2

Non-Perturbative Quantum Field Theory

An Introduction to Topological and Semiclassical Methods

Roberto Percacci

SISSA Medialab

ICTP The Abdus Salam International Centre for Theoretical Physics

This page has intentionally left blank.

This page has intentionally left blank.

Non-Perturbative Quantum Field Theory

An Introduction to Topological and Semiclassical Methods

Roberto Percacci

Published by SISSA Medialab S.r.l.
Via Bonomea 265
34136 Trieste, Italy
https://medialab.sissa.it/

Cover: Giacomo Sanna — Dotik

Typesetting: Elia A. Calderan and Giorgia del Bianco — SISSA Medialab S.r.l.

This eBook was published Open Access with funding support from the Sponsoring Consortium for Open Access Publishing in Particle Physics (SCOAP3).

Title: Non-Perturbative Quantum Field Theory. An Introduction to Topological and Semiclassical Methods

Authors: Roberto Percacci

Keywords: 1. Quantum field theory 2. Topology 3. Solitons 4. Vortices 5. Magnetic monopoles 6. Instantons 7. Anomalies

First published 2024
ISBN: 9788898587056 (electronic edition)
DOI: 10.22323/9788898587056

Foreword

This book is a slightly extended and improved version of the lecture notes of a course I have given for several years to PhD students at SISSA. In its most recent form, it was called "Quantum Field Theory II", or QFT2 for short, and covered mostly non-perturbative phenomena. To a student who has struggled through some complicated one- or two-loop calculation in perturbative QFT, the notion that one may be able to compute some nonperturbative effect could sound at first preposterous. In fact, if some phenomenon can be explained in perturbation theory, the first few perturbative orders (often just the first) are enough. Any non-perturbative effect is practically negligible in that case. The only cases when one may have to study non-perturbative effects are those when the effect itself is absent in perturbation theory. Then, even a non-perturbative contribution may give rise to some detectable effect. Such cases are more common in condensed matter physics than in particle physics. For this reason the course contained, in addition to the material presented here, also a part on critical phenomena. Only a few sections of that part remain here, in those places where they serve to illustrate some specific applications of the general notions.

The study of the non-perturbative effects that we will be concerned with relies to a large extent on the use of semiclassical methods. The simplest example is spontaneous symmetry breaking. One finds in that case that the zero field, though solving the classical field equations, is not stable and the lowest energy state involves a non-zero field. This classical solution is a first approximation to the vacuum expectation value of the quantum field. One can then study the effect of quantum fluctuation (or, in statistical field theory, thermal fluctuations) around this state. The case of spontaneous symmetry breaking involves only homogeneous vacuum states. We will be interested in cases when there exist classical solutions of the equations of motion that have a nontrivial shape in space (solitons) or in spacetime (instantons). These give rise to more complicated effects. The existence of solitons in the classical theory signals that the quantum theory contains, in addition to the usual

particle states that are accessible in perturbation theory, also other heavier particles. Instantons give rise to tunnelling in the quantum theory, also a nonperturbative phenomenon. In both cases, the topology of the configuration space plays an important role. There is a third type of topological effect, leading to the quantization of a parameter that appears in the Lagrangian. It is to all these effects that this book is devoted. There are other nonperturbative ways of studying QFT, such as lattice field theory, or conformal field theory. They are not covered here.

The notions discussed in this book played an important role in the development of theoretical physics in the second half of the 20th century, with a peak of activity in the late 1970's and early 1980's. (It is probably not a coincidence that those were also the author's formative years.) They involved a large number of contributors, but a few names recur frequently: A. Polyakov, (various solitons and instantons); R. Jackiw (theta vacua, anomalies); G. 't Hooft (monopoles, quantum fluctuations around the instanton); E. Witten (anomalous actions). One notable feature of these topics is that they gave rise to an intense dialog between physicists and mathematicians, that would have been unthinkable just a few decades earlier. M. Atiyah and I. Singer were particularly instrumental in explaining to physicists the deeper meaning of what they were doing and opening the mathematical community to ideas coming from physics.

There exist already several excellent resources on this subject.

- Between 1966 and 1979 Coleman gave several courses in Erice on various topics, including solitons and instantons, that have become classics. His lecture notes, written in an inimitable style, are collected in the book "Aspects of Symmetry" [Col85].
- R. Jackiw wrote two extended review papers: "Field theoretic investigations in current algebra" [Jac72] and "Topological investigations of quantized gauge theories" [Jac83]. Both are reprinted, together with several important papers by Treiman, Zumino and Witten, in [Tre86].
- Rajaraman's "Solitons and instantons" [Raj82] was the earliest book on the subject and in spite of its age is still a very good introduction.
- Shifman's "Advanced topics in quantum field theory" [Shi12] is a more modern and more detailed book, including a second part on supersymmetry.
- Finally, the readers who want to look more deeply into the dynamics of solitons will find a lot of information in Manton and Sutcliffe's book "Topological solitons," [Man04].

While relying heavily on these sources, this book differs mainly in the way the material is presented. Whereas often one tends to organize the topological effects on the basis of the dimension of the space they occur in, it is somewhat more natural to group them according to the connectedness properties of the configuration space $\mathcal{Q}$. A vector space is an example of a space without topological features. The simplest nontrivial thing that can happen is that a space consists of several connected components. These are counted by the zeroth homotopy set $\pi_0(\mathcal{Q})$. The next thing that can happen is that the space is not simply connected. This is measured by the first homotopy group $\pi_1(\mathcal{Q})$. Going still further, a space can contain non-contractible two-dimensional spheres, a property that is measured by the second homotopy group $\pi_2(\mathcal{Q})$. In each of these cases there exists a simple, finite dimensional, quantum mechanical system that exhibits peculiar properties due to this topology, and the quantum field theoretic phenomena that we will be interested in are direct generalizations of these phenomena to the case when the configuration space is infinite dimensional.

The organization of the book reflects this logic. Chapter 1 is introductory and contains miscellaneous material that many readers will already be at least partly familiar with. It could be skipped and consulted only when needed. Chapters 2, 3 and 4 are the core of the book and present phenomena that happen when the configuration space has nontrivial zeroth, first or second homotopy group, respectively, while Chapter 5 is devoted to situations when more than one of these groups is nontrivial. Finally, Chapter 6 is devoted to anomalies. Even though they form a separate subject, they have many connections with the topics of the preceding chapters that are worth pointing out.

It should be clear that in this whole subject a central role is played by topology, a topic that is not part of the traditional toolkit of particle physicists. In the main text I make free use of notions of geometry, homotopy and cohomology. The geometry is mostly of the type that one learns in a course on General Relativity. Homotopy is in many cases just a natural way to encode the boundary conditions imposed on the fields, and cohomology is a generalization of notions that every physicist has encountered in the study of thermodynamics and electromagnetism. On the other hand I have tried to avoid or at least minimize the use of fiber bundles: even though they are the natural mathematical framework for gauge theories, a proper treatment would put too much emphasis on the geometry. Since this is not a mathematical textbook, I have collected the main mathematical notions and some useful results in several appendices, which can be consulted whenever needed. For a much more extended introduction aimed specifically at physicists, see for example [Nak03]. However, readers who want to properly learn these notions

should turn directly to the mathematical literature. Excellent pedagogical treatments can be found in [SiT67] or, more avanced, in [BoT82].

It would be impossible to properly acknowledge all the people that have, directly or indirectly, contributed to this book. Much of my understanding of this subject has been shaped during collaborations with R. Floreanini, J. Mickelsson, N. Pak, R. Rajaraman and E. Sezgin. The preparation of these notes has taken place over a long period of time and I am especially indebted to L. Griguolo for help in an earlier attempt at systematizing the material, in particular Chapter 6. Of course, this book would not exist were it not for all the students who followed my course over the years and whose questions helped me improve the presentation. In addition I wish to thank M. Fabrizio, V. Naso, M. Serone, G.P. Vacca, O. Zanusso for comments and suggestions during the final preparation of the manuscript, and especially D. Buccio for much help with the exercises. Last but not least, many thanks to Cristiana Prever of SISSA Medialab for her gentle prodding. Some projects never get finished unless someone sets deadlines.

Trieste, November 2024

Contents

List of Exercises

Chapter 1

Fields and symmetries

Here we begin by recalling some basic notions, setting up the notation and introducing the models whose topological properties will be discussed in the following chapters: linear and nonlinear scalar theories and gauge fields, possibly coupled to fermions. For each model we shall also give some examples of physical application, either in particle physics or in condensed matter physics. There is much more to be said on each of these models. The treatment here is very concise, giving only the essential notions that are needed in the following chapters.

Whereas in quantum field theory one almost always works in a covariant formulation, in the end of the chapter we shall introduce the canonical approach, where (quantum) field theories are viewed as infinite dimensional (quantum) mechanical systems. This will allow us, in later chapters, to categorize different types of topological effects on the basis of the homotopy or cohomology groups of their configuration space, and to understand them as infinite dimensional versions of simple quantum mechanical phenomena.

We shall focus mostly on bosons, since the nontrivial topology resides entirely in the bosonic sector. There is a nontrivial interplay between the effects of topology and fermions, that will emerge later, in particular in Section 2.1.4, and in Chapter 6.

1.1 Noether's theorems

There are two Noether theorems. The first applies to finite dimensional invariance groups, the second to infinite dimensional invariance groups. Here we discuss the first theorem. The second will be the subject of Section 1.5.2.

The action of a theory is the time integral of the Lagrangian:

$$S = \int dt\, L$$

and for a field theory the Lagrangian is the space integral of a Lagrangian density:

$$L = \int d^d x\, \mathcal{L}.$$

This is more often written in covariant form

$$S(\phi) = \int d^n x\, \mathcal{L}(\phi, \partial_\mu \phi)$$

with $n = d + 1$ the dimension of spacetime. Notice that the action is viewed as a functional of the field, whereas in the Lagrangian density it is customary to indicate separately the dependence on the field and its derivatives. For our purposes it will be sufficient to consider Lagrangian densities that depend on the field and its first derivatives only.

The nature of the field is unspecified at this stage: it could be a fermion or a boson, and carry any number of internal or spacetime indices, that we need not write. We define the conjugate momentum vector

$$\pi^\mu = \frac{\partial \mathcal{L}}{\partial \partial_\mu \phi}, \tag{1.1}$$

whose time component is the usual canonical momentum

$$\pi \equiv \pi^0 = \frac{\partial \mathcal{L}}{\partial \dot{\phi}} \tag{1.2}$$

where, as usual, $\dot{\phi} = \partial_0 \phi$. By considering variations of the field that vanish at infinity, so that integrations by parts are allowed, and demanding invariance of the action under such transformations, leads to the Euler–Lagrange equations, that can be written in the form

$$\frac{\partial \mathcal{L}}{\partial \phi} = \partial_\mu \pi^\mu. \tag{1.3}$$

On the other hand let $\delta_\epsilon \phi$ be an infinitesimal transformation of the field with constant transformation parameter ϵ. The variation of the Lagrangian density is

$$\delta_\epsilon \mathcal{L} = \frac{\partial \mathcal{L}}{\partial \phi} \delta_\epsilon \phi + \frac{\partial \mathcal{L}}{\partial \partial_\mu \phi} \delta_\epsilon (\partial_\mu \phi). \tag{1.4}$$

One assumes that $\delta_\epsilon(\partial_\mu\phi) = \partial_\mu\delta_\epsilon\phi$. Then, if one can show, without using the equations of motion, that this variation is a total derivative $\partial_\mu\Omega^\mu_\epsilon$, the action will be invariant and we call the transformation a *symmetry*.

Using the Euler–Lagrange equation (1.3), the variation (1.4) can be written in another form that is true irrespective of $\delta\phi$ being a symmetry:

$$\delta_\epsilon\mathcal{L} = \partial_\mu(\pi^\mu\delta_\epsilon\phi). \tag{1.5}$$

If $\delta_\epsilon\phi$ *is* a symmetry, we also have $\delta_\epsilon\mathcal{L} = \partial_\mu\Omega^\mu_\epsilon$, so equating these two expressions we arrive at the conclusion that

$$\partial_\mu(\pi^\mu\delta_\epsilon\phi - \Omega^\mu_\epsilon) = 0, \tag{1.6}$$

in other words the current

$$j^\mu_\epsilon = \pi^\mu\delta_\epsilon\phi - \Omega^\mu_\epsilon \tag{1.7}$$

is conserved. One can define the current also in the case when $\delta_\epsilon\phi$ is not a symmetry, in which case its divergence is

$$\partial_\mu j^\mu_\epsilon = \delta_\epsilon\mathcal{L}. \tag{1.8}$$

Integrating the time component of the current on a constant-time surface we obtain a charge

$$Q_\epsilon = \int d^dx\, j^0_\epsilon \tag{1.9}$$

that does not change in time:

$$\frac{dQ_\epsilon}{dt} = 0. \tag{1.10}$$

This is the statement of Noether's first theorem: to each one-parameter continuous group of symmetries there corresponds a conserved quantity.

The finite transformations form a Lie group G and the corresponding infinitesimal transformations span its Lie algebra $\mathfrak{g}$. Thus the transformation itself has as many parameters as the dimension of G. We can write $\epsilon = \epsilon^a T_a$, where T_a are a basis of $\mathfrak{g}$, then the current and charge can be written

$$j^\mu_\epsilon = \epsilon^a j^\mu_a, \qquad Q = \epsilon^a Q_a.$$

It is common to write the components of the current with the parameter ϵ stripped off.

For internal transformations one has $\Omega_\epsilon = 0$ and the Noether current is just $j_a^\mu = \pi^\mu \delta_a \phi$. Let us consider the case of a bosonic field carrying a linear representation of G, with infinitesimal transformation[1]

$$\delta_a \phi^m = -(T_a)^m{}_n \phi^n. \tag{1.11}$$

Then the Noether current is

$$j_a^\mu = -\pi_m^\mu (T_a)^m{}_n \phi^n, \tag{1.12}$$

where we now make the index explicit in $\pi_m^\mu = \frac{\partial \mathcal{L}}{\partial \partial_\mu \phi^m}$. In particular, the time component of the current is

$$j_a^0 = -\pi_m (T_a)^m{}_n \phi^n, \tag{1.13}$$

where π_m is the momentum conjugate to ϕ^m. We have the canonical equal time commutation relations

$$[\phi^m(t,x), \pi_n(t,y)] = i\delta_n^m \delta^{(d)}(x-y) \tag{1.14}$$

(we work in natural units) and the commutation relations of the generators

$$[T_a, T_b] = f_{ab}{}^c T_c. \tag{1.15}$$

(The generators are assumed antihermitian and the structure constants $f_{ab}{}^c$ are real). Using these, one easily finds that

$$[j_a^0(t,x), j_b^0(t,y)] = -if_{ab}{}^c j_c^0(t,x)\delta^{(d)}(x-y). \tag{1.16}$$

In particular, the charges $Q_a = \int d^d x J_a^0$ satisfy

$$[Q_a, Q_b] = -if_{ab}{}^c Q_c. \tag{1.17}$$

The additional factor of i in the r.h.s. of (1.17), compared to (1.15), comes from the quantum theory: it is required because the charges are physical observables and must be hermitian operators. If we computed classical Poisson brackets of currents we would find that they satisfy the same algebra as the generators T_a themselves. The $-$ sign is conventional: it follows from the choice of sign in (1.11).

[1] We use indices from the beginning of the latin alphabet for the Lie algebra and the adjoint representation, and from the middle of the latin alphabet for other representations.

For a fermion multiplet ψ^m carrying the same representation of G, from the free Lagrangian density $\mathcal{L} = -\bar{\psi}\gamma^\mu\partial_\mu\psi$ one finds that $\pi^\mu = -\bar{\psi}\gamma^\mu$, and the Noether current is given by the formula

$$j_a^\mu = \pi_m^\mu (T_a)^m{}_n \psi^n = \bar{\psi}_m \gamma^\mu (T_a)^m{}_n \psi^n. \tag{1.18}$$

In the case of global $U(1)$ transformations $\delta\psi = -i\epsilon\psi$ the abelian current is

$$j^\mu = i\bar{\psi}\gamma^\mu\psi. \tag{1.19}$$

The factor i makes the current real. In fact, it is not too obvious that these currents are hermitian operators, so this is left as Exercise 1.1

The momentum canonically conjugated to ψ is $\pi^0 = -\bar{\psi}\gamma^0 = i\psi^\dagger$. Therefore

$$\{\psi^{Am}(x,t), \bar{\psi}_{Bn}(y,t)\} = i(\gamma^0)^A_B \delta^m_n \delta^{(d)}(x-y), \tag{1.20}$$

where A, B are spinor indices. Using these canonical anticommutation relations, one finds that the currents (1.18) also satisfy the current algebra (1.16).

Notice that these calculations do not require knowledge of the Lagrangian but follow purely from canonical arguments. Furthermore, these relations hold independently of whether the current is conserved or not. While very general, they are also formal and do not take into account the difficulties that may arise, for example, when we multiply two field operators at the same point. For this reason, one should not be too surprised that relations (1.17) can be violated in certain situations. We will discuss this point in Section 1.6 and then more extensively in Chapter 6.

1.2 Linear scalar theories

A scalar field is a map from space, or spacetime, to some target space. If the target space is a linear space, we shall call the field a linear scalar field; if the target space is some other manifold, we shall call the field a nonlinear scalar field. The distinction is purely kinematical. In both cases we will be interested in interacting fields, so the field equations will be nonlinear, but the type of interactions that arise in the two cases are quite different. In this section we discuss linear scalar theories. In Section 1.3.1 we shall see how they give rise to nonlinear theories in some limiting situations, and then discuss the nonlinear theories *per se*.

