**Chapter 1**

$\pi_0(Q)$ 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. Further, more restrictive technical conditions may also be imposed, but will not concern us here. 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, $Q$, to be the space of smooth, finite energy configurations of the field at some instant of time. Note that $Q$ 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 $\pi_0(Q)$ (the reason for this notation is explained in Appendix A):

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

where $Q_i$ are connected. Having determined the structure of the configu-
ration 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 characterizes the 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 fields. 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 procedure that would require much more technical work to be made precise, but holds great heuristic power. 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 solitons

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

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

(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 $y_i$, $i \in J$, the minimum points. For definiteness one can think of the quartic potential

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

(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:

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

(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

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

(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 $\phi$ 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 $\phi$ 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 than one minimum, $\mathcal{Q}$ will not be connected. In fact, let

$$\mathcal{Q} = \bigcup_{i,j} \mathcal{Q}_{ij}, \quad \mathcal{Q}_{ij} = \left\{ \phi \in \mathcal{Q} \mid \phi \xrightarrow{x \to -\infty} y_i, \phi \xrightarrow{x \to +\infty} y_j \right\}.$$ 

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 $\phi$ 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 $\psi$ which tends asymptotically to zero at $\pm \infty$. The function $\psi$ 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 $\infty$) 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 are going to be absolute minima of the static energy also in the sectors $\mathcal{Q}_{-+}$ and $\mathcal{Q}_{+-}$. Let us denote $\ell$ 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 $f^2/\ell$, and hence decreases with $\ell$, while the potential energy is of order $\lambda f^4 \ell$, and hence increases with $\ell$. 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 $Q_{-+}$ and $Q_{+-}$ 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 $\phi/f$ 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

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

and there is equipartition between elastic and potential energy (i.e. each of the two terms in (1.4) contributes exactly $E_S/2$).

Figure 1.1: The kink of $\phi^4$ theory.
In the theory with potential (1.2), consider the current

\[ J^\mu_T = \frac{1}{2f} \varepsilon^{\mu\nu} \partial_\nu \phi ; \]  

(1.8)

clearly we have

\[ \partial_\mu J^\mu_T = 0 . \]  

(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

\[ Q_T = \int_{-\infty}^{\infty} dx J^0_T = \frac{1}{2f} [\phi(+\infty) - \phi(-\infty)] \]  

(1.10)

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

Another interesting potential is

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

(1.11)

This corresponds to the so called “sine-Gordon” model. The indexing set of

\[ \frac{\phi}{4f} \]

\[ mx \]

Figure 1.2: The kink of the Sine-Gordon model.
1.1. SCALAR SOLITONS IN 1 + 1 DIMENSIONS

minima is the set of the integers $J = \mathbb{Z}$, so there is a double infinity ($\mathbb{Z} \times \mathbb{Z}$) of connected components. The topological current and the topological charge are given again by (1.9) and (1.10), where $f$, which is half the distance between two 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 $Q_{01}$ and $Q_{0 - 1}$

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

This solution is plotted in the Figure (1.2).

Just adding $2nf$ we get the soliton and antisoliton, still with $Q_T = \pm 1$, which minimize the energy in $Q_{nn+1}$ and $Q_{nn - 1}$. Note that if in the field equation (1.5) with the potential (1.11) we reinterpret $x$ as time and $\phi$ 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 solitons

In this section we consider the quantum version of the $\phi^4$ 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 $\phi \in Q_{++}$ 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 $\phi = 0$ gives a Fock space of scalar particles, that we shall call “pions” with mass

$$m_\pi = \sqrt{V''(f)} = \sqrt{2}m .$$
CHAPTER 1. $\pi_0(Q)$ AND SOLITONS

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

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

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

where we employed a simple momentum cutoff $\Lambda$.

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

$$m_k = \frac{m^3_\pi}{3\lambda}.$$  \hspace{1cm} (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 $Q_{++}$, $Q_{+-}$, $Q_{-+}$, $Q_{--}$. 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 $Q_{--}$ 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

$$\phi(x) = \tilde{\phi}(x) + \eta(x),$$
where $\phi$ is the static solution (1.6), treated as a classical background, and $\eta$ is the quantum field.

Let us expand the action around the background:

$$S(\phi) = \int dt \left[ \frac{1}{2} \int dx \left( \frac{d\phi}{dt} \right)^2 - E_S(\phi) \right]$$

$$= S(\phi) + \int dt dx \left[ \frac{1}{2} \left( \frac{d\eta}{dt} \right)^2 - \frac{1}{2} \eta L \eta - \lambda \left( \phi \eta^3 + \frac{1}{4} \eta^4 \right) \right]$$

(1.15)

where

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

(1.16)

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

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

$$V''(\phi) = \lambda (3\phi^2 - f^2)$$

$$= m^2 \left( -1 + 3 \tanh^2 \left( \frac{m x}{\sqrt{2}} \right) \right) .$$

(1.17)

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

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

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

(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 $\eta_0 = \frac{1}{\cosh^2 \left( \frac{m_\pi}{\sqrt{2}} \right)}$;

• an isolated eigenvalue $\omega_1^2 = \frac{3}{2} m^2$ with eigenfunction $\eta_1 = \frac{\sinh \left( \frac{m_\pi}{\sqrt{2}} \right)}{\cosh^2 \left( \frac{m_\pi}{\sqrt{2}} \right)}$

describing an excited state of the kink;

• a continuous spectrum with $\omega_p^2 = m_\pi^2 + p^2$, 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 $p$. The presence of the kink deforms the spectrum but in a rather simple way. A “right-moving” mode with momentum $p > 0$ is given for large negative $x$ by a plane wave $\eta_p(x) \approx e^{ipx}$. Solving the eigenvalue equation (1.18) is like studying the quantum mechanical scattering of this incoming wave on the potential (1.17). \(^1\) It turns out that there is no reflected wave, and the transmitted wave, for large positive $x$ is simply

$$\eta_p(x) \approx e^{ipx+i\delta_p}$$

(1.19)

where the phase shift is given by

$$e^{i\delta_p} = \frac{1 + ip/m_\pi}{1 - ip/m_\pi} \frac{1 + 2ip/m_\pi}{1 - 2ip/m_\pi} \frac{1}{1 - 2ip/m_\pi}.$$

(1.20)

\(^1\)See Morse and Feschbach, eq (12.3.22) and following.
1.1. SCALAR SOLITONS IN 1 + 1 DIMENSIONS

This is left as an exercise (see Exercise XXX). There is another eigenfunction with the same \( p \), 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:

\[
A \eta_p(x) + B \eta_p(-x) .
\]  

(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 \( \eta_p(L/2) = \pm \eta_p(-L/2) \) and 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 \( \bar{\omega}_n^2 = m_k^2 + \bar{p}_n^2 \) the corresponding eigenvalues. We denote

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

the momenta, and \( \omega_n^2 = m_k^2 + p_n^2 \) the eigenvalues, in the absence of the kink.

It is natural to expand the quantum field \( \eta \) on the basis of eigenfunctions of \( L \), instead of ordinary Fourier modes:

\[
\eta(t, x) = a_0(t) \eta_0(x) + a_1(t) \eta_1(x) + \sum_{n=0}^{\infty} a_n(t) \eta_n(x) ,
\]  

(1.22)

where the first two terms correspond to the isolated modes and the sum to the “continuous” spectrum. At order \( \lambda^0 \) the Hamiltonian of the fluctuation field is then a sum over independent oscillators:

\[
H = \int dx \left[ \frac{1}{2} \eta^2 + \frac{1}{2} \eta L \eta \right] = \frac{1}{2} \sum_n (a_n^2 + \bar{\omega}_n^2 a_n^2) .
\]  

(1.23)

where the sum extends over all modes. By choosing to work in the basis of eigenfunctions of \( L \) we have decomposed the system into infinitely many decoupled oscillators.

There only one odd degree of freedom, 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 semiclassical approximation rests on the assumption that the quadratic term in the Lagrangian is dominant with respect to the quartic one:

\[ \omega^2 \langle q^2 \rangle \gg \lambda \langle q^4 \rangle . \]

For all the oscillator states, this is true provided the wave function remains sufficiently localized near the origin. This cannot be said of the zero mode.

Its origin can be understood as follows. Among all possible deformations of the kink field, there is one that corresponds simply to an infinitesimal translation of the kink by \( \delta x \):

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

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

This analysis suggests that instead of the zero mode \( a_0 \), 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 \( x_0(t) \):

\[ \phi(x, t) = \bar{\phi}(x - x_0(t)) . \]

Inserting in the action we find

\[ S = \int dt \int dx \left( \frac{1}{2} \dot{\phi}^2 \right) - \frac{1}{2} \phi'^2 - V \]

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

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

\[ S = \int dt \left[ \frac{1}{2} m_k \dot{x}_0^2 + m_k \right] . \quad (1.24) \]

---

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.
1.1. **SCALAR SOLITONS IN 1 + 1 DIMENSIONS**

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

\[
H = m_k + \frac{p^2}{2m_k}.
\]  
\[ (1.25) \]

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 must remove the zero mode from the list of the degrees of freedom.

The energy of a kink at rest, with the pion field in the Fock vacuum, is then

\[
H = m_k + \frac{\sqrt{3}}{4} m_\pi + \sum_{n=0}^{\infty} \frac{1}{2} \tilde{\omega}_n,
\]  
\[ (1.26) \]

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 \( n \), \( \tilde{\omega}_n \sim n \), 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 \( Q_{-\ldots} \). 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

\[
\delta m_k = \frac{\sqrt{3}}{4} m_\pi + \frac{1}{2} \sum_{n=1}^{\infty} (\tilde{\omega}_n - \omega_n)
\]

\[
= \frac{\sqrt{3}}{4} m_\pi + \frac{1}{2} \sum_{n=1}^{\infty} \frac{p_n \delta p}{L \omega_n},
\]  
\[ (1.27) \]

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

\[
\tilde{p}_n^2 = p_n^2 + 2 \frac{\delta p_n}{L} + O(1/L^2).
\]

\(^3\)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.
At this point we can take the limit $L \to \infty$ and we return to continuous momenta:

$$\delta m_k = \frac{\sqrt{3}}{4} m_\pi + \frac{1}{2\pi} \int dp \frac{p \delta_p}{\sqrt{m_\pi^2 + p^2}}$$

$$= \frac{\sqrt{3}}{4} m_\pi + \frac{1}{2\pi} \lim_{\Lambda \to \infty} \delta_p \sqrt{m_\pi^2 + p^2} \left|_0^\Lambda - \frac{1}{2\pi} \int dp \sqrt{m_\pi^2 + p^2} \frac{d\delta_p}{dp} \right.$$}

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.

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

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

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

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

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 $\lambda/m_\pi^2$, 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:

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

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 $\psi$ with Lagrangian

$$L_F = \bar{\psi} (\gamma^\mu \partial_\mu + g \phi) \psi$$

(1.30)

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

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

as well as the discrete transformation

$$\phi \to -\phi ; \quad \psi \to \gamma A \psi ; \quad \bar{\psi} \to -\bar{\psi} \gamma A$$

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 $m_F = g f$.  

In the sectors $Q_{- -}$ and $Q_{++}$, i.e. in scalar vacuum, the fermion field can be decomposed in plane waves

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

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

$$u_p(x) = e^{ipx} \left( \frac{\sqrt{E}}{-p-im_F} \right) ; \quad v_p(x) = e^{-ipx} \left( \frac{\sqrt{E}}{-p+im_F} \right).$$

(1.31)

The field is quantized by imposing the canonical anticommutation relations

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

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

For the fermion current it is best to use the definition

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

(1.32)

where $\psi^c = \psi^*$ is the charge conjugate field, obeying the same equation as $\psi$. This expression has the advantage of avoiding the infinite charge of the

---

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

---
Dirac sea that is present in the more familiar expression \( j^\mu = \bar{\psi} \gamma^\mu \psi \). Indeed we have
\[
Q = \int \frac{dp}{2\pi} \left( b_p^\dagger b_p - d_p^\dagger d_p \right)
\]  
(1.33)
whereas the Hamiltonian is given by
\[
H = \int \frac{dp}{2\pi} E_p \left( b_p^\dagger b_p + d_p^\dagger d_p \right).
\]  
(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
\[
\begin{pmatrix}
  P^\dagger & \partial_t \\
  -\partial_t & P
\end{pmatrix}
\]
where \( P = \partial_x + g\bar{\phi} \), \( P^\dagger = -\partial_x + g\phi \).

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
\[
P^\dagger P = -\partial_x^2 + g^2\bar{\phi}^2 - g\partial_x\phi, \quad PP^\dagger = -\partial_x^2 + g^2\phi^2 + g\partial_x\phi.
\]
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 $\partial^2_t + L$. 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 $C_1 C_2 = 1/E^2$.

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

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

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

$$\psi = b_0 \begin{pmatrix} \bar{0} \\ u_0(x) \end{pmatrix} + \int \frac{dp}{2\pi \sqrt{2E}} \frac{1}{\sqrt{2E}} \left[ b_p e^{-iEt} u_p(x) + d^*_p e^{iEt} v_p(x) \right].$$

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 (1.1.3) 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 $\pm 1/2$.

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

$$\sum_i \partial_i^2 \phi = V',$$

(1.35)
1.2. SCALAR FIELDS IN OTHER DIMENSIONS

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:

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

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 \( \lambda \) 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

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

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

\[
\mathcal{Q} = \bigcup_{i \in J} \mathcal{Q}_i, \quad \mathcal{Q}_i = \{ \phi \in \mathcal{Q} \mid \phi \underset{r \to \infty}{\longrightarrow} y_i \}
\]
and $\mathcal{J}$ is the set of the minima of $V$. The absolute minimum of $E_S$ in each $Q_i$ 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” $S^0_\infty$ 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” $S^{d-1}_\infty$ is connected. By continuity the value of the field at infinity must be constant and there cannot be solutions with nontrivial boundary conditions.

Let us next consider the case of $N > 1$ scalar fields $\phi = \phi^a$ ($a = 1, \ldots, N$) 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

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

(1.36)

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

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

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

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

(1.37)

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

One can ask whether it is necessary to allow $\phi$ 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 $\phi$ and $\phi'$ be two field configurations such that $\phi \rightharpoonup y$ and $\phi' \rightharpoonup y'$, 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 $0 \leq \tau \leq 1$, 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 $\infty$ instead from 0 to 1. For example, we can define

$$\tau = \frac{1}{2} + \frac{1}{\pi} \arctan t .$$

(1.38)
1.2. SCALAR FIELDS IN OTHER DIMENSIONS

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 \( E_K = \int d^d x \frac{1}{2} \left( \frac{d\hat{\phi}}{dt} \right)^2 \) 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 \( \phi \) to \( \phi' \) one must go through configurations with infinite kinetic energy, so the boundary value of \( \phi \) cannot change in the course of the time evolution.

Using the \( SO(N) \) invariance of the theory, we can assume without loss of generality that the value of \( \phi \) as \( r \to \infty \) be \( y_0 = (0, 0, \ldots, 0, f) \). The limit \( r \to \infty \) defines a “sphere at infinity” \( S_{d-1}^\infty \); since the map \( \phi \) must be constant on \( S_{d-1}^\infty \), all its points may be identified to a single point \( \infty \). Then \( \phi \) 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” \( \infty \) of \( S^d \) to the “basepoint” \( y_0 \). Therefore \( Q = \Gamma_\ast(S^d, \mathbb{R}^N) \). All maps with these properties are homotopic to one another, so the space \( Q \) 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 1.36 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 (1.37) 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 \( \phi_\lambda \) be a one-parameter family of configurations defined by \( \phi_\lambda(x) = \phi_1(\lambda x) \). We have

\[
E_1(\phi_\lambda) = \lambda^{2-d} E_1(\phi_1), \quad E_2(\phi_\lambda) = \lambda^{-d} E_2(\phi_1).
\]

In order for \( \phi_1 \) to be a stationary point of \( E_S \) it is necessary that

\[
0 = \frac{d}{d\lambda} E_S(\phi_\lambda) \bigg|_{\lambda=1} = (2-d) E_1(\phi_1) - d E_2(\phi_1). \tag{1.39}
\]

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 \( E_2(\phi_1) = 0 \) which implies \( \frac{\partial V}{\partial \phi^a} = 0 \). 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 (1.36). 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 $\lambda$.

