(x) \ ,$$

$$S_F(\psi, \bar\psi, A) = \sum_n \bar a_n a_n (\lambda_n - im) \ ; \qquad (d\psi d\bar\psi) = \prod_n da_n d\bar a_n$$

The functional integrals in (5.32) can be performed using Berezin's rules for the integration over fermion fields:

$$\int da_n a_m = \delta_{nm} \ ; \ \int d\bar a_n \bar a_m = \delta_{nm}$$

The numerator of (5.32) is

$$\begin{aligned}&\prod_n \int da_n d\bar{a}_n \left(1-\bar{a}_n a_n(\lambda_n - im)\right) \left[\sum_{rs} \bar{a}_r a_s \int d^4y\, \bar{\psi}_r(y)\gamma^A \psi_s(y)\right]\\ &= \sum_r \int d^4y \bar{\psi}_r(y)\gamma^A\psi_r(y) \prod_{s\neq r}(\lambda_s - im)\ , \qquad (5.33)\end{aligned}$$

while the denominator is

$$\begin{aligned}\int (d\psi d\bar{\psi})e^{-S_F} &= \prod_n \int d\bar{a}_n da_n (1-(\lambda_n - im)\bar{a}_n a_n)\\ &= \prod_n (\lambda_n - im) = \det(D-im)\ . \qquad (5.34)\end{aligned}$$

In the above formulas the formal determinant of the Dirac operator appears. It can be given a meaning by choosing a specific regularization procedure. In any case, for the calculation we are interested in we do not need its details, as we shall see. Therefore,

$$\int d^4y\, \langle\bar{\psi}(y)\gamma^A\psi(y)\rangle = \sum_r \frac{\int d^4y\, \bar{\psi}_r(y)\gamma^A\psi_r(y)}{\lambda_r - im}$$

Since D anticommutes with γ^A, if ψ_n is an eigenfunction with eigenvalue λ_n, $\gamma^A\psi_n$ is an eigenfunction with eigenvalue $-\lambda_n$. Therefore, using the orthogonality (5.2) we find that if $\lambda_s \neq 0$, $\int d^4y\, \bar{\psi}_s(y)\gamma^A\psi_s(y) = 0$. On the other hand, if ψ_n is a zero mode, also $\gamma^A\psi_n$ is a zero mode, and we have chosen the zero modes to have definite chirality. Therefore, if $\lambda_n = 0$, $\int d^4y\, \bar{\psi}_s(y)\gamma^A\psi_s(y) = \pm 1$, depending on the chirality. So we find that

$$\int d^n y\, \langle\bar{\psi}(y)\gamma^A\psi(y)\rangle = -\frac{1}{im}(n_+ - n_-)\ ,$$

and inserting back in (5.31) we obtain the index theorem (5.30). We have shown in this way that with certain assumptions about the boundary conditions, the index theorem follows from the existence of the axial anomaly. Conversely we can understand the apperence of the axial anomaly not as an accident of the quantum field theory but as a consequence of the dependence of the spectrum of the Dirac operator on the topology of the gauge field.

5.2.2 Consequences for the theta sectors

In the previous discussion the gauge field was treated as a fixed background. Let us now see the implications of these results for the quantization of the full theory, with dynamical gauge field. We are specifically interested in the tunnelling amplitude through the noncontractible path in $\mathcal{Q}$, since this amplitude was responsible for the θ–dependence of the vacuum energy, as we have seen in Sections 2.7-9. This amplitude is given by the Euclidean functional integral

$$\int_{c_{n/2}\neq 0} (dAd\bar{\psi}d\psi)e^{-S(\psi,\bar{\psi},A)} = \int_{c_{n/2}\neq 0} (dA)e^{-S_{\rm YM}(A)}{\rm det}(D)\ ,$$

where we have used (5.34). The integral is restricted to those gauge field configurations which have nonvanishing topological invariant. Now the index theorem (5.30) implies that for these fields there must be at least one zero mode, and therefore the determinant of the Dirac operator is identically zero, for all A. This implies that the tunnelling amplitude is zero, and therefore the theta vacua are all degenerate, in sharp contrast to what happened without massless fermions. Note that in deriving this result we did not have to make use of the WKB approximation.

There is also a formal argument that directly relates the existence of the anomaly to the degeneracy of the theta vacua. Consider a gauge- and chiral-invariant operator $\mathcal{O}$. The v.e.v. of $\mathcal{O}$ in the vacuum specified by a value θ can be computed as

$$\langle\mathcal{O}\rangle = \frac{\delta \log Z_\theta(J)}{\delta J}$$

where

$$Z_\theta(J) = \int (dAd\bar{\psi}d\psi)e^{-S_{YM}(A)+i\theta c_2(A)-S_F(\psi,\bar{\psi},A)+\int d^4x\, J\mathcal{O}}\ .$$

To simplify the notation we do not write explicitly the gauge fixing and ghost terms, since they are irrelevant for what follows. A priori, the v.e.v. of $\mathcal{O}$ seems to depend upon θ. Let us now examine how Z_θ behaves under the axial transformations (5.4). Since the action is invariant under chiral transformations, if the measure was also invariant, the whole functional integral would be invariant. But this is incompatible with the statement that there is an anomaly, so the measure cannot be invariant. Without discussing this in detail, we can infer how the measure has to transform, from our knowledge of the anomaly. Under an infinitesimal axial transformation $\delta\psi = \delta\alpha\gamma^A\psi$ the transformation of the measure can be written as $(d\bar{\psi}d\psi) = (d\bar{\psi}'d\psi')e^{\delta S}$. The variation of the action can be inferred from Noether's theorem to be

$$\delta S = \int d^4x\, \delta\mathcal{L} = \int d^4x\, \partial_\mu j^\mu_V = 2i\delta\alpha c_2(A)\ .$$

Therefore the effect of an axial transformation on the fermion fields is equivalent to a shift of θ by $2\delta\alpha$:

$$Z_\theta(J) = \int (dAd\bar{\psi}'d\psi')e^{-S_{YM}(A)+i(\theta+2\alpha)c_2(A)-S_F(\psi',\bar{\psi}',A)+\int d^4x\, J\mathcal{O}} = Z_{\theta+2\alpha}(J)\ .$$

In the last step we have replaced ψ' and $\bar{\psi}'$ by ψ and $\bar{\psi}$, since these are integration variables. The conclusion is that the value of θ is irrelevant: the expectation value of every gauge and chiral invariant observable is independent of θ.