1.2.1 The $O(N)$ models

We consider a multiplet of $N \geq 1$ scalar fields ϕ^m ($m = 1, \ldots, N$) in n spacetime dimensions. Assuming symmetry under $O(N)$, the action is

$$S = \int d^n x \left[-\frac{1}{2}\partial_\mu \phi^m \partial^\mu \phi^m - V(|\phi|) \right], \tag{1.21}$$

where $|\phi| = \sqrt{\phi^m \phi^m}$ and repeated indices are summed over. The case $N = 1$ is in some respect different from the others because the group $O(1)$ is discrete, being isomorphic to the group $\mathbb{Z}_2$ generated by the reflection $\phi \to -\phi$. Here we consider only quartic potentials that can be parametrized as

$$V(|\phi|) = \frac{1}{2} m^2 |\phi|^2 + \frac{\lambda}{4} |\phi|^4 + U, \tag{1.22}$$

where U is an arbitrary constant (at least as long as we are not interested in gravity). The minimum of the potential is the classical vacuum state of the system and in the following we shall loosely refer to it as the vacuum, or, in the quantum context, the vacuum expectation value (VEV) of the field.[2]

The system can be in two phases, depending on the minimum of the potential, which in turn depends on the sign of the mass term. When $m^2 > 0$ the minimum is in the origin and we can put $U = 0$. The vacuum is unique and the linearization of the action around the vacuum, which correponds to just dropping the quartic term in the potential, shows that there are N free scalar fields, with mass matrix

$$\left. \frac{\partial^2 V}{\partial \phi^r \partial \phi^s} \right|_{\phi=0} = \begin{pmatrix} m^2 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots \\ 0 & \dots & m^2 & 0 \\ 0 & \dots & 0 & m^2 \end{pmatrix}. \tag{1.23}$$

This is called the symmetric phase, and in this phase the symmetry $O(N)$ is said to be *linearly realized*.

When $m^2 < 0$ the minimum is not unique: the minima form an $N-1$-dimensional sphere S^{N-1} of radius

$$f = \sqrt{\frac{|m^2|}{\lambda}}. \tag{1.24}$$

In this case we put $U = \frac{m^4}{4\lambda}$ and the potential can be rewritten in the convenient form

$$V(|\phi|) = \frac{\lambda}{4} \left(|\phi|^2 - f^2 \right)^2. \tag{1.25}$$

[2]We shall follow here the standard presentation of this topic, but the reader should be aware that in this way we overlook some subtleties. Both in QFT and in statistical mechanics the potential one usually minimizes is not the one appearing in the classical action, but rather the potential in the effective action and in the free energy, respectively. Both functionals are defined by Legendre transform and therefore must be convex, which is not the case with (1.22) when $m^2 < 0$. The point is that the potential (1.22) gives the energy for homogeneous configurations, whereas the free energy is minimized by spatially inhomogeneous configurations. Furthermore, for $m^2 < 0$ the perturbative evaluation yields an effective potential that is complex near the origin. The physical significance of this fact is clarified in [Wei87].

Since all the minima have zero energy, the vacuum state is not unique. Each point on the sphere defines a different vacuum state, and once that state has been chosen, it will remain the same throughout the history of the system. Even though the action of the system is $O(N)$-invariant, any one of the vacuum states is only invariant under a subgroup $O(N-1)$ (and in fact the locus of the minima is the coset space $O(N)/O(N-1) = S^{N-1}$). For this reason, the symmetry is said to be *spontaneously broken* (as opposed to the case of a symmetry that is explicitly broken by the presence of some non-invariant term in the action) and this phase is called the broken symmetry phase. That the breaking is only due to the choice of state is confirmed by the fact that the Noether currents are conserved also in the broken phase, see Exercise 1.2.

Without loss of generality we can choose the vacuum to be the "north pole"

$$\phi_* = (0, \dots, 0, f).$$

Linearizing the action around this vacuum we find again N free scalars, but this time the mass matrix is

$$\left.\frac{\partial^2 V}{\partial\phi^r \partial\phi^s}\right|_{\phi=\phi_*} = \begin{pmatrix} 0 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots \\ 0 & \dots & 0 & 0 \\ 0 & \dots & 0 & 2|m^2| \end{pmatrix}. \tag{1.26}$$

We thus have $N-1$ massless particles called the *Goldstone bosons*, that correspond simply to the fields $\phi^1,\dots\phi^{N-1}$, and one massive particle that is described by the shifted field $\phi^N - f$. Note that the mass of this particle is $\sqrt{2}$ times the mass of the particles in the symmetric phase. Also note that the identification of the Goldstone bosons with the fields $\phi^1,\dots\phi^{N-1}$ is only valid in an infinitesimal neighborhood of the north pole, where we can think of them as belonging to the tangent space to the sphere at the north pole. The true configuration space of the Goldstone bosons is the sphere S^{N-1}. Thus a more appropriate description of the broken phase would use spherical coordinates in $\mathbb{R}^N$, with the angular coordinates describing the Goldstone bosons and the (shifted) radial coordinate being the massive field. We shall therefore refer to the massive particle as the radial mode. The appearance of a number of massless particles equal to the number of broken generators of the symmetry group is the content of Goldstone's theorem.

The model we have just described is called the $O(N)$ model in statistical physics and the linear sigma model in nuclear physics. We will discuss these two applications in more detail in the next two sections. The $O(N)$ model is suitable to describe any situation in which a system exhibits a global (as

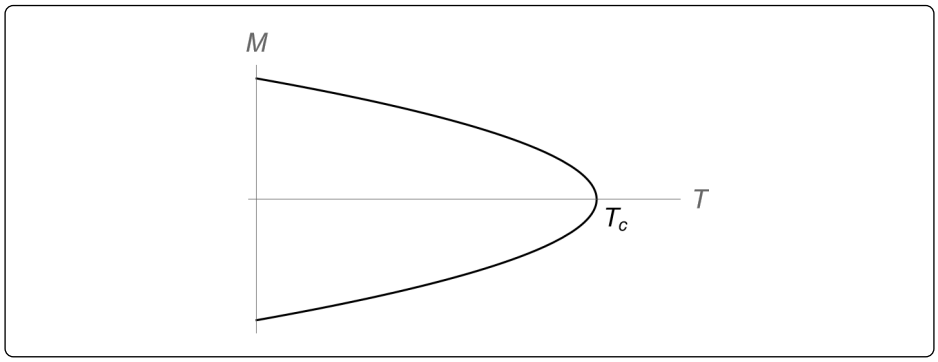


Figure 1. Spontaneous magnetization as a function of temperature. At a given $T < T_c$, the $\mathbb{Z}_2$ symmetry of the action is spontaneously broken: there are two possible, energetically degenerate states, and which one the system is in depends on it earlier history.

opposed to local, or gauged) $O(N)$ symmetry, that is spontaneously broken to $O(N-1)$. It can be generalized in a straightforward way to the case of any global symmetry G spontaneously broken to a subgroup H. In this case the multiplet of fields ϕ^a carries some representation of G, the potential has to be G-invariant and the minima form a coset space G/H. For the construction of G-invariant quartic potentials see e.g. [Mic79].

1.2.2 The Ginzburg–Landau theory of phase transitions

If we replace n-dimensional spacetime by n-dimensional Euclidean space, the functional (1.21), with the overall sign changed, is used as an approximate form of the free energy of a statistical system in n space dimensions. In this context, this theory is known as the Ginzburg–Landau model, and has important applications to the description of phase transitions.

To illustrate this type of application we consider the theory of magnetism and we focus on the simplest case of the uniaxial ferromagnet as a paradigmatic example. It corresponds to the $N = 1$ case of the models of the previous section, and we recall that $O(1) = \mathbb{Z}_2$. The macroscopic state of the material is described by the average magnetization M, which depends on the temperature T and on the external magnetic field H. When $H = 0$ and the temperature is below a critical temperature T_c, called the Curie temperature, there is a residual (or "spontaneous") magnetization. This is called the ferromagnetic state. The spontaneous magnetization depends upon the earlier history of the system (hysteresis). Thus for example, if at some earlier time the magnetic field had been pointing up, and then had been monotonically reduced to

zero, then the residual magnetization will point up. If one continuously turns on a magnetic field pointing down, the magnetization also follows and has a discontinuous jump. Actually, the material will remain for some time in a metastable state with the direction of magnetization opposite to that of H, but if one waits long enough it will align itself with the magnetic field. These transitions between the two macroscopic states of magnetization are discontinuous transitions, or *first order transitions*.[3] As the temperature increases, the discontinuity in magnetization across the transition decreases, until the Curie temperature is reached, where the discontinuity vanishes, see Figure 1.

Above the Curie temperature the system is in a paramagnetic state, where the magnetization changes smoothly as a function of the magnetic field, and $M = 0$ when $H = 0$. Exactly at the critical temperature there is still a phase transition, but a continuous one. At this point, the specific heat has a discontinuity and the magnetic susceptibility

$$\chi = \left.\frac{\partial M}{\partial H}\right|_{H=0}$$

diverges. So, we have a line of first order transitions that ends at a second order transition. A remarkable property of systems near a critical point is that many quantities exhibit a characteristic power law behavior. It is convenient to define a reduced temperature

$$t = \frac{|T - T_c|}{T_c}$$

that is zero at the critical point. Then, the residual magnetization at $H = 0$ scales as

$$M \sim (-t)^{\beta}, \tag{1.27}$$

the magnetic susceptibility diverges as

$$\chi \sim |t|^{-\gamma} \tag{1.28}$$

and at $t = 0$

$$M \sim H^{1/\delta}. \tag{1.29}$$

The exponents β, γ, δ (and others that we shall not discuss) are called *critical exponents*.

[3] In the classification proposed by Ehrenfest, a transition is of n-th order if an n-th derivative of the free energy is discontinuous. This terminology is now considered obsolete but is still widely used.

These phenomena are due to the interaction of spins on the lattice with the external magnetic field. The microscopic model for the uniaxial ferromagnet is the Ising model. It consists of spins pointing in a fixed direction, with values $S_a = \pm 1$, on a d-dimensional lattice, with Hamiltonian

$$\mathcal{H} = -J \sum_{a,b} S_a S_b - \mu H \sum_a S_a \tag{1.30}$$

where the first sum extends over all nearest neighbors. When $H = 0$, the Hamiltonian has a global $\mathbb{Z}_2$ symmetry that flips the signs of all spins. If $J > 0$, it is energetically favorable for the neighboring spins to have the same sign, while the second term tends to align the spins with the magnetic field.

Then, the various states of the system result from the competition between these interactions, that tend to bring the system into a state where all spins are aligned, and the effect of thermal fluctuations, that tend to randomize them. In the paramagnetic state thermal fluctuations dominate, whereas in the ferromagnetic state they are small. Thus, the paramagnetic state is called the *disordered* state and is also a state in which the $\mathbb{Z}_2$ symmetry of the Ising model is unbroken. The ferromagnetic state is an *ordered* state in which the $\mathbb{Z}_2$ symmetry of the Ising model is spontaneously broken.

At a macroscopic level, all this is well described by the Ginzburg–Landau theory, that is just a linear scalar theory of the type discussed in the previous section. In terms of the Ising model, the average magnetization in a given volume is the sum $M = \langle \sum_a S_a \rangle$ over the sites in that volume, but we need not refer to the microscopic spins at all. The important feature is that for $H = 0$ both models have $\mathbb{Z}_2$ symmetry. Although the true free energy may be a very complicated function, given that we are interested in the macroscopic (infrared) properties of the system, it is enough to retain the leading terms in a derivative expansion. Including the linear coupling to the magnetic field, that acts as an external source, these are

$$F = \int d^n x \left[\frac{1}{2} r M^2 + \frac{u}{4} M^4 - HM + \frac{1}{2} \partial_i M \partial_i M + \ldots \right], \tag{1.31}$$

This is just a Euclidean version of (1.21) with $N = 1$, where the field has been renamed M, the squared mass has been renamed r and the quartic coupling u. The average magnetization is dictated by the potential and the kinetic term is needed only for the two point function. One can derive this form of the free energy in a mean field treatment of the Ising model, see Exercise 1.3, but here we treat it as a valid model *per se*.

One of the crucial properties of the critical point is that the correlation length ξ, which is defined by the behavior of the two point function

$$G(x - y) = \frac{e^{-|x-y|/\xi}}{|x - y|^{n-2}},$$

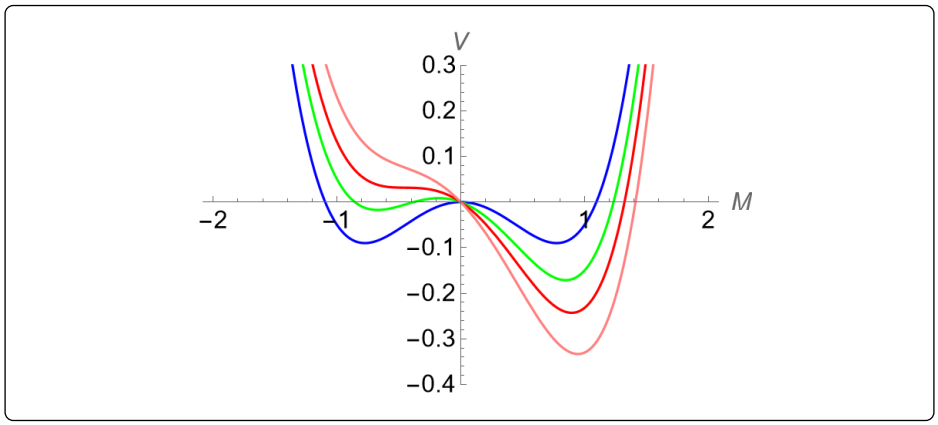


Figure 2. Potentials corresponding to $u = 1$, $r = -0.6$ and $H = 0, 0.1, 0.182, 0.28$ in blue, green, red, pink, respectively.

tends to infinity when $T \to T_c$. The correlation length is the inverse of the mass, so this suggests that we think of r and u as analytic functions of temperature, with r having a simple zero at the critical temperature:

$$r(T) = r_0(T - T_c) + \dots, \qquad u(T) = u_0 + \dots$$

The ground state of the system is a stationary point of the free energy, so it satisfies

$$M(r + uM^2) = H. \tag{1.32}$$

Above the critical temperature, $r > 0$ and we are in the symmetric phase, that corresponds to the paramagnetic state, where $M = 0$ is the unique minimum for $H = 0$. Below the critical temperature, $r < 0$ and we are in the broken phase. For $H = 0$ the ground states are

$$M = \pm\sqrt{-\frac{r}{u}}. \tag{1.33}$$

For small positive H the positive minimum is the stable ground state whereas the negative minimum is a metastable state (green curve in Figure 2). For sufficiently large H this metastable state ceases to exist (pink curve in Figure 2). The plot of the solutions for M as function of H correctly reproduces the hysteresis curves of actual ferromagnets, see Figure 3.

At a quantitative level, the Ginzburg–Landau model gives predictions for the critical exponents. From (1.33), since $r \sim t$, we see that $\beta = 1/2$. In order to calculate the magnetic susceptibility, we differentiate (1.32) and find that

$$\frac{\partial M}{\partial H} = -\frac{1}{r + 3uM^2}.$$

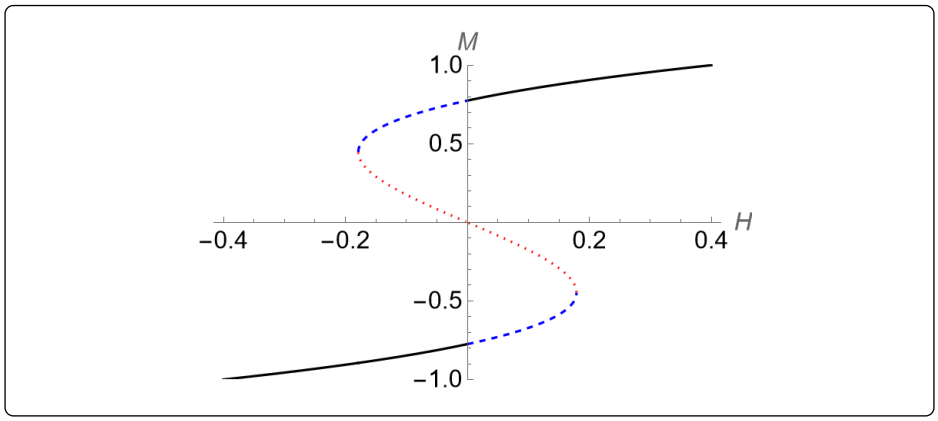


Figure 3. The solution of (1.32) for $u = 1, r = -0.6$. The stable branches are drawn in black, the metastable branches in dashed blue and the unstable branch in dotted red.

At $H = 0$, M is given by (1.33), the susceptibility is

$$\left.\frac{\partial M}{\partial H}\right|_{H=0} = \frac{1}{2r} \sim t^{-1},$$

so $\gamma = 1$. At the critical temperature $r = 0$ and (1.32) gives $M \sim H^{1/3}$, which implies that $\delta = 3$. The following table compares these values to the ones measured in actual systems, in three space dimensions. The agreement is not impressive, but it is reasonably good, considering the simplicity of the model.

	Ginzburg–Landau	measured ($d = 3$)
β	0.5	0.326419(3)
γ	1	1.237075(10)
δ	3	4.78984(1)

One of the most striking features of critical phenomena is *universality*. This means that different materials have the same critical exponents, as long as they are in the same dimension and have the same global symmetries. Consider for example the critical point of a fluid, that is found at the end of the line separating liquid and vapor. The order parameter of this system, whose value defines the phase, (and is analogous to the magnetization) is $\Delta\rho$, the difference in densities between the liquid and vapor. The free energy is an even function of $\Delta\rho$ near the critical point, so the fluid also has a $\mathbb{Z}_2$ symmetry that consists of changing the sign of $\Delta\rho$. Every thermodynamic quantity of the magnetic system has a fluid analog (for example, the analog of the magnetic susceptibility is isothermal compressibility) and it is observed

experimentally that the critical exponents of the two systems are the same, in spite of their very different nature. One says that these critical systems are in the same *universality class*.

The Ginzburg–Landau model explains universality, but it is too coarse: for example it predicts critical exponents that do not depend on the dimension of the system, whereas experimentally one finds that they do. In fact, the predictions of the Ginzburg–Landau models are valid in (space) dimension four or higher, and become progressively worse as the dimension decreases. This is because the Ginzburg–Landau models ignore fluctuations, and fluctuations are more important in lower dimensions.

A better quantitative understanding of critical phenomena came with Wilson's use of the renormalization group [Wil73]. Recall that in the Ginzburg–Landau model the mass is zero at the critical temperature, in such a way that the system becomes scale invariant. In Wilson's description, the critical state is described by an infrared-attractive fixed point of the renormalization group, where scale invariance is realized also in the presence of fluctuations, and where the features of the system become largely independent of the microscopic Hamiltonian: the only properties that are remembered are those that determine the universality class, namely the dimension and the global symmetries. A proper discussion of this would take us too far afield. The interested reader is directed towards the excellent books [Gol92] or [Car96].