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

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

(1.40)

The equation of motion for $\Lambda$ is $\Lambda = \sqrt{\lambda V}$ and when this equation is used in (1.40) it gives back (1.36). Thus (1.40) is classically equivalent to (1.36). The advantage of the action (1.40) is that it remains well defined in the limit $\lambda \to \infty$. In fact, it reduces to

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

(1.41)
1.2. SCALAR FIELDS IN OTHER DIMENSIONS

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, where $G = SO(3)$, $H = SO(2)$ and $G/H = S^2$. We can solve the constraint $\phi^a \phi^a = f^2$ expressing the three fields $\phi^a$ in terms of only two independent fields $\varphi^a$. There are infinitely many ways of doing this. For example we could choose $\varphi^a$ to be the spherical coordinates ($\varphi^1 = \Theta$, $\varphi^2 = \Phi$):

$$\begin{align*}
\phi^1 & = f \sin \Theta \cos \Phi \\
\phi^2 & = f \sin \Theta \sin \Phi \\
\phi^3 & = f \cos \Theta
\end{align*}$$

Introducing into (1.41), we find the action

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

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

$$\begin{align*}
\phi^1 & = f \frac{4\omega_1}{\omega_1^2 + \omega_2^2 + 4} \\
\phi^2 & = f \frac{4\omega_2}{\omega_1^2 + \omega_2^2 + 4} \\
\phi^3 & = f \frac{\omega_1^2 + \omega_2^2 - 4}{\omega_1^2 + \omega_2^2 + 4}
\end{align*}$$

Introducing in (1.41),

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

In any case the action has the form

$$S = -\frac{f^2}{2} \int d^{d+1}x \partial_\mu \varphi^a \partial^\mu \varphi^b h_{ab}(\varphi),$$

where $h_{ab}(\varphi)$ is the standard metric on the sphere $S^2$ of unit radius, written in the chosen coordinate system, and $f^2$ is a constant with dimension of mass.
We recall that the original linear model (1.36) is a standard example of the Goldstone theorem. Picking a vacuum state, for example \( \phi = (0, \ldots, 0, f) \) 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 (1.48).

This discussion can be generalized to scalar fields carrying a representation of any Lie group \( G \). If \( \phi_0 \) is a minimum of the potential, every other point in the orbit of \( G \) through \( \phi_0 \) is also a minimum. We assume that all the minima belong to a single orbit. If \( H \) is the stabilizer of \( \phi_0 \), 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 (1.48) 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

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

(1.49)

Assume that \( \delta \varphi^\gamma = \epsilon^a K_a^\gamma(\varphi) \), where \( \epsilon^a \) 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

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

Then it is 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

\[
\partial_\mu \partial^\mu \varphi^\gamma + \partial_\mu \varphi^\alpha \partial^\mu \varphi^\beta \Gamma_{\alpha\beta}^\gamma(\varphi) = 0.
\]

(1.50)

where \( \Gamma_{\alpha\beta}^\gamma \) 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.

---

6See 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)
1.2. SCALAR FIELDS IN OTHER DIMENSIONS

1.2.4 Power counting

We conclude this section with some remarks on the quantization of non-linear sigma models, with the action (1.48). Since the metric is in general a non-polynomial function, the fields have to be dimensionless. Therefore the constant \( f^2 \) multiplying the action must have dimension \( L^{2-n} \), where \( n = d + 1 \) is the dimension of spacetime. In two spacetime dimensions, and only in two, we can choose \( f^2 = 1 \). In order to give the scalar fields their canonical dimension we absorb first the constant \( f^2 \) in the fields defining \( \bar{\varphi}^\alpha = f \varphi^\alpha \). The dimension of \( \bar{\varphi} \) is then \([ \bar{\varphi}^\alpha ] = L^{2-n} \). Now the action reads

\[
S = -\frac{1}{2} \int d^n x \, \partial_\mu \bar{\varphi}^\alpha \partial^\mu \bar{\varphi}^\beta h_{\alpha\beta} \left( \frac{\bar{\varphi}}{f} \right).
\]

(1.51)

The metric \( h_{\alpha\beta} \left( \frac{\bar{\varphi}}{f} \right) \) is still dimensionless. In order to separate the kinetic term from the interaction terms we have to fix some constant background \( \bar{\varphi}_0^\alpha \), write \( \bar{\varphi}^\alpha = \bar{\varphi}_0^\alpha + \eta^\alpha \), and expand the metric in Taylor series in \( \eta \):

\[
h_{\alpha\beta} \left( \frac{\bar{\varphi}}{f} \right) = h_{\alpha\beta} \left( \frac{\bar{\varphi}_0}{f} \right) + \frac{1}{f} \partial_\gamma h_{\alpha\beta} \left( \frac{\bar{\varphi}_0}{f} \right) \eta^\gamma + \frac{1}{2 f^2} \partial_\gamma \partial_\delta h_{\alpha\beta} \left( \frac{\bar{\varphi}_0}{f} \right) \eta^\gamma \eta^\delta + \cdots
\]

(1.52)

where we write \( \partial_\gamma \) for \( \partial / \partial \bar{\varphi}^\gamma \). 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 \( \partial_\gamma \partial_\eta \eta^m \), is \([ \left( \frac{1}{f} \right)^m ] = L^{m-2} \). 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 coupling constant \( g = 1/f \), of mass dimension \( 1 - n/2 \). If one is interested in the perturbative treatment of this theory, it is more appropriate to write the prefactor of the action (1.48) as \(-1/(2g^2)\), instead of \(-f^2/2\).

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 (1.39) 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 $\varphi$ representing a map from $\mathbb{R}^2$ to $S^2$. Finiteness of the static energy

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

(1.53)

demands that $\partial_i \varphi \to 0$ as $r \to \infty$. Thus $\varphi$ 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 $\infty$, 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 $Q = \Gamma_*(S^2, S^2)$. This space consists of infinitely many connected components: $\pi_0(Q) = \pi_2(S^2) = \mathbb{Z}$ (see Appendix XXXXXX). So we can write $Q = \bigcup_{i \in \mathbb{Z}} Q_i$; the number $i$ labelling the homotopy classes is known as the winding number. A general formula for this quantity is given in Eq.(A.3). In spherical and stereographic coordinates it has the expression

$$W(\Theta, \Phi) = \frac{1}{4\pi} \int d^2 x \sin \Theta \varepsilon^{ij} \partial_i \Theta \partial_j \Phi ,$$

(1.54)

$$W(\omega^1, \omega^2) = \frac{1}{4\pi} \int d^2 x \frac{16}{(\omega_1^2 + \omega_2^2 + 4)^2} \varepsilon^{ij} \partial_i \omega^1 \partial_j \omega^2 ,$$

(1.55)
1.3. NONLINEAR SIGMA MODEL IN $D=2$

respectively. In any coordinate system

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

One can think of $W$ as a function on $Q$: it is constant and equal to $i$ on each connected component $Q_i$. It is intuitively clear that since the time evolution is a continuous curve in $Q$, the value of the winding number cannot change, so $W$ must be a constant of motion of the theory. This conclusion can be substantiated by the following calculation. We define a topological current

$$J^\lambda_T = \frac{1}{8\pi} \varepsilon^{\lambda\mu\nu} \partial_\mu \varphi^\alpha \partial_\nu \varphi^\beta \sqrt{\text{det} h} \varepsilon_{\alpha\beta},$$

which is identically conserved:

$$\partial_\lambda J^\lambda_T = \frac{1}{8\pi} \varepsilon^{\lambda\mu\nu} \partial_\lambda \varphi^\gamma \partial_\mu \varphi^\alpha \partial_\nu \varphi^\beta \varepsilon_{\alpha\beta} \frac{\partial \sqrt{\text{det} h}}{\partial \varphi^\gamma} = 0,$$

because the indices $\alpha, \beta, \gamma$ run from 1 to 2 and at least one index must be repeated. One sees immediately that the topological charge is equal to the winding number:

$$Q_T = \int d^2x J^0_T = \frac{1}{8\pi} \int d^2x \varepsilon^{ij} \partial_i \varphi^\alpha \partial_j \varphi^\beta \varepsilon_{\alpha\beta} \sqrt{\text{det} h} = W(\varphi). \quad (1.56)$$

There follows that $Q_T=W$ is a constant of motion. (See also exercise 1.3.1).

1.3.2 Dynamics

Let us look at the absolute minimum of the static energy (1.53) in each topological sector $Q_i$. Consider the following inequality $^7$

$$0 \leq \int d^2x h_{\alpha\beta} \left( \partial_i \varphi^\alpha \pm \varepsilon_{ik} \partial_k \varphi^\gamma \varepsilon_{\gamma\delta} h^{\delta\alpha} \sqrt{\text{det} h} \right) \left( \partial_i \varphi^\beta \pm \varepsilon_{ij} \partial_j \varphi^\epsilon \varepsilon_{\epsilon\alpha} h^{\alpha\beta} \sqrt{\text{det} h} \right) =$$

$$= \int d^2x \left[ 2h_{\alpha\beta} \partial_i \varphi^\alpha \partial_i \varphi^\beta \pm 2\varepsilon_{ij} \partial_i \varphi^\alpha \partial_j \varphi^\epsilon \varepsilon_{\epsilon\alpha} \sqrt{\text{det} h} \right] = \frac{4}{f^2} E_S \mp 16\pi W$$

where we used $h_{\gamma\epsilon} = (\text{det} h) \varepsilon_{\gamma\delta} \varepsilon_{\epsilon\phi} h^{\delta\phi}$. If $W > 0$ (resp. $W < 0$) the inequality with the upper sign (resp. lower sign) is stronger. There follows that

$$E_S \geq 4\pi f^2 |W| .$$

Furthermore, equality holds if and only if

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

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 (1.57) are first order equations, and therefore simpler than the second order Euler-Lagrange equations. It is convenient to specialize the discussion to stereographic coordinates $\omega^1$ and $\omega^2$. Equation (1.57) reduces to

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

and spelling these out

$$\partial_1 \omega^1 = \pm \partial_2 \omega^2 ,$$
$$\partial_2 \omega^1 = \mp \partial_1 \omega^2 .$$

If we define $\omega = \omega^1 + i \omega^2$ and $z = x^1 + i x^2$ we recognize (1.59) 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 (1.59). For example $\omega(z) = z^n$ and $\omega(z) = (z^*)^n$, with $n \geq 0$, are solutions of (1.59). Note that for large $|z|$, $\omega$ 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 $\varphi \in \Gamma_\ast(S^2, S^2)$ with winding number $W = n$ and $W = -n$ respectively. They are absolute minima of the static energy in the sectors $Q_n$ and $Q_{-n}$ respectively ($n \geq 0$).

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

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

(1.60)

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

(1.61)
where the complex number $z_0$ gives the position of the center of the soliton, the angle $\alpha$ its “internal orientation” and the positive real number $\lambda$ its scale. The parameters $z$, $\alpha$ and $\lambda$ will be called the “collective coordinates” of the soliton.

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 $\varphi$ 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.

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 (1.48) 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 \( Q = \Gamma(S^d, N) \). Therefore, there is room for the existence of topological solitons whenever \( \pi_0(Q) = \pi_d(N) \neq 0 \). 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 \( A_\mu \), 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 \( q_\alpha \), where \( \alpha \) is an index that runs over the six flavors. The quark part of the QCD action is

\[
S_q = \sum_\alpha \int d^4x \bar{\psi}_\alpha (i\gamma^\mu D_\mu - m_\alpha) \psi_\alpha.
\]

where \( D_\mu \) 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

\[
\delta_{V\alpha} \psi = i\alpha \psi \quad ; \quad \delta_{V\alpha} \bar{\psi} = -i\alpha \bar{\psi}
\]

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

\[
\delta_{V, \epsilon} \psi = \epsilon^a T_a \psi \quad ; \quad \delta_{V, \epsilon} \bar{\psi} = -\bar{\psi} \epsilon^a T_a
\]

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:

$$\delta_{A\alpha}\psi = i\alpha\gamma^A\psi \quad ; \quad \delta_{A\alpha}\bar{\psi} = i\alpha\bar{\psi}\gamma^A \quad U(1)_A ; \quad (1.63)$$

$$\delta_{A\epsilon}\psi = \epsilon^a T_a \gamma^A \psi \quad ; \quad \delta_{A\epsilon}\bar{\psi} = \bar{\psi}\epsilon^a T_a \gamma^A \quad SU(N)_A . \quad (1.64)$$

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:

$$j_V^\mu = \bar{\psi}\gamma^\mu \psi \quad \text{for } U(1)_V \quad (1.65)$$

$$j_A^\mu = \bar{\psi}\gamma^\mu \gamma^A \psi \quad \text{for } U(1)_A \quad (1.66)$$

$$j_{V\epsilon}^\mu = \bar{\psi}\gamma^\mu \epsilon^a T_a \psi \quad \text{for } SU(N)_V \quad (1.67)$$

$$j_{A\epsilon}^\mu = \bar{\psi}\gamma^\mu \gamma^A \epsilon^a T_a \psi \quad \text{for } SU(N)_A \quad (1.68)$$

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

$$\{\psi^{a\alpha}(\vec{x}, t), \bar{\psi}^{\beta\beta}(\vec{y}, t)\} = \delta^{a\alpha}_\beta \delta^\beta_\gamma (\vec{x} - \vec{y}) , \quad (1.69)$$

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

$$[j_V^0, j_V^0] = [j_A^0, j_A^0] = [j_{V\epsilon}^0, j_{V\epsilon}^0] = [j_{A\epsilon}^0, j_{A\epsilon}^0] = 0 ; \quad (1.70)$$

$$[j_{V\epsilon1}^0, j_{V\epsilon2}^0] = j_{V\epsilon1,\epsilon2}^0 ; \quad (1.71)$$

$$[j_{A\epsilon1}^0, j_{A\epsilon2}^0] = j_{A\epsilon1,\epsilon2}^0 ; \quad (1.72)$$

$$[j_{V\epsilon1}^0, j_{A\epsilon2}^0] = j_{V\epsilon1,\epsilon2}^0 ; \quad (1.73)$$

One can verify that these are the algebras implied by (1.64).

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 $\gamma^A$ satisfies $(\gamma^A)^2 = 1$, the operators

$$P_\pm = \frac{1 \pm \gamma^A}{2} \quad (1.74)$$
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

\[
\begin{align*}
\tilde{j}^\mu_{Le} &= \frac{j^\mu_{Ve} - j^\mu_{Ae}}{2} = \bar{\psi}\gamma^\mu P_- \epsilon^a T_a \psi ; \\
\tilde{j}^\mu_{Re} &= \frac{j^\mu_{Ve} + j^\mu_{Ae}}{2} = \bar{\psi}\gamma^\mu P_+ \epsilon^a T_a \psi ;
\end{align*}
\]

we can rewrite (1.73) as

\[
\begin{align*}
[j^0_{Le_1}, j^0_{Le_2}] &= j^0_{L[e_1, e_2]} ; \\
[j^0_{Le_1}, j^0_{Re_2}] &= 0 ; \\
[j^0_{Re_1}, j^0_{Re_2}] &= j^0_{R[e_1, e_2]} ;
\end{align*}
\]

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 (1.79) 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 \( SU(2) \). 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 \( SU(3) \).

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 \( U(x) \in SU(N) \).

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

\[
S = \frac{f^2}{4} \int d^4x \text{tr}(U^{-1} \partial_\mu U U^{-1} \partial^\mu U) .
\]  

(1.80)

(We are using an inner product in the Lie algebra such that \( \text{tr}(T_a T_b) = -\frac{1}{2} \delta_{ab} \). In the case \( N = 2 \), \( T_a = -\frac{i}{2} \sigma_a \), where \( \sigma^a \) are the Pauli matrices). If we choose a coordinate system on the group and call \( \varphi^a(x) \) the coordinates of the group element \( U(x) \), the action (1.80) can be shown to be identical to the nonlinear sigma model action (1.48) with riemannian metric \( h_g(v,w) = \text{tr}(g^{-1} v g^{-1} w) \).

(See Exercise 1.4.1). The advantage of the form (1.80) is that it makes the \( SU(N)_L \times SU(N)_R \) invariance of the theory very transparent: if we transform

\[
U \to g_L^{-1} U g_R ,
\]

the (constant) group elements \( g_L \) and \( g_R \) cancel (the latter using ciclicity of the trace). 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 with \( g_L = g_R \).