We emphasize once again that this does not mean that that there are no theta sectors anymore. The topological arguments remain valid. One can also argue that the theta sectors have to be still distinct in order for the cluster property be satisfied. All that has happened is that the the sectors are now completely degenerate.

5.3 Gauge anomalies

Next we consider anomalies in a current that couples to gauge fields. We will call these "gauge anomalies". Let us consider quite generally a fermionic current J_a^μ coupled to a gauge field A_μ^a via an interaction term $\mathcal{L}_I = J_a^\mu A_\mu^a$. To begin with, we do not specify whether J is a vector, axial or other current. All we assume is that the classical action S is gauge invariant. The classical current can be defined as

$$J_a^\mu = \frac{\delta S}{\delta A_\mu^a} \ .$$

Functional integration over the fermions yields a contribution to the action for the gauge fields:

$$W\left[A\right] = -i \ln \int (d\psi d\bar{\psi}) e^{iS_F\left[\psi,\bar{\psi},A\right]} \ .$$

We will loosely refer to W as the effective action. The expectation value of the current in the fermionic vacuum is given by

$$\langle J_\mu^a \rangle = \frac{\delta W}{\delta A_\mu^a} \ . \tag{5.35}$$

For an infinitesimal gauge transformation parameter $\epsilon = \epsilon^a T^a$, define the operator

$$\delta_\epsilon = \int d^{2n}x \, D_\mu \epsilon^a(x) \frac{\delta}{\delta A_\mu^a(x)} \ .$$

It can be thought of as a vector tangent to the gauge orbit through A in the space of all gauge fields. The derivative of W in the direction of this vector is

$$\begin{aligned} \delta_\epsilon W\left[A\right] &= \int d^{2n}x \, D_\mu \epsilon^a(x) \frac{\delta W}{\delta A_\mu^a(x)} \\ &= \int d^{2n}x \, D_\mu \epsilon^a(x) \langle J_a^\mu \rangle \\ &= -\int d^{2n}x \, \epsilon^a(x) \langle D_\mu J_a^\mu(x) \rangle \ . \end{aligned} \tag{5.36}$$

Since ϵ is arbitrary, we see that gauge invariance of the effective action W is equivalent to the covariant conservation of the current. This is a very important property in the full quantum gauge theory: in perturbation theory, it ensures unitarity and renormalizability.

In section 5.1 we discussed the case when $J^\mu = j_V^\mu$. We proved that there exists a quantization scheme that preserves the conservation of this current, violating the conservation of the axial current. Since the axial current was not coupled to gauge fields, no problem arose in that case.

The situation is different if the coupling is not purely vectorial. The most general situation is to have the vector and axial currents coupled to two different gauge fields

$$\mathcal{L}_I = j_V{}^\mu_a A_V{}^a_\mu + j_A{}^\mu_a A_A{}^a_\mu \ .$$

In order to ensure the invariance of the action under the vector and axial gauge transformations (5.22) and (5.24) the gauge fields have to transform as follows:

$$\begin{aligned} \delta_{V\epsilon} A_{V\mu} &= D_\mu \epsilon \ ; \qquad \delta_{V\epsilon} A_{A\mu} = [A_{A\mu}, \epsilon] \ \ ; && (5.37) \\ \delta_{A\epsilon} A_{V\mu} &= [A_{A\mu}, \epsilon] \ ; \qquad \delta_{A\epsilon} A_{A\mu} = D_\mu \epsilon \ \ . && (5.38) \end{aligned}$$

These transformations obey the following algebra:

$$[\delta_{V\epsilon_1}, \delta_{V\epsilon_2}] = \delta_{V[\epsilon_1,\epsilon_2]} \ ; \qquad (5.39)$$

$$[\delta_{V\epsilon_1}, \delta_{A\epsilon_2}] = \delta_{A[\epsilon_1,\epsilon_2]} \ ; \qquad (5.40)$$

$$[\delta_{A\epsilon_1}, \delta_{A\epsilon_2}] = \delta_{V[\epsilon_1,\epsilon_2]} \ ; \qquad (5.41)$$