There are magnetic systems where the magnetization lies in a plane, or points anywhere in three dimensional space.[4] These systems are described at microscopic level by the so-called XY model and Heisenberg model respectively, but on scales that are much larger than the lattice spacing they have effective descriptions by Ginzburg–Landau models that are just the Euclidean versions of the $O(2)$ and $O(3)$ linear scalar models of the previous section. They differ from the uniaxial ferromagnet in having a continuous symmetry group, so that when symmetry breaking occurs, Goldstone bosons appear. Apart from this, the phase transition in these theories has the same general features described above, but different values of the critical exponents. These models define the $O(2)$ and $O(3)$ universality classes.

1.2.3 The linear sigma model

One of the most important applications of the linear scalar model in particle physics is in the theory of the strong interactions at low energy. The lightest baryons are the proton and neutron, that can be put together into a "nucleon"

[4] The dimension of the space spanned by the spins should not be confused with the dimension of the lattice.

doublet $N = \binom{p}{n}$, transforming in the fundamental (spinor) representation of the isospin group $SU(2)$:

$$\delta_v N = -v^a \tau_a N, \tag{1.34}$$

where v^a is the infinitesimal transformation parameter and $\tau_a = \frac{i}{2}\sigma_a$, where σ_a are the Pauli matrices (see Appendix A for notation and conventions). The lightest mesons are the pions $\pi^\pm$ and π^0, that can be put together into a triplet (π^1, π^2, π^3) (with $\pi^\pm = \pi^1 \pm i\pi^2$), transforming in the adjoint representation of the isospin group $SU(2)$:

$$\begin{aligned} \delta_v \pi^a &= -v^b \, ad(T_b)^a{}_c \pi^c \\ &= -v^b \epsilon_{abc} \pi^c. \end{aligned} \tag{1.35}$$

The Noether current associated to these transformations is

$$j^\mu_{Va} = \bar{N}\gamma^\mu \tau_a N + \epsilon_{abc} \pi^b \partial^\mu \pi^c. \tag{1.36}$$

The isospin, or "vector" current j^μ_{Va} is conserved in the strong interactions, and the masses of the two nucleons and those of the three pions are almost degenerate. Thus, isospin is a symmetry of the strong interactions at low energy.

If the nucleons were massless, their free Lagrangian

$$\mathcal{L} = -\bar{N}\gamma^\mu \partial_\mu N \tag{1.37}$$

would additionally be invariant under the "axial $SU(2)$" transformations

$$\delta_\alpha N = -\alpha^a \tau_a \gamma^A N, \tag{1.38}$$

where α^a is the transformation parameter and $\gamma^A = -i\gamma^0\gamma^1\gamma^2\gamma^3$ is the chirality operator. The associated current is

$$j^\mu_{Aa} = \bar{N}\gamma^\mu \gamma^A \tau_a N, \tag{1.39}$$

where A stands for "axial". Note that axial transformations do not close. However, if we put together the vector and axial currents of the nucleon, they form a closed algebra:

$$[j^0_{Va}(x,t), j^0_{Vb}(y,t)] = i\epsilon_{abc} j^0_{Vc}(x,t)\delta^{(3)}(x-y). \tag{1.40a}$$

$$[j^0_{Va}(x,t), j^0_{Ab}(y,t)] = i\epsilon_{abc} j^0_{Ac}(x,t)\delta^{(3)}(x-y). \tag{1.40b}$$

$$[j^0_{Aa}(x,t), j^0_{Ab}(y,t)] = i\epsilon_{abc} j^0_{Vc}(x,t)\delta^{(3)}(x-y). \tag{1.40c}$$

The vector and axial transformations are entangled and it is convenient to reshuffle them in a different way. Since the chirality operator γ^A satisfies $(\gamma^A)^2 = \mathbf{1}$, the operators

$$P_\pm = \frac{1 \pm \gamma^A}{2} \tag{1.41}$$

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

$$j^\mu_{La} = \frac{j^\mu_{Va} - j^\mu_{Aa}}{2} = \bar{N}\gamma^\mu P_- \tau_a N, \tag{1.42a}$$

$$j^\mu_{Ra} = \frac{j^\mu_{Va} + j^\mu_{Aa}}{2} = \bar{N}\gamma^\mu P_+ \tau_a N, \tag{1.42b}$$

we can rewrite (1.40) as

$$[j^0_{La}(x,t), j^0_{Lb}(y,t)] = i\epsilon_{abc} j^0_{Lc}(x,t)\delta^{(3)}(x-y)\,; \tag{1.43a}$$

$$[j^0_{La}(x,t), j^0_{Rb}(y),t] = 0\,; \tag{1.43b}$$

$$[j^0_{Ra}(x,t), j^0_{Rb}(y,t)] = i\epsilon_{abc} j^0_{Rc}(x,t)\delta^{(3)}(x-y), \tag{1.43c}$$

showing that the global symmetry group is the so-called *chiral group* $SU(2)_L \times SU(2)_R$, whose two factors act separately on the left and right components of the fermions:

$$N_L \to g_L^{-1} N_L, \qquad \bar{N}_L \to \bar{N}_L g_L \tag{1.44a}$$

$$N_R \to g_R^{-1} N_R, \qquad \bar{N}_R \to \bar{N}_R g_R. \tag{1.44b}$$

The axial transformations are badly broken by the mass term of the nucleon $-m_N \bar{N}N$. However, let us focus on their action on the pions. A priori, it is not clear what this action should be, but historically some hints came from the weak decay of the charged pions. In the Fermi theory the weak interactions are of the form

$$\frac{G_F}{\sqrt{2}}(V_{\mu a} - A_{\mu a})(V^\mu_a - A^\mu_a), \tag{1.45}$$

where V and A are suitable vector and axial currents. Because the pion is a pseudoscalar, its decay $\pi^- \to \mu^- + \bar{\nu}_\mu$ must be mediated by an axial current. Since the only vectorial property of a pion is its momentum, the relevant matrix element must be of the form

$$\langle 0|A^\mu_a(x)|\pi_b(p)\rangle = iF_\pi p^\mu \delta_{ab} e^{ip\cdot x} \tag{1.46}$$

where $F_\pi = 92.4\text{MeV}$ is called the pion decay constant. Taking the divergence and then using the Klein–Gordon equation for the pion we obtain that

$$\langle 0|\partial_\mu A_a^\mu(x)|\pi_b(p)\rangle = F_\pi m_\pi^2 \delta_{ab} e^{ip\cdot x}. \tag{1.47}$$

Since the squared mass of the pions is considerably smaller than that of the nucleons, this relation suggests that it may be a reasonable approximation to assume that the axial current is conserved. This is referred to as *partial conservation of the axial current* or PCAC. These relations also suggested that the axial current of the pion could be identified with

$$A_\mu^a = F_\pi \partial_\mu \pi^a. \tag{1.48}$$

The divergence of the nucleon axial current follows from (1.8)

$$\partial_\mu(\bar{N}\gamma^\mu\gamma^A\tau_a N) = -2m_N \bar{N}\gamma^A\tau_a N, \tag{1.49}$$

and being proportional to the nucleon mass, is large. However, given that the nucleon interacts strongly with the pions, it may be more significant to consider the total axial current[5]

$$\bar{N}\gamma^\mu\gamma^A\tau_a N + F_\pi\partial_\mu\pi^a. \tag{1.50}$$

If we demand that this current be exactly conserved, we obtain

$$\Box\pi^a = \frac{2m_N}{F_\pi}\bar{N}\gamma^A\tau_a N, \tag{1.51}$$

which is the Klein–Gordon equation for a massless pseudoscalar triplet coupled to the nucleon. If, as required by the PCAC relation, we demand only that this current be almost conserved, we obtain

$$(\Box - m_\pi^2)\pi^a = \frac{2m_N}{F_\pi}\bar{N}\gamma^A\tau_a N, \tag{1.52}$$

which is the Klein–Gordon equation for a massive pseudoscalar triplet interacting with the nucleons. If the interaction is written conventionally in the form

$$\mathcal{L}_{\pi NN} = -2g_{\pi NN}\pi^a\bar{N}\gamma^A\tau_a N, \tag{1.53}$$

the pion-nucleon coupling is predicted to be

$$g_{\pi NN} = \frac{m_N}{F_\pi}. \tag{1.54}$$

[5]The nucleon axial current has a correction factor $g_A \approx 1.25$ due to the weak interactions, that we ignore here for simplicity. See [Wei95], Section 19.4 for more details.

This is known as the Goldberger–Treiman relation. It relates the mass of the nucleons to the pion-nucleon coupling (both properties of the strong interactions) via the pion decay constant (a property of the weak interactions). It is satisfied with a 5% precision, which is somewhat surprising, and supports the notion that the dynamics of the strong interactions is, within some approximaton, invariant under the chiral group.

On the other hand, chiral invariance would require that for each multiplet of baryons and mesons there should exist another multiplet with the same masses but opposite parity. Since such particles do not exist, the spectrum of the strongly interacting particles does not exhibit chiral symmetry. One concludes that $SU(2)_L \times SU(2)_R$ 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 3 massless particles (Goldstone bosons) with negative parity, zero spin and unit isospin. Since the pions are much lighter than the other strongly interacting particles and have all the right quantum numbers, it is reasonable to identify them with the Goldstone bosons, but since they are not exactly massless, they are usually referred to as *pseudo-Goldstone bosons*.

All this can be understood better from the more modern perspective of QCD, from which, in principle, all the above properties could be derived. There are six known types (or *flavors*) of quarks: u (up), d (down), s (strange), c (charm), b (bottom) and t (top), in order of increasing mass. Each of them is described by a Dirac spinor. We can collect these quark fields into a column vector q_α, where α is an index that runs over the six flavors. The free quark part of the QCD Lagrangian is

$$-\sum_\alpha \bar{q}_\alpha \left(\gamma^\mu \partial_\mu + m_\alpha\right) q_\alpha. \tag{1.55}$$

The masses of the quarks are distributed over a large range, but for many purposes one can pretend that the lightest ones 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 focus on the isospin doublet $q = \binom{u}{d}$. For $m_u = m_d = 0$, the Lagrangian (1.55) is invariant under the chiral group, generated by currents that have the form (1.42), except for the replacement of N by q. By the same argument used above, the chiral symmetry must be broken spontaneously to its vector (isospin) subgroup, and again we end up concluding that the pions can be regarded as the resulting pseudo-Goldstone bosons. In this case we learn that the smallness of the pion masses is related to the smallness of the u and d masses.

In QCD we can also deduce the transformation properties of the pions from those of their constituent quarks. Since the quarks have the same isospin

transformation properties as the nucleons, and the pions π^a have the quantum numbers of $\bar{q}\gamma^A\tau_a q$, from the vector transformation (1.34) we obtain (1.35), whereas from the axial transformation (1.38) we obtain

$$\delta_\alpha(\bar{q}\gamma^A\tau_a q) = \alpha^a \frac{1}{2}\bar{q}q. \tag{1.56}$$

Thus the pions transform into a hypothetical scalar isosinglet meson $\sigma = \frac{1}{2}\bar{q}q$, which would be a singlet under vector transformations. In turn,

$$\delta_\alpha\sigma = \frac{1}{2}\delta_\alpha(\bar{q}q) = -\alpha^b\bar{q}\gamma^A\tau_b q = -\alpha^b\pi^b. \tag{1.57}$$

The spontaneous breaking of the chiral symmetry must occur in QCD via the formation of a condensate $\langle\bar{q}q\rangle$. This is a difficult nonperturbative problem, because the dynamics of QCD at low energy is strongly coupled, so the spontaneous breaking of chiral symmetry is usually described by simple ad hoc models. One of the earliest and most successful ones was proposed in 1960 by Gell–Mann and Levy [GML60]. It describes the interactions of the pions and σ with the heavy nucleons. The group $SU(2)_L \times SU(2)_R$ is locally isomorphic to $SO(4)$ and the unbroken group $SU(2)_V$ is locally isomorphic to $SO(3)$, thus the bosonic part of the model is simply the $SO(4)$ model described in Section 1.2.1. The perturbative description of the pions is by a triplet of pseudoscalar fields π^a in the adjoint representation of $SU(2)_V$, transforming under $SU(2)_L \times SU(2)_R$ by

$$\pi^a\tau_a \longmapsto g_L^{-1}\,(\pi^a\tau_a)\,g_R. \tag{1.58}$$

The model contains another scalar field σ that is a singlet of $SU(2)_V$, such that $(\pi^1, \pi^2, \pi^3, \sigma)$ form a vector of $SO(4)$. The Lagrangian is[6]

$$\begin{aligned}\mathcal{L}_{GML} = &-\frac{1}{2}\partial_\mu\pi^a\partial^\mu\pi^a - \frac{1}{2}\partial_\mu\sigma\partial^\mu\sigma - \bar{N}\gamma^\mu\partial_\mu N \\ &- \frac{\lambda}{4}(\pi^a\pi^a + \sigma^2 - f^2)^2 - g\bar{N}(\sigma + 2\pi^a\tau_a\gamma^A)N.\end{aligned} \tag{1.59}$$

The first line gives the free part of the Lagrangian. The nucleons are a priori massless. The purely bosonic terms are manifestly $SO(4)$ invariant, with a potential that has the standard form leading to symmetry breaking. The pions couple linearly to the axial charge density with a coupling of the form (1.53) and σ couples to the baryon charge density. The invariance of the last term is

[6] It is also instructive to see this Lagrangian written in terms of "spinorial" variables, see Exercise 1.4.

not so evident. For convenience we write here the complete form of the chiral transformations:

$$\delta_v N = -v^a \tau_a N, \qquad \delta_v \bar{N} = \bar{N} v^a \tau_a, \qquad \delta_v \pi^a = \epsilon_{abc} v^b \pi^c, \quad \delta_v \sigma = 0, \tag{1.60a}$$

$$\delta_\alpha N = -\alpha^a \tau_a \gamma^A N, \quad \delta_\alpha \bar{N} = -\bar{N} \gamma^A \alpha^a \tau_a, \quad \delta_\alpha \pi^a = \alpha^a \sigma, \qquad \delta_\alpha \sigma = -\alpha^a \pi^a. \tag{1.60b}$$

The vector transformations correspond to the subgroup $SO(3) \subset SO(4)$ while the axial transformations correspond to the generators T_{m4} with $m = 1, 2, 3$, rotating the pions into σ and σ to the pions. Note that the axial transformations of the bosons correctly reproduce (1.56) and (1.57), that we expect from QCD.

The notation has been chosen so that it is natural to choose the vacuum in the σ direction:

$$\langle \pi^a \rangle = 0, \qquad \langle \sigma \rangle = f.$$

Then expanding $\sigma = f + \chi$ the Lagrangian becomes

$$\begin{aligned} \mathcal{L} = &-\frac{1}{2}\partial_\mu \pi^a \partial^\mu \pi^a - \frac{1}{2}\partial_\mu \chi \partial^\mu \chi - m^2 \chi^2 - \bar{N}(\gamma^\mu \partial_\mu + gf)N \\ &- \frac{\lambda}{4}(\pi^a \pi^a + \chi^2)^2 - \lambda f \chi (\pi^a \pi^a + \chi^2) - g\bar{N}(\chi + 2\pi^a \tau_a \gamma^A)N\,. \end{aligned} \tag{1.61}$$

Only the $SU(2)_V$ symmetry is still manifest. We see that the pions are exactly massless, the χ field has squared mass $2m^2$, as usual for the radial mode in the $O(N)$ models, and the fermions have acquired mass $m_N = gf$ from the Yukawa interaction with σ. The Goldberger–Treiman relation (1.54) can be satisfied by simply choosing $f = F_\pi$. The bosonic Noether axial current is

$$-\partial_\mu \pi^a \sigma + \pi^a \partial_\mu \sigma\,. \tag{1.62}$$

whose vacuum expectation value reproduces the PCAC relation (1.48).

We can make the model a bit more realistic by introducing a new term in the Lagrangian

$$\Delta\mathcal{L} = a\sigma, \tag{1.63}$$

that breaks the chiral symmetry and leaves the vector subgroup unbroken, exactly like the VEV of σ. The difference is that $\Delta\mathcal{L}$ is an *explicit* breaking because it spoils the symmetry of the Lagrangian, whereas the VEV does not.

For small a the VEV of σ gets shifted to

$$\sigma_* = \langle \sigma \rangle = f + \frac{a}{2\lambda f^2} + O(a^2) \tag{1.64}$$

We can demand that this is still equal to F_π, so as to preserve the Goldberger–Treiman relation, by choosing

$$f = F_\pi - \frac{a}{2\lambda f^2}. \tag{1.65}$$

The σ mass gets shifted slightly, and the pion now becomes massive, with

$$m_\pi^2 = \frac{\partial V}{\partial \pi^2}\bigg|_{\sigma=F_\pi} = \frac{a}{F_\pi}, \tag{1.66}$$

which fixes $a = F_\pi m_\pi^2$.

1.3 Nonlinear scalar theories

1.3.1 From linear to nonlinear theories

Let us return to the general $O(N)$ models of Section 1.2.1. In the broken phase its excitations consist of $N-1$ massless Goldstone bosons and one massive radial mode with mass $2m$. Suppose we probe the system at energies much below $2m$. Then, we will not be able to excite the radial mode and all we will see are the massless Goldstone bosons. In such a situation, we can dispense with the radial mode altogether and remain with a theory that describes only the dynamics of the Goldstone bosons. Formally, this can be achieved by taking the limit $\lambda \to \infty$ with f kept constant, in which case the potential becomes sharply peaked around 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 in Figure 4.