For the present purposes, the most useful coordinates on \( SU(N) \) are the normal coordinates \( \pi^\alpha \), defined by:

\[
U(x) = e^{2\pi^\alpha(x)T_\alpha/f} \quad \text{(1.81)}
\]

where \( T_\alpha \) is a basis in the Lie algebra of \( SU(N) \), satisfying \([T_\alpha, T_\beta] = f_{abc}T_\gamma\). Note that the coordinates have been scaled as in (1.51) so as to have the canonical dimension of mass.

Using (1.81), (1.80) can be expanded as

\[
\int d^4x \left[ -\frac{1}{2} \partial_\mu \pi^\alpha \partial^\mu \pi^\alpha + \frac{1}{f^2} \varepsilon^{abc} \partial_\mu \pi^\alpha \varepsilon^{ade} \pi^\beta \partial_\nu \pi^\gamma + \ldots \right]. \quad \text{(1.82)}
\]

This corresponds to the expansion (1.52) 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 \( \pi \) 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 1.4.1):

\[
W(U) = -\frac{1}{24\pi^2} \int d^3x \varepsilon^{\lambda\mu\nu} \text{tr} \left( U^{-1} \partial_\lambda UU^{-1} \partial_\mu UU^{-1} \partial_\nu U \right). \quad \text{(1.83)}
\]

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

\[
W(U_1U_2) = W(U_1) + W(U_2) \quad ; \quad W(U^{-1}) = -W(U). \quad \text{(1.84)}
\]
A field configuration of the form

\[ U(\vec{x}) = \exp[T_\alpha \hat{x}^\alpha g(r)] \] (1.85)

where \( \hat{x}^\alpha = \frac{x^\alpha}{r} \) and \( g \) is a function which is \(-2\pi\) 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 (1.80). In fact, from (1.39) we get

\[
\left. \frac{dE(\phi_\lambda)}{d\lambda} \right|_{\lambda=1} = -E(\phi_1) < 0,
\]

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. 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

\[
S = \int d^4x \left[ \frac{f^2}{4} \text{tr}(U^{-1} \partial_\mu U U^{-1} \partial^\mu U) + \frac{1}{32e^2} \text{tr}[U^{-1} \partial_\mu U, U^{-1} \partial_\nu U][U^{-1} \partial^\mu U, U^{-1} \partial^\nu U] \right] (1.86)
\]

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.

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. Numerical approximations are

---

necessary. However, as in section (1.1), we can apply qualitative arguments to infer the existence of a solution and derive some of its properties. Suppose that the function $g$ goes from $-\pi$ to zero within a distance $\ell$ 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 size of the soliton results from a balance between these two terms, and turns out to be of order $1/fe$. Note that for $e \to \infty$, $\ell$ tends to zero, in accordance with the argument in the end of section (1.2). The mass of the soliton is of the order $f/e$.

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

$$F^a_{\mu\nu} = \partial_\mu A^a_\nu - \partial_\nu A^a_\mu + e f^{abc} A^b_\mu A^c_\nu,$$

(1.87)

where $[T^a, T^b] = f^{abc} T^c$ and $e$ is the coupling constant of the theory. The Yang–Mills action is

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

It is invariant under local gauge transformations

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

where $g : \mathbb{R}^{d+1} \to G$ 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

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

where the curvature is now defined by

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

The nonabelian gauge transformations then read

$$A_\mu \rightarrow g^{-1} A_\mu g + g^{-1} \partial_\mu g, \quad F_{\mu\nu} \rightarrow g^{-1} F_{\mu\nu} g. \quad (1.91)$$

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

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

where $E_i^a = F_{0i}^a = \partial_0 A_i^a - D_i A_0^a$ is the nonabelian “electric” field (we have used the notation $D_i A^a_0 = \partial_i A_0^a + e f_{abc} A^b_0 A^c_0$; 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 fields are

$$P_0^a \equiv \frac{\partial \mathcal{L}_{YM}}{\partial A_a^0} = 0, \quad P_i^a \equiv \frac{\partial \mathcal{L}_{YM}}{\partial_0 A_i^a} = E_i^a.$$

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

$$H_c = \int d^d x \left[ \frac{1}{2} P_i^a P_i^a + \frac{1}{4} F^a_{ij} F^a_{ij} - A_0^a G_a \right], \quad (1.92)$$
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 \( A^0_0 \) 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 \( g(x,t) = P \exp \left( -e \int^t dt' A_0(x,t') \right) \), where \( P \) stands for path ordering). This leaves the freedom of performing time-independent gauge transformations. In this gauge \( E^a_i = A^a_i \), 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 \( A_\mu \) 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). \(^9\)

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

\[
E_1 = \frac{1}{2} \int d^d x \left( D_i A_0 \right)^2 > 0 \quad \text{and} \quad E_2 = \frac{1}{4} \int d^d x \left( F_{ij}^a \right)^2 > 0.
\]

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

\[
A^{(\sigma,\lambda)}_{(\sigma,\lambda)}(x) = \sigma \lambda A^0_0(\lambda x), \quad (1.93)
\]

\[
A^{(\sigma,\lambda)}_1(x) = \lambda A^1_0(\lambda x). \quad (1.94)
\]

We have \( E_1(A_{(\sigma,\lambda)}) = \sigma^2 \lambda^{4-d} E_1(A_{(1,1)}) \) and \( E_2(A_{(\sigma,\lambda)}) = \lambda^{4-d} E_2(A_{(1,1)}) \). For

1.6. VORTICES

$A_{(1,1)}$ to be a solution of the field equations we must have

\[
0 = \frac{d}{d\lambda} L \bigg|_{\lambda = \sigma = 1} = (4 - d)L(A_{(1,1)}),
\]
\[
0 = \frac{d}{d\sigma} L \bigg|_{\lambda = \sigma = 1} = 2E_1(A_{(1,1)}),
\]

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

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.\footnote{H.B. Nielsen and P. Olesen, Nucl. Phys. B61 45 (1973).} The dynamical variables are a $U(1)$ gauge field $A_\mu$ coupled to a complex scalar field $\phi$, with action

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

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 $A = A^1 T = -iA^1$. The field $A_\mu$ used in this section is $A^1_\mu$ 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

\[
A_\mu \rightarrow A'_\mu = A_\mu + \frac{i}{e} g^{-1} \partial_\mu g = A_\mu - \frac{1}{e} \partial_\mu \alpha, \quad \phi \rightarrow \phi' = g^{-1} \phi = e^{-i\alpha(x)} \phi.
\]

(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
\[
E_K = \int d^2x \left[ \frac{1}{2} \dot{\phi}^2 + \frac{1}{2} |\phi|^2 \right].
\]
and $E_S$ is the static energy
\[
E_S = \int d^2x \left[ \frac{1}{2} B^2 + \frac{1}{2} |D_i\phi|^2 + \frac{\lambda}{4} \left( |\phi|^2 - f^2 \right)^2 \right]. \tag{1.99}
\]
where $B = F_{12}$. The absolute minimum of $E_S$, the classical vacuum, occurs for
\[
B = 0, \quad D_i\phi = 0, \quad |\phi| = f. \tag{1.100}
\]
A particular solution of these conditions is
\[
A_i = 0, \quad \phi = f. \tag{1.101}
\]
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 $m_A = e f$ and a scalar field with mass $m_S = \sqrt{2\lambda f}$. Any gauge transformation of (1.101) $A_i = \frac{1}{e} g^{-1} \partial_i g$, $\phi = g^{-1} f$ 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 $(A, \phi)$ will have finite energy only if the conditions (1.100) are satisfied asymptotically as $r \to \infty$. This requires that
\[
\begin{align*}
\phi(r, \theta) & \to \phi_\infty \quad = f e^{-i\alpha_\infty}, \quad (1.102) \\
A_i(r, \theta) & \to - \frac{1}{e} \partial_i \alpha_\infty, \quad (1.103)
\end{align*}
\]
where $\alpha_\infty$ depends only on the angular variable $\theta$ parameterizing the “circle at infinity” $S^1_\infty$. We see that unlike the case of the sigma model, the condition $D_i\phi \to 0$ does not imply that $\phi$ tends to a constant at infinity: as long as $|\phi| \to f$, any dependence of $\phi$ on the angle $\theta$ is permitted, because one can always compensate for this dependence by choosing $A_i = \frac{1}{ie} \partial_i \phi$. 


The asymptotic behaviour of the field $\phi$ 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

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

The field $\phi$ 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 $S^1_\infty$.

![Figure 1.6: A homotopy in fields space. The circle of unit radius is the locus of the minima of the potential.](image)

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 $\phi_\infty$, 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 $\phi_\infty$, so there must be in the theory a topological conservation law. In fact, consider the topological current

$$J_T^\lambda = \frac{1}{2\pi i} \epsilon^{\lambda \mu \nu} \partial_\mu \hat{\phi} \star \partial_\nu \hat{\phi},$$
CHAPTER 1. $\pi_0(Q)$ AND SOLITONS

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

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

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

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

where $\Phi$ 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:

$$\Phi = \frac{2\pi}{e} n. \quad (1.105)$$

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

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

where $A$ and $F$ are functions of the radius such that $A(r) \to \frac{1}{cr}$ 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.
1.6. VORTICES

What we called scalar electrodynamics is called in condensed matter physics 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

$$\phi(x) \rightarrow \exp(i 2e_\alpha(x)/\hbar)\phi(x),$$

where $-e$ is the electron charge and $\alpha$ is identified mod $2\pi$. 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

$$\varphi \rightarrow \varphi + \alpha.$$  \hfill (1.107)

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

$$\tilde{\psi}(x) = \psi(x) \exp(i q \varphi(x)/\hbar),$$ \hfill (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

$$D_\mu \varphi = \partial_\mu \varphi - A_\mu.$$ \hfill (1.109)

We can define the current

$$J^\mu = \frac{\delta \mathcal{L}_m}{\delta A^\mu_\mu}.$$ \hfill (1.109)

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

$$\partial_0 J^0 = \partial_i \frac{\delta \mathcal{L}_m}{\delta A^0_i} = -\partial_0 \frac{\delta \mathcal{L}_m}{\delta \varphi} = \partial_i \frac{\delta \mathcal{L}_m}{\delta \partial_0 \varphi} = -\partial_i \frac{\delta \mathcal{L}_m}{\delta A_i} = -\partial_i J^i.$$
The main assumption that is needed to describe superconductivity is that the Goldstone boson is covariantly constant:

\[ D_\mu \varphi = 0 \quad \text{(1.110)} \]

This condition will follow if we make the rather reasonable assumption that in the Lagrangian for \( \varphi \), 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

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

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

\[ A_i = \partial_i \varphi \quad \text{(1.111)} \]

implies that the magnetic field must be zero:

\[ B_i = 0 \quad \text{.} \]

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, \( \varphi \) can be set to any fixed constant by a transformation (1.107). Now consider a thick torus made of superconductor and let \( \ell \) be a closed loop deep in the material. Integrating (1.111) on this loop and using Stokes’ theorem we find that

\[ \Delta \varphi = \int_\ell A = \int_S B = \Phi \quad , \]

where \( \Phi \) is the magnetic flux through a surface \( S \) bounded by the loop \( \ell \). 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:

\[ \Phi = \frac{\pi \hbar}{e} n \quad . \quad \text{(1.112)} \]

Because of this, the current in the superconductor cannot decay continuously. The fact that the current is not affected by ordinary electrical resistance can
1.6. VORTICES

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 $\rho$ 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

$$\lambda = \frac{1}{ef} = \frac{1}{m_V}, \quad (1.113)$$

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

$$\xi = \frac{1}{\sqrt{2}m_S}. \quad (1.114)$$

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

$$\kappa = \frac{\lambda}{\xi}$$

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 penetrates the superconductor in the form of thin tubes, which in Landau-Ginzburg theory are described by 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. The density of vortices increases with the external magnetic field, up to a second, higher, critical field, where superconductivity is lost.
For our purposes, type II superconductors are more interesting. Since
the core of the tube is not superconducting, the topology of a piece of su-
perconductor that is pierced by a vortex line is the same as that of the thick
torus discussed earlier and 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

\[
\begin{align*}
\partial_\mu F^{\mu\nu} &= 4\pi J'_\nu(E), \\
\partial_\mu \ast F^{\mu\nu} &= 0,
\end{align*}
\]

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

\[
F_{\mu\nu} = \begin{pmatrix}
0 & E_1 & E_2 & E_3 \\
-E_1 & 0 & +B_3 & -B_2 \\
-E_2 & -B_3 & 0 & B_1 \\
-E_3 & B_2 & -B_1 & 0
\end{pmatrix}
\quad \ast F_{\mu\nu} = \begin{pmatrix}
0 & -B_1 & -B_2 & -B_3 \\
B_1 & 0 & E_3 & -E_2 \\
B_2 & -E_3 & 0 & E_1 \\
B_3 & E_2 & -E_1 & 0
\end{pmatrix}
\]

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

\[
\begin{align*}
F &\rightarrow \cos \theta F + \sin \theta \ast F \\
\ast F &\rightarrow - \sin \theta F + \cos \theta \ast F.
\end{align*}
\]
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 (1.116) is identically zero. They could be made symmetric under duality transformations by introducing a
\[ \partial_\mu F^{\mu\nu} = 4\pi J^\nu_M \] (1.119)
and postulating the transformation
\[ J_E \to \cos \theta J_E + \sin \theta J_M , \]
\[ J_M \to -\sin \theta J_E + \cos \theta J_M . \] (1.120)
That \( J^\nu_M \) is a magnetic current is seen by observing for example that the time component of (1.119) would read \( \text{div}\, B = 4\pi \rho_M \), 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 \( A_\mu \) 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
\[ B_i = \frac{Q_M}{r^2} \hat{x}^i , \] (1.121)
describing a static pointlike magnetic monopole in the origin, solves the equation \( \text{div}\, B = 4\pi Q_M \delta(r) \). Since the field is singular in the origin, one can remove this point from space and regard (1.121) as a smooth field on \( \mathbb{R}^3 \setminus \{0\} \). Since the field \( B \) given in (1.121) is divergence free on \( \mathbb{R}^3 \setminus \{0\} \), 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 particular 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 \( \phi^a \) in the adjoint (triplet) representation.
We use the unscaled gauge fields, with curvature (1.87) and action (1.88). The total Lagrangian density is

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

(1.122)

where $D_\mu \phi^a = \partial_\mu \phi^a + e \varepsilon_{abc} A^b_\mu \phi^c$. The structure constants of the Lie algebra of $SU(2)$ are $f_{abc} = \varepsilon_{abc}$ (in the adjoint representation the generators are $(T^a)_b{}^c = -\varepsilon_{abc}$). 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 $A^a_0 = 0$. Then $F^a_{0i} = \partial_0 A^a_i$, $D_0 \phi^a = \partial_0 \phi^a$. The static energy is

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

(1.123)

Its absolute minimum is obtained for

$$
F^a_{ij} = 0,
$$

$$
D_i \phi^a = 0,
$$

$$
\phi^a \phi^a = f^2,
$$

(1.124)

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 $A^a_i = 0$, $\phi^a = \bar{\phi}^a = (0, 0, f)$ 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 (1.125) 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 $\phi$ defines a map $\phi_\infty : S^2_\infty \to S^2_{\text{int}}$, where $S^2_\infty$ denotes the “sphere at infinity” in $\mathbb{R}^3$ and $S^2_{\text{int}}$ 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 (1.125) does not restrict the map $\phi$ itself. The asymptotic field $\phi^a_\infty$ can depend on the angles in an arbitrary way; the condition $D_i \phi \to 0$ can then be solved by

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

(1.125)
for an arbitrary constant $\alpha_i$.

The scalar fields $\phi$ fall into classes, labelled by the winding number of the map $\phi_\infty$. 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 $\phi_\infty$. 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

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

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

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

The last equality can be proven by choosing a particular coordinate system on $S^2$, for example the spherical coordinates (1.44), and comparing with (1.55).

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 $F_{\mu\nu} = \partial_\mu A_3^\nu - \partial_\nu A_3^\mu$.