The vector and axial transformations are deeply entangled. It is convenient to define left and right currents

$$j_L{}^\mu_a = \frac{j_V{}^\mu_a - j_A{}^\mu_a}{2} = \bar\psi T_a \gamma^\mu \left(\frac{1-\gamma^A}{2}\right)\psi \qquad (5.42)$$

$$j_R{}^\mu_a = \frac{j_V{}^\mu_a + j_A{}^\mu_a}{2} = \bar\psi T_a \gamma^\mu \left(\frac{1+\gamma^A}{2}\right)\psi \qquad (5.43)$$

and left and right gauge fields

$$A_L{}^\mu_a = A_V{}^a_\mu - A_A{}^a_\mu \qquad (5.44)$$

$$A_R{}^\mu_a = A_V{}^a_\mu + A_A{}^a_\mu \qquad (5.45)$$

In term of these new variables the interaction reads

$$\mathcal{L}_I = j_L{}^\mu_a A_L{}^a_\mu + j_R{}^\mu_a A_R{}^a_\mu \ ,$$

and defining $\delta_L = \delta_V - \delta_A$ and $\delta_R = \delta_V + \delta_A$ the algebra becomes

$$[\delta_{L\epsilon_1}, \delta_{L\epsilon_2}] = \delta_{L[\epsilon_1,\epsilon_2]} \ ; \qquad (5.46)$$

$$[\delta_{L\epsilon_1}, \delta_{R\epsilon_2}] = 0 \ ; \qquad (5.47)$$

$$[\delta_{R\epsilon_1}, \delta_{R\epsilon_2}] = \delta_{R[\epsilon_1,\epsilon_2]} \ . \qquad (5.48)$$

In terms of these variables the left and right gauge transformations are completely decoupled. The left and right gauge fields transform in the usual way under the left and right gauge transformations and are coupled to the left and right currents respectively.

In discussing the possible anomalies of this theory it is therefore more convenient to use the left–right decomposition than the vector–axial decomposition. Since the left and right sectors of the theory are classically decoupled, it will be enough to study only one of them. From now on we will assume that only the left handed component of the fermion is coupled to a gauge field (henceforth denoted A); this is equivalent to setting $A_R = 0$.

We are now going to consider anomalies for the local gauge transformations in this chirally coupled theory. Some new features, related to dynamical character of the anomalous current, come into play. We start with a chirally modified non-abelian version of (5.1):

$$S_F^L(\psi, \bar{\psi}, A) = \int d^{2n}x\, \bar{\psi}\left(i\gamma^\mu D_\mu^L\right)\psi \ , \tag{5.49}$$

where the new operator is defined as:

$$D_\mu^L = \partial_\mu - ie\left(\frac{1-\gamma^A}{2}\right)A_\mu . \tag{5.50}$$

This action has a local symmetry G_L,

$$\begin{aligned}
\psi'_L &= g^{-1}\psi_L \quad ; \qquad \bar{\psi}'_L = \bar{\psi}_L g, & (5.51)\\
\psi'_R &= \psi_R \qquad\qquad \bar{\psi}'_R = \bar{\psi}_R, & (5.52)\\
A'_\mu &= \frac{i}{e} g\partial_\mu g^{-1} + gA_\mu g^{-1} & (5.53)
\end{aligned}$$

where $\psi_L = \left(\frac{1-\gamma^A}{2}\right)\psi$, $\psi_R = \left(\frac{1-\gamma^A}{2}\right)\psi$ and $g = e^{-i\alpha^a T_a}$. Interactions of this type actually occur in the Standard Model.

We omit the calculation of the gauge anomaly. The result is

$$\begin{aligned}
[D_\mu\langle {j_L}^\mu\rangle]^a &= \pm\frac{e}{4\pi}\varepsilon^{\mu\nu}\mathrm{tr}\ T^a\partial_\mu A_\nu \quad \text{for} \quad n=2\ ;\\
[D_\mu\langle {j_L}^\mu\rangle]^a &= \pm\frac{ie^2}{24\pi^2}\varepsilon^{\mu\nu\lambda\rho}\mathrm{tr}\ T^a\partial_\mu\left(A_\nu\partial_\lambda A_\rho + \frac{1}{2}A_\nu A_\lambda A_\rho\right) \ \text{for } n=4\ , (5.54)
\end{aligned}$$

where the overall sign depends on the chirality of the fermions.

The form (5.54) of the anomaly is by no means unique: it depends on the chosen regularization of the fermionic determinant. Another regularization could result in another form of the effective action differing by a local functional of the gauge field and its derivatives. The determinant is given by a sum of one loop graphs with any number of insertions of the external field A. The first term contains one power of A and diverges like Λ^{n-1}, where Λ is some ultraviolet cutoff; the second contains two powers of A and diverges like Λ^{n-2} and so on. The n-th term contains A^n and is logarithmically divergent. All subsequent terms are finite. Divergent terms give rise to ambiguities in the effective action. One is free to change the renormalization conditions so as to add to the effective action finite terms proportional to the coefficients of these divergences. Therefore, one is free to modify the effective action by adding a polynomial in A and its derivatives of order n (and containing terms of dimension n). If the expression (5.54) was itself the variation of such a polynomial, then by a different choice of renormalization one could obtain zero anomaly. It can be shown that this is not the case, so the anomaly is a genuine physical phenomenon.

Independently of this, one can redefine the current in such a way that its transformation property is covariant. One defines a new current:

$$\hat{j_L}{}_a^{\mu} = j_L{}_a^{\mu} + X_a^{\mu}$$

where

$$\begin{aligned} X_a^{\mu} &= \qquad \text{for } n=2, && (5.55)\\ X_a^{\mu} &= \pm\frac{i}{24\pi^2}\varepsilon^{\mu\nu\rho\sigma}\mathrm{tr}\left(A_\nu\partial_\rho A_\sigma + \partial_\nu A_\rho A_\sigma + \frac{3}{2}A_\nu A_\rho A_\sigma\right) \ \text{for } n=4 && (5.56)\end{aligned}$$

This current has the property

$$\delta_\epsilon\langle\hat{j_L}{}_a^{\mu}(x)\rangle = -[\epsilon, \langle\hat{j_L}{}^{\mu}\rangle]_a$$

thereby recovering the classical tensorial transformation. The polynomial X_a^μ is called the Bardeen-Zumino counterterm and the covariant divergence of $\langle\hat{j_L}{}_a^{\mu}(x)\rangle$ is known as the covariant anomaly. In $d = 2, 4$ we get

$$\begin{aligned} \langle D_\mu \hat{j}_L^\mu\rangle_a &= \pm\frac{e}{2\pi}\varepsilon^{\mu\nu}\mathrm{tr}T_a F_{\mu\nu} \quad \text{for } n=2\\ \langle D_\mu \hat{j}_L^\mu\rangle_a &= \pm i\frac{e^2}{16\pi^2}\varepsilon^{\mu\nu\lambda\rho}\mathrm{tr}T^a F_{\mu\nu}F_{\lambda\rho} \quad \text{for } n=4. && (5.57)\end{aligned}$$

We remark that this redefinition does not correspond to the addition of local terms in the effective action, therefore the physical meaning of this current is not linked directly to the local gauge invariance.