It is not very elegant to have a potential that diverges almost everywhere. A mathematically more sensible way of studying the limit is to introduce an auxiliary field Λ and consider the action

$$S = \int d^n x \Big[-\frac{1}{2}\partial_\mu \phi^a \partial^\mu \phi^a - \frac{2\Lambda}{\sqrt{\lambda}}\sqrt{V} + \frac{\Lambda^2}{\lambda} \Big]. \tag{1.67}$$

The equation of motion for Λ is $\Lambda = \sqrt{\lambda V}$ and when this equation is used in (1.67) it gives back (1.21), with the potential (1.25). Thus (1.67) is classically equivalent to (1.21). The advantage of the action (1.67) is that it remains well defined in the limit $\lambda \to \infty$. In fact, it reduces to

$$S = \int d^n x \Big[-\frac{1}{2}\partial_\mu \phi^a \partial^\mu \phi^a - \Lambda(|\phi|^2 - f^2) \Big]. \tag{1.68}$$

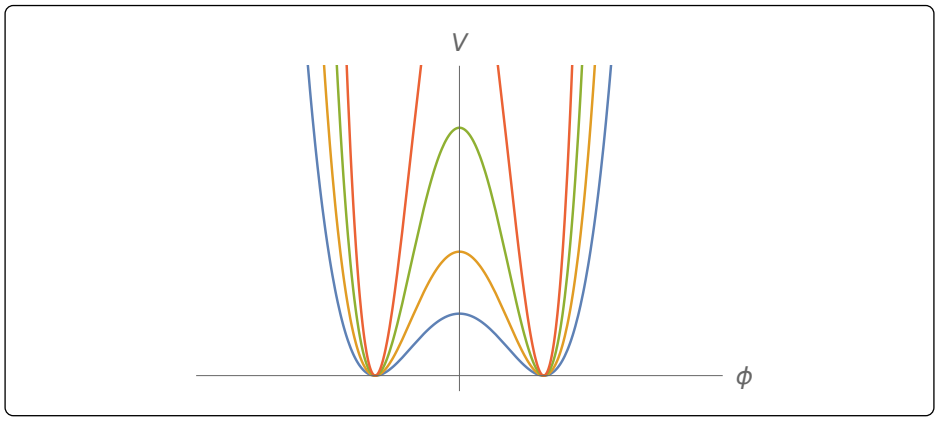


Figure 4. The potential with increasing λ.

The auxiliary field Λ has become a Lagrange multiplier enforcing the constraint $\phi^2 = f^2$. This is called a *nonlinear sigma model with values in* S^{N-1}, or a $O(N)$*-nonlinear sigma model*. Since the action of $O(N)$ on the Goldstone bosons is nonlinear, one says that in this model $O(N)$ is *nonlinearly realized*, but the unbroken group $O(N-1)$ is still linearly realized on the tangent space of the sphere at the north pole.

This procedure can be generalized to linear scalar theories that are invariant under any Lie group G, whose potential has minima on a coset space G/H. Then, taking the strong coupling limit will give a nonlinear sigma model with values in G/H. In the physical literature these are often described as nonlinear realizations of G, linear for H.

1.3.2 Geometric formulation

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 $O(3)$ model. We can solve the constraint $\phi^a\phi^a = f^2$ expressing the three fields ϕ^a in terms of only two independent fields φ^α. There are infinitely many ways of doing this, that correspond to choosing coordinates on the sphere, see Exercise 1.5. For example we could choose φ^α to be the spherical coordinates ($\varphi^1 = \Theta$, $\varphi^2 = \Phi$):

$$\phi^1 = f \sin\Theta \cos\Phi, \tag{1.69a}$$
$$\phi^2 = f \sin\Theta \sin\Phi, \tag{1.69b}$$
$$\phi^3 = f \cos\Theta. \tag{1.69c}$$

Introducing into (1.68), we find the action

$$S = -\frac{f^2}{2}\int d^n x\,(\partial_\mu\Theta\partial^\mu\Theta + \sin^2\Theta\partial_\mu\Phi\partial^\mu\Phi). \tag{1.70}$$

Another way to proceed is to solve the constraint $|\phi|^2 = f^2$ by

$$\phi^3 = \sqrt{(\phi^1)^2 + (\phi^2)^2} \tag{1.71}$$

and use $\varphi^\alpha = \phi^\alpha/f$, $\alpha = 1, 2$, as coordinates. (This coordinate system covers only one hemisphere.) In this case the action reads

$$S = -\frac{f^2}{2}\int d^n x\left(\delta_{\alpha\beta} + \frac{\varphi_\alpha\varphi_\beta}{1-\varphi_1^2-\varphi_2^2}\right)\partial_\mu\varphi^\alpha\partial^\mu\varphi^\beta. \tag{1.72}$$

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

$$\phi^1 = f\frac{4\omega_1}{\omega_1^2+\omega_2^2+4}, \tag{1.73a}$$

$$\phi^2 = f\frac{4\omega_2}{\omega_1^2+\omega_2^2+4}, \tag{1.73b}$$

$$\phi^3 = f\frac{\omega_1^2+\omega_2^2-4}{\omega_1^2+\omega_2^2+4}. \tag{1.73c}$$

Introducing in (1.68),

$$S = -\frac{f^2}{2}\int d^n x\,\frac{\partial_\mu\omega_1\partial^\mu\omega_1 + \partial_\mu\omega_2\partial^\mu\omega_2}{\left(1+\frac{\omega_1^2}{4}+\frac{\omega_2^2}{4}\right)^2}. \tag{1.74}$$

We see that in each case the action has the form

$$S = -\frac{f^2}{2}\int d^n x\,\partial_\mu\varphi^\alpha\partial^\mu\varphi^\beta h_{\alpha\beta}(\varphi), \tag{1.75}$$

where $h_{\alpha\beta}(\varphi)$ is the standard metric on the sphere S^2 of unit radius, written in the chosen coordinate system. The following features have to be noted: since fields have the meaning of coordinates of S^2 and the metric is a nonpolynomial function of the coordinates, the fields have to be dimensionless. In dimensions $n > 2$ the prefactor f is needed for dimensional reasons, to compensate the noncanonical dimension of the fields. Even though the action looks very much

like a kinetic term, it is really a closed package containing both kinetic term and interactions, as we shall discuss in Section 1.4.1. This is another reason for the appearance of the constant f, that in the linear theory was closely related to the mass, but here plays the role of coupling constant. Finally, the form of the action (1.75) is invariant under coordinate transformations, but the metric is not, as is evident from the preceding examples. Coordinate transformations change the form of the interactions and therefore are not symmetries of the theory. The only symmetries are those transformations that leave the form of the metric invariant, i.e. the isometries of $h_{\alpha\beta}$.

We have derived the action for a nonlinear $O(N)$ model but it is now obvious that one can generalize the models by considering a completely arbitrary target manifold N, endowed with a metric $h_{\alpha\beta}$. This degree of generality is relevant for example in string theory, where the nonlinear sigma model describes the dynamics of the worldsheet, and the target space is reinterpreted as spacetime. However, we will not need to consider such models here: for us the target space will always be a coset space, and the nonlinear sigma model will always be a theory of Goldstone bosons.

In this intrinsic formulation where the fields are coordinates on the target space manifold, one has to use tools of differential geometry that should be familiar to anyone who has studied General Relativity. For example, in the case of a G/H model, G-invariance of the action can be proven as follows. Let us first consider a general variation of the field. We have

$$\delta S = -\frac{f^2}{2}\int d^n x \left[2\partial_\mu \delta\varphi^\alpha \partial^\mu \varphi^\beta h_{\alpha\beta} + \partial_\mu \varphi^\alpha \partial^\mu \varphi^\beta \partial_\gamma h_{\alpha\beta}\delta\varphi^\gamma\right]. \tag{1.76}$$

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

$$K_a^\gamma \partial_\gamma h_{\alpha\beta} + h_{\alpha\gamma}\partial_\beta K_a^\gamma + h_{\beta\gamma}\partial_\alpha K_a^\gamma = 0. \tag{1.77}$$

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

$$\partial_\mu \partial^\mu \varphi^\gamma + \partial_\mu \varphi^\alpha \partial^\mu \varphi^\beta \Gamma^\gamma_{\alpha\beta}(\varphi) = 0, \tag{1.78}$$

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

The manifest nonpolynomiality is the price we pay in order to minimize the number of fields.

1.3.3 The nonlinear chiral models

Tvectohe main drawback of the linear sigma model of Section 1.2.3 is that the σ field cannot be identified with any one of the existing mesons. For this reason we can send its mass to infinity by taking the $\lambda \to \infty$ limit with f fixed and we remain with the so-called *chiral nonlinear sigma model*, that describes just the pions and their interactions with the nucleons.[7] Instead of trying to rewrite the linear sigma model action in this limit, we will just take the basic idea that the model has to describe the low energy dynamics of Goldstone bosons with values in the coset

$$\frac{SU(2)_L \times SU(2)_R}{SU(2)_V}$$

and try to write a general Lagrangian consistent with these symmetries. We note that, even though it lacks a group structure, the coset is diffeomorphic to the group $SU(2)$. The identification comes by picking an arbitrary point in the coset, namely the point that we shall take as the vacuum state, and identifying it with the identity of the group. It is then natural to parametrize the Goldstone bosons by a $SU(2)$-valued field U, and the action of the chiral group on this field is

$$U \to g_L^{-1}\, U\, g_R, \tag{1.79}$$

in such a way that the field $U = \mathbb{I}$ (that we identify as the vacuum state) is invariant under the vector transformations $g_L = g_R$.

This is supposed to be a description of strong interactions at low energy, and at low energy 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

$$S = \frac{F_\pi^2}{4} \int d^4x\, \mathrm{tr}\big(U^{-1}\partial_\mu U U^{-1}\partial^\mu U\big). \tag{1.80}$$

The invariance of this action under (1.79) is immediately clear. To relate this to the geometric description of the preceding section, given a coordinate system on $SU(2)$, we call $\varphi^\alpha(x)$ the coordinates of the group element $U(x)$ and we decompose

$$U^{-1}\partial_\mu U = \partial_\mu \varphi^\alpha L^a_\alpha(\varphi)\tau_a, \tag{1.81}$$

where L^a_α are the components of the Maurer–Cartan form on $SU(2)$, given explicitly in Appendix D. The basis in the Lie algebra, in the fundamental representation, is given by $\tau_a = \frac{i}{2}\sigma_a$, where σ^a are the Pauli matrices, and

[7] In the case $N = 3$, when also the s-quark is treated as if it were massless, the model describes the dynamics of the octet of mesons, including the pions, the kaons and the η meson.

we have $\mathrm{tr}\,\tau_a\tau_b = -\frac{1}{2}\delta_{ab}$. Then we define a Riemannian metric on $SU(2)$ by declaring one half of the Maurer–Cartan forms to be an orthonormal field of co-frames:

$$h_{\alpha\beta} = \frac{1}{4}L^a_\alpha L^b_\beta \delta_{ab}. \tag{1.82}$$

In this way we see that the action (1.80) is identical to the nonlinear sigma model action (1.75). The advantage of the form (1.80) is that it makes the chiral invariance of the theory very transparent. When the action is written in the form (1.75), its invariance is less evident. It follows from the fact that the metric $h_{\alpha\beta}$ is both left- and right-invariant, which can be proven by showing that the vector fields with components L^α_a and R^α_a, that generate right- and left-multiplications, respectively, are Killing vectors for h.

In order to recover the perturbative description in terms of pion fields it is useful to adopt normal coordinates π^α, defined by:

$$U(x) = e^{2\pi^a(x)\tau_a/F_\pi} = \mathbf{1} + 2\pi^a(x)\tau_a/F_\pi + \dots . \tag{1.83}$$

Note that the coordinates have been scaled so as to have the canonical dimension of mass. The action (1.80) has an expansion

$$\int d^4x \left[-\frac{1}{2}\partial_\mu\pi^a\partial^\mu\pi^a + \frac{1}{6F_\pi^2}\left(\pi^a\pi^a\partial_\mu\pi^b\partial^\mu\pi^b - \pi^a\partial_\mu\pi^a\pi^b\partial^\mu\pi^b\right) + \dots\right], \tag{1.84}$$

where the leading term is a canonically normalized kinetic term, that is followed by infinitely many interaction terms. One observes again that in this model the pions are massless and all interactions contain derivatives of the fields: this is as it should be, since a potential for π would certainly break the global invariance of the theory.

In the linear model the mass of the nucleons is generated when the VEV of the field σ is inserted in the Yukawa coupling. Now that σ is frozen to its VEV, this mechanism is less visible, but we can still recover the Goldberger–Treiman relation in the following way. The nucleon mass term

$$-m_N\bar{N}N = -m_N(\bar{N}_L N_R + \bar{N}_R N_L) \tag{1.85}$$

is invariant under isospin transformations, namely those transformations (1.44) with $g_L = g_R$. We can rewrite it in a way that is invariant under the full set of chiral transformations as follows:

$$-m_N(\bar{N}_L U N_R + \bar{N}_R U^{-1} N_L). \tag{1.86}$$

The leading term in the expansion (1.83) then reproduces the mass term, and the term linear in π generates a pion-nucleon interaction

$$-\frac{m_N}{F_\pi}\pi^a \bar{N}\tau_a\gamma^A N \tag{1.87}$$

that automatically satisfies the Goldberger–Treiman relation.

1.3.4 Sigma models with gauge invariance

We have discussed two types of formulations of nonlinear scalar theories: a geometric formulation where the fields are coordinates in the target space, and another where the target space is embedded in a linear space and the linear fields are subjected to some constraints. There is a third way that can be used when the target space N is the quotient of another space P by the action of a group K.[8] Then, one can write an action for fields that are coordinates on P and if the action is invariant under local K transformations, only the coordinates of N are physical degrees of freedom. So this is another equivalent way of describing the dynamics of an N-valued model. One should not confuse this with the coupling of the nonlinear sigma model to independent gauge fields, that we shall discuss in Section 1.5.3.

A large subclass of models of this type have values in a Lie group G and are invariant under local transformations of a subgroup H, and thus are equivalent to a G/H-valued model. In this form, these models have been studied in two dimensions, where, in many cases, they are integrable dynamical systems [DLD78, Eic79]. In four dimensions, they are known as phenomenological Lagrangians for nonlinear realizations [CWZ69, CCWZ69].

Let $g(x)$ be a function with values in G. We define a composite gauge potential to be the pullback of the component of the left-invariant Maurer–Cartan form in the subalgebra $\mathfrak{h}$ of H:

$$B_\mu = g^{-1}\partial_\mu g\Big|_{\mathfrak{h}}. \tag{1.88}$$

Consider a local gauge transformation

$$g'(x) = g(x)h(x). \tag{1.89}$$

Under such a transformation, B_μ transforms as a gauge potential

$$B'_\mu = h^{-1}B_\mu h + h^{-1}\partial_\mu h. \tag{1.90}$$

[8] Thus P is a principal bundle over N, see Appendix C.

We can therefore define a covariant derivative

$$D_\mu g = \partial_\mu g - g B_\mu \tag{1.91}$$

that transforms in the same way as g:

$$D_\mu g' = h^{-1} D_\mu g. \tag{1.92}$$

As in the chiral models, we can then write an action

$$S = \frac{f^2}{4} \int d^n x \, \mathrm{tr}(g^{-1} D_\mu g \, g^{-1} D^\mu g). \tag{1.93}$$

This action is manifestly invariant under the global left action of G and under local transformations of H acting on the right.

A closely related class of models where the space P is not a group, are the $\mathbb{CP}^N$ models. Recall that the real projective space $\mathbb{RP}^{N-1}$ is the space of lines through the origin in $\mathbb{R}^N$. Since every line crosses the unit sphere in exactly two points, it can also be thought of as $S^{N-1}/\mathbb{Z}_2$. Thinking of the sphere as the coset $O(N)/O(N-1)$ and $\mathbb{Z}_2 = O(1)$,

$$\mathbb{RP}^{N-1} = \frac{O(N)}{O(N-1) \times O(1)}.$$

In a similar way one defines the complex projective space $\mathbb{CP}^{N-1}$ to be the space of complex lines in $\mathbb{C}^N$. By a similar reasoning as above, it is the quotient

$$\mathbb{CP}^{N-1} = \frac{U(N)}{U(N-1) \times U(1)} = S^{2N-1}/U(1),$$

where we used the fact that odd-dimensional spheres are cosets of unitary groups. It is a complex manifold of complex dimension $N-1$. A special case is $\mathbb{CP}^1 = S^2$. We are in the situation described in the beginning of this section, with $P = S^{2N-1}$ and $K = U(1)$. One could describe the model with geometric S^{2N-1}-valued fields φ^α, but it is more common in the literature to embed S^{2N-1} in $\mathbb{C}^N$ and then take the quotient by the action of $U(1)$. Thus, let $z = (z_1, \ldots, z_N) \in \mathbb{C}^N$, where z_i are complex scalar fields. We impose the constraint

$$z^\dagger z = 1, \tag{1.94}$$

so z defines a map into S^{2N-1}. The group $U(1)$ acts on z^a multiplying all fields by the same phase. The group $U(N)$ also acts on the multiplet z^a in the standard way. Define the covariant derivative

$$D_\mu z = \partial_\mu z + i B_\mu z \tag{1.95}$$

where B_μ is an auxiliary $U(1)$ gauge field. Consider the action

$$S(z,B) = -\frac{f^2}{2}\int d^n x\, D_\mu z^\dagger D^\mu z. \tag{1.96}$$

The equation of motion for B says that

$$B_\mu = i z^\dagger \partial_\mu z. \tag{1.97}$$

Inserting back this solution, the action becomes that of a nonlinear sigma model with hermitian metric (the barred index multiplies a dz^*)

$$g_{m\bar{n}} = \delta_{m\bar{n}} - z^*_m z_{\bar{n}}. \tag{1.98}$$

In its most general form, the construction of nonlinear sigma models with gauge symmetry works as follows. We start with coordinates y^α on N, $\tilde{y}^{\tilde{\alpha}}$ on P and a basis T_a in the Lie algebra $\mathfrak{k}$ of K. Let $F_a^{\tilde{\alpha}}$ be a set of vectorfields satisfying the Lie algebra of K and generating the right action of K on P. They are called the fundamental vectorfields. Let ω be a one-form on P with values in $\mathfrak{k}$, transforming in the adjoint representation, in the sense that $\omega(p\cdot k) = Ad(k^{-1})\omega(p)$, and such that $\omega(F_a) = T_a$. The space spanned by the fundamental vectorfields at a point p is called the vertical space at p. The kernel of $\omega(p)$ is the complement of the vertical space, and is called the horizontal space at p.