Following 't Hooft, we can generalize this to position-dependent Higgs fields. Let $A_\mu = A_\mu^{\hat{\phi}^a}$. We define an abelian electromagnetic field $F_{\mu\nu}$ by

$$F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu - \frac{1}{e} \varepsilon_{abc} \partial_\mu \hat{\phi}^a \partial_\nu \hat{\phi}^b \hat{\phi}^c.$$

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

$$F_{\mu\nu} = \hat{\phi}^a F_{\mu\nu}^a - \frac{1}{e} \varepsilon_{abc} \hat{\phi}^a D_\mu \hat{\phi}^b D_\nu \hat{\phi}^c.$$
which is manifestly invariant under $SU(2)$ gauge transformations. This tensor does not obey the Bianchi identities. Instead
\[
\partial_\nu \star F_{\mu\nu} = -\frac{4\pi}{e} J^\mu_T .
\]  
(1.130)
Comparing with (1.119), we see that we can interpret $\frac{1}{e} J^\mu_T$ as a magnetic current. The corresponding magnetic charge is
\[
Q_M = \frac{1}{e} Q_T = \frac{1}{e} W .
\]  
(1.131)
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
\[
\begin{align*}
\phi^a &= \frac{2}{r} F(r) , \\
A^a_i(x) &= \varepsilon_{aij} \frac{x_j}{r} A(r) , \\
A^a_0 &= 0 ,
\end{align*}
\]  
(1.132)
where $F(r) \to f$ and $A(r) \to \frac{1}{er}$ for $r \to \infty$. 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
\[
B_i = \frac{1}{2} \varepsilon_{ijk} F_{jk} \to \frac{\hat{r}^i}{e r^2}
\]  
(1.133)
while $E_i = F_{0i} = 0$. Therefore, for large $r$, the abelian field strength becomes identical to the one of the Dirac monopole.

When the ansatz (1.132) 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 $\lambda$
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

$$E = \int d^3x \left[ \frac{1}{4} F^a_{ij} F^a_{ij} + \frac{1}{2} D_i \phi^a D_i \phi^a \right]$$

$$= \frac{1}{4} \int d^3x \left( F^a_{ij} + \varepsilon_{ijk} D_k \phi^a \right)^2 \pm \frac{1}{2} \int d^3x \varepsilon_{ijk} F^a_{ij} D_k \phi^a . \quad (1.134)$$

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

$$\frac{1}{2} \int d^3x \partial_k \left( \varepsilon_{ijk} F^a_{ij} \phi^a \right) = f \int d\sigma^k B_k = 4\pi f Q_M = \frac{4\pi f}{e} W , \quad (1.135)$$

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

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

with equality holding if and only if

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

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, the explicit form of the functions appearing in (1.132) is

$$F(r) = \frac{f}{\tanh(efr)} - \frac{1}{er} ,$$

$$A(r) = \frac{1}{er} - \frac{f}{\sinh(efr)} . \quad (1.138)$$

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

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.

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 $\epsilon_I^a$ transforms

$$\delta_I \phi^a = \epsilon_{abc} \epsilon_I^b \phi^c .$$

(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 \( \delta_R x^i = \epsilon_R^a x^i \), so

\[
\delta_R \phi^a = \epsilon_{abc} R^b \phi^c. \tag{1.140}
\]

Combining (1.139) and (1.140) we see that \( \phi \) 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 \( \phi \) at infinity is fixed and we do not allow transformations that change it. Since the field \( \phi^a \) 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

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

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.
CHAPTER 1. \( \pi_0(\mathcal{Q}) \) AND SOLITONS
Chapter 2

\( \pi_1(Q) \) and theta sectors

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. We will give several examples of this phenomenon in quantum mechanics and quantum field theory.

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 \( \Phi \) 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 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 \( Q \) which is all of \( \mathbb{R}^3 \) with the solenoid removed and the magnetic field vanishes on \( Q \). The space \( Q \) is multiply connected, with \( \pi_1(Q) = \mathbb{Z} \). Consider the magnetic potential

\[
\mathcal{A} = \theta \frac{\hbar}{2\pi e} d\varphi ,
\]

where \( \theta \) is an arbitrary real parameter, and \( \varphi \) 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 \( Q \). To find the meaning of the parameter \( \theta \), 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 \( F = d\mathcal{A} \) 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, \( \Phi \). 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

\[
H = -\frac{\hbar^2}{2m} \mathcal{D}_i \mathcal{D}^i ,
\]

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

\[
\psi(q) = \psi_0(q) e^{i\frac{\varphi}{\hbar} \int^q \mathcal{A} } ,
\]

obey 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
2.1. THETA SECTORS

difference in the absence of magnetic flux, plus $\xi \oint \mathcal{A} = \theta$. This phase, and hence the interference pattern, varies linearly with flux. When $\theta$ changes by $2\pi$, the phase repeats itself. So the interference pattern has to be periodic in $\Phi$ with period $\frac{2\pi \hbar}{e}$, as observed. This concludes the theoretical explanation of the Aharonov-Bohm effect.

Let us now discuss the mathematical interpretation of this phenomenon. In mathematics, a gauge potential is called a connection and its field strength is called the curvature. Their geometrical interpretation is in terms of fiber bundles, but this language is not strictly necessary for what follows. In the Aharonov-Bohm effect we are interested in connections with zero curvature, that are called flat connections.

The effect of a gauge transformation on the wave function and on the connection is

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

where $g(x) = e^{i\alpha(x)}$ is a function from $Q$ into $U(1)$. We assume that the wavefunctions $\psi$ are periodic both before and after the gauge transformation, and therefore also $g$ has to be a well-defined, single valued function into $U(1)$.

We say that two gauge potentials $\mathcal{A}$ and $\mathcal{A}'$ are $U(1)$-gauge related only if the function $g$ in (3.7) 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 (3.7) 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. There follows that the gauge equivalence classes of $U(1)$ gauge potentials on $Q$ are parameterized by the values of $\theta$ belonging to the domain $0 \leq \theta < 2\pi$. The field strength is not enough to identify the connection entirely, and the Aharonov–Bohm effect shows that quantum physics is sensitive to this additional information.

\footnote{It is not strictly necessary to assume that $\psi$ is periodic. See the discussion below. However we will see there that one does not lose any generality by making this assumption.}
2.1.2 Generalization

One way of interpreting the Aharonov-Bohm effect is to say that there is an ambiguity in the quantization of the electron on the multiply connected space $Q = \mathbb{R}^3 \setminus \{\text{line}\}$. In classical physics a charged particle feels the effect of the electromagnetic field only through the Lorentz force and since in the present case the magnetic field on $Q$ vanishes, the classical electron is completely insensitive to the flux. The Aharonov-Bohm experiment shows that the quantum electron is sensitive to the flux, i.e. it is sensitive to changes of the gauge potentials $A$, even when the field strength $F$ remains the same (in this case, zero). One concludes that to a single classical theory there correspond infinitely many quantum theories, parametrized by the angle $\theta$.

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

$$L = \frac{1}{2} m g_{ij}(q) \dot{q}^i \dot{q}^j + e A_i(q) \dot{q}^i - V(q) . \quad (2.5)$$

The momentum conjugate to $q^i$ is

$$p_i = m g_{ij}(q) \dot{q}^j + e A_i(q) , \quad (2.6)$$

and the canonical hamiltonian is

$$H = \frac{1}{2m} g^{ij}(q) (p_i - e A_i)(p_j - e A_j) + V(q) ,$$

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 $\hat{q}^i(q)$ and $p_i$ with the derivative operator $\hat{p}_i = -i \hbar \frac{\partial}{\partial q^i}$. Then we have

$$\hat{p}_i - e \hat{A}_i = -i \hbar \left( \frac{\partial}{\partial q^i} - \frac{i e}{\hbar} A_i \right) = -i \hbar \mathcal{D}_i , \quad (2.7)$$

where $\mathcal{D}_i$ is the covariant derivative with respect to $A_i$, acting now on wavefunctions $\psi(q)$. Under local phase transformations (3.7) we have $\mathcal{D}_i \psi' = e^{-i \alpha} (\mathcal{D}_i \psi)$. The hamiltonian becomes the operator

$$\hat{H} = -\frac{\hbar^2}{2m} \sqrt{g} \mathcal{D}_i \sqrt{g} \mathcal{D}_j + V , \quad (2.8)$$
where \( g = \det(g_{ij}) \). (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.)

Now let us consider the special case when the connection is flat: \( \mathcal{F} = 0 \). There exists at least locally a function \( \Lambda \) such that

\[
A_i = -\hbar e \partial_i \Lambda \quad (2.9)
\]

so the second term in (2.5) is a total derivative:

\[
L_T = e q^i A_i(q) = -\hbar \frac{d\Lambda}{dt} . \quad (2.10)
\]

This term does not affect the equations of motion and therefore can be neglected in the classical theory. It is called a “topological term”.

Looking at (2.9) one may be tempted to say that \( A \) is a pure gauge. This would be correct if \( A \) was a gauge field for the group of translations, \( \mathbb{R} \). However, in quantum mechanics, \( A \) has to be interpreted as a gauge field for the gauge group \( U(1) \). This is because in first quantization the gauge transformations of electromagnetism are phase transformations of the wave function. The abelian groups \( U(1) \) and \( \mathbb{R} \) have the same Lie algebra and since \( A \) is a Lie-algebra-valued one-form, there is no way to discriminate locally between the two cases. The distinction appears at the global level, in the allowed gauge transformations. In fact, a function \( \alpha : Q \to \mathbb{R} \) defines a true (i.e. single-valued) \( U(1) \) gauge transformation \( g = e^{i\alpha} \) only if \( \alpha \), restricted to any loop in \( Q \), has a polydromy which is an integral multiple of \( 2\pi \). Note that if the loop is homotopic to a constant, \( \alpha \) is necessarily single valued. Therefore, nontrivial \( U(1) \) flat connections can only exist if \( Q \) is multiply connected.

Classically, the Lagrangian (2.10) is completely immaterial. However, at the quantum level, the topological term matters: gauge inequivalent potentials give rise to different quantum theories, as proven by the Aharonov-Bohm effect. Thus, for a given classical theory there will be as many inequivalent quantum theories as there are gauge equivalence classes of flat \( U(1) \) connections on \( Q \).

The set of flat \( U(1) \) connections modulo gauge transformations is classified by the characters of the fundamental group:

\[
\text{Hom}(\pi_1(Q), U(1)) .
\]
To show this we recall that all the gauge invariant information about a connection is contained in its holonomies ("Wilson loops")

\[ \chi(\ell) = e^{i \frac{2\pi}{\theta} \int_{\ell} A} . \]  

(2.11)

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 \( \chi([\ell_1] \cdot [\ell_2]) = \chi([\ell_1])\chi([\ell_2]) \), so \( \chi \) defines a homomorphism from \( \pi_1(Q) \) into \( U(1) \). Conversely, given any character \( \chi \), it can be shown that there exists a flat connection \( A \) such that (2.11) holds.

In the following we shall encounter only two cases: \( \pi_1(Q) = \mathbb{Z} \) and \( \pi_1(Q) = \mathbb{Z}_2 \). In the former case the characters are given by \( \chi_\theta(n) = e^{i\theta n} \).

Since \( \theta \) and \( \theta + 2\pi m \), with \( m \in \mathbb{Z} \) define the same character, we have

\[ \text{Hom}(\mathbb{Z}, U(1)) = U(1) , \]

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

\[ \text{Hom}(\mathbb{Z}_2, U(1)) = \mathbb{Z}_2 . \]

To summarize: if \( \pi_1(Q) \neq 0 \), the correspondence between classical and quantum theories is not unique. All classical theories with lagrangian (2.5) and \( F = dA = 0 \) are equivalent. Quantum theories with lagrangian (2.5) and \( F = dA = 0 \) are only equivalent when the potentials are \( U(1) \)-related. The set of inequivalent quantum theories with the same classical limit is parameterized by the characters of \( \pi_1(Q) \).

There is an alternative description of the \( \theta \) sectors that does not rely on the existence of a topological term in the Lagrangian. This is possible if we observe that on a multiply connected space we could allow the wave functions to be periodic up to a phase. ⁢ 

For example, in the case \( \pi_1(Q) = \mathbb{Z} \), denoting \( \varphi \) a parameter along a non-contractible loop,

\[ \psi(\varphi + 2\pi) = e^{-i\theta} \psi(\varphi) . \]  

(2.12)

³In geometrical language this is possible because the wave functions are not ordinary complex functions but rather sections of a complex line bundle. What we normally call a wave function (a complex-valued function \( \psi \)) is merely the (local) representative of a section relative to a (local) choice of gauge. The sections are always continuous objects,
2.1. \textit{THETA SECTORS}

Such wave functions form a Hilbert space $\mathcal{H}_\theta$ and it is clear that values of $\theta$ differing by integer multiples of $2\pi$ correspond to the same quantum theory, so it is only $\theta \mod 2\pi$ that counts. We could regard the wave functions in $\mathcal{H}_\theta$ as the result of acting on the single-valued wave functions with a singular gauge transformation with parameter $\alpha(x) = -\Lambda(x)$, as we see from (3.7) and (2.9). This is not a $U(1)$ gauge transformation in a strict sense that we have used previously, but it is continuous almost everywhere on $Q$ and we can formally implement it in the quantum theory by an operator $U = e^{-i\Lambda(q)}$ acting on wave functions in such a way that

$$\psi \in \mathcal{H}_0 \mapsto \psi' = U^{-1}\psi \in \mathcal{H}_\theta.$$ 

Such a transformation on states would have to be accompanied also by a transformation on operators. In particular, the transformation would remove the gauge potential from the definition of the Hamiltonian:

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

which becomes independent of $\theta$. In the primed picture the Hamiltonian is independent of $\theta$ and the information on the different quantum theories is carried by the states: for every value of $\theta \mod 2\pi$ one has a different Hilbert space and therefore a different quantum theory.

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 $\theta$ dependence is in the Hamiltonian, in the second in the states. (In this sense, the relation between these descriptions is similar to the relation between the Heisenberg and the Schrödinger picture of quantum mechanics.) 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.

but the local representatives need not be. In the case of flat connections, the bundle is always locally trivializable (as we shall see in the next chapter) and therefore it is possible to choose the gauge globally in a continuous way. This is the gauge in which the wave functions are also continuous, but the gauge potential is generally non-zero. We can locally undo the gauge potential, but this can only be achieved locally. If we try to extend this gauge choice to the whole space $Q$, we find that the gauge transformation needed to achieve it is not single-valued.
CHAPTER 2. $\pi_1(Q)$ AND THETA SECTORS

In the next four sections we shall consider increasingly complicated systems with multiply connected configuration spaces. In most cases they will have $\pi_1(Q) = \mathbb{Z}$. 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 The pendulum

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

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

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 order to recognize the existence of inequivalent quantum theories, we use the freedom of adding to the lagrangian a total time derivative $d\Lambda/dt$ where $\Lambda$ is a given function of $\varphi$. So we add to $L_0$ a term

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

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

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

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 $L_T$ represents the interaction of the particle (carrying charge $e = 1$) with the magnetic potential $A_\theta = \theta \frac{\hbar}{2\pi} d\phi$. 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). So the discussion in the previous section goes through unchanged and we are led to conclude that values of $\theta$ differing by integers multiples of $2\pi$ correspond to gauge equivalent potentials and therefore give equivalent quantum theories.

For each of these theories the Hilbert space is $\mathcal{H} = L^2(S^1)$, the space of complex functions $\psi(\varphi)$ such that

$$\psi(\varphi + 2\pi) = \psi(\varphi) \quad (2.15)$$

and $\int_0^{2\pi} d\varphi \psi^* \psi < \infty$. The Hamiltonian operator is

$$\hat{H}_\theta = -\frac{\hbar^2}{2} D_\varphi D_\varphi + V(\varphi), \quad (2.16)$$

where

$$D_\varphi = \frac{\partial}{\partial \varphi} - i \frac{\theta}{2\pi}. \quad (Note that since the metric on $S^1$ is independent of $\varphi$ in this case there are no ordering ambiguities).

The “$\theta$-Heisenberg” description of the theory is based on a Hilbert space $\mathcal{H}_\theta$ of wave functions satisfying

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

and Hamiltonian

$$\hat{H}_0 = -\frac{\hbar^2}{2} \partial^2_\varphi + V(\varphi)$$

The transformation from the “$\theta$-Schrödinger” to the “$\theta$-Heisenberg” picture is given by the operator $U = e^{-i \frac{\theta}{2\pi} \hat{\varphi}}$. It is easy to see that if $\psi$ is periodic, then $\psi' = U \psi$ satisfies (2.2.1) and $\hat{H}_0 = U \hat{H}_\theta U^{-1}$. Therefore $U$ is an isomorphism of $\mathcal{H}$ to $\mathcal{H}_\theta$ mapping $\hat{H}_\theta$ to $\hat{H}_0$. 
2.2.2 Spin and statistics

Before coming to the field theoretic examples it is worthwhile mentioning that quantum spin and statistics can also be seen as a manifestation of the same type of ambiguity that leads to the existence of theta sectors.