5.4 Gauge anomalies and cohomology

We have seen in the previous sections that the anomalies appear as effects of a regularization procedure; we have also remarked in sect XXX that the gauge anomaly is just the manifestation of the impossibility to construct a gauge-invariant functional of gauge fields, when integrating out chiral fermions. It is therefore clear that the underlying group structure must play an important role in determining the form of the anomaly itself.

At the level of infinitesimal gauge transformations, a simple integrability condition, known as the Wess-Zumino consistency condition, is strong enough to nearly completely determine the anomaly itself. At the level of finite transformations, the same comdition leads to the definition of the Wess-Zumino functional. We shall discuss here the interrelation between these notions.

5.4.1 The WZ consistency condition

Define the anomaly $\mathcal{A}$ as the anomalous divergence of the gauge current. Then from equation (5.36),

$$\delta_\epsilon W[A] = -\mathcal{A}(\epsilon, A) \tag{5.58}$$

The operators δ_ϵ form a representation of the gauge algebra:

$$[\delta_{\epsilon_1}, \delta_{\epsilon_2}] = \delta_{[\epsilon_1,\epsilon_2]}. \tag{5.59}$$

If we now apply the above operatorial relation to the vacuum functional $W[A]$ we get an equation for $\mathcal{A}(\epsilon, A)$:

$$\delta_{\epsilon_1}\mathcal{A}(\epsilon_2, A) - \delta_{\epsilon_2}\mathcal{A}(\epsilon_1, A) = \mathcal{A}([\epsilon_1, \epsilon_2], A)\ . \tag{5.60}$$

This is called the Wess-Zumino (WZ) consistency condition. If the anomaly is defined as gauge variation of the effective action W, as in (5.58), then it must satisfy the above constraint. Such anomalies are called *consistent* anomalies. If on the other hand one defines the anomaly as $\langle D_\mu J^\mu \rangle$, then the result of a calculation may or may not satisfy this condition, depending on the regularization procedure.

For example, multiplying (5.54) by an infinitesimal gauge parameter ϵ^a and integrating over spacetime we obtain the expressions [3]

$$\begin{aligned}
\mathcal{A}(A,\epsilon) &= \mp\frac{e}{4\pi}\int d^2x\,\varepsilon^{\mu\nu}\,\mathrm{tr}\,\partial_\mu\epsilon A_\nu \quad \text{for} \quad n=2;\\
\mathcal{A}(A,\epsilon) &= \mp\frac{ie^2}{24\pi^2}\int d^4x\,\varepsilon^{\mu\nu\lambda\rho}\,\mathrm{tr}\,\partial_\mu\epsilon\left(A_\nu\partial_\lambda A_\rho + \frac{1}{2}A_\nu A_\lambda A_\rho\right) \ \text{for } n=4 \quad (5.61)
\end{aligned}$$

One can verify by explicit calculation that they satisfy the WZ consistency condition. On the other hand, if we do the same with the covariant anomalies (5.57)

$$\begin{aligned}
\mathcal{A}(A,\epsilon) &= \frac{e}{2\pi}\int d^2x\,\varepsilon^{\mu\nu}\,\mathrm{tr}\,\epsilon F_{\mu\nu} \quad \text{for } n=2\\
\mathcal{A}(A,\epsilon) &= i\frac{e^2}{16\pi^2}\int d^4x\,\varepsilon^{\mu\nu\lambda\rho}\,\mathrm{tr}\,\epsilon F_{\mu\nu}F_{\lambda\rho} \quad \text{for } n=2.
\end{aligned} \tag{5.62}$$

we find that they do not satisfy it. Therefore, these anomalies are not the variation of a functional $W(A)$, not even of a locally defined functional, as we shall now discuss.

The definitions in this section have a clear geometrical meaning. Let $\mathcal{C}$ be the space of connections A, and $\mathcal{G}$ the gauge group. [4] For a fixed infinitesimal gauge transformation ϵ, δ_ϵ is a first order (functional) differential operator

[3] For the present purposes it proves convenient to perform an integration by parts so that one derivative acts on the gauge parameter. This is legitimate, since ϵ vanishes at infinity.

[4] Unlike earlier sections, here we consider connections and gauge transformations on spacetime, not just space.

corresponding to the directional derivative along a vector field tangent to the orbits of the gauge group in the space of connections. We can think of it as a vertical vectorfield on $\mathcal{C}$, *i.e.* a vectorfield that is in the kernel of the projection $\mathcal{C} \to \mathcal{C}/\mathcal{G}$. Fix a reference gauge field A and consider its gauge orbit $\mathcal{O}_A$. It is diffeomorphic to the gauge group $\mathcal{G}$. The anomaly $\mathcal{A}$ is a linear functional that maps vectorfields on $\mathcal{O}_A$ to real numbers. Thus, we can think of it as a one-form on $\mathcal{O}_A$. Equation (5.60) is the statement that $\mathcal{A}$ is a closed form. If W was a globally well-defined functional on $\mathcal{O}_A$, equation (5.58) would say that $\mathcal{A}$ is an exact form, and $\mathcal{A}$ would be in the trivial cohomology class in $H^1(\mathcal{O}_A)$, or equivalently of $H^1(\mathcal{G})$. However, at this stage we do not really know W well enough. Equation (5.58) must be interpreted as saying that $\mathcal{A}$ is *locally* exact, i.e. the differential a locally-defined functional $-W$. In the next section we will consider the global properties of this functional and we shall see that it is not globally well-defined.