Let $\varphi^\alpha(x)$ be a field with values in N and let $\tilde{\varphi}^{\tilde{\alpha}}$ be a field with values in P such that $p\circ\tilde{\varphi} = \varphi$, where $p : P \to N$ is the natural projection. We say that $\tilde{\varphi}$ is a *lift* of φ. The group K acts on the lifted field by

$$\delta\tilde{\varphi}^{\tilde{\alpha}} = \epsilon^a F_a^{\tilde{\alpha}}. \tag{1.99}$$

We define the covariant derivative of the lifted field by

$$D_\mu\tilde{\varphi}^{\tilde{\alpha}} = \partial_\mu\tilde{\varphi}^{\tilde{\alpha}} - B^a_\mu F_a^{\tilde{\alpha}}, \tag{1.100}$$

where

$$B^a_\mu = \partial_\mu\tilde{\varphi}^{\tilde{\beta}}\omega^a_{\tilde{\beta}}. \tag{1.101}$$

Next, define a metric $\tilde{h}_{\tilde{\alpha}\tilde{\beta}}$ in P as follows.[9] Assume that the basis T_a is orthonormal, so that it defines an invariant metric in K. The inner product in the vertical spaces is given by the metric in K, the one in the horizontal

[9]This construction is used in Kaluza–Klein theories, where P is interpreted as a higher dimensional spacetime and N is four-dimensional spacetime.

spaces is given by the metric in N and the vertical and horizontal spaces are orthogonal. We can now write the action for the lifted fields:

$$S(\tilde{\varphi}) = -\frac{f^2}{2}\int d^n x\, D_\mu \tilde{\varphi}^{\tilde{\alpha}} D^\mu \tilde{\varphi}^{\tilde{\beta}} \tilde{h}_{\tilde{\alpha}\tilde{\beta}}. \tag{1.102}$$

In order to see that this is equivalent to the action of the N-valued field, consider a coordinate system on P that is adapted to its bundle structure, in the sense that a subset of coordinates y^α is in one-to-one correspondence with coordinates on N and is constant on each orbit of K, while another subset $\hat{y}^{\hat{\alpha}}$ are coordinates in the orbits. Then we have $F^\alpha_a = 0$, $\omega^a_\beta = 0$ and $\omega^a_{\hat{\beta}} F^{\hat{\alpha}}_a = \delta^{\hat{\alpha}}_{\hat{\beta}}$. Thus

$$D_\mu \tilde{\varphi}^\alpha = \partial_\mu \tilde{\varphi}^\alpha - \partial_\mu \tilde{\varphi}^\beta \omega^a_\beta F^\alpha_a - \partial_\mu \tilde{\varphi}^{\hat{\beta}} \omega^a_{\hat{\beta}} F^\alpha_a = \partial_\mu \varphi^\alpha, \tag{1.103}$$

$$D_\mu \tilde{\varphi}^{\hat{\alpha}} = \partial_\mu \tilde{\varphi}^{\hat{\alpha}} - \partial_\mu \tilde{\varphi}^\beta \omega^a_\beta F^{\hat{\alpha}}_a - \partial_\mu \tilde{\varphi}^{\hat{\beta}} \omega^a_{\hat{\beta}} F^{\hat{\alpha}}_a = 0. \tag{1.104}$$

In these coordinates (1.102) reduces to (1.75).

1.4 Fundamental vs. effective field theories

In the application of quantum field theory to particle physics, the criterion of renormalizability has played historically an important role. Some of the models that we discussed in this chapter are renormalizable in four dimensions, and some are not. The linear scalar theories with quartic potential, the scalar-fermion systems with Yukawa couplings (such as the linear sigma model) and Yang–Mills theories are renormalizable. The proof of renormalizability of gauge theories in the Higgs phase by 't Hooft was one of the main reasons that led to widespread acceptance of the Weinberg–Salam model, even before the direct detection of the W and Z bosons. On the other hand, the nonlinear sigma models are non-renormalizable. For a long time it was believed that such models could not make sense as quantum field theories, and their relative success in describing low energy physics was a bit of a mystery. It eventually emerged that also these theories can be used as quantum field theories, as long as they are applied only in a finite energy range. We will now see briefly how the nonlinear sigma models can be treated in perturbation theory, and then how they constitute paradigmatic examples of *effective field theories* (EFTs).

1.4.1 Power counting in nonlinear sigma models

We consider a generic nonlinear sigma model in the geometric formulation, and define $f = 1/g$. Since the metric $h_{\alpha\beta}$ is in general a nonpolynomial

function, the fields have to be dimensionless. Therefore the constant g^2 must have dimension L^{n-2}, where n is the dimension of spacetime. In two spacetime dimensions, and only in two, we can choose $g^2 = 1$. In order to give the scalar fields their canonical dimension, we first absorb the constant g^2 in the fields, defining $\tilde{\varphi}^\alpha = \varphi^\alpha / g$. The dimension of $\tilde{\varphi}$ is then $[\tilde{\varphi}^\alpha] = L^{\frac{2-n}{2}}$ and the action reads

$$S = -\frac{1}{2} \int d^n x \, \partial_\mu \tilde{\varphi}^\alpha \partial^\mu \tilde{\varphi}^\beta h_{\alpha\beta}(g\tilde{\varphi}). \tag{1.105}$$

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

$$h_{\alpha\beta}(g\tilde{\varphi}) = h_{\alpha\beta}(g\tilde{\varphi}_0) + g\partial_\gamma h_{\alpha\beta}(g\tilde{\varphi}_0)\eta^\gamma + \frac{1}{2} g^2 \partial_\gamma \partial_\delta h_{\alpha\beta}(g\tilde{\varphi}_0)\eta^\gamma \eta^\delta + \cdots \tag{1.106}$$

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

The dimension of the coupling constant in the m-th term, i.e. the coefficient of $\partial\eta \, \partial\eta \, \eta^m$, is $[g^m] = L^{\frac{m}{2}(n-2)}$. In spite of the infinite number of couplings, this theory is renormalizable in a generalized sense for $n = 2$ [Fri80] and non-renormalizable for $n > 2$. We note that in some cases, such as the sphere, the metric is entirely determined, up to an overall scale, by symmetry requirements. In these cases there is really only one independent coupling constant g: all the coefficients of the expansion of the metric are determined by the requirement of G-invariance.

The non-renormalizability of the theory means that if we compute a one loop diagram with the Lagrangian (1.105) we will have to introduce counterterms of the form

$$T_{\alpha\beta\gamma\delta} \partial_\mu \varphi^\alpha \partial^\mu \varphi^\beta \partial_\nu \varphi^\gamma \partial^\nu \varphi^\delta,$$

and at higher loops also terms with still higher derivatives. It seems that such a theory is completely out of control.

1.4.2 Chiral Perturbation Theory

The key to managing theories of this type is to use them only in a low energy/momentum expansion, that is equivalent to the expansion in derivatives

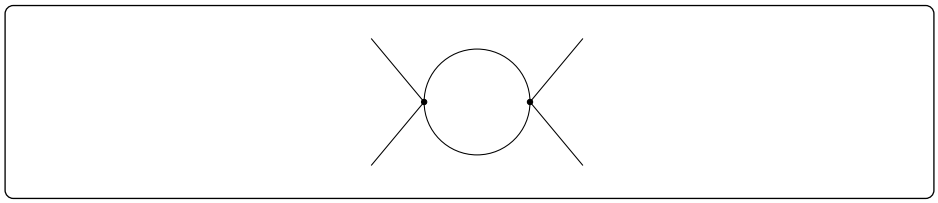


Figure 5. One loop contribution from $\mathcal{L}_2$.

of the Lagrangian. Let us see this in the paradigmatic example of the chiral nonlinear sigma model, in which case the expansion is known as *chiral perturbation theory* (χPT). The first two terms of the derivative expansion of the chiral Lagrangian are

$$S = \int d^4x \left[\mathcal{L}_2 + \mathcal{L}_4 + O(\partial^6)\right] \tag{1.107a}$$

$$\mathcal{L}_2 = F_\pi^2 \mathrm{tr}(U^{-1}\partial U)^2 \tag{1.107b}$$

$$\mathcal{L}_4 = \ell_1 \mathrm{tr}(((U^{-1}\partial U)^2)^2) + \ell_2 (\mathrm{tr}(U^{-1}\partial U)^2)^2. \tag{1.107c}$$

This theory describes well the dynamics of pions at low energy, but at high energy it is superseded by QCD. Even if we didn't know QCD we could still make a good guess of the scale at which the chiral model must break down (what is often called the "scale of new physics"): by dimensional analysis it must be related to the pion decay constant. A more accurate diagnostic for the breakdown of pion theory is the violation of unitarity by the tree level scattering cross section, occurring at the scale $M = 16\pi F_\pi$ which is of the order of the GeV.

Let us make some rough estimates for the contribution of various terms in (1.107) to a $2\pi \to 2\pi$ scattering process. A crucial point is that all interaction terms contain derivatives. Assuming that all the momenta of the external particles are of order p, $\mathcal{L}_2$ will give at tree level a contribution of order

$$g^2 p^2 \approx (p/M)^2$$

while $\mathcal{L}_4$ gives a contribution of order

$$\ell g^4 p^4 \approx \ell (p/M)^4$$

which is evidently subleading at low momenta.

Now we may try to estimate the effect of the diagram in Figure 5, constructed with vertices taken from $\mathcal{L}_2$. The integrand is of the form $g^4 \int d^4q F(q,p)$ where $F(q,p)$ is a fraction involving combinations of q or p to fourth power

in the numerator (coming from the vertices) and combinations of q or p to fourth power in the denominator (coming from the propagators). It is at most quartically divergent. When the integral is regulated, for example by means of dimensional regularization, it leaves behind something that for dimensional reasons can only involve p^4. Thus the diagram gives a contribution to the process of order $g^4 p^4 = (p/M)^4$. The important point here is that loop effects involving the leading term of the expansion are of the same order as tree level effects involving the subleading term.

It turns out that this is all one needs for low-energy meson physics. Terms with more derivatives, or higher loop effects involving $\mathcal{L}_2$ and $\mathcal{L}_4$, would give effects that are unmeasurably small for the current experiments. Thus we need to know only the parameters F_π, ℓ_1, ℓ_2 and a bunch of others that are related to the quark masses. Calculations at one loop in F_π and at tree level in ℓ_1, ℓ_2 successfully describe a rich phenomenology [GaL83].

The same formalism can be applied to electroweak physics, see section 1.5.5. Before the discovery of the Higgs particle in 2012, the Higgs sector of the SM could be described by a Lagrangian of the form (1.107), with suitable couplings to the gauge fields and fermions [ApB80, Lon80]. The reason is that the Higgs doublet can be parametrized by four real fields, and suppressing the hitherto unobserved radial mode leaves one with three scalars parametrizing a three-sphere. The three-sphere is both topologically and geometrically equivalent to $SU(2)$. These three degrees of freedom are the electroweak Goldstone bosons, which, via the Higgs mechanism, manifest themselves as the longitudinal components of the W and Z bosons. The existence of these degrees of freedom had been known since the discovery of the W and Z in 1983. The main difference between the electroweak chiral model and the QCD one is the value of the coupling F_π, that in the electroweak case is replaced by the Higgs VEV, $v \approx 246\text{GeV}$.

A still simpler low energy description of the weak interactions is Fermi's non-renormalizable current-current interaction, that comes with a coupling $G_F = 1.16 \times 10^{-5}\text{GeV}^{-2}$. In the Weinberg–Salam model there is no four-fermion interaction but there is a renormalizable interaction between the fermions and gauge bosons. In this case the heavy state is the W, and integrating it out one is left with the current-current interaction, where the Fermi constant is now related to the mass of the W by $G_F = \sqrt{2} g^2 / 8 m_W^2$, where g is the gauge coupling. As long as the fermion momenta are much smaller than 80GeV, the Fermi theory is a good description.

1.4.3 The Effective Field Theory paradigm

Chiral perturbation theory and Fermi theory are prototypical examples of a general philosophy that has deeply changed the way we think about quantum field theories and now pervades the field of particle physics. It is based on the so-called decoupling theorem [ApC74], stating that the effect of heavy particles, say with mass M, on the dynamics of lighter ones, is to contribute to the effective action of the latter with generally non-renormalizable terms that are suppressed by inverse powers of M. These non-renormalizable interactions are not problematic, because this effective action should only be used at energy below M, so that M acts in practice as an UV cutoff for the effective theory of the light states.

The contribution of the non-renormalizable terms to a low-energy scattering process of the light particles is suppressed, relative to the contribution of renormalizable interactions, by powers of E/M. This suggests using the ratio E/M as an expansion parameter. At sufficiently low energy, it will be enough to keep the leading term in this ratio, when one approaches M more terms in the expansion will be needed and above M one has to consider the full theory involving also the massive fields. This may be a renormalizable theory, or perhaps it will be another effective theory where still heavier states have been integrated out. There are thus two complementary ways to learn more about the physics at high energies: one is simply to "go there" by increasing the energy of the accelerators and the other is to increase the precision, in order to measure the coefficients of the non-renormalizable terms. These considerations are the basis of the *Effective Field Theory* (EFT) approach.

Suppose for example that we have to compute some cross section for a process involving only the light particles that will be measured in a new accelerator. As in the examples mentioned above, the QFT describing the light particles contains a hint of the "scale of new physics" via some large mass scale M that appears in its non-renormalizable interactions. The informations we need about the experiment are the energy of the beam, E, and the precision of the apparatus. Since $E \ll M$, we use the small ratio E/M as an expansion parameter. For example, if $E = M/10$ and the cross-section is going to be measured with a 1% precision, we will need to compute the cross section in the EFT at order $(E/M)^2$. Power-counting arguments show that at any finite order in E/M there will be only a finite number of terms contributing to the process. Generalizing what we saw above for chiral perturbation theory, a systematic analysis [Wei78] shows that at order n in $(E/M)^2$ one must take into account diagrams with $n-1$ loops constructed from L_2, $n-2$ loops constructed from L_4, down to tree diagrams from L_n.

If the underlying fundamental theory is known, one may try to calculate the couplings of the EFT from first principles. (See Exercise 1.3 for a simple case where this works.) It is more frequently the case that either the fundamental theory is unknown or if it is known, this calculation proves too hard. In these cases, the coefficients of these terms can be measured by a finite number of experiments and these values can then be used in the formula for the cross-section. The theoretical prediction for the cross section can be compared to the result of the experiment. The cross section is only measured at finitely many data points, but it is clear that in principle there can be many more data points than undetermined coefficients. In this way even a non-renormalizable EFT can be predictive.

The EFT logic puts the notion of renormalizability in a different perspective. For example, in comparing the linear and the nonlinear sigma model, one may be tempted to think that the former is preferable because it is renormalizable, but applying the EFT logic we see that the latter is equally useful in practice, and insofar as it does not contain the spurious σ meson it better reflects the physics of strong interactions at low energy. Conversely, the fact that a theory is renormalizable does not guarantee that it is really fundamental. Indeed, if the separation between the scale of the observations and the UV scale M is very large, all the non-renormalizable terms will be very tiny and may escape observation. This may very well be the case in the Standard Model.

All these considerations concerning the applicability of quantum field theory to particle physics have a parallel in condensed matter physics, as we have already seen to some extent. The main difference is that in condensed matter physics one never tries to make sense of a theory up to infinite energy, because these phenomena have a clear lenght scale below which the very notion of field makes no sense. This is generically the size of molecules, but in particular in solids, the lattice spacing provides a natural UV cutoff for all momenta. In spite of this, also in condensed matter physics one makes a clear distinction between *microscopic* theories, describing the interactions of the molecular degrees of freedom at the lattice scale, and effective theories that apply at mesoscopic or macroscopic scales. Examples of the former are the Ising model and BCS theory, while examples of the latter are the various Ginzburg–Landau theories. Readers are referred to [Wei96, Wei09] and [Gol23] for further thoughts on these issues.

1.5 Gauge theories

1.5.1 Yang–Mills theories

We limit ourselves here to a *local* description of gauge fields interacting with matter, i.e. a description that is valid in some open neighborhood of a point. A more geometrical description, that is valid globally on a manifold with arbitrary topology, would use the language of fiber bundles. Although useful in some cases (such as the monopole of Section 4.1), it is mostly not necessary and we shall try to avoid it as much as possible.

A Yang–Mills (YM) field for a Lie group G is a one-form with values in the Lie algebra $\mathfrak{g}$ of G:

$$A = A^a_\mu \, dx^\mu \otimes T_a.$$

It is sometimes convenient to exhibit or hide some of the indices, for example $A_\mu = A^a_\mu T_a$ or $A^a = A^a_\mu dx^\mu$. We refer to Appendix B for notation and conventions regarding Lie groups and Lie algebras.

Let V be a vectorspace carrying a representation ρ of the gauge group G. The generators T_a have an explicit representation as matrices $\rho(T_a)$ acting on V and satisfying

$$[\rho(T_a), \rho(T_b)] = f_{ab}{}^c \rho(T_c). \tag{1.108}$$

A matter field ψ in the representation ρ is a field with values in V. Also in this case indices may be shown or hidden. If e_m is a basis in V, the field carries the index m and the action of the algebra consists of matrices with indices mn, for example

$$(\rho(T_a)\psi)^m = \rho(T_a)^m{}_n \psi^n.$$

Unscaled version. In perturbation theory it is convenient to have the coupling constant appearing explicitly in the definition of the covariant derivative and curvature. The covariant derivative of a matter field ψ^A in the representation ρ is

$$D_\mu \psi^m = \partial_\mu \psi^m + e A^a_\mu \rho(T_a)^m{}_n \psi^n. \tag{1.109}$$

Indices of the matter representation have been made explicit here, but they are often omitted to avoid clutter. Furthermore, when there are no ambiguities the symbol ρ specifying the representation can also be omitted and the covariant derivative becomes simply

$$D_\mu \psi = \partial_\mu \psi + e A_\mu \psi. \tag{1.110}$$

The nonabelian field strength is

$$F^a_{\mu\nu} = \partial_\mu A^a_\nu - \partial_\nu A^a_\mu + e f_{bc}{}^a A^b_\mu A^c_\nu \tag{1.111}$$

and the YM action is

$$S_{YM} = -\frac{1}{4}\int d^n x\, F^a_{\mu\nu} F^{a\mu\nu}. \tag{1.112}$$

In the notation of (1.110), where A_μ is thought of as a matrix in a given representation of G, one can also write

$$F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu + e[A_\mu, A_\nu] \tag{1.113}$$

and using the normalization (B.2) for the generators

$$S_{YM} = \frac{1}{2}\int d^n x\, \mathrm{tr}\, F_{\mu\nu} F^{\mu\nu}. \tag{1.114}$$

The theory is invariant under the *local gauge transformations*

$$\psi' = g^{-1}\psi, \tag{1.115a}$$

$$A'_\mu = g^{-1} A_\mu g + \frac{1}{e} g^{-1}\partial_\mu g, \tag{1.115b}$$

which imply

$$D_\mu \psi' = g^{-1} D_\mu \psi, \tag{1.116a}$$

$$F'_{\mu\nu} = g^{-1} F_{\mu\nu} g. \tag{1.116b}$$

Note that "local" can have here a double meaning: the transformation is local in the sense that the transformation parameter g is a function of spacetime, as opposed to a constant, and also in the sense that g is only defined in the neighborhood of some point. This second meaning is sometimes relevant when discussing topological properties.