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 \( \text{Hom}(\mathbb{Z}_2, U(1)) = \mathbb{Z}_2 \) 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:

\[
\psi(t + 2\pi) = e^{i\theta} \psi(t),
\]

where \( t \) is some parameter along the loop. In the case of the rotation group, the fundamental noncontractible loop consists of a rotation of the body by an angle \( 2\pi \) about some axis. Therefore (2.17) describes the behaviour of the wave function under a \( 2\pi \) 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 \( 2\pi \). So we learn that \( \theta \) is equal to \( 2\pi s \mod 2\pi \). In \( d > 2 \) the spin can be integer, corresponding to single-valued wave functions, if \( \theta = 2\pi n \), or half integer, corresponding to wave functions that change sign under \( 2\pi \) rotations, if \( \theta = \pi n \) with \( n \) odd. In two dimensions the spin can take any real value and the corresponding particles are called anyons.

For a multiparticle system, the statistical parameter \( \sigma \) is defined by

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

(2.18)

The usual Bose–Einstein and Fermi–Dirac statistics correspond to \( \sigma \) 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, because in this case the statistics could only be bosonic ((2.18) is compatible with \( \vec{x}_1 = \vec{x}_2 \) only for integer \( \sigma \)). The configuration space is then \( \mathcal{Q} = (\mathbb{R}^{2d} \setminus \Delta)/S_2 \), where \( \Delta \) is the subset of \( \mathbb{R}^{2d} \) for which the particle positions 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 \((x_{CM}, \Delta \vec{x}) = (\frac{\vec{x}_1 + \vec{x}_2}{2}, \frac{\vec{x}_2 - \vec{x}_1}{2})\) shows that the topology of
2.3. SPHERICAL SIGMA MODELS

the space $\mathbb{R}^d \setminus \Delta$ is $\mathbb{R}^d \times \mathbb{R}^+ \times S^{d-1}$ (here $\mathbb{R}^d$ is parametrized by $\vec{x}_{CM}$, $\mathbb{R}^+$ 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 $(x_{CM}, \Delta \vec{x}) \to (x_{CM}, -\Delta \vec{x})$ and therefore acts antipodally on $S^{d-1}$; the quotient has topology $\mathbb{R}^d \times \mathbb{R}^+ \times \mathbb{RP}^{d-1}$ where $\mathbb{RP}^{d-1} = S^{d-1}/\mathbb{Z}_2$ 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. For $d = 2$, already $\mathbb{R}^d \setminus \Delta$ has a nontrivial fundamental group, equal to $\mathbb{Z}$, and $\pi_1(Q) = \mathbb{Z}$ too. In this case the inequivalent quantizations are labelled by the angle $\sigma$; one then speaks of fractional statistics. These considerations can be generalized to the case of $N$ indistinguishable particles.

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 $\varphi^\alpha$ 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 $Q = \Gamma_+(S^1, S^2)$, where the constant time spacelike surfaces $\mathbb{R}$ have been compactified to $S^1$ due to the requirement of finiteness of the energy \(^4\). This space is called the loop space of $S^2$. Its fundamental group is $\pi_1(Q) = \pi_2(S^2) = \mathbb{Z}$ (see Appendix F). So this theory will admit theta sectors, labelled by an angle $0 \leq \theta < 2\pi$. The fundamental non-contractible loop in $Q$ (i.e. the loop whose homotopy class generates $\pi_1(Q)$) can be described as follows. Points on $Q$ are loops in $S^2$ beginning and ending at some basepoint $y_0$, i.e. maps $c : [0, 1] \to S^2$ such that $c(0) = c(1) = y_0$. The basepoint of $Q$ itself is the constant loop

\(^4\)Strictly speaking finiteness of the energy requires only that $\varphi(x) \to N_\pm$ for $x \to \pm \infty$ where $N_+$ could be different from $N_-$. This would not change the following results. We assume $N_+ = N_-$ in the following.
which maps all of \([0,1]\) into \(y_0\). Consider the one-parameter family of loops \(c_t\) depicted in fig. XXX. 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 \(\pi_0(Q)\) and \(\pi_2(S^2)\) can be described as follows: if \(c : I \to Q\) is a loop in \(Q\) we define \(\hat{c} : I \times I \to S^2\) by \(\hat{c}(t,s) = (c(t))(s)\), 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 \(\varphi\) (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 \(\tau\). Then \(W(\varphi) = \int d^2x \partial_\mu \omega^\mu\), where

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

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

\[
\pi_\alpha = f^2 h_{\alpha\beta} \partial_0 \varphi^\beta + A_{\alpha},
\]

where

\[
A_{\alpha}(x) = \frac{\theta}{4\pi} \partial_1 \varphi^\beta \sqrt{h} \varepsilon_{\alpha\beta}.
\]

Comparing with equation (2.6) we see that \(A_{\alpha}\) can be regarded as a “functional magnetic potential” on \(Q\). In fact we can write the action \(S = S_0 + S_T = \int dt(L_0 + L_T)\), with

\[
L_0 = \frac{f^2}{2} g_\varphi(\dot{\varphi}, \dot{\varphi}) - V(\varphi) \quad ; \quad L_T = A_\varphi(\dot{\varphi})
\]

This is an infinite dimensional version of the form (2.5) where we replaced the index \(i\) with the infinite indexing set \((\alpha, x)\). The potential is \(V(\varphi) = \)
\[ f^2 g_{\phi}(\partial_1 \varphi, \partial_1 \varphi), \]

the magnetic potential is the one-form \[ A = \int dx A_\alpha(x) \delta \varphi^\alpha(x) \]

and the riemannian metric is \[ g = \int dx h_{\alpha\beta}[\varphi(x)] \delta \varphi^\alpha(x) \delta \varphi^\beta(x) \]. In these formulæ \[ \delta \varphi^\alpha(x) \] play the role of the differentials \[ dq^\alpha \] in the finite dimensional case. This terminology is further explained in appendix E.

In this way the theory can be interpreted as the motion of a particle with mass \( m = f^2 \) and charge \( e = 1 \) on the manifold \( Q \) in a background metric \( g \) and background magnetic field with magnetic potential \( A \). Since the topological term (i.e. the magnetic field) does not appear in the equation of motion, we expect that \( F = dA = 0 \). This is what one gets from a direct calculation based on formula (E.16)

\[ dA(v, w) = v(A(w)) - w(A(v)) - A([v, w]) \]

There follows that, at least locally on \( Q \),

\[ A = d\Lambda. \]

In fact we have

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

The function \( \Lambda \) is not single valued. The polidromy of \( \Lambda \) on the fundamental loop in \( Q \) is

\[
\oint d\Lambda = \oint A = \int d\tau \left[ \frac{\theta}{4\pi} \int dx \partial_1 \varphi^\alpha \frac{d\varphi^\beta}{d\tau} \sqrt{h} \epsilon_{\alpha\beta} \right] = \frac{\theta}{4\pi} \int d^2x \epsilon^\lambda \partial_\lambda \hat{\varphi}^\alpha \partial_\mu \hat{\varphi}^\beta \frac{1}{2} \sqrt{h} \epsilon_{\alpha\beta} = \theta W(\hat{\varphi}) = \theta. \]

Therefore, \( \Lambda \) 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 \( A_{\theta + 2\pi n} \) and \( A_\theta \) are gauge-related in the strict sense. The gauge inequivalent magnetic potentials, and hence the inequivalent quantizations, are labelled by \( 0 \leq \theta < 2\pi \).

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 \( Q = \Gamma_*(S^d, S^{d+1}) \) and \( \pi_1(Q) = \pi_d(S^d) = \mathbb{Z} \). The topological term is given again by the winding number.
2.4 QED in 1+1 dimensions

Next consider a $U(1)$ gauge field $A_\mu$ 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

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

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

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

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

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

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 $A_\mu$ 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

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

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

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 $\rho$ 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 $E_1 = A_1$, so the energy of the gauge
field \( E = \int dx \frac{1}{2} E_i^2 \) is seen to be of purely kinetic character: the static energy is zero.

The configuration space \( Q \) 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 \( \mathcal{G} = \Gamma_s(S^1, U(1)) \) 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 \( Q = \mathcal{C}/\mathcal{G} \). 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, \( Q \) is an infinite dimensional manifold and \( \mathcal{C} \) is a principal bundle over \( Q \) 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 \((A_1, \phi)\).

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

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

The fact that \( \pi_1(Q) \) 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 \( A_{(0)}^g = g^{-1}dg \) 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 \( p : \mathcal{C} \to Q \) 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 \( Q \). It is natural to take \( A_{(0)} \) as the basepoint in \( \mathcal{C} \), \([A_{(0)}]\) as a basepoint in \( Q \). 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 \( A_{(0)}^g = g^{-1}dg \) to \( A_{(0)} \). However, the space \( \mathcal{C} \) is connected and so there is some path \( \tilde{\ell} \) in \( \mathcal{C} \), with \( t \in [0, 1] \) such that \( \tilde{\ell}_0 = A_{(0)} \) and \( \tilde{\ell}_1 = A_{(0)}^g \). For instance one can take \( c_t = tg^{-1}dg = t\,d\alpha \). The natural projection \( p \) maps this
path in $C$ to a path $\ell_t = [\tilde{\ell}_t]$ in $Q$ beginning and ending at $[A_{(0)}]$. The desired isomorphism $\pi_1(Q) \to \pi_0(G)$ is obtained by mapping the homotopy class of the loop $\ell_t$ in $Q$ to the homotopy class of $g$. See fig. XXX.

Returning to equations (2.24) and (2.25) we see that the topological term $\theta c_1$ in the action can be written, in the gauge $A_0 = 0$, as $\int dt \int dx \tilde{A}_1 \frac{\theta}{2\pi}$ 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 $C$)

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

Since the components of the vector potential are constant, it is easy to verify that the corresponding magnetic field $\tilde{F} = d\tilde{A} = 0$. 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 $\tilde{F}$. Since $d\tilde{A} = 0$, we can write at least locally $\tilde{A} = d\tilde{\Lambda}$. The functional $\tilde{\Lambda}$ on $C$ that has this property is

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

All this is on the contractible space $C$.

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

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

$$\tilde{\Lambda}(A^g) - \tilde{\Lambda}(A) = \frac{\theta}{2\pi i} \int dx g^{-1} dg = \frac{\theta}{2\pi} \int dx \frac{d\alpha}{dx} = \theta W(g) . \quad (2.26)$$

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 $\Lambda$ on $C/G$, but only a function which is defined up to integer multiples of $\theta$.

Similarly, we can ask if $\tilde{A} = p^* \mathcal{A}$ for some one-form $\mathcal{A}$ on $C/G$. This is true provided:
2.5. NONABELIAN YANG–MILLS THEORY IN 3+1 DIMENSIONS

\begin{itemize}
  \item 1) \(\tilde{A}\) is gauge invariant;
  \item 2) \(\tilde{A}(v) = 0\) when \(v\) is a vertical vector (i.e. \(v\) is tangent to the orbit).
\end{itemize}

(See reference KOBAYASHI NOMIZU vol. II, p. 294, lemma 1). 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 \(\epsilon\) is an infinitesimal gauge parameter; then

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

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

According to the general discussion in section 3.1, inequivalent quantizations correspond to the gauge inequivalent magnetic potentials \(A\). The magnetic potential \(A(\theta)\) will be gauge equivalent to \(A(\theta = 0)\) if the function \(e^{i\Lambda}\) is single-valued, i.e. if the polydromy of \(\Lambda\) is an integral multiple of \(2\pi\).

From the construction of the fundamental loop \(\ell\) in \(Q\) we see that the polydromy of \(\Lambda\) on \(\ell\) is equal to \(\oint_\ell \Lambda = \int_{\tilde{\ell}} \tilde{A}\), where \(\tilde{\ell}\) is a lift of \(\ell\), i.e. a path joining \(A(0)\) to \(A^g(0)\), with \(W(g) = 1\). But then \(\int_{\tilde{\ell}} \tilde{A} = \hat{\Lambda}(A^g) - \hat{\Lambda}(A) = \theta\), by equation (2.26). So, whenever \(\theta = 2\pi n\), \(A(\theta)\) is a pure gauge. The classes of gauge inequivalent \(A\)'s are parameterized again by \(0 \leq \theta < 2\pi\).

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

\[
c_2 = \frac{1}{64\pi^2} \int d^4 x \varepsilon^{\mu\nu\rho\sigma} F^{a}_{\mu\nu} F^{a}_{\rho\sigma},
\]

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

\[
\frac{1}{64\pi^2} \varepsilon^{\mu\nu\rho\sigma} F^{a}_{\mu\nu} F^{a}_{\rho\sigma} = \partial_\mu C^\mu ,
\]

(2.27)
where
\[ C^\mu = \frac{1}{16\pi^2} \varepsilon^{\mu\nu\rho\sigma} \left( A^a_\nu \partial_\rho A^a_\sigma + \frac{1}{3} f_{abc} A^a_\nu A^b_\rho A^c_\sigma \right) \] (2.28)
is known as the (dual of the) three dimensional Chern-Simons form. Thus \( c_2 \) is invariant under infinitesimal variations of \( A^a_\mu \). However, it changes the relation between velocities and momenta. We have
\[ P^0_a = \frac{\partial L}{\partial \partial_0 A^0_a} = 0 \] (2.29)
\[ P_a = \frac{\partial L}{\partial \partial_0 A^a_i} = \frac{1}{e^2} E^a_i + \frac{\theta}{8\pi^2} B^a_i \] (2.30)
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
\[ H = \int d^3x \left[ \frac{e^2}{2} \left( P^a_i - \theta \frac{e^2}{8\pi^2} B^a_i \right)^2 + \frac{1}{2e^2} (B^a_i)^2 - A^0_a D_i E^a_i \right]. \] (2.31)

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 \( g(x) \rightarrow 1 \) for \( |\vec{x}| \rightarrow \infty \). With these boundary conditions, \( \mathbb{R}^3 \) can be compactified to \( S^3 \) and \( \mathcal{G} = \Gamma_s(S^3, G) \). 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 \( g^{-1} A g + g^{-1} d g = A \). Thus \( g \) satisfies the equation \( d g + [A, g] = 0 \), 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 \( Q = \mathcal{C}/\mathcal{G} \), and the projection \( p : \mathcal{C} \rightarrow Q \) is a smooth infinite dimensional bundle.\(^5\)

Since \( \mathcal{C} \) is topologically trivial we have, following again the arguments of Appendix D, \( \pi_1(Q) = \pi_0(\mathcal{G}) = \pi_3(G) = \mathbb{Z} \). The isomorphism between \( \pi_1(Q) \) 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 \( \ell \) which is obtained by projecting to \( Q \) a curve \( \tilde{\ell} \) joining \( A = 0 \) to \( A^g = g^{-1} d g \).

Comparing equations (2.30) and (2.31) with (2.8) and (2.7) we see that the topological term has given rise to a magnetic potential $\tilde{A}$ on $C$ defined by

$$\tilde{A}(A) = \frac{\theta}{8\pi^2} \int d^3x B^a_i \delta A^a_i .$$

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

$$\tilde{\Lambda} = \frac{\theta}{16\pi^2} \int d^3x C^0 = \frac{\theta}{16\pi^2} \int d^3x \epsilon^{ijk} \left( A^a_i \partial_j A^a_k + \frac{1}{3} f_{abc} A^a_i A^b_j A^c_k \right) .$$

(2.32)

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

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

(2.33)

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

To see if $\tilde{A}$ projects, we have to verify whether the conditions given in the preceding section are satisfied. Given an infinitesimal gauge transformation parameter $\epsilon$, 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 $C$

$$\nu_\epsilon = \int d^3x D_i \epsilon^a \frac{\delta}{\delta A^a_i} .$$

Then we have:

- 1) $\tilde{A}$ is gauge invariant ($B^a_i$ and $\delta A^a_i$ both transform homogeneously);

- 2) $\tilde{A}(\nu_\epsilon) = \frac{\theta}{8\pi^2} \int d^3x B^a_i D_i \epsilon^a = 0$ upon integrating by parts, using Bianchi’s identity and the fact that $\epsilon \to 0$ for $|\vec{x}| \to \infty$.