5.4.2 The WZ functional

In the preceding section we have considered the effect of infinitesimal gauge transformations on the fermionic determinant. Let us now consider the effect of finite gauge transformations. Define the Wess-Zumino functional $\Gamma_{WZ}(A, g)$ to be (minus) the change in the fermionic effective action under a gauge transformation:

$$W(A^g) - W(A) = -\Gamma_{WZ}(A, g) \ . \tag{5.63}$$

When g differs infinitesimally from the identity, Γ_{WZ} becomes the anomaly:

$$\Gamma_{WZ}(A, 1 + \epsilon) = \mathcal{A}(A, \epsilon) \ .$$

From the definition one finds that

$$\Gamma_{WZ}(A^{g_1}, g_2) - \Gamma_{WZ}(A, g_1 g_2) + \Gamma_{WZ}(A, g_1) = 0 \ .$$

This condition has a cohomological significance which is the analogue of the WZ consistency condition for finite transformations. A functional satisfying it is said to be a one-cocycle for the action of the gauge group with coefficients in the smooth functionals of A_μ.

The way it was derived, Γ_{WZ} depends on a connection A and a gauge transformation g. However, g is just a map from spacetime to the group G and we can also think of it as a configuration for a chiral model. In this case we denote it as U and we can think as $\Gamma_{WZ}(A, U)$ as a possible term in the action for a chiral model coupled to gauge fields. In this case it is more convenient to rewrite (5.4.2) in the form

$$\Gamma_{WZ}(A^g, U^g) - \Gamma_{WZ}(A, U) = -\Gamma_{WZ}(A, g) \ . \tag{5.64}$$

where $U^g = g^{-1}U$ can be thought of as the gauge transform of U by g. Comparing with (A.2), this formula shows that the WZ functional has the same anomalous transformation property as the fermionic determinant. The important difference, that we shall now see, is that whereas W is a non-local functional, Γ_{WZ} is a local functional.

We can compute Γ explicitly by integrating the anomaly. We begin by fixing a reference gauge field A_μ. Then we can identify the orbit through A_μ with $\mathcal{G}$ (mapping g to A^g_μ) and we can regard W as a function on $\mathcal{G}$. As above, we think of the anomaly $\mathcal{A}$ as one-form on $\mathcal{G}$. Since $\mathcal{A}$ is closed, its integral along a curve does not change under continuous deformations of the curve, as long as the endpoints remain fixed. It can only change in a discontinuous way if we change the homotopy class of the curve. Let $g(r)$ be a one-parameter family of gauge transformations interpolating between g and the identity:

$$\hat{g}(r) = e^{r\epsilon^a T_a} \; ; \qquad \hat{g}(0) = e \; ; \qquad \hat{g}(1) = g \; . \tag{5.65}$$

and let

$$\hat{A}_\mu(r) = \hat{g}^{-1} A_\mu \hat{g} + \hat{g}^{-1}\partial_\mu \hat{g} \tag{5.66}$$

be the gauge transform of A_μ at the point r along the path. Then the WZ functional can be written

$$\Gamma_{WZ}[A,g] = \int_0^1 dr\, \mathcal{A}(A^{g(r)}, \hat{g}^{-1}\partial\hat{g}) \; . \tag{5.67}$$

Let us perform the integral explicitly in $d = 2$. Using the anomaly (5.61) we have to compute

$$\frac{1}{4\pi}\int_0^1 dr \int d^2x\, \varepsilon^{\mu\nu} \mathrm{tr}\, \hat{g}^{-1}\partial_r \hat{g}\, \partial_\mu(\hat{g}^{-1}A_\nu\hat{g} + \hat{g}^{-1}\partial_\nu\hat{g})$$
$$\frac{1}{4\pi}\int_0^1 dr \int d^2x\, \varepsilon^{\mu\nu} \mathrm{tr}\, \partial_r \hat{g}\hat{g}^{-1} \left[\partial_\mu A_\nu - \partial_\mu\hat{g}\hat{g}^{-1}A_\nu + A_\nu\partial_\mu\hat{g}\hat{g}^{-1} - \partial_\mu\hat{g}\hat{g}^{-1}\partial_\nu\hat{g}\hat{g}^{-1}\right] .$$

Now we rewrite this in a covariant form in the three coordinates r, x_1, x_2, which parametrize a three-dimensional ball with boundary S^2 (it is assumed that the r-component of A_μ is zero):

$$\frac{1}{4\pi}\int d^3x\, \varepsilon^{\lambda\mu\nu} \mathrm{tr} \left[\partial_\lambda\hat{g}\hat{g}^{-1}\partial_\mu A_\nu - \partial_\lambda\hat{g}\hat{g}^{-1}\partial_\mu\hat{g}\hat{g}^{-1}A_\nu - \frac{1}{3}\partial_\lambda\hat{g}\hat{g}^{-1}\partial_\mu\hat{g}\hat{g}^{-1}\partial_\nu\hat{g}\hat{g}^{-1}\right] .$$