An infinitesimal gauge transformation is a function $\epsilon = \epsilon^a T_a$ with values in $\mathfrak{g}$ and the infinitesimal version of (1.115) is

$$\delta_\epsilon \psi' = -\epsilon^a \rho(T_a)\psi, \tag{1.117a}$$

$$\delta_\epsilon A_\mu = \frac{1}{e} D_\mu \epsilon, \tag{1.117b}$$

where $D_\mu \epsilon = \partial_\mu \epsilon + e[A_\mu, \epsilon]$ or, more explicitly

$$D_\mu \epsilon^a = \partial_\mu \epsilon^a + e f_{bc}{}^a A^b_\mu \epsilon^c. \tag{1.118}$$

Rescaled version. In many cases, and in particular to discuss geometrical properties, it is more convenient to rescale the field A by a factor $1/e$. Then,

the covariant derivative of the matter field is (in the streamlined notation of (1.110))

$$D_\mu \psi = \partial_\mu \psi + A_\mu \psi \tag{1.119}$$

and the nonabelian field strength is

$$F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu + [A_\mu, A_\nu]. \tag{1.120}$$

In this case the YM action reads

$$S_{YM} = -\frac{1}{4e^2} \int d^n x \, F^a_{\mu\nu} F^{a\mu\nu}, \tag{1.121}$$

and the nonabelian gauge transformations are

$$\psi' = g^{-1}\psi, \tag{1.122a}$$
$$A'_\mu = g^{-1} A_\mu g + g^{-1} \partial_\mu g. \tag{1.122b}$$

Abelian case. Electromagnetism is a special case of YM theory. The Lie algebra of $U(1)$ consists of the purely imaginary numbers and one can take as a basis element $T = -i$. In this case A_μ is usually meant to represents the real and unscaled field, not the Lie algebra-valued field $-iA_\mu$. Thus the abelian covariant derivative is

$$D_\mu \psi = \partial_\mu \psi - ieA_\mu \psi. \tag{1.123}$$

Finite gauge transformations are $g(x) = e^{i\alpha(x)}$. The gauge transformations of $-iA_\mu$ are still given by (1.115), thus we have

$$A'_\mu = A_\mu - \frac{1}{e} \partial_\mu \alpha, \tag{1.124a}$$
$$\psi' = e^{-i\alpha(x)} \psi. \tag{1.124b}$$

These formulas are valid in natural units, as commonly used in particle physics. In the discussion of quantum mechanical systems we will use Heaviside units. Then, the second term in the r.h.s. of (1.123) has an additional factor $1/\hbar c$ and the gauge transformation is $A'_\mu = A_\mu - \frac{\hbar c}{e} \partial_\mu \alpha$.

1.5.2 Gauge currents

Noether's first theorem associates a conserved current to each generator of a finite dimensional Lie group that leaves the action invariant. There is a second theorem of Noether that applies to infinite dimensional invariance groups. It says that if a theory is invariant under transformations parameterizd by N

functions $\epsilon^a(x)$, with $a = 1, \dots, N$, then there exist N differential relations between the Euler–Lagrange equations, and furthermore there exist N currents that are conserved without using the Euler–Lagrange equations.

We will prove this theorem in a special case when the variation of the field is purely internal and depends only on the transformation parameter $\epsilon^a(x)$ and its first derivatives:

$$\delta_\epsilon \phi = R_a \epsilon^a + R_a^\mu \partial_\mu \epsilon^a \tag{1.125}$$

and furthermore $\Omega_\epsilon = 0$ (i.e. $\delta_\epsilon \mathcal{L} = 0$ rather than a total derivative). As in the discussion of the first theorem, we omit to write all the indices that the field carries. The coefficients R_a and R_a^μ also carry the same set of hidden indices.

Following the steps of the first theorem, the variation of the Lagrangian is

$$0 = \delta_\epsilon \mathcal{L} = E \cdot \delta_\epsilon \phi + \partial_\mu K_\epsilon^\mu \tag{1.126}$$

where $K_\epsilon^\mu = \pi^\mu \delta_\epsilon \phi$, E is the l.h.s. of the Euler–Lagrange equation $E = 0$ (carrying indices contravariant to those of ϕ) and the sign $\cdot$ means that all hidden indices are contracted. Notice that for constant transformation parameter ($\partial_\mu \epsilon^a = 0$), K_ϵ^μ is Noether's current as defined in the first theorem.

The first term on the r.h.s. can be rewritten:

$$\begin{aligned} E \cdot \delta_\epsilon \phi &= E \cdot (R_a \epsilon^a + R_a^\mu \partial_\mu \epsilon^a) \\ &= \epsilon^a (R_a \cdot E - \partial_\mu (R_a^\mu \cdot E)) + \partial_\mu (\epsilon^a R_a^\mu \cdot E). \end{aligned} \tag{1.127}$$

Then we define

$$Q_a = R_a \cdot E - \partial_\mu (R_a^\mu \cdot E) \tag{1.128}$$

in such a way that (1.126) becomes

$$0 = \epsilon^a Q_a + \partial_\mu (K_\epsilon^\mu + \epsilon^a R_a^\mu \cdot E). \tag{1.129}$$

Let us assume that the transformation parameter and its derivatives go to zero sufficiently fast that the integral of the total derivative on the r.h.s. is zero. Then we find that $\int d^n x \epsilon^a Q_a = 0$, and since $\epsilon^a(x)$ is arbitrary,

$$Q_a = 0. \tag{1.130}$$

This means that there are differential relations between the Euler–Lagrange equations, that hold identically also off-shell. In other words, not all Euler–Lagrange equations are independent

Furthermore, without using the equations of motion, the current

$$k_\epsilon^\mu = K_\epsilon^\mu + \epsilon^a R_a^\mu \cdot E \tag{1.131}$$

is identically conserved:

$$\partial_\mu k_\epsilon^\mu = 0. \tag{1.132}$$

We leave it to Exercise 1.6 to verify that in the case of YM theory

$$Q_a = D_\mu D_\nu F^{\mu\nu a}, \tag{1.133}$$

which is identically zero, and furthermore the current conservation relation (1.132) reduces to

$$\partial_\mu \partial_\nu (\epsilon^a F^{\mu\nu a}) = 0, \tag{1.134}$$

which is also identically true. Thus, these currents are not particularly useful in this context.

Of greater significance are covariantly conserved currents. Recall that in YM theories the matter current that couples to the YM field can be defined as

$$J_a^\mu = \frac{\delta S_m(\phi, A)}{\delta A_\mu^a}, \tag{1.135}$$

where S_m is the matter action. When the free bosonic and fermionic actions considered in Section 1.1 are minimally coupled to gauge fields, and one puts $A = 0$ after the variation, the currents defined in this way agree with (1.12) and (1.18). From the invariance of the matter action under gauge transformations, with a parameter ϵ that tends to zero at infinity, one deduces

$$\begin{aligned} 0 = \delta_\epsilon S &= \int d^n x \left(\frac{\delta S_m}{\delta \phi} \delta_\epsilon \phi + \frac{\delta S_m}{\delta A_\mu^a} \delta_\epsilon A_\mu^a \right) \\ &= \int d^n x \, J_a^\mu D_\mu \epsilon^a \\ &= -\int d^n x \, \epsilon^a D_\mu J_a^\mu, \end{aligned} \tag{1.136}$$

where we used the matter equation of motion $\frac{\delta S_m}{\delta \phi} = 0$. Since ϵ is arbitrary, we derive that the current must be *covariantly conserved*:

$$D_\mu J_a^\mu = 0. \tag{1.137}$$

This is consistent with the YM equation of motion

$$D_\mu F^{\mu\nu a} = J^{\nu a} \tag{1.138}$$

and the relation $D_\mu D_\nu F^{\mu\nu a}$ that is Noether's second theorem. The relations (1.137) do not give rise to conserved charges: this is because the YM field

also carries charge, and there can be transfer of charge between the YM field and matter. In fact, suppressing the algebra indices, we can rewrite the YM equation as

$$\partial_\mu F^{\mu\nu} = J^\nu - [A_\mu, F^{\mu\nu}], \tag{1.139}$$

and the expression on the r.h.s. must then be conserved in the normal sense. In fact, $j^\nu = J^\nu - [A_\mu, F^{\mu\nu}]$ is just the Noether current deriving from invariance of the full action (gauge fields plus matter) under global gauge transformation, see Exercise 1.6.

1.5.3 The Higgs phenomenon

We work at the semiclassical level and focus on the simple example of the gauged $O(N)$ model, that consists of a $O(N)$ gauge field A_μ coupled to a multiplet of scalar fields ϕ^m in the fundamental representation of $O(N)$. Since the Lie algebra of $O(N)$ is isomorphic to the algebra of antisymmetric $N \times N$ matrices, the Lie algebra index a consists here of an antisymmetric pair of indices m, n, as in General Relativity:

$$A_\mu = \frac{1}{2} \sum_{m,n=1}^{N} A_\mu^{mn} T_{mn} = \sum_{m<n} A_\mu^{mn} T_{mn}.$$

The Lagrangian of this theory is the sum of the YM Lagrangian, and the Lagrangian of the $O(N)$ model, where we replace ordinary derivatives by covariant derivatives:

$$\mathcal{L} = -\frac{1}{4} \sum_{m<n} F_{\mu\nu}^{mn} F^{mn\mu\nu} - \frac{1}{2} D_\mu \phi^m D^\mu \phi^m - \frac{1}{2} m^2 |\phi|^2 + \frac{\lambda}{4} |\phi|^4 + U, \tag{1.140}$$

where

$$D_\mu \phi^m = \partial_\mu \phi^m + e A_\mu^{mn} \phi^n. \tag{1.141}$$

As in the ungauged case, this theory can be in two phases, depending on the sign of m^2. If $m^2 > 0$ the state of minimum energy has $F_{\mu\nu}^{mn} = 0, D_\mu \phi^a = 0$ and $|\phi| = 0$. The small excitations around this state are $\frac{N(N-1)}{2}$ massless vectors and N massive scalar fields. If $m^2 < 0$ the state of minimum energy has $F_{\mu\nu}^{mn} = 0$, $D_\mu \phi^m = 0$ and $|\phi| = f \equiv \sqrt{-m^2/\lambda}$ everywhere. This does not imply that ϕ^m is necessarily constant, because the covariant derivative of ϕ^m can be zero without the field being constant. However, the action of $O(N)$ on the S^{N-1} is transitive, meaning that every point can be moved to any other point by some element of the group. This can be done continuously

throughout spacetime, which means that we can find a gauge transformation $g(x)$ that aligns the field, for example, in the N-th direction:

$$g(x)^{-1}\phi(x) = (0, \dots, 0, f).$$

This is called the *unitary gauge* and it leaves a residual local gauge freedom $O(N-1)$. In this gauge one can most easily understand the spectrum of the theory. All the Goldstone boson (angular) degrees of freedom of the scalar have been fixed; only the radial mode remains and it has mass $m_S = \sqrt{2}m = \sqrt{2\lambda}f$, as in the ungauged case. As for the YM field, we note that in this gauge the kinetic term of ϕ becomes

$$-\frac{1}{2}ef^2 \sum_m A_\mu^{mN} A^{\mu mN}. \tag{1.142}$$

Thus, the (mN) components of the gauge field have a mass $m_A = \sqrt{e}f$, whereas the components (mn) with $m, n = 1, \dots, N-1$, remain massless. This is called the Higgs phase of the theory. It is a misnomer to call it a spontaneously broken phase, as in the ungauged case, because gauge invariance is never broken, unless we decide to do so by fixing the gauge.

As in Section 1.3.1, let us now imagine probing the system at energies that are much below the mass of the radial mode, so that it cannot be excited. If $e \ll 2\lambda$, there is a regime in which the massive vectors will still be present, along with the massless $O(N-1)$ YM field. Instead of a linear scalar ϕ^a coupled to gauge fields, one now has a nonlinear, S^{N-1}-valued scalar coupled to gauge fields. In the intrinsic formulation where the scalars φ^α are coordinates on S^{N-1}, the covariant derivatives are defined by

$$D_\mu \varphi^\alpha = \partial_\mu \varphi^\alpha - eA_\mu^a K_a^\alpha(\varphi), \tag{1.143}$$

where K_a^α are the vector fields that generate the action of $O(N)$ on S^{N-1}. This formula can be made plausible by noting that if the action of the group on the target space was linear,

$$K_a^\alpha(\varphi) = -(T_a)^\alpha{}_\beta \varphi^\beta, \tag{1.144}$$

where $(T_a)^\alpha{}_\beta$ are the matrices representing the Lie algebra generators, we would recover the usual definition of covariant derivative of a linearly transforming field.[10] In the case of the two-sphere, one can easily prove this formula

[10]The minus sign in the definition of K_a ensures that the Lie brackets of these vectorfields are isomorphic to the commutators of the generators T_a.

by starting from the covariant derivative of the fields ϕ^m and transforming to spherical coordinates (see Exercise 1.7).

The Lagrangian of the gauged nonlinear sigma model is

$$\mathcal{L} = -\frac{1}{4}\sum_{a<b} F^{ab}_{\mu\nu}F^{ab\mu\nu} - \frac{f^2}{2}D_\mu\varphi^\alpha D^\mu\varphi^\beta h_{\alpha\beta}(\varphi). \tag{1.145}$$

Because the action of $O(N)$ on the sphere is transitive, the Goldstone bosons φ^α are gauge degrees of freedom. By going to the unitary gauge $\varphi = \varphi_*$ (where φ_* are the coordinates of the north pole) and using

$$K^\alpha_a(\varphi_*)K^\beta_b(\varphi_*)h_{\alpha\beta}(\varphi_*) = \delta_{ab}, \qquad \text{for } a, b = 1, \dots, N-1$$

the Lagrangian (1.145) reduces to the mass term (1.142). All the Goldstone bosons have disappeared and the only propagating particles are vectors, of which $N-1$ are massive. Unlike the standard Higgs phenomenon, in this case there is no Higgs particle. Thus one may call this a *Higgsless Higgs phenomenon*.

It is worth mentioning the connection between this construction and another one that goes under the name of *Stückelberg construction*. It was first proposed in the context of electrodynamics, so let us discuss it first in that context. While the Maxwell Lagrangian is invariant under $U(1)$ gauge transformations (1.124), the Proca Lagrangian for a massive spin 1 field

$$\mathcal{L}_P = -\frac{1}{4}F_{\mu\nu}F^{\mu\nu} - \frac{1}{2}m_A^2 A_\mu A^\mu \tag{1.146}$$

is not. However, it can be made invariant by introducing a $U(1)$-valued field φ that transforms by a shift

$$\varphi \longmapsto \varphi - \frac{1}{e}\alpha. \tag{1.147}$$

One can define

$$D_\mu\varphi = \partial_\mu\varphi - A_\mu, \tag{1.148}$$

that is $U(1)$ invariant. Then the Lagrangian

$$\mathcal{L} = -\frac{1}{4}F_{\mu\nu}F^{\mu\nu} - \frac{1}{2}m_A^2 D_\mu\varphi D^\mu\varphi \tag{1.149}$$

is invariant under local $U(1)$ transformations and reduces to (1.146) in the unitary gauge φ =constant. Thus the Stückelberg construction is a way of rewriting a gauge non-invariant action in gauge invariant form. For a Proca field, the result is a gauged nonlinear sigma model with values in $U(1)$.

Similarly, the gauged S^{N-1}-model action (1.145) can be seen as the result of having applied the Stückelberg construction to the Lagrangian

$$\mathcal{L}_P = -\frac{1}{4}\sum_{m<n} F^{mn}_{\mu\nu}F^{mn\mu\nu} - \frac{1}{2}\sum_n m_A^2 A^{nN}_\mu A^{nN\mu}, \tag{1.150}$$

where $N-1$ of the YM fields are Proca-like rather than Maxwell-like.

All the constructions in this section can be repeated with minimal changes for any group G and subgroup H. In the next sections we discuss two prominent examples of Higgs phenomenon: superconductivity, and the Weinberg–Salam theory of electroweak interactions.

1.5.4 Superconductivity

This was one of the earliest applications of the Higgs mechanism.[11] The microscopic description of superconductivity is due to Bardeen–Cooper–Schrieffer (BCS). In the BCS theory the charge carriers are weakly bound pairs of electrons. Such pairs can be described by a complex scalar field transforming under $U(1)$ as

$$\phi(x) \to e^{i\frac{2e}{\hbar}\alpha(x)}\phi(x), \tag{1.151}$$

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

$$\varphi \to \varphi + \alpha. \tag{1.152}$$

At low energy, the BCS description can thus be replaced by an effective Ginzburg–Landau theory, whose free energy is

$$F = \int d^3x \left[\frac{\lambda}{4}(|\phi|^2 - f^2)^2 + \frac{1}{4}F_{\mu\nu}F^{\mu\nu} + \frac{1}{2}D_\mu\phi^* D^\mu\phi + ...\right], \tag{1.153}$$

The potential has already been written in a form that is suitable for the Higgs phase, where the radial Higgs mode has mass $m_S = \sqrt{2\lambda}f$, the photon has mass $m_A = \sqrt{e}f$ and the phase of ϕ (the Goldstone boson) is a pure gauge degree of freedom that can be set to zero by going to the unitary gauge. Still further in the infrared, namely at energies much below m_S, the radial mode can be ignored and we remain just with the Lagrangian (1.149) that describes

[11] For this discussion we follow mostly [Wei86], also reported in [Wei95], Volume II.

a massive photon. The dynamics of this theory is extremely simple: we have Maxwell's equations

$$\partial_\mu F^{\mu\nu} = J^\nu \tag{1.154}$$

with

$$J^\nu = -m_A^2 D^\nu \varphi, \tag{1.155}$$

while the equation of motion of φ is just the statement of current conservation $\partial_\mu J^\mu = 0$.