So $\tilde{A}$ satisfy the two conditions which are needed for it to be the pullback of a one-form $A$ on $Q$. The relation between $A$ and $\Lambda$ is again, locally, $A = d\Lambda = \frac{1}{i} e^{-i\Lambda} d e^{i\Lambda}$. The polydromy of $\Lambda$ on the loop $\ell$ which generates $\pi_1(Q)$ is $\oint_\ell A^0 = \oint_\ell \tilde{A} = \tilde{\Lambda}(A^0) - \tilde{\Lambda}(A) = \theta$. So we come again to the conclusion
that there is a \( U(1) \)'s worth of quantum Yang-Mills theories, parameterized by the angle \( 0 \leq \theta < 2\pi \).

Before closing this section we note for future reference the following interpretation of the Gauss law of the theory. Let \( G_\epsilon = \int d^3 x \epsilon^a G_a \). 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: \( G_\epsilon \psi_{\text{phys}} = 0 \) for all \( \epsilon \). Upon using the quantization rule \( P^a_i = -i \frac{\delta}{\delta A^a_i} \), we find

\[
G_\epsilon \psi = e^2 \int d^3 x \epsilon^a D_i \left( P^a_i - \frac{\theta}{8\pi^2} B^a_i \right) \psi
= ie^2 \int d^3 x \epsilon^a \left( \frac{\delta \psi}{\delta A^a_i} - i \frac{\theta}{8\pi^2} B^a_i \psi \right)
= ie^2 \left( \psi \epsilon^a + i \tilde{A} (v_e) \psi \right) = ie^2 v_e \epsilon \psi .
\]

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 exist, 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 visualize alternatively a field configuration as a function on space or as a point in the infinite-dimensional configuration space $Q$. It will now be useful to visualize field histories alternatively as functions on spacetime or as paths in $Q$. 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 $Q$.

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 spatial 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 Examples

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

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

(2.35)

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:

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

(2.36)
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 (1.53). The instantons of this model in 1+1 Euclidean dimensions are the functions given in (1.61). 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 \( Q \).

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

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

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 non-contractible loop in \( Q \). It follows from the discussion of section 3.4 that this loop in \( Q \) is the projection of a path in \( C \) joining the classical vacuum \( (A_{(0)}, \phi_{(0)}) = (0, f) \) to \( (A_{g_1}^{(0)}, \phi_{g_1}^{(0)}) = (g_1^{-1}dg_1, g_1^{-1}f) \), 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 \( r = \sqrt{x^2 + \tau^2} \to \infty \), \( A_i \to i e^{-i\alpha} \) and \( \phi \to f \), where \( g_1(\theta) \) is a map from \( S^1_{\infty} \) 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.
2.6. INSTANTONS

We can now understand better in what sense the integral $c_1$ in equation (2.22) is a “topological number”. We restrict our attention to spacetime fields with finite euclidean action. This demands that when $r = \sqrt{x^2 + \tau^2} \to \infty$, $A_i \to \frac{1}{r} g_\infty^{-1} dg_\infty$ and $\phi \to g_\infty^{-1} f$, where $g_\infty(\theta)$ is a map from $S^4_\infty$ 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 (2.23) and the asymptotic form of $A$ we get

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

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

2.6.2 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.

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

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

where $g_\infty$ is a function of the angles or equivalently a function from the sphere at infinity $S^3_\infty$ to the gauge group $SU(2)$. Since $\pi_3(SU(2)) = \mathbb{Z}$, we find that the finite action gauge potentials $A^a_\mu$ fall into topologically distinct classes distinguished by their asymptotic behaviour. The topological invariant $c_2$ precisely measures these classes. In fact using (2.27) we can write

$$c_2 = \int_{\mathbb{R}^4} d^4 x \, \partial_\mu C^\mu = \frac{1}{16\pi^2} \int_{S^3_\infty} d^3 x \varepsilon^{ijk} \left( A^a_i \partial_j A^a_k + \frac{1}{3} f_{abc} A^a_i A^b_j A^c_k \right) = W(g_\infty) .$$
The last equality is obtained by noting that on $S_\infty^3$ we can replace $A_\mu^a$ by its asymptotic form (2.38); the result then follows from Exercise 2.5.2.

This calculation gives the precise meaning of the statement that $c_2$ is a topological invariant. An even more sophisticated understanding of the topology of Yang–Mills fields requires the use of fiber bundles and characteristic classes.

The instanton has to represent the motion of the system through the fundamental noncontractible loop in $Q$. 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 $A_i = g^{-1} \partial_\tau g$, where $g$ is a time-independent gauge transformation with winding number one. We take the parameter on this path to be euclidean time $\tau$. 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 $g_\infty$ 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

$$0 \leq \int d^4x \left( F_{\mu\nu}^a \pm \ast F_{\mu\nu}^a \right) \left( F^{\mu\nu a} \pm \ast F^{\mu\nu a} \right)$$

$$= 2 \int d^4x F_{\mu\nu}^a F^{\mu\nu a} \pm 2 \int d^4x F_{\mu\nu}^a \ast F^{\mu\nu a} ,$$

which implies

$$S_E \geq \frac{8\pi^2}{e^2} |c_2| . \quad (2.39)$$

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

$$F_{\mu\nu}^a = \pm \ast F_{\mu\nu}^a . \quad (2.40)$$

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 (2.40).

To find the explicit form of the instanton in the sector $c_2 = \pm 1$ we make an ansatz of the form

$$A_\mu(x) = f(r^2)g^{-1} \partial_\mu g \quad (2.41)$$

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

$$g(x) = \begin{bmatrix} \hat{x}^4 + i\hat{x}^3 & \hat{x}^2 + i\hat{x}^1 \\ -\hat{x}^2 + i\hat{x}^3 & \hat{x} - i\hat{x}^3 \end{bmatrix} = \hat{x}^\mu r_\mu ,$$
where \( \tau_k = i \sigma_k \) for \( k = 1, 2, 3 \) and \( \tau_4 = 1 \). This function clearly has \( W(g_\infty) = 1 \). Note that we can write
\[
g(x)^{-1} = g(x)^\dagger = \hat{x}^\mu \bar{\tau}_\mu ,
\]
where \( \bar{\tau}_k = -i \sigma_k \) for \( k = 1, 2, 3 \) and \( \tau_4 = 1 \). From here one finds
\[
g^{-1} \partial_\mu g = -2i \Sigma_{\mu \rho} \frac{\hat{x}^\rho}{r} ; \quad g \partial_\mu g^{-1} = -2i \Sigma_{\mu \rho} \frac{\hat{x}^\rho}{r} ,
\]
where
\[
\Sigma_{\mu \nu} = \frac{1}{2} \begin{pmatrix}
0 & \sigma_3 & -\sigma_2 & \sigma_1 \\
-\sigma_3 & 0 & \sigma_1 & \sigma_2 \\
\sigma_2 & -\sigma_1 & 0 & \sigma_3 \\
-\sigma_1 & -\sigma_2 & -\sigma_3 & 0
\end{pmatrix} ; \quad \bar{\Sigma}_{\mu \nu} = \frac{1}{2} \begin{pmatrix}
0 & \sigma_3 & -\sigma_2 & -\sigma_1 \\
-\sigma_3 & 0 & \sigma_1 & -\sigma_2 \\
\sigma_2 & -\sigma_1 & 0 & -\sigma_3 \\
\sigma_1 & \sigma_2 & \sigma_3 & 0
\end{pmatrix} .
\]
These matrix-valued tensors are self-dual and anti-self-dual respectively.

The function \( f \) in (2.41) 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 (2.41):
\[
F_{\mu \nu} = 4i \left( \bar{\Sigma}_{\mu \rho} \hat{x}^\rho \hat{x}^\nu - \Sigma_{\mu \rho} \hat{x}^\rho \hat{x}^\mu \right) \left( f' + \frac{1}{r^2} f(f - 1) \right) - 41 \Sigma_{\mu \nu} \frac{1}{r^2} f(f - 1) .
\]
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
\[
*f_{\mu \nu} = 4i \left( \bar{\Sigma}_{\mu \rho} \hat{x}^\rho \hat{x}^\nu - \Sigma_{\nu \rho} \hat{x}^\rho \hat{x}^\mu \right) \left( f' + \frac{1}{r^2} f(f - 1) \right) - 4i \Sigma_{\mu \nu} f' .
\]
The anti-self-duality equation \( 0 = F_{\mu \nu} + * F_{\mu \nu} \) implies \( f' + \frac{1}{r^2} f(f - 1) = 0 \), which it is solved by
\[
f(r^2) = \frac{r^2}{\lambda^2 + r^2}
\]
where \( \lambda \) is an arbitrary constant. A similar procedure starting from the map \( \hat{x}^\mu \bar{\tau}_\mu \) with winding number \(-1\), and solving the self-duality equation
0 = F_{\mu\nu} - * F_{\mu\nu} leads to the same function f. Altogether the regular unit instanton and anti-instanton solutions can be written in the form

\[ A_{\mu} = -2i \frac{\Sigma_{\mu\nu}(x - x_0)^\nu}{\lambda^2 + (x - x_0)^2} ; \quad A_{\mu} = -2i \frac{\Sigma_{\mu\nu}(x - x_0)^\nu}{\lambda^2 + (x - x_0)^2} \]

where we have allowed the center of the solutions to be located at an arbitrary position \( x_0 \). The respective field strengths are

\[ F_{\mu\nu} = -4i \frac{\Sigma_{\mu\nu}\lambda^2}{(\lambda^2 + (x - x_0)^2)^2} ; \quad F_{\mu\nu} = -4i \frac{\Sigma_{\mu\nu}\lambda^2}{(\lambda^2 + (x - x_0)^2)^2} . \]

Note that the solution depends of 5 free parameters: the location of the center \( x_0 \), reflecting translation invariance of the action, and the scale \( \lambda \) reflecting the conformal invariance of the action.

Note also that if we impose self-duality on the \( W = -1 \) configuration or anti-self-duality on the \( W = 1 \) configuration we are led to the equation

\[ f' - \frac{1}{r^2} f (f - 1) = 0, \]

which it is solved by

\[ f(r^2) = \frac{\lambda^2}{\lambda^2 + r^2} . \]

This gives rise to solutions that are singular in the origin. They are related to the regular solutions given above by a singular gauge transformation with parameter \( g \).

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

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

\[ K(q_2, t_2 \mid q_1, t_1) = \int_{q_1, t_1}^{q_2, t_2} (dq) e^{i S_0(q)} , \quad (2.42) \]
where the integral is performed over all paths joining \( q_1 \) to \( q_2 \). We assume that the action has the form
\[
S_0 = \int dt \left[ \frac{1}{2} m g_{ij}(q) \dot{q}^i \dot{q}^j - V(q) \right].
\] (2.43)

To make the integral convergent we perform a Wick rotation to imaginary time \( t_E = it \). The euclidean action is given by
\[
S_{0E} = -i S_0(t = -it_E) = \int dt_E \left[ \frac{1}{2} m g_{ij} \left( \frac{dq^i}{dt_E} \right) \left( \frac{dq^j}{dt_E} \right) + V(q) \right].
\] (2.44)

We will omit the subscript \( E \) from now on for notational simplicity. The euclidean amplitude is
\[
K_E(q_2, \tau_2 | q_1, \tau_1) = \int_{q_1, \tau_1}^{q_2, \tau_2} (dq) e^{-\frac{i}{\hbar} S_{0E}(q)}.
\]

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:
\[
Z_E(T) = K_E(q_0, T/2; q_0, -T/2).
\] (2.45)

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 \( e^{-\frac{i}{\hbar} \hat{H} t} \), and we have
\[
Z_E(T) = \langle q_0 | e^{-\frac{i}{\hbar} \hat{H} T} | q_0 \rangle = \sum_n |\langle q_0 | E_n \rangle|^2 e^{-\frac{i}{\hbar} E_n T},
\]
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
\[
\lim_{T \to \infty} Z_E(T) = \lim_{T \to \infty} |\langle q_0 | E_0 \rangle|^2 e^{-\frac{i}{\hbar} E_0 T}.
\] (2.46)

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
\[
\left( \frac{\omega}{\pi \hbar} \right)^{1/2} e^{-\frac{\omega T}{2}}.
\] (2.47)

Comparing with (2.46) one finds the ground state energy \( E_0 = \frac{\hbar}{2} \omega \).
2.7.1 Path integrals on multiply connected spaces

In sections 2.3-4-5 we have established that certain quantum field theories with a multiply connected configuration space exhibit a quantization ambiguity. In each case we have added a suitable topological term to the action that can be interpreted as a "magnetic potential" $A$ on $Q$. Inequivalent quantizations correspond to Gauge inequivalent magnetic potentials correspond to inequivalent quantum theories and in the cases we considered these were labelled by an angle $0 \leq \theta < 2\pi$. In each case, an alternative definition based on wave functions that are only periodic up to a phase, and without topological term in the action, would also be possible.

A proper definition of these quantum field theories within the Hamiltonian context would be very difficult, but insofar as the results depend only on the topology of the configuration space, they are part of the models at a very basic level. Any proper definition of these theories would have to reflect these features.

Now we would like to find the ground state in each theta sector and to see whether the vacuum expectation value of the energy and other observables depends on $\theta$ or not. It is not possible to do this without approximations. We will use the method of the path integrals, which can be easily generalized from quantum mechanics to quantum field theory.

Our first task will be to understand the appearance of theta sectors from the path integral point of view. Recall that for a system with configuration space $Q$ 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

$$K(q_2, t_2 \mid q_1, t_1) = \int_{q_1, t_1}^{q_2, t_2} (dq)e^{\frac{i}{\hbar}S_0(q)},$$  \hspace{1cm} (2.48)

where the integral is performed over all paths joining $q_1$ to $q_2$. We assume that the action has the form

$$S_0 = \int dt \left[ \frac{1}{2} m g_{ij}(q) \dot{q}^i \dot{q}^j - V(q) \right].$$  \hspace{1cm} (2.49)

Forgetting all that has been said in the preceding sections, suppose we want to compute the amplitude (2.48) in a theory with action (2.49). 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
However, the correspondence between homotopy classes of paths and elements of $\pi_1(Q)$ 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

$$K_\alpha(q_2, t_2 \mid q_1, t_1) = \int_{q_1, t_1}^{q_2, t_2} (dq)_\alpha e^{i\frac{\hbar}{\ell} S_0(q)},$$

where the subscript $\alpha$ 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 $\alpha \in \pi_1(Q)$. 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

$$K(q_2, t_2 \mid q_1, t_1) = \sum_{\alpha \in \pi_1(Q)} \chi(\alpha) K_\alpha(q_2, t_2 \mid q_1, t_1).$$

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

$$K(q_2, t_2 \mid q_1, t_1) = \int dq K(q_2, t_2 \mid q, t) K(q, t \mid q_1, t_1)$$

for $t_1 < t < t_2$.

It can be shown 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 defined in Appendix A. Thus $\chi$ has to be a character of $\pi_1(Q)$. Each choice of $\chi$ defines an inequivalent quantum theory, so we have reached again the conclusion that inequivalent quantizations are labelled by $\text{Hom}(\pi_1(Q), U(1))$.

How is this related to the previous discussion based on the addition of topological terms to the action? As mentioned in Section 3.1 there is a one-to-one correspondence between the characters of $\pi_1(Q)$ and the gauge equivalence classes of flat $U(1)$ connections on $Q$. The correspondence can be described as follows: if $A$ is a flat connection, the corresponding character is given by

$$\chi(\alpha) = e^{i \frac{\hbar}{\ell} \int A},$$
where $\ell$ is a loop in the homotopy class $\alpha$ ($\chi$ depends only on the homotopy class of $\ell$ since $\mathcal{A}$ is flat). There follows that if we define a “topological term”

$$ S_T = e \int_{t_1}^{t_2} dt \dot{q}^i \mathcal{A}_i , $$

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 $\alpha$. Thus we can absorb this term in the functional integral and write:

$$ \sum_{\alpha} \chi(\alpha) K_\alpha(q_2, t_2; q_1, t_1) = \sum_{\alpha} \int_{q_1, t_1}^{q_2, t_2} (dq)_\alpha e^{i(S_0 + S_T)} = \int_{q_1, t_1}^{q_2, t_2} (dq) e^{i(S_0 + S_T)} . $$

So the effect of performing the functional integral with the action $S_0$ and weighting the partial amplitudes with characters of $\pi_1(Q)$ is exactly the same as performing the functional integral with the action $S_0 + S_T$.