The first two terms are a total derivative, and can be rewritten as a two-dimensional integral, so we obtain

$$\Gamma_{WZ}(A,g) = -\frac{1}{4\pi}\int d^2x\, \varepsilon^{\mu\nu} \mathrm{tr}\, R_\mu A_\nu - \frac{1}{12\pi}\int d^3x\, \varepsilon^{\lambda\mu\nu} \mathrm{tr}\, \hat{R}_\lambda \hat{R}_\mu \hat{R}_\nu \; , \tag{5.68}$$

where $\hat{R}_\mu = \partial_\mu\hat{g}\hat{g}^{-1}$. We note that if $g = 1 + \epsilon$, $R_\mu = \partial_\mu\epsilon$, so the first term gives back $\mathcal{A}(A,\epsilon)$, as expected. Also, we recognize that the second term is the correctly normalized WZW action S_{WZW}, defined in (**??**), depending on the extension of the field g from the compactified two-dimensional spacetime S^2 to the interior of a ball. Recall that, even though the integrand of $S_{WZW}(\hat{g})$ is the same as that of the winding number, $S_{WZW}(\hat{g})$ is not a topological invariant, since it depends on the boundary values of $\hat{g}$. The WZW action that appears in the WZ functional corresponds to the choice $n=1$, or $c=2\pi$, for the coefficient. Thus, the WZ functional can be viewed as a left-gauged extension of the WZW functional.

As an exercise, let us check that this functional satisfies the condition (5.64). We compute

$$-\frac{1}{4\pi}\int d^2x\,\varepsilon^{\mu\nu}\mathrm{tr}\left[\partial_\mu(g^{-1}U)(g^{-1}U)^{-1}(g^{-1}A_\nu g+g^{-1}\partial_\nu g)\right]+S_{WZW}(g^{-1}U)\ .$$

The first integral gives

$$\frac{1}{4\pi}\int d^2x\,\varepsilon^{\mu\nu}\mathrm{tr}\left[R^g_\mu A_\nu\ +R^g_\mu R^g_\nu-R^U_\mu A_\nu-R^U_\mu R^g_\nu\right] \tag{5.69}$$

where we denote $R^g_\mu=\partial_\mu g g^{-1}$ and $R^U_\mu=\partial_\mu U U^{-1}$. The second term gives

$$-\frac{1}{12\pi}\int d^3x\,\varepsilon^{\lambda\mu\nu}\mathrm{tr}\,\left[-\hat{R}^g_\lambda\hat{R}^g_\mu\hat{R}^g_\nu+3\hat{R}^g_\lambda\hat{R}^g_\mu\hat{R}^U_\nu-3\hat{R}^g_\lambda\hat{R}^U_\mu\hat{R}^U_\nu+\hat{R}^U_\lambda\hat{R}^U_\mu\hat{R}^U_\nu\right]\ , \tag{5.70}$$

The first and last terms in this expression are equal to $-S_{WZW}(\hat{g})$ and $S_{WZW}(\hat{U})$ respectively. The two middle terms add up to a total derivative that exactly cancels the last term in (5.69). The second term in (5.69) vanishes identically. The remaining terms exactly reconstruct $-\Gamma_{WZ}(A,g)+\Gamma_{WZ}(A,U)$.

One can proceed in the same way in higher dimensions. Integrating the anomaly (5.61) one arrives at the following expression for the Wess-Zumino functional in four dimensions

$$\begin{aligned}\Gamma(A_\mu,g) \quad = \quad & -\frac{i}{48\pi^2}\int d^4x\ \varepsilon^{\mu\nu\rho\sigma}\mathrm{tr}\Big[\big(A_\mu\partial_\nu A_\rho+\partial_\mu A_\nu A_\rho+A_\mu A_\nu A_\rho\big)R_\sigma \\ & \qquad -\frac{1}{2}A_\mu R_\nu A_\rho R_\sigma-A_\mu R_\nu R_\rho R_\sigma\Big] \\ & -\frac{i}{240\pi^2}\int_B d^5x\ \varepsilon^{\lambda\mu\nu\rho\sigma}\mathrm{tr}\ R_\lambda R_\mu R_\nu R_\rho R_\sigma\ .\end{aligned} \tag{5.71}$$

Once again we recognize that the last term is the WZW functional with the correctly normalized coefficient $c=2\pi$. (FACTORS!!!)

Finally, let us return to the question whether W is a globally well-defined functional on the orbits of the gauge group. We will discuss this in the two-dimensional case, with gauge group $SU(2)$. Assuming that spacetime has been compactified to S^2, the gauge group is $\mathcal{G}=\Gamma_*(S^2,SU(2))$ and $\pi_1(\mathcal{G})=\pi_3(SU(3))=\mathbb{Z}$. Thus we need to worry whether W is single-valued along a non-contractible path in the orbit. As observed above, the gauge variations of W is the same as the gauge variation of the WZ functional. Therefore, up to an additive constant, these two functionals are the same, when restricted to a gauge orbit. We therefore ask whether Γ_{WZ} is single-valued along a non-contractible path in the orbit. To answer this question one just has to integrate the anomaly along a closed loop, in which case in (5.65) we have to set $\hat{g}(1) = e$.

Now consider the four-dimensional case, with gauge group $SU(N)$, $N > 2$. Assuming that spacetime has been compactified to S^4, the gauge group is $\mathcal{G} = \Gamma_*(S^4, SU(N))$ and $\pi_1(\mathcal{G}) = \pi_5(SU(N)) = \mathbb{Z}$. The situation is the same as in the two-dimensional case.