This very simple theory of a Goldstone boson coupled to the electromagnetic field is sufficient to explain all the main features of a superconductor.[12] As usual, the ground state is defined by the conditions $F_{\mu\nu} = 0$ and $D_\mu \varphi = 0$. Now assume that there are some external currents that produce a magnetic field with potential A_i, and that a superconducting sample is placed in it. While now $F_{\mu\nu}$ is no longer zero, it is reasonable to assume that, at least for sufficiently weak external sources, the energy will still be minimized by

$$D_\mu \varphi = 0. \tag{1.156}$$

(We shall give more quantitative conditions for this to happen in Section 2.6.2.) In a static situation with $\partial_0 \varphi = 0$ and $A_0 = 0$, the space component of (1.156)

$$A_i = \partial_i \varphi \tag{1.157}$$

implies that the magnetic field in the sample must be zero:

$$B_i = 0. \tag{1.158}$$

This is known as the *Meissner effect*. If there is an external magnetic field, it can only be tangential to the surface of the sample and it decays exponentially in the interior with a characteristic length called the *London penetration depth* $\lambda_L \sim 1/m_A$ (see Exercise 1.8). This is just a consequence of the photon being massive in the bulk of the superconductor. In the gauge invariant formalism one can interpret the effect by saying that the Goldstone boson gives rise to currents that flow in the boundary layer so as to exactly cancel the magnetic field. Either way, the net effect is that field lines will be deformed so as to avoid going through the bulk of the sample.[13]

Absence of electrical resistance can be gleaned from the following argument. In any simply connected piece of superconductor, φ can be set to any

[12] Actually, it is not even necessary to assume that the Lagrangian has exactly the form (1.149): it is enough to assume that it is gauge invariant and that the energy is minimized by $D\varphi = 0$.

[13] It is worth emphasizing that in the full Ginzburg–Landau model, the condition $D_\mu \phi = 0$ implies the Meissner effect only when the VEV of ϕ is nonzero.

fixed constant by a transformation (1.152). Now consider a thick torus made of superconductor and let ℓ be a closed loop deep in the material. Integrating (1.157) on this loop and using Stokes' theorem we find that

$$\Delta\varphi = \int_\ell A = \int_S B = \Phi, \tag{1.159}$$

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

$$\Phi = \frac{\pi\hbar}{e} n. \tag{1.160}$$

Because of this, the current in the superconductor cannot decay continuously.

A more general way of showing this is to note that the momentum conjugate to φ is $\pi = m_A^2 D_0\varphi = -J^0$, so that Hamilton's equation for φ is

$$\dot\varphi = \frac{\partial H}{\partial \pi} = -\frac{\partial H}{\partial J^0}. \tag{1.161}$$

The change of the energy due to a change in the charge density at a point is just the electrostatic potential V at that point, so that

$$\dot\varphi(x) = -V(x). \tag{1.162}$$

Now consider a superconducting wire carrying a steady current, with time-independent electromagnetic field. From (1.155), if the current and electromagnetic potential are time-independent, also the gradient of φ must be time-independent. But then, from (1.162) we see that

$$0 = \partial_t(\partial_i\varphi) = \partial_i\dot\varphi = -\partial_i V \tag{1.163}$$

which implies that the potential must be constant. A nonzero current with zero voltage difference is the definition of zero resistance.

Finally, we can get an understanding of the Josephson effect. Suppose two superconductors are separated by a thin gap (this is called a Josephson junction). If there are no gauge potentials and no gradients along the gap, the Goldstone boson in the gap gives a contribution to the Lagrangian of the form

$$L = AF(\Delta\varphi),$$

where A is the area of the junction, and F is a function whose exact form we need not know, but that must be periodic with period $\pi\hbar/e$.[14] If there

[14]The leading quadratic term $(D_\mu\varphi)^2$ that is written in (1.149) would lead to $F(\Delta\varphi) = (\Delta\varphi)^2$, but this has to be interpreted as the first term in the expansion of a cosine or some other periodic function.

is a magnetic potential A_i, by gauge invariance the argument of F must be replaced by $\Delta_A\varphi$, which is defined as the line integral of $D\varphi$ across the gap. Then from (1.155), the current is equal to $F'(\Delta_A\varphi)$ and points across the gap. This means that for $A_i = 0$ the current must be proportional to $F'(\Delta\varphi)$. Suppose that the two superconductors are kept at a fixed potential difference ΔV. By (1.162), the difference of the Goldstone boson field across the gap must grow linearly with time:

$$\Delta\varphi = -t\Delta V + \text{const.}$$

Since F is periodic, this implies that also the current must be periodic with a frequency

$$\nu = \frac{e}{\pi\hbar}|\Delta V|. \tag{1.164}$$

This exact linear relation between the frequency of the Josephson current and the potential difference gives a direct and very accurate method to measure the ratio $e/\hbar$.

In Section 2.6.2 we shall discuss in more detail the phase diagram of superconductivity and the role of vortices.

1.5.5 Electroweak theory

The bosonic sector of the Weinberg–Salam (WS) model of the electroweak interactions consists of YM fields W^a_μ and B_μ for the group $SU(2)_L \times U(1)_Y$ and of a Higgs field $H = \begin{pmatrix}\varphi^+\\ \varphi^0\end{pmatrix}$ in the fundamental representation of $SU(2)_L$. The charge associated to $SU(2)_L$ is called weak isospin and that of $U(1)_Y$ weak hypercharge. The dynamics for this sector is given by the Lagrangian[15]

$$\mathcal{L}_{WS} = -\frac{1}{4}W^a_{\mu\nu}W^{a\mu\nu} - \frac{1}{4}B_{\mu\nu}B^{\mu\nu} - D_\mu H^\dagger D^\mu H - m^2 H^\dagger H - \lambda(H^\dagger H)^2, \tag{1.165}$$

with the curvatures

$$W^a_{\mu\nu} = \partial_\mu W^a_\nu - \partial_\nu W^a_\mu - g_2\epsilon_{abc}W^b_\mu W^c_\nu, \tag{1.166a}$$

$$B_{\mu\nu} = \partial_\mu B_\nu - \partial_\nu B_\mu \tag{1.166b}$$

and the covariant derivative

$$D_\mu H = \partial_\mu H + \frac{1}{2}ig_1 B_\mu H + g_2 W^a_\mu \tau_a H. \tag{1.167}$$

[15] We follow here a common convention for the parametrization of the Higgs potential, that differs slightly from our standard.

Here g_1 is the $U(1)_Y$ coupling constant and g_2 is the $SU(2)_L$ YM coupling.

When $m^2 < 0$ the system is in the Higgs phase, with

$$H^\dagger H = \frac{1}{2}v^2, \qquad v = \frac{m}{\sqrt{\lambda}}. \tag{1.168}$$

We can choose the gauge

$$H = \begin{pmatrix} 0 \\ v/\sqrt{2} \end{pmatrix} \tag{1.169}$$

Inserting in the kinetic term of H, we find the following mass matrix for the gauge bosons

$$\frac{1}{4}g_2 v^2 W^+_\mu W^{\mu -} + \frac{1}{8}v^2 \begin{pmatrix} W^3_\mu & B_\mu \end{pmatrix} \begin{pmatrix} g_2^2 & -g_1 g_2 \\ -g_1 g_2 & g_1^2 \end{pmatrix} \begin{pmatrix} W^{\mu 3} \\ B^\mu \end{pmatrix}. \tag{1.170}$$

The mass eigenstates are the charged gauge bosons

$$W^\pm_\mu = \frac{1}{\sqrt{2}}(W^1_\mu \mp W^2_\mu) \tag{1.171}$$

and the following linear combinations of neutral gauge bosons

$$Z_\mu = \cos\theta_W W^3_\mu - \sin\theta_W B_\mu \tag{1.172}$$

$$A_\mu = \sin\theta_W W^3_\mu + \cos\theta_W B_\mu \tag{1.173}$$

where $\tan\theta_W = g_1/g_2$. Their masses are

$$m_W = \frac{1}{2}v g_2, \qquad m_Z = \frac{1}{2}v\sqrt{g_1^2 + g_2^2}, \qquad m_A = 0 \tag{1.174}$$

and $m_W = M_Z \cos\theta_W$.

If we decompose φ^+ and φ_0 in their real and imaginary parts, we see that

$$H^\dagger H = (\mathrm{Re}(\varphi^+))^2 + (\mathrm{Im}(\varphi^+))^2 + (\mathrm{Re}(\varphi^0))^2 + (\mathrm{Im}(\varphi^0))^2$$

is invariant under $SO(4)$, and so is the Higgs potential. This is a larger symmetry than the one of the gauge and fermion sector. It can be better understood if, as in Exercise 1.4, we construct a matrix

$$\Sigma = \begin{pmatrix} \varphi^+ & -\varphi^{0*} \\ \varphi^0 & \varphi^{+*} \end{pmatrix} \tag{1.175}$$

in terms of which

$$H^\dagger H = \frac{1}{2}\mathrm{tr}\,\Sigma^\dagger\Sigma. \tag{1.176}$$

Writing the potential in terms of Σ shows explicitly that it is invariant under $SU(2)_L \times SU(2)_R \sim SO(4)$. Here $SU(2)_L$ is the weak isospin group and $SU(2)_R$ is a global symmetry of the Higgs sector only. It makes it very similar to the linear sigma model of the strong interactions. Then, at energies below the mass of the Higgs particle, we can remove the radial mode from the spectrum and we remain with a nonlinear sigma model coupled to the gauge fields. This is the basis of an *electroweak chiral perturbation theory* (EWχPT), that is very similar to the one of the strong interactions, except for the replacement of F_π by v and for the coupling to the gauge fields [Fer93, HeR94].

In this theory the ground state is characterized by

$$D_\mu \varphi^\alpha = 0, \tag{1.177}$$

where φ^α are the electroweak Goldstone bosons, that have values in the coset $SU(2)_L \times U(1)_Y / U(1)_Q \approx S^3$. As in superconductivity, this implies that the field strengths $W_{\mu\nu} = B_{\mu\nu} = 0$, which is an analog of the Meissner effect.

1.6 Status of symmetries

Symmetries play a crucial role in modern physics, but the terminology that is used is sometimes contradictory and possibly confusing, so at the cost of appearing pedantic, we shall review here various ways in which symmetries appear in physical systems, and fix our terminology.

One has to start at a kinematical level with a group G of transformations acting on the dynamical variables ϕ of a system. For notational simplicity we shall simply write this action as $\phi \longmapsto g\phi$. Then one defines the dynamics by giving an action $S(\phi)$. We will say that G is an *invariance* of the classical theory if S is invariant under the transformations of G, meaning that

$$S(g\phi) = S(\phi).$$

This is a purely mathematical statement.

At a physical level, we need to distinguish the situation when the transformations of G correspond to physical operations on the system, from the situation when they do not. In other words, we have to ask whether a configuration ϕ and the transformed configuration $g\phi$ are physically distinguishable or not. Somewhat surprisingly, there does not seem to be a universally agreed upon terminology for this important distinction. In Hamiltonian mechanics

the term *gauge transformations* is used to denote unphysical transformations, irrespective whether they form a finite or infinite dimensional group. See Section 1.7. In particle physics the term "gauge transformation" is generally used for all the internal transformations of YM theories. However, one distinguishes between *global gauge transformations* (also referred to as *rigid gauge transformations* or *gauge transformations of the first kind*), that have a constant transformation parameter, and *local* tranformations (also referred to as *gauge transformations of the second kind*) whose parameters are functions on spacetime.

More precisely, if M is spacetime, let $\mathcal{G}$ be the infinite dimensional group of all maps from M to the YM group G and let $\mathcal{G}_*$ be the normal subgroup of maps that tend asymptotically to the identity at infinity. The quotient $\mathcal{G}/\mathcal{G}_*$ is isomorphic to G and can be identified with the global, or rigid, gauge transformations. It is a group of physical transformations, because changing the fields at infinity changes the charges, that are physically observable. The group $\mathcal{G}_*$, instead, consists of unphysical transformations.

Since there is a broad tendency to identify "gauge transformations" with "unphysical transformations", it would be better not to call gauge transformations the global YM transformations and to reserve this name only for the transformations in $\mathcal{G}_*$. Without going so far, we will refer to an invariance group as a *symmetry group* if it consists of "physical" transformations, and as a *gauge invariance* if it consists of "unphysical" transformations. Note that if one adopts this definition, the term "gauge symmetry" becomes an oxymoron.

Let us now focus on a group of "physical" transformations of some system. It may or may not be a symmetry, depending on whether S is invariant or not. Sometimes there are terms in the action that are invariant and others that are not. In this case one says that the latter *explicitly break* the invariance. This notion is useful when there is only one symmetry breaking term, or when the symmetry breaking terms are, in some appropriate sense, small compared to the rest. A typical example is the term (1.63) in the chiral model.

Assuming that the action is invariant, we next consider a specific instance of the system that finds itself in a given state, and ask whether it is left invariant by the transformations of G or not. If it is, we say that the symmetry in unbroken, and if it is not we say that the symmetry is broken *in that state*. One expects that the latter situation is quite generic. For example, in a translation invariant theory, translation invariance will be broken in any state containing a particle in some position. The question is of greater significance when referred to the ground state of the system: if it is not invariant under G, one says that G is *spontaneously broken*. The subgroup $H \subset G$ that leaves the system invariant is called the *unbroken group*.

In quantum mechanics, the states of the system are rays in a Hilbert space and groups of transformations are linearly realized as operators acting on this space:

$$\psi \to \psi' = \mathcal{U}(g)^{-1}\psi.$$

A symmetry group is unbroken if the vacuum state ψ_0 is invariant and spontaneously broken otherwise. Invariance means that for every generator T of the group's algebra,

$$T\psi_0 = 0.$$

Thus both in classical and quantum mechanics, one must first have a symmetry group leaving the action invariant, and the question whether it is broken or not depends on the vacuum.

The term "spontaneously broken" is often applied also to YM gauge transformations, but this is misleading. When such transformations are present, it means that the variables ϕ we are using to describe the system are redundant: in classical mechanics the physical state of a system with a gauge invariance is not given by ϕ but by the equivalence class of ϕ under the action of $\mathcal{G}_*$. Similarly in quantum mechanics, the Hilbert space of physical states can be defined as the quotient of some pre-Hilbert space, on which the gauge transformations act, by the gauge group. But then, in both cases, the transformations of $\mathcal{G}_*$ leave the physical states invariant by definition, and hence this group can never be broken. The better terminology for these situations is to say that the theory is in the Higgs phase. On the other hand, when the Higgs phenomenon occurs, the global gauge group $\mathcal{G}/\mathcal{G}_* \approx G$, that consists of physical transformations, is indeed broken.

Finally, there are the anomalies. These are situations when a group of transformations leaving the dynamics of a classical system invariant cannot be realized in the corresponding quantum mechanical system. Again one must distinguish between the case of physical and unphysical transformations. For example, for a global symmetry group, an anomaly implies that the Noether current is conserved in the classical theory but not in the quantum theory. This can have interesting physical implications and is not a pathology. On the other hand in the case of a group of gauge transformations, breaking the invariance in the quantum theory typically spoils its good properties such as unitarity and renormalizability. This can manifest itself at various levels. Very generally, an anomaly in a gauge group means that certain degrees of freedom that are unphysical at the classical level, become physical in the quantum theory. These new degrees of freedom then lead to pathological behavior. In the covariant formalism, the anomaly manifests itself as a failure of the covariant conservation law (1.137). In the canonical formalism, the anomaly

manifests itself in extensions in the algebra of the group generators, as already remarked in the end of Section 1.1. In the simplest cases these are just central extensions, but more generally they involve functions of the field operators

$$[Q_a, Q_b] = i f_{ab}{}^c Q_c + \Omega_{ab}(\phi). \tag{1.178}$$

Then, if one defines physical states to be the ones annihilated by the gauge generators

$$Q_a \psi_{phys} = 0, \tag{1.179}$$

the presence of the nontrivial extension implies that $\psi_{phys} = 0$. We shall discuss anomalies in Chapter 6.

* * *

This may be a good place to highlight the different ways in which symmetries appear in various theories. This will also serve as a very high level summary of the models we have introduced in this chapter. We consider three classes of phenomena: the strong force, superconductivity and the electroweak forces. These are related to the chiral group $SU(2) \times SU(2)$ (a proper symmetry, not gauged), $U(1)$ (gauged) and $SU(2) \times U(1)$ (gauged), respectively. The following table summarizes the ways these groups are realized in each case and emphasizes the (necessarily imperfect) similarities between these phenomena.

	Strong	Superconductivity	ElectroWeak
Fundamental description	QCD	BCS	Weinberg–Salam
Linear EFT	Linear σ model	Ginzburg–Landau	Weinberg–Salam
Low energy EFT of Goldstone bosons	Chiral model	Ginzburg–Landau with $\lvert\phi\rvert = f$	Electroweak chiral model

In the fundamental description, holding at short length scales, the group is always realized linearly, and in the strong and electroweak case, the theory is even renormalizable. At some energy scale, the theories in the second and third columns give rise dynamically to a condensate that breaks the symmetry (in the case of a true symmetry, as in QCD) or puts it in Higgs phase (in the case of a gauge invariance, in the last two columns). The formation of the condensate can be described by an effective field theory where the group is still linearly realized. This is a model containing an order parameter, whose VEV decides what phase the theory is in. For the strong interactions, this is the linear sigma model of Section 1.2.3. In condensed matter models it is a Ginzburg–Landau theory. The regime in which these effective models work

best is near the phase transition. The Weinberg–Salam model is already in Ginzburg–Landau form: its order parameter is regarded as a fundamental field. In fact the table suggests that at high energy the WS model may be replaced by a more fundamental model where the order parameter is a fermionic condensate. In the past this idea went under the name of Technicolor, but these theories are now ruled out and there is currently no experimental evidence for a composite Higgs.