### 2.8 The dilute instanton gas approximation

We begin by discussing the example of the pendulum $Q = S^1$ as presented in section 2.2. 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 $\theta$ (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 mod $2\pi$, whereas in the latter they are not. We will see that this will lead to a different physical interpretation of the results.
2.8. THE DILUTE INSTANTON GAS APPROXIMATION

2.8.1 The $n = \pm 1$ contributions

We are going to study the vacuum energy of the pendulum as a function of $\theta$ (where $\theta$ is the coefficient which appears in (2.14)) using the method of path integrals. We begin by observing that the classical “vacuum state” of the pendulum is $\varphi = 0 \mod 2\pi$ (independent of time). The vacuum-to-vacuum transition amplitude is

$$K(0 \mod 2\pi, T/2 | 0, -T/2) = \sum_{n=-\infty}^{\infty} e^{i\theta n} K_n(2\pi n, T/2 | 0, -T/2),$$

where $n \in \mathbb{Z} = \pi_1(S^1)$ labels the homotopy class of the loop $\varphi(t)$ 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 (2.13). 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

$$e^{i\theta n} K_n(2\pi n, T/2 | 0, -T/2) = \int_{0, -T/2}^{2\pi n, T/2} (d\varphi)_n e^{\frac{i}{\hbar} S_0 + i\theta n} = \int_{0, -T/2}^{2\pi n, T/2} (d\varphi)_n e^{i(S_0 + \theta W)} = \tilde{K}_n(2\pi n, T/2 | 0, -T/2).$$

In this way the topological term introduced in section 3.2 has made its reappearance.

Next we perform the Wick rotation. Due to the fact that it contains only one time derivative, the euclidean topological term is imaginary:

$$S_{T,E} = -iS_T(t = -i\tau) = -i\theta W.$$

The euclidean amplitude is

$$\tilde{K}_E(0 \mod 2\pi, T/2 | 0, -T/2) = \sum_n \tilde{K}_{E,n}(2\pi n, T/2 | 0, -T/2),$$

$$\tilde{K}_{E,n}(2\pi n, T/2 | 0, -T/2) = \int_{0, -T/2}^{2\pi n, T/2} (d\varphi)_n e^{-\frac{1}{\hbar} S_E},$$

where $S_E = S_{0E} + S_T = S_{0E} - i\theta W$. Note that the topological term retains oscillatory character also in the Euclidean regime. 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. Such paths can be regarded
as maps from the interval \([-T/2, T/2]\) to \(\mathbb{R}\), and with the given boundary
conditions they form a functional space which is diffeomorphic to the space \(Q_{01}\) that we considered in section 2.3 for the sine-Gordon theory. Furthermore, the euclidean action \(S_{0E}\) for the pendulum is equal to the static energy (1.4) for the sine-Gordon theory. Therefore the absolute minimum of \(S_{0E}\) is
the trajectory given by (1.1.11), with \(\phi\) and \(x\) replaced by \(\phi\) and \(\tau\). Within
the context of the present theory, a trajectory of this type is called an
instanton with topological number one centered around \(\tau_0\) (the reason for the
name is that for most of the time the trajectory lies near the classical vacuum
\(\phi = 0 \mod 2\pi\), except for a brief “instant” of time around \(\tau_0\).

Next we expand the action around this particular classical trajectory \(\varphi_{cl}(t)\). We get
\[
S_E(\varphi) = S_E(\varphi_{cl}) + \frac{1}{2} \int d\tau d\tau' \eta(t) O(t, t') \eta(t'),
\]
where \(\eta = \varphi - \varphi_{cl}\) and
\[
O(t, t') = \left. \frac{\delta^2 S_E}{\delta \varphi(t) \delta \varphi(t')} \right|_{\varphi_{cl}} = \delta(t - t') \left( -\frac{d^2}{dt^2} + V''(\varphi_{cl}) \right).
\]
In the WKB approximation
\[
K_{E1}(2\pi, T/2 | 0, -T/2) = e^{-\frac{i}{\hbar}S_E(\varphi_{cl})} \int (d\eta) e^{-\frac{i}{2} \int \eta O \eta} = e^{-\frac{i}{\hbar}S_E(\varphi_{cl})} B(T) [\text{Det} O]^{-1/2}
\]
where \(B(T)\) is a measure factor.

The operator \(-\frac{d^2}{dt^2} + V''(\varphi_{cl})\) 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” \(t_0\), 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 (2.52) can be rewritten
\[
e^{-\frac{i}{\hbar}S_E(\varphi_{cl})} B(T)JT [\text{Det}' O]^{-1/2},
\]
where $\text{Det}'$ is the product of the nonzero eigenvalues. The evaluation of the determinant is difficult because $\varphi_{\text{cl}}$, which appears in the operator (2.51) 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” $\varphi_{\text{cl}} = 0 \mod 2\pi$ and therefore $V''(\varphi_{\text{cl}}) = \omega^2$. We can then write

$$[\text{Det}' \mathcal{O}]^{-1/2} = K \left[ \text{Det} \left( -\frac{d^2}{dt^2} + \omega^2 \right) \right]^{-1/2}$$

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 (2.47).

We thus find

$$K_{E,1}(2\pi, T/2 \mid 0, -T/2) = e^{-\frac{1}{\hbar}S_{0E} + i\theta} K J T e^{-\frac{\omega T}{\pi \hbar}} \left( \frac{\omega}{\pi \hbar} \right)^{1/2}$$

where we have written $S_{E}(\varphi_{\text{cl}}) = S_{0E}(\varphi_{\text{cl}}) - i\theta \hbar W(\varphi_{\text{cl}}) = S_{0E} - i\theta \hbar$. This is the contribution of the one-instanton sector to the total amplitude. The one anti-instanton sector gives

$$K_{E,-1}(-2\pi, T/2 \mid 0, -T/2) = e^{-\frac{1}{\hbar}S_{0E} - i\theta} K J T e^{-\frac{\omega T}{\pi \hbar}} \left( \frac{\omega}{\pi \hbar} \right)^{1/2}$$

### 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 $-\frac{d^2 \varphi}{dx^2} + \frac{dV}{d\varphi} = 0$ in the sectors $Q_{0i}$ 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 $K_E((m_1 - m_2)2\pi, T/2 \mid 0, -T/2)$. 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 $S_{E}(\varphi_{cl})$ 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 $V''(\varphi_{cl})$ is replaced by $\omega^{2}$ in the determinant. Altogether the contribution to the total amplitude due to configurations containing $m_{1}$ instantons and $m_{2}$ anti-instantons is

$$\frac{1}{m_{1}!m_{2}!} \exp \left[ -\frac{1}{\hbar}(m_{1} + m_{2})S_{0E} + i(m_{1} - m_{2})\theta \right] \left( KJT \right)^{m_{1}+m_{2}} \left( \frac{\omega}{\pi \hbar} \right)^{1/2} e^{-\frac{\omega T}{2}}$$

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 $\tau_{1}$ and instanton 2 is in position $\tau_{2}$ is physically the same as when instanton 1 is in position $\tau_{2}$ and instanton 2 is in position $\tau_{1}$).

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

$$Z_{\theta}(T) = \exp \left( KJT e^{-\frac{1}{\hbar}S_{0E}+i\theta} \right) \exp \left( KJT e^{-\frac{1}{\hbar}S_{0E}-i\theta} \right) \left( \frac{\omega}{\pi \hbar} \right)^{1/2} e^{-\frac{\omega T}{2}}$$

$$= \left( \frac{\omega}{\pi \hbar} \right)^{1/2} \exp \left[ -\frac{1}{\hbar}T \left( \frac{1}{2} \hbar \omega - 2\hbar KJ e^{-\frac{1}{\hbar}S_{0E} \cos \theta} \right) \right]. \quad (2.53)$$

Comparing with (2.46) we find that the energy of the vacuum in the presence of the $\theta$-term in the action is

$$E_{\theta} = \frac{1}{2} \hbar \omega - 2\hbar KJ e^{-\frac{1}{\hbar}S_{0E} \cos \theta} . \quad (2.54)$$

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 $\theta$ are permitted. Here every value of $\theta$ defines a different theory and no transition between different $\theta$-states can occur.

This way of computing the functional integral for a theory with multiply connected $Q$ 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 $Q$, has its minimum at some point $\varphi_{0}$
that we call the classical vacuum. We take $\varphi_0$ as “basepoint” in $Q$. 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 $Q_0$ parameterized by $-\infty \leq t \leq \infty$, 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 which 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 $\varphi$ onto the zero mode $\eta_0$ 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

$$S_{cl} = \int dt \left[ \frac{1}{2} \dot{\varphi}_{cl}^2 + V(\varphi_{cl}) \right] = \int dt \dot{\varphi}_{cl}^2 . \quad (2.55)$$

There follows that the normalized zero mode is

$$\eta_0(t - t_0) = \frac{1}{\sqrt{S_{cl}}} \frac{d\varphi_{cl}(t - t_0)}{dt} . \quad (2.56)$$

In what follows both the instanton $\varphi_{cl}$ and the zero mode $\eta_0$ are located at a particular time $t_0$. Thus using (2.56) we find that the projection of the
instanton on the zero mode is

\[
(\varphi_{cl}, \eta_0) = \int_{-\infty}^{\infty} dt \eta_0(t - t_0) \varphi_{cl}(t - t_0)
\]

\[
= \frac{1}{2\sqrt{S_{cl}}} \int_{-\infty}^{\infty} dt \frac{d\varphi_{cl}(t - t_0)^2}{dt}
\]

\[
= \frac{1}{2\sqrt{S_{cl}}} \varphi_{cl}^{2} \bigg|_{-\infty}^{\infty} = \frac{2\pi^2}{\sqrt{S_{cl}}}
\]  

(2.57)

and is independent of \(t_0\).

Then, the following relation holds

\[
\Delta[\varphi] \int dt_0 \delta \left[ (\varphi, \eta_0) - (\varphi_{cl}, \eta_0) \right] = 1
\]  

(2.58)

where

\[
(\varphi, \eta_0) = \int_{-\infty}^{\infty} dt \eta_0(t - t_0) \varphi_{cl}(t)
\]

will in general depend on \(t_0\). The quantity \(\Delta\) is

\[
\Delta[\varphi] = \frac{d}{dt_0} [(\varphi, \eta_0) - (\varphi_{cl}, \eta_0)]
\]

evaluated at a point (here assumed unique) where the argument of the delta function is zero. We have

\[
\Delta[\varphi] = \int_{-\infty}^{\infty} dt \frac{d\eta_0(t - t_0)}{dt_0} \varphi(t)
\]

\[
= \int_{-\infty}^{\infty} dt \eta_0(t - t_0) \frac{d\varphi(t)}{dt}.
\]  

(2.59)

Now we insert the identity (2.58) in the functional integral

\[
Z = \int dt_0 \int (d\varphi)e^{-\frac{1}{\hbar}(S_0 + i\delta)} \Delta[\varphi] \delta \left[ (\varphi, \eta_0) - (\varphi_{cl}, \eta_0) \right]
\]

When we expand the action around the instanton located at \(t_0\) the argument of the delta function becomes \((\eta, \eta_0)\), with \(\eta_0\) 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 \( \eta \) 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 \( J = \sqrt{S_{cl}} \), 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 \( \theta \neq 0 \text{mod} 2\pi \), there is no Higgs phase.

As already discussed in section XXX, the instanton of scalar QED is the vortex of QED 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 \( e^{-T E_\theta} \), and by evaluating \( Z_\theta \) we shall obtain \( E_\theta \) 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 \( Z_\theta \) 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

\[ \lim_{T \to \infty} Z_\theta(T) = Af^{-LT(C - e^{-S_{0E} 2B \cos \theta})} \quad (2.60) \]
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 (2.54).

The physical meaning of the parameter $\theta$ 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

$$\langle F_{01}(x,\tau) \rangle_\theta = \frac{1}{LT} \left\langle \int dxd\tau F_{01} \right\rangle_\theta = \frac{1}{2LT} \left\langle \int dxd\tau \varepsilon^{\mu\nu} F_{\mu\nu} \right\rangle_\theta = \frac{2\pi}{LT} \langle c_1 \rangle_\theta$$

We have

$$\langle c_1 \rangle_\theta = i \frac{d}{d\theta} \ln Z_\theta = -i \frac{d}{d\theta} (E_\theta T) = -iLT e^{-S_{0E}} 2B \sin \theta .$$

Therefore

$$\langle E_1(x,\tau) \rangle_\theta = 4\pi e^{-S_{0E}} B \sin \theta .$$

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
2.9. THE ABELIAN HIGGS MODEL


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:

\[
\int (dA d\phi d\phi^*) e^{-S_{0E} + i\theta c_1} W = \exp(-TE_\theta - \tilde{T}\Delta E_\theta(\tilde{L})) .
\] (2.61)

We have then

\[
\lim_{\tilde{T} \to \infty} \langle W \rangle = 1
\]

\[
\int (dA d\phi d\phi^*) e^{-S_{0E} + i\theta c_1 - iq \oint A} = e^{-\tilde{T}\Delta E_\theta(\tilde{L})} .
\] (2.62)

Therefore we can compute the interaction between the charges from the Wilson loop:

\[
\Delta E_\theta(\tilde{L}) = -\lim_{T \to \infty} \frac{1}{T} \ln \langle W \rangle_\theta .
\] (2.62)

We use again the dilute instanton gas approximation. We divide \( n_\pm = n_\pm^{(in)} + n_\pm^{(out)} \), 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:

\[
W = e^{i\theta \oint A} = e^{i\theta \int_U d^2x \epsilon^{\mu\nu} F_{\mu\nu}} = e^{2\pi i(qn_{(in)}^{(in)} - n_{(in)}^{(out)})}
\]

where \( U \) is the region enclosed by the loop. Then, the functional integral (2.61) can be evaluated as follows:

\[
A e^{-LTC} \sum_{n_\pm^{(in)}, n_\pm^{(out)}} \frac{1}{n_\pm^{(in)}! n_\pm^{(out)}!} \exp \left( - (n_\pm^{(in)} + n_\pm^{(out)})S_{0E} \right.
\left. + i\theta (n_\pm^{(in)} - n_\pm^{(out)}) \right)
\times \left[ B(\tilde{L}T - \tilde{L}\tilde{T}) \right]^{n_\pm^{(out)} - n_\pm^{(in)}} e^{-S_{0E} [\tilde{L}\tilde{T} \cos(\theta + 2\pi q) + (L \tilde{T} - \tilde{L}T) \cos \theta]} .
\] (2.63)

Using (2.60) and (2.63) in (2.62) we get

\[
\Delta E_\theta(\tilde{L}) = 2B e^{-S_{0E}} [\tilde{L} \cos(\theta - \cos(\theta + 2\pi q)] .
\] (2.64)
CHAPTER 2. \( \pi_1(Q) \) AND THETA SECTORS

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 force vanishes exponentially in the classical limit.