5.5 The descent equations

The axial anomaly, the gauge anomaly and the Schwinger terms of gauge theories in different dimensions, are strictly related. In fact, we will see that one can obtain the solution of the WZ consistency condition, i.e. the consistent anomaly, including the correct normalization, by a series of purely geometrical operations, starting from the axial anomaly in a space of two more dimensions. We will define the cocycles ω^k_r, for $k = 0, 1, 2$, where $r = 2n - k - 1$ is the degree of ω as a form (and hence also the dimension of the space over which ω is to be integrated), and k is its degree as a form in the space of connections (more precisely, in an orbit of the gauge group in the space of connections). This means that, given k infinitesimal gauge transformation parameters $\epsilon_1, \ldots, \epsilon_k$, and r vectorfields $v_1, \ldots, v_r$, $\omega^k_r(\epsilon_1, \ldots, \epsilon_k, v_1, \ldots, v_r)$ is a real number. The fact that these are cocycles means that they are closed, both as r-forms on space(time) and as k-forms on the orbit if the gauge group. The relations between all these forms constitute the so-called "descent equations".

In order to minimize the index clutter it is convenient to use the algebra of differential forms. Hence we write $A = A^a_\nu dx^\nu T_a$, $F = \frac{1}{2} F^a_{\mu\nu} dx^\mu \wedge dx^\nu T_a$. The exterior derivative acting on a p-form ω can be defined in a coordinate-independent way by specifying the result of acting with $d\omega$ on $p+1$ vectorfields:

$$\begin{aligned} d\omega(v_1, \ldots, v_{p+1}) = & \sum_{1<i<p+1} (-1)^{i+1} v_i(\omega(v_1, \ldots, \hat{v}_i, \ldots, v_{p+1})) \\ & + \sum_{1<i<j<p+1} (-1)^{i+j} \omega([v_i, v_j], v_1, \ldots, \hat{v}_i, \ldots, \hat{v}_j, \ldots, v_{p+1}) \ , \end{aligned} \tag{5.72}$$

where a hat over a vector means that it is missing. If the components of ω are defined by

$$\omega = \frac{1}{p!} \omega_{\mu_1, \ldots, \mu_p} dx^{\mu_1} \wedge \ldots \wedge dx^{\mu_p} \tag{5.73}$$

we can also write

$$d\omega = \frac{1}{p!} d\omega_{\mu_1, \ldots, \mu_p} \wedge dx^{\mu_1} \wedge \ldots \wedge dx^{\mu_p} \ , \tag{5.74}$$

where the differential acts only on the components. Thus we can write $F = dA + A \wedge A$, and to further condense the notation also wedge products will not be written explicitly, so $F = dA + A^2$.

We are going to define a sequence of cocycles ω_k^p, which are simultaneously k-forms on space(time) (hence k-cocycles in the sense of de Rham cohomology) and p-cocycles in the Lie algebra of the gauge group $\mathcal{G}$, i.e. antisymmetric multilinear functions of p infinitesimal gauge transformations. In physical applications one is mostly interested in the cases $p = 0, 1, 2$.

One begins from the expression for the Chern character, a $2n$-form in $2n$ dimensions:

$$c_n = k_n \mathrm{tr}\, F^n \ , \tag{5.75}$$

where k_n is a normalization constant, depending on the group G, such that the integral of c_n is an integer. It is gauge invariant

$$\delta_\epsilon c_n = 0 \tag{5.76}$$

and closed

$$dc_n = 0 \ . \tag{5.77}$$

Thus we can write c_n, at least locally, as the exterior differential of an $2n-1$ form ω_{2n-1}^0, called the Chern–Simons form

$$d\omega_{2n-1}^0 = c_n \ . \tag{5.78}$$

A general formula can be given in any dimension, but we limit ourselves to the cases $n = 2, 3, 4$, where we have

$$\begin{aligned}
\omega_3^0(A) &= k_2 \mathrm{tr}\bigg(FA - \frac{1}{3}A^3\bigg) \ , &(5.79)\\
\omega_5^0(A) &= k_3 \mathrm{tr}\bigg(F^2A - \frac{1}{2}FA^3 + \frac{1}{10}A^5\bigg) \ , &(5.80)\\
\omega_7^0(A) &= k_4 \mathrm{tr}\bigg(F^3A - \frac{2}{5}F^2A^3 - \frac{1}{5}FAFA^2 + \frac{1}{5}FA^5 - \frac{1}{35}A^7\bigg) \ . &(5.81)
\end{aligned}$$

The gauge variation of the Chern-Simons form is closed, because

$$d\delta_\epsilon \omega_{2n-1}^0 = \delta_\epsilon d\omega_{2n-1}^0 = \delta_\epsilon c_n = 0 \ ,$$

therefore it is locally the differential of a $(2n-2)$-form:

$$\delta_\epsilon \omega_{2n-1}^0(A) = d\omega_{2n-2}^1(A, \epsilon) \ . \tag{5.82}$$

This form can be written as

$$\omega_{2n-2}^1(A, \epsilon) = \mathrm{tr}\, d\epsilon\, \phi_{2n-3}(A) \ . \tag{5.83}$$

where the $2n-3$-form $\phi_{2n-3} = \phi_{2n-3}^a T_a$ is a polynomial in A and F. For $n = 2, 3, 4$ this polynomial is given by

$$\begin{aligned}
\phi_1 &= -k_2 A \ , &(5.84)\\
\phi_3 &= -\frac{k_3}{2}(FA + AF - A^3) \ , &(5.85)\\
\phi_5 &= -\frac{k_4}{3}\big[(F^2A + FAF + AF^2)\\
&\qquad -\tfrac{4}{5}(A^3F + FA^3) - \tfrac{2}{5}(A^2FA + AFA^2) + \tfrac{3}{5}A^5\big] \ . &(5.86)
\end{aligned}$$

From these formulae we recognize that $\int \omega^1_{2n-2}(A,\epsilon) = \mathcal{A}(A,\epsilon)$ is the consistent anomaly in dimension $2n-2$.

The coboundary of ω^1_{2n-2} (in the sense of Lie algebra cohomology) is a closed $(2n-2)$-form, thus locally it is the differential of a $(2n-3)$-form

$$\delta_{\epsilon_1}\omega^1_{2n-2}(A,\epsilon_2) - \delta_{\epsilon_2}\omega^1_{2n-2}(A,\epsilon_1) - \omega^1_{2n-2}(A,[\epsilon_1,\epsilon_2]) = d\omega^2_{2n-3}(A,\epsilon_1,\epsilon_2)\ . \tag{5.87}$$

We note that integrating this relation on a closed manifold without boundary, this is just the WZ consistency condition (5.60). For $n = 2,3,4$ we find

$$\begin{aligned} \omega^2_1(A,\epsilon_1,\epsilon_2) &= 2k_2 \mathrm{tr}\, \epsilon_1 d\epsilon_2\ , && (5.88)\\ \omega^2_3(A,\epsilon_1,\epsilon_2) &= k_3 \mathrm{tr}\, \{d\epsilon_1, d\epsilon_2\} A\ , && (5.89)\\ \omega^2_5(A,\epsilon_1,\epsilon_2) &= \frac{k_4}{15}\mathrm{tr}\, (5F-3A^2)\, [2A\{d\epsilon_1,d\epsilon_2\} - d\epsilon_1 A d\epsilon_2 + d\epsilon_2 A d\epsilon_1] && (5.90) \end{aligned}$$

Remark 1. These 2-cocycles appear as "Schwinger terms" in the algebra of the gauge generators $G_\epsilon = \int d^{2n-3}x\, \epsilon^a G_a$ for an anomalous gauge theory in $2n-2$ spacetime dimensions:

$$[G_{\epsilon_1}, G_{\epsilon_2}] = G_{[\epsilon_1,\epsilon_2]} + \int d^{2n-3}x\, \omega^2(\epsilon_1,\epsilon_2)\ . \tag{5.91}$$

In the case $n = 2$ they define a central extension of the gauge algebra (a Kac-Moody algebra). In higher dimension the extension is by a function of A. The presence of the Schwinger term obstructs the definition of physical states as those states that are annihilated by the Gauss operator $G_\epsilon|\psi_{phys}\rangle = 0$. This is the manifestation of the anomaly at the canonical level.

Remark 2. It is clear from (5.82) and (5.87) that ω^1_{2n-2} and ω^2_{2n-3} are only defined up to a closed form. In particular one could add to ω^1_{2n-2} the closed form $-d(\mathrm{tr}\epsilon\phi(A))$ and get

$$\hat\omega^1_{2n-2}(A,\epsilon) = -\mathrm{tr}\, \epsilon d\phi_{2n-3}\ , \tag{5.92}$$

which is another form of the consistent anomaly. Applying the coboundary to $\hat\omega^1_{2n-2}$ defines a different 2-cocycle $\hat\omega^2_p$:

$$\delta_{\epsilon_1}\hat\omega^1_{2n-2}(A,\epsilon_2) - \delta_{\epsilon_2}\hat\omega^1_{2n-2}(A,\epsilon_1) - \hat\omega^1_{2n-2}(A,[\epsilon_1,\epsilon_2]) = d\hat\omega^2_{2n-3}(A,\epsilon_1,\epsilon_2)\ . \tag{5.93}$$

For $n = 2,3,4$

$$\begin{aligned} \hat\omega^2_1(A,\epsilon_1,\epsilon_2) &= k_2 \mathrm{tr}\, [\epsilon_1,\epsilon_2]\, A\ , && (5.94)\\ \hat\omega^2_3(A,\epsilon_1,\epsilon_2) &= \tfrac{1}{2}k_3 \mathrm{tr}\, \big[[\epsilon_1,\epsilon_2](FA+AF-A^3) \\ &\qquad -\epsilon_1 dA\epsilon_2 A - \epsilon_1 A\epsilon_2 dA\big]\ , && (5.95)\\ \hat\omega^2_5(A,\epsilon_1,\epsilon_2) &= \tfrac{1}{3}k_4 \mathrm{tr}\big\{[\epsilon_1,\epsilon_2]\big[(F^2A+FAF+AF^2) \\ &\qquad -\tfrac{4}{5}\{A^3,F\} - \tfrac{2}{5}\{A,AFA\} + \tfrac{3}{5}A^5\big] \\ &\quad -\tfrac{1}{5}\,[\epsilon_1,d\epsilon_2][F,A^2] - \tfrac{3}{5}(d\epsilon_1 A\epsilon_2 + \epsilon_2 A d\epsilon_1)(FA+AF-A^3) \\ &\quad +\tfrac{1}{5}\,[\epsilon_2,d\epsilon_1][F,A^2] - \tfrac{3}{5}(d\epsilon_2 A\epsilon_1 + \epsilon_1 A d\epsilon_2)(FA+AF-A^3)\big\} && (5.96) \end{aligned}$$

These are the cocycles one gets in the Gauss law algebra of an anomalous fermionic theory using the Bjorken-Johnson-Low procedure [5] or in the gauged Wess-Zumino-Witten model at the canonical level. [6] They differ from the cocycles ω^2_{2n-3} by a redefinition of the current. Note that even in two dimensions these cocycles do not define a central extension.

[5] S. Jo, Phys. Lett. **163B** (1985) 353.

[6] R. Percacci and R. Rajaraman, Phys. Lett. **201B** (1988) 256; Int. J. Mod. Phys. **A 4** (1989) 4177.