To get a physical intuition for the \( \theta \)- and \( q \)-dependence, we can expand for small \( \theta \) and \( q \) and find

\[
\langle E_1 \rangle_\theta = 4\pi B e^{-S_{0E}} \theta
\]

\[
E_\theta = 2B L e^{-S_{0E}} \theta^2
\]

\[
\Delta E_\theta(\bar{L}) = 2B L e^{-S_{0E}} \left[ (\theta + q)^2 - \theta^2 \right].
\]

In the \( \theta \) 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 (2.64), 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 particle-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 screening cannot be complete, leaving a residual force which is independent of distance.
Chapter 3

$H^2(Q)$ and the quantization of parameters

Let us consider again the motion of a charged particle on a manifold $Q$, in a background magnetic field $F$, with a potential $A$, and lagrangian (2.5). In chapter 2 we have discussed at length the case in which $Q$ is multiply connected and $F = 0$. We will now consider situations in which $Q$ is simply connected and $F \neq 0$. As discussed in section 2.1, quantization demands that $A$ 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 $F$ to be the field strength of a $U(1)$ connection is that

$$\int_m F = \frac{2\pi \hbar}{e} n, \quad n \in \mathbb{Z}, \quad (3.1)$$

for any two-dimensional submanifold $m$ of $Q$ without boundary. In mathematical terms, $F$ has to define an integral cohomology class in $H^2(Q)$ (see Appendix ??). A general proof of this result is given in Appendix ???. We will give a proof only in the simplest case, namely when $Q = S^2$. 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 non-trivial 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 \( \theta \) on a sphere of radius \( r \) (fig. ??). 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

\[
\mathcal{F} = Q_M \sin \theta \, d\theta \wedge d\varphi .
\]  

A solution of the equation \( \mathcal{F} = d\mathcal{A} \) is given by

\[
\mathcal{A} = \mathcal{A}^{(+)} = Q_M (1 - \cos \theta) d\varphi .
\]
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

\[
A^-(\theta) = -Q_M(1 + \cos \theta) d\varphi ,
\]

which is singular on the positive axis. Let \( U^+ \) and \( U^- \) be the subsets of \( Q \) with \( \theta \neq \pi \) and \( \theta \neq 0 \) respectively. Even though one cannot introduce a magnetic potential everywhere on \( Q \), it is still possible to give a satisfactory description of the monopole field by giving the potential \( A^+ \), which is regular on \( U^+ \) and the potential \( A^- \), which is regular on \( U^- \). Together, these two open sets cover all of \( Q \). On the intersection \( U^+ \cap U^- = \mathbb{R}^3 \setminus \{z\text{-axis}\} \), the two potentials are related by a gauge transformation:

\[
A^+ - A^- = 2Q_M d\varphi .
\]

As emphasized in section 2.1, quantization of the particle with lagrangian (2.5) demands that \( A \) 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:

\[
\psi^+(\theta, \varphi) = g(\theta, \varphi)^{-1} \psi^-(\theta, \varphi) \quad \text{on } U^+ \cap U^- .
\]

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

\[
\mathcal{A}^+ = \mathcal{A}^- - \frac{\hbar}{ie} g^{-1} dg = \mathcal{A}^- - \frac{\hbar}{e} d\alpha ,
\]

so comparing with (3.5) we see that the appropriate gauge transformation is

\[
g(\theta, \varphi) = e^{-i\frac{\varphi}{2e}Q_M} .
\]

This will be single-valued if the magnetic charge satisfies the following Dirac quantization condition:

\[
Q_M = \frac{\hbar}{2e} n .
\]
CHAPTER 3. $H^2(Q)$ AND THE QUANTIZATION OF PARAMETERS

Noting that $Q_M = \frac{1}{4\pi} \int_{S^2} F$, 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

$$I = \int_{-\infty}^{\infty} dt \left[ \frac{m}{2} \left( \frac{dq^i}{dt} \right)^2 + e \frac{dq^i}{dt} A_i \right].$$

(3.10)

This action suffers from two related problems. First, in the case of the monopole, $A$ 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), $I' = I - \hbar (\alpha(\infty) - \alpha(-\infty))$.

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

$$e \int_c A = e \int_U F,$$

(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 $dF = 0$, 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 $\Delta I$ in the actions $e \int_{U_1} F$ and $e \int_{U_2} F$ is equal to the integral of $F$ 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 $e^{\pm \Delta I} = 1$, 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 $\oint p dq = 2\pi \hbar n$. From (2.6) we have $\oint p dq = \int (mg_{ij}\dot{q}^i \dot{q}^j + eq^i A_i) dt$. 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 $e \oint A = e \int_{S^2} F$, 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 $Q = \Gamma(S^1, S^3)$. Using the, by now familiar, technique of Appendix ??? we find that $\pi_0(Q) = \pi_1(SU(2)) = 0$, $\pi_1(Q) = \pi_2(SU(2)) = 0$ and $\pi_2(Q) = \pi_3(SU(2)) = \mathbb{Z}$. The generator of $\pi_2(Q)$ is a map $m : S^2 \to Q$ which is defined by $(m(t_1, t_2))(t_3) = \hat{m}(t_1, t_2, t_3)$, 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 $H^0(Q, \mathbb{Z}) = 0$, $H^1(Q, \mathbb{Z}) = 0$ and $H^2(Q, \mathbb{Z}) = \mathbb{Z}$. The low homotopy and cohomology groups of $Q$ 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 $\varphi$ 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 $\varphi$ 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 $\varphi$ and not on the value of the fields in the interior of
the ball. We therefore have

$$S_{WZW} = c \int \varphi^* \tau = \frac{c}{2} \int d^2 x \varepsilon^{\mu\nu} \partial_\mu \varphi^\alpha \partial_\nu \varphi^\beta \tau_{\alpha\beta}$$

(3.12)

where $\omega = d\tau$ or, in components,

$$\omega_{\alpha\beta\gamma} = 3(\partial_\alpha \tau_{\beta\gamma} + \partial_\beta \tau_{\gamma\alpha} + \partial_\gamma \tau_{\alpha\beta}) \, .$$

(3.13)

and $\omega$ 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 $\theta$. For example, suppose we choose on $SU(2)$ a coordinate system given by the Euler angles $(\Theta, \Phi, \Psi)$ (see Appendix G). The volume form is given by

$$\omega = \frac{1}{8\pi} \sin \Theta \, d\Theta \wedge d\Phi \wedge d\Psi \, .$$

(3.14)

and a choice of $\tau$ is

$$\tau = -\frac{1}{8\pi} \cos \Theta \, d\Phi \wedge d\Psi \, .$$

(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 $\tau$ is not uniquely defined: if $\tau$ 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 $\tau$ is not globally defined. If it was, $\omega$ 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 $\tau$ is singular somewhere on $SU(2)$. For example the form $\tau$ defined in (3.15) is singular for $\Theta = 0$ or $\pi$.

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, $\varphi$ will not meet the singular points of $\tau$. Thus there will be an open subset $U$ of $\Gamma_*(S^2, SU(2))$ where $\text{Im} \varphi \cap \{\text{singular set}\} = \emptyset$, and the WZW action (3.12) will be well defined on $U$. However, there are also maps $\varphi$ whose image intersects the singular set of $\tau$. 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 $\tau$, and the action $c \int \varphi^* \tau' = c \int \varphi^* \tau + c \int d(\varphi^* \beta)$ will be well defined on another open subset $U'$. 
3.2. WESS–ZUMINO–WITTEN TERMS

of $\Gamma_*(S^2, SU(2))$. Since the image of $\varphi$ has dimensions 2 and the singular set of any form $\tau$ has dimension zero, it is clear that for every $\varphi \in \Gamma_*(S^2, SU(2))$ one can find a one-form $\beta$ such that $\tau'$ does not have any singularity on the image of $\varphi$. In this way one can cover $\Gamma_*(S^2, SU(2))$ 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 $\varphi$ in the ordinary sense. Instead, it is a section of a certain line bundle over $\Gamma_*(S^2, SU(2))$.

Let us consider the non-linear sigma model with the action given by

$$S = S_0 + S_{WZW},$$

where

$$S_0 = -\frac{1}{2} \int d^2 x \partial_\mu \varphi^\alpha \partial^\mu \varphi^\beta h_{\alpha\beta}. \quad (3.16)$$

The equation of motion reads

$$h_{\alpha\beta} \partial_\mu \varphi^\beta + \Gamma_{\alpha\beta\gamma} \partial_\mu \varphi^\beta \partial^\mu \varphi^\gamma + \frac{c}{2!} \varepsilon^{\mu\nu} \partial_\mu \varphi^\beta \partial_\nu \varphi^\gamma \omega_{\alpha\beta\gamma} = 0 \quad (3.17)$$

where $\Gamma_{\alpha\beta\gamma} = \frac{1}{2} (\partial_\beta h_{\alpha\gamma} + \partial_\gamma h_{\alpha\beta} - \partial_\alpha h_{\beta\gamma})$ 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 $\int dt \dot{\varphi}^\alpha A_\alpha(\varphi)$, where

$$A = -c \int dx \partial_\nu \varphi^\alpha \tau_{\alpha\beta} \delta \varphi^\beta \quad (3.18)$$

is a one-form on $Q$. When we think of the sigma model as a particle moving on $Q$, $A$ 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

$$F = dA = \frac{c}{2} \int dx \partial_\nu \varphi^\alpha \omega_{\alpha\beta\gamma} \delta \varphi^\beta \delta \varphi^\gamma. \quad (3.19)$$

To confirm our interpretation of $A$ and $F$ note that the last term in (3.17) can be written $c \varphi^\beta F_{\alpha\beta}$ and therefore can be interpreted as the Lorentz force due to $F$. The one-form $A$ is only well defined on a subset $V$ of $Q$ such that the image of $\varphi(x)$ (a loop in $SU(2)$) does not intersect the singular set of $\tau$. By contrast, $F$ is well defined everywhere on $Q$. 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 $A$. 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

$$\int_{m} \mathcal{F} = \int^{1}_{0} dt_{1} \int^{1}_{0} dt_{2} \left\{ c \int dx \frac{\partial \varphi^{\alpha}_{1}}{\partial t_{1}} \frac{\partial \varphi^{\gamma}_{2}}{\partial t_{2}} \right\}$$

$$= \frac{c}{3!} \int d^{3}x \varepsilon^{\lambda\mu\nu} \frac{\partial \varphi^{\alpha}_{1}}{\partial x^{\lambda}} \frac{\partial \varphi^{\beta}_{2}}{\partial x^{\mu}} \frac{\partial \varphi^{\gamma}_{1}}{\partial x^{\nu}} \omega_{\alpha\beta\gamma}$$

$$= cW(\tilde{\varphi}) = c.$$  \hspace{2cm} (3.20)

Using (3.1) we find that the theory can be quantized only for

$$c = 2\pi n.$$  \hspace{2cm} (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 $\tilde{\varphi}$ is not unique. Consider two extensions $\tilde{\varphi}_{1} : B_{1}^{3} \rightarrow SU(2)$ and $\tilde{\varphi}_{2} : B_{2}^{3} \rightarrow SU(2)$, with $\tilde{\varphi}_{1}|_{S^{2}} = \tilde{\varphi}_{2}|_{S^{2}} = \varphi$. Since they coincide on $\partial B_{1} = \partial B_{2} = S^{2}$, we can think of them as a single map $\tilde{\varphi} : S^{3} \rightarrow SU(2)$, where $S^{3}$ is obtained by gluing 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 $\varphi_{1}$ and $\varphi_{2}$ together define a map $\tilde{\varphi} : S^{3} \rightarrow SU(2)$. 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 $\omega$ representing a nontrivial cohomology class of the target space. This form could be written locally as the exterior differential of a two-form $\tau$. The Wess–Zumino–Witten action was the integral of the pullback of $\tau$. In 3+1 dimensions $\omega$ would have to be a closed five-form and $\tau$ 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
3.3. CHERN–SIMONS TERMS

of $-\frac{i}{240\pi^2}\text{tr}R^5$. Therefore, the Wess–Zumino–Witten term can be written in either one of the following two forms:

$$S_{WZW} = c\int d^4x \varepsilon^{\mu\nu\rho\sigma} \partial_\mu \varphi^\alpha \partial_\nu \varphi^\beta \partial_\rho \varphi^\gamma \partial_\sigma \varphi^\delta \tau_{\alpha\beta\gamma\delta}$$

$$= -\frac{ic}{240\pi^2} \int d^5x \varepsilon^{\lambda\mu\nu\rho\sigma} \text{tr}(R_\lambda R_\mu R_\nu R_\rho R_\sigma), \quad (3.22)$$

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

$$A = -c\int d^3x \varepsilon^{ijk} \partial_i \varphi^\alpha \partial_j \varphi^\beta \partial_k \varphi^\gamma \tau_{\alpha\beta\gamma\delta} \delta \varphi^\delta \quad (3.23)$$

on $Q$. The corresponding field strength is

$$F = dA = c\int d^3x \varepsilon^{ijk} \partial_i \varphi^\alpha \partial_j \varphi^\beta \partial_k \varphi^\gamma \omega_{\alpha\beta\gamma\delta\eta} \delta \varphi^\delta \delta \varphi^\eta. \quad (3.24)$$

One can now repeat the arguments given above leading to the quantization of the parameter $c$.

Finally we observe that the relation between $\omega$ and $F$ 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 $A_\mu = A_\mu^a T_a$, where $T_a$ are matrices satisfying $[T_a, T_b] = \varepsilon_{abc} T_c$ and $\text{tr}(T_a T_b) = -\frac{1}{2} \delta_{ab}$ (for example in the spinor representation, $T_a = -\frac{i}{2} \sigma_a$, where $\sigma_a$ are the Pauli matrices). In this notation the Yang–Mills action (1.89) reads

$$S_{YM} = \frac{1}{2e^2} \int d^3x \text{tr} F_{\mu\nu} F^{\mu\nu}. \quad (3.25)$$
As in Sections 3.4 and 3.5, we choose the gauge $A_0 = 0$; then the static energy reads

$$E_S = -\frac{1}{e^2} \int d^2 x \text{tr} B^2,$$

(3.26)

where $B = F_{12}$ is the nonabelian magnetic field. The configuration space is then $Q = C/G$, where $C$ is the space of connections $A_i(x)$, $i = 1, 2$, such that $E_S$ is finite, and $G = \Gamma_*(S^2, SU(2))$ is the residual gauge group consisting of time independent gauge transformations.

This configuration space is connected and furthermore has $\pi_1(Q) = \pi_0(G) = \pi_2(SU(2)) = 0$ and $\pi_2(Q) = \pi_1(G) = \pi_3(SU(2)) = \mathbb{Z}$. The generator of the group $\pi_2(Q)$ can be described as follows. The gauge group $G$ is connected but not simply connected. Let $\ell(t)$ be a loop whose homotopy class generates $\pi_1(G)$. Fix a reference point $A(0)$ in $C$ 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 $C$. Thus there is a map $\tilde{m}$ from a two dimensional ball $B^2$ to $C$ which is equal to $A^{\ell(t)}(0)$ on the boundary. Now compose this map with the projection $C \rightarrow Q$. Since all points on the boundary of the disk are mapped to the same orbit, we get a map $m$ from $S^2$ to $Q$ which is not homotopic to a constant (see Fig. ??). The isomorphism between $\pi_1(G) = \mathbb{Z}$ and $\pi_2(Q)$ is the map that sends (the homotopy class of) $\ell$ 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 $A_\mu$ on a three sphere are topologically trivial ($\pi_2(SU(2)) = 0$) and therefore can be thought of as boundary values of gauge fields $A_\mu$ 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 $c_2(A)$ 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 (2.28), the appropriate topological term to be added to $S_{YM}$ in three dimensions is the Chern–Simons term

$$S_{CS} = \mu \frac{8\pi^2}{e^2} \int d^3 x \Omega,$$

(3.27)
where
\[ \Omega = -\frac{1}{8\pi^2} \varepsilon^{\lambda\mu\nu} \text{tr} \left( A_\lambda \partial_\mu A_\nu + \frac{2}{3} A_\lambda A_\mu A_\nu \right) . \] (3.28)

(Note that apart from replacing the coefficient \( \theta \) by the coefficient \( \mu \frac{8\pi^2}{e^2} \), \( S_{\text{CS}} \) is identical to the functional \( \tilde{\Lambda} \) defined in (2.32).) The constant \( \mu \) 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 \( \mu \), 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
\[ \Omega(A^g) = \Omega(A) - \frac{1}{8\pi^2} \varepsilon^{\lambda\mu\nu} \text{tr} \left( \partial_\lambda (\partial_\mu g^{-1} A_\nu) + \frac{1}{24\pi^2} \varepsilon^{\lambda\mu\nu} \text{tr} \left( g^{-1} \partial_\lambda g^{-1} \partial_\mu g^{-1} g_\nu \right) \right) , \] (3.29)
and since we assume \( g \) to tend to the identity at infinity, upon integration we find
\[ S_{\text{CS}}(A^g) = S_{\text{CS}}(A) + \mu \frac{8\pi^2}{e^2} W(g) . \] (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 \( e^{i\Delta S_{\text{CS}}} = 1 \), or
\[ \mu = \frac{e^2}{4\pi} n , \quad n \in \mathbb{Z} . \] (3.31)

Note that in the Euclidean path integral one would demand \( e^{-\Delta S_{\text{CS,E}}} = 1 \), where \( S_{\text{CS,E}} \) is the Euclidean Chern–Simons action. Since \( S_{\text{CS}} \) is linear in the time derivative, \( S_{\text{CS,E}} = i S_{\text{CS}} \), 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
\[
Z = V_0 \sum_n e^{-\mu \frac{g_0^2}{4\pi}} \int (dA) \Delta_{FP}(A) \delta(f(A)) e^{iS(A)},
\]
where \( V_0 \) is the volume of one connected component of the gauge group. The sum in front of the integral gives zero unless \( \mu \) 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.