In all these linear models, in the broken (or Higgs) phase the order parameter is a massive field. When one looks at the theory at energies below its mass, the order parameter is frozen at its VEV and only the Goldstone boson degrees of freedom are still active. In this domain the group is nonlinearly realized and the relevant models are nonlinear sigma models. For the strong interactions, this is the chiral perturbation theory, discussed in Sections 1.3.3 and 1.4.2. In the other two models the Goldstone boson is a gauge degree of freedom and is only needed if one wants a gauge-invariant description of physics. By choosing the unitary gauge, the Goldstone boson is eliminated and one remains just with a massive gauge field. One can also see the gauge invariant description as a result of the Stückelberg construction applied to the massive gauge theory. Either way, the theory is non-renormalizable, but one does not need that for a description that has a clear upper bound in energy.

1.7 Canonical formalism

1.7.1 Field theory as infinite dimensional mechanics

In our discussions of topological effects in quantum field theory, we will study the configuration space of the theory, defined as the (infinite dimensional, generally nonlinear) space of field configurations at a fixed time t.

In a field theory, the field is usually regarded as a function on spacetime of the form $\phi(t, x_i)$, where x_i are the coordinates of the hypersurface $t = \text{const}$. The tautological redefinition

$$\phi(x_i, t) = (\phi(t))(x_i) \tag{1.180}$$

allows us to think of ϕ as a map from time into the space of maps from space to N. Let us say this in a more formal way. We must assume that spacetime has, at least locally, the structure $\Sigma \times \mathbb{R}$, where Σ is space, with coordinates x_i, and $\mathbb{R}$ is the time axis. Then we define $\mathcal{Q} = \Gamma(\Sigma, N)$ the space of maps from Σ to N. We call this the *configuration space* of the theory.[16] Then, (1.180) means

[16] Actually, one has to impose boundary conditions on the fields. We will specify $\mathcal{Q}$ more precisely on a case by case basis.

that we can think of ϕ as a map from $\mathbb{R}$ (time) to $\mathcal{Q}$.

At a given $t = \bar{t}$, $\phi(\bar{t}) : \Sigma \to N$ is an instantaneous state of the field, and can be thought of as a point in $\mathcal{Q}$. As time runs, $\phi(t)$ traces out a curve in $\mathcal{Q}$. Recall that in Lagrangian and Hamiltonian mechanics, the configuration space of a system is parametrized by generalized coordinates q^i. In the field theory the role of the q^i is played by the map $\phi(x_j)$, and the role of the index i is played by the coordinates x_j and by any other Lorentz or internal indices that ϕ may be carrying, that we have suppressed here for notational simplicity. The fact that there are infinitely many values for the coordinates x_j is a precise way of saying that a field theory is an infinite dimensional mechanical system.

When a classical field theory is thought of as a mechanical system, one can think of quantum field theory as the Schrödinger quantization of this system. In other words, the quantum mechanical state of the system would be given by a wave functional $\psi(\phi)$ satisfying a functional Schrödinger equation

$$i\hbar \frac{\partial \psi}{\partial t} = H\psi, \tag{1.181}$$

where H is the field's Hamiltonian. This formalism is not convenient for the calculation of typical quantum field theoretic observables. Furthermore, a mathematically rigorous formulation of quantum field theory along these lines would require overcoming several technical difficulties. For example, in order to turn the space of wave functonals $\mathcal{H}$ into a Hilbert space, one would have to define an inner product that would already involve a functional integral:

$$(\psi_1, \psi_2) = \int_{\mathcal{Q}} (d\phi)\psi_1^*(\phi)\psi_2(\phi). \tag{1.182}$$

We will not need any of this because we shall only use this formulation at a heuristic level. The topological effects we shall encounter have finite dimensional analogs that are clearly related to the lowest homotopy or cohomology groups of the configuration space. Then, whenever the configuration space of a field theory has the same nontrivial homotopy or cohomology groups, a similar effect is expected to take place. Any rigorous definition of the inner product, for example, must reflect the topological properties of $\mathcal{Q}$. This approach will allow us to have a clear classification of various types of topological effects.

1.7.2 Constrained Hamiltonian dynamics

Let us consider a mechanical system with generalized coordinates q^i, with $i = 1, \ldots N$, and a Lagrangian quadratic in $\dot{q}^i$. As we have just discussed,

this could also be a field theory. In order to calculate the time evolution of the system, one has to give initial conditions consisting of the positions and velocities at some time t_0: $q^i(t_0)$ and $\dot{q}^i(t_0)$. The equations of motion contain $\ddot{q}^i$ and allow us to calculate $q^i(t_0+dt)$ and $\dot{q}^i(t_0+dt)$. In many cases of interest, it happens that some equations do not contain $\ddot{q}^i$. Such equations do not help in solving the time evolution, and have to be regarded as constraints that must be satisfied by the initial conditions. They are therefore called *constraint equations*.

The Hamiltonian formulation of systems with constraints can be dealt with in general using the so-called Dirac–Bergmann algorithm. A detailed discussion is outside the scope of this book, but it will be useful to review some of the main points and to see how they apply to the theories we are interested in.

We consider dynamical systems with Lagrangians of the general form

$$L = \frac{1}{2} g_{ij}(q)\dot{q}^i\dot{q}^j + \mathcal{A}_i(q)\dot{q}^i - V(q), \tag{1.183}$$

where q^i are coordinates in some configuration space $\mathcal{Q}$. The system is said to be regular if g_{ij} is nondegenerate, and singular otherwise. For regular systems, the standard Hamiltonian formalism applies. Let us therefore assume that g_{ij} has rank $r < N$. We assume that the coordinates are labeled in such a way that the first r coordinates correspond to the nondegenerate part of the metric and $g_{ij} = 0$ for $i, j > r$. We split the indexing set so that indices a, b ... (from the first part of the alphabet) run from 1 to r and indices m, n, ... (from the second part of the alphabet) run from r to N. Our singular Lagrangian then reads

$$L = \frac{1}{2} g_{ab}(q)\dot{q}^a\dot{q}^b + \mathcal{A}_a(q)\dot{q}^a + \mathcal{A}_m(q)\dot{q}^m - V(q), \tag{1.184}$$

with the summation convention applying to repeated indices running over the appropriate ranges.

To begin with, the phase space is the cotangent bundle $\mathcal{F} = T^*\mathcal{Q}$, a $2N$-dimensional manifold parametrized by the coordinates q^i and momenta p_i.[17] We compute the momenta $p_i = \frac{\partial L}{\partial \dot{q}^i}$. For the first r momenta we have

$$p_a = g_{ab}(q)\dot{q}^b + \mathcal{A}_a(q) \tag{1.185}$$

and this relation can be inverted

$$\dot{q}^a = g^{ab}(p_a - \mathcal{A}_a(q)). \tag{1.186}$$

[17] The cotangent bundle of a manifold M is just the space of all covariant vectors.

On the other hand, the remaining momenta $p_m = \mathcal{A}_m(q)$ do not depend on the velocities. Let us define the functions of phase space

$$\Phi_m(q^i, p_n) = p_m - \mathcal{A}_m(q). \tag{1.187}$$

The conditions

$$\Phi_m = 0 \tag{1.188}$$

are called the *primary constraints*. They define a $N + r$-dimensional subspace $\mathcal{F}_P \subset \mathcal{F}$. We can define the canonical Hamiltonian

$$\begin{aligned} H_C &= p_a \dot{q}^a + p_m \dot{q}^m - L(q^i, \dot{q}^a, \dot{q}^m) \\ &= \frac{1}{2} g^{ab}(p_a - \mathcal{A}_a(q))(p_b - \mathcal{A}_b(q)) + V(q) + \dot{q}^m \Phi_m. \end{aligned} \tag{1.189}$$

Since the velocities $\dot{q}^m$ cannot be expressed as functions of the momenta, H_C has to be thought of as a function of all the coordinates q^i, of the first momenta p_a and of the remaining velocities $\dot{q}^m$. However, these velocities appear multiplied by the constraint Φ_m and therefore this dependence is absent on $\mathcal{F}_P$.

With a little work, and using as usual the Euler–Lagrange equations in the form $\frac{\partial L}{\partial q^i} = \dot{p}_i$, one finds the system of equations

$$\dot{q}^a = \frac{\partial H_C}{\partial p_a}, \tag{1.190a}$$

$$\dot{p}_i = -\frac{\partial H_C}{\partial q^i} + \dot{q}^m \frac{\partial \mathcal{A}_m}{\partial q^i}. \tag{1.190b}$$

These look like Hamilton's equations, but there are only $N + r$ of them, instead of $2N$: the $N - r$ velocities $\dot{q}^m$ remain undetermined. Furthermore, there is an additional term in the r.h.s. of the second set of equations.

Let us say that a function $F(q^i, p^i)$ is *weakly zero*, and write $F \approx 0$, if it vanishes on $\mathcal{F}_P$. It can be shown (see Exercise 1.9) that if a function is weakly zero, it is a linear combination of the primary constraints:

$$F = \frac{\partial F}{\partial p_m} \Phi_m. \tag{1.191}$$

In order to have a full system of equations we need a Hamiltonian that is defined on all of $\mathcal{F}$, does not depend on velocities and is weakly equal to the canonical Hamiltonian. This is achieved by defining the *primary Hamiltonian*

$$H_P = H_C + \mu^m \Phi_m, \tag{1.192}$$

where μ^m are $N-r$ functions that we shall try to determine and the canonical Hamiltonian is the one given in (1.189) with the last term set to zero (since it is already weakly zero). The time evolution of any function on phase space is now given by the Poisson bracket of that function with the primary Hamiltonian. In particular,

$$\dot{q}^i = \frac{\partial H_P}{\partial p_i}, \tag{1.193a}$$

$$\dot{p}_i = -\frac{\partial H_P}{\partial q^i}. \tag{1.193b}$$

These now look much more like Hamilton's equations, but they still depend on the arbitrary functions μ^m so that the time evolution of the variables q^m is still undetermined.

To proceed, we observe that consistency of the dynamics requires that the primary constraints, once imposed at some initial time, remain valid at all later times. Thus we must impose

$$0 \approx \{\Phi_m, H_P\}. \tag{1.194}$$

At this point several things may happen: these relations may either be an identity $0 \approx 0$, or determine some of the multipliers μ^m, or give rise to new functional relations between the phase space variables, that are called *secondary constraints*. When secondary constraints arise, they must also be conserved by the time evolution. This may again determine some of the μ^m or give rise to further constraints. The procedure has to be iterated until no new constraints arise.

Let Ψ_i be all the constraints, with $i = 1, \dots, n < 2N$. They all satisfy

$$\Psi_i \approx 0, \quad \text{and} \quad \{\Psi_i, H_P\} \approx 0.$$

We now say that a constraint is *first class* if its Poisson bracket with all other constraints is weakly zero, and *second class* otherwise. These two classes of constraints have very different meaning, so let us discuss them in turn.

The number of second class constraints is even and they represent inessential pairs of canonical variables that can be readily eliminated. If the system has only $2s$ second class constraints Ψ_i, the physical phase space $\mathcal{F}' \subset \mathcal{F}$ defined by these constraints is $2(N-s)$-dimensional. The question arises of how to define a Poisson bracket in $\mathcal{F}'$. Let f and f' be two functions on $\mathcal{F}$ that are equal on $\mathcal{F}'$, so that $f' = f + \sum_i \rho_i \Psi_i$. The Poisson bracket of these functions with any other function g are not the same:

$$\{f', g\} = \{f, g\} + \sum_i \{\rho_i, g\}\Psi_i + \sum_i \rho_i \{\Psi_i, g\}.$$

The second term vanishes on $\mathcal{F}'$, but the third, in general, does not, so this bracket does not depend only on the values of the functions in $\mathcal{F}'$. The way around this difficulty is to define a new bracket that "respects the constraints". One can show that the matrix of Poisson brackets of the second class constraints $M_{ij} = \{\Psi_i, \Psi_j\}$ is nondegenerate. We define the *Dirac bracket* of two functions on phase space

$$\{f, g\}_D = \{f, g\} - \{f, \Psi_i\} M^{-1}_{ij} \{\Psi_j, g\}. \tag{1.195}$$

It has the property that the Dirac bracket of any second class constraint with any other function on phase space is zero. Therefore, in the example discussed above, the Dirac bracket is $\{f', g\}_D = \{f, g\}_D$, and it depends only on the values of f and g in the constrained subspace.

This property is also useful when the system has to be quantized. Under the normal correspondence that maps the Poisson bracket of two classical observables to $i\hbar$ times the commutator of the corresponding quantum operators, the quantization of a constrained system would run into difficulties. For example, if $f(q, p)$ is any classical observable and $\Psi(q, p) = 0$ is a constraint, and $\widehat{f}$ and $\widehat{\Psi}$ are the corresponding quantum operators, it is natural to assume that $\widehat{\Psi} = 0$. But then $[\widehat{f}, \widehat{\Psi}] = 0$, whereas in general $\{\widehat{f}, \widehat{\Psi}\} \neq 0$. On the other hand, if we postulate that the correspondence maps the Dirac bracket of two classical observables to $i\hbar$ times the commutator of the corresponding quantum operators, then this difficulty does not arise.

Let us come to the first class constraints. It can be shown that, after all the constraints have been found, the number of multipliers μ^m that remain undetermined is equal to the number of primary first class constraints. So, when there are primary first class constraints, there are still residual ambiguities in the time evolution: given an initial condition for the q^i and p_i, their values at a later time will be different if they are evolved with two Hamiltonians that have different multipliers μ^m. The only way to reconcile this with a deterministic evolution is to assume that the variables whose evolution is affected by this ambiguity are unphysical. We conclude that the primary first class constraints generate transformations that leave the physics unchanged, or, in the terminology of Section 1.6, they generate gauge transformations. (Note that this term is used even when the system has finitely many degrees of freedom.)

It is then important to understand the structure of the gauge group $\mathcal{G}$: its parameters, the gauge algebra etc. The gauge parameters associated to the primary first class constraints are just the multipliers μ^m. However, there may exist other gauge transformations. Since infinitesimal gauge transformations must form a closed algebra, the Poisson bracket of two primary first class

constraints must be a gauge generator, but the Poisson bracket of two primary first class constraints need not be a primary first class constraint: it may well be a secondary first class constraint. In fact it turns out that under certain conditions, that are satisfied in the cases of most interest to us, all first class constraints are generators of gauge transformations. We shall therefore assume that this is the case.

To summarize, a generic constrained system may contain both first and second class constraints. Then, the canonical variables can be grouped into three disjoint sets:

1. $2s$ functions that can be eliminated by the second class constraints and do not carry physical information.

2. m first class constraints and, for each of these a corresponding gauge variable, whose time evolution is arbitrary. Also these $2m$ variables do not carry physical information.

3. $2(N - s - m)$ independent functions that have weakly zero Poisson bracket with all the constraints. These are the physical variables of the system.

Under suitable regularity conditions, the second class constraints define a $2(N - s)$-dimensional submanifold $\mathcal{F}' \subset \mathcal{F}$ whose symplectic structure is given by the Dirac brackets. Likewise, the first class constraints define a $2(N - s) - m$-dimensional submanifold $\mathcal{F}''$ on which the gauge group acts. The gauge variables parameterize the orbits of the gauge group in $\mathcal{F}''$. The final *reduced phase space* is the quotient $\mathcal{F}''' = \mathcal{F}''/\mathcal{G}$. If this action is suitably regular, it is a smooth $2(N - s - m)$-dimensional manifold. Thus we observe that every second class constraint reduces the dimension by one, and each first class constraint reduces the dimension by two. An alternative way to see this is to construct a copy of the reduced phase space by choosing a representative in each gauge orbit, i.e. to choose a gauge.[18] A gauge condition is a set of m functions such that, together with the m first class constraints, they form a second class system. At this point we have a total of $2(s + m)$ second class constraints and the symplectic structure in the gauge fixed manifold is given by the Dirac brackets. Since each first class constraint must have a corresponding gauge condition, we see again that each first class constraint reduces the dimension of phase space by two.

[18]The projection $p : \mathcal{F}'' \to \mathcal{F}'''$ given by taking gauge equivalence classes, in the best situation, is a principal $\mathcal{G}$ bundle (see Appendix C). The choice of gauge is then a (local) section of this bundle, i.e. a map $s : \mathcal{F}''' \to \mathcal{F}''$ such that $p \circ s$ is the identity of $\mathcal{F}'''$.

To make this discussion more concrete, we will now look at two examples of constrained systems: a formulation of the nonlinear sigma model where only second class constraints appear, and YM theory, where only first class constraints appear.

1.7.3 The $O(3)$-nonlinear sigma model

The $O(3)$-nonlinear sigma model in the geometric formulation discussed in Section 1.3.2 is a Lagrangian system of the form (1.183), where $\mathcal{A}_i = 0$, $V = 0$ and the metric g_{ij} is nondegenerate. It is therefore a regular system. Here we consider instead the $O(3)$-nonlinear sigma model in the form (1.68). Since in this form the Lagrangian already contains a constraint, it is hardly surprising that the same happens in the canonical formulation. All the essential features can already be seen when we consider what happens at a single point. This is equivalent to choosing a space of dimension zero (i.e. $d = 0$ or $n = 1$). The Lagrangian is then

$$L = \frac{1}{2}\sum_{m=1}^{3}\left(\frac{d\phi^m}{dt}\right)^2 - \Lambda\left(\sum_{m=1}^{3}\phi^m\phi^m - f^2\right). \tag{1.196}$$

When space has dimension $d \geq 1$, the nonlinear sigma model has one such set of degrees of freedom per space point. The only new ingredient in the analysis is that coordinates and momenta at different points commute, which leads to the presence everywhere of delta functions of $x - y$, say, but otherwise the structure of the constraints is the same.

There are four Lagrangian variables ϕ^m and Λ and the corresponding momenta are

$$\pi_m = \dot{\phi}^m \tag{1.197}$$

and

$$\Pi = 0. \tag{1.198}$$

This is a case where the Lagrangian does not depend on $\dot{\Lambda}$ at all, so the primary constraint is $\Psi_1 = \Pi$. The canonical Hamiltonian is (summation convention is now used)

$$H_C = \frac{1}{2}\pi_m\pi_m + \Lambda(\phi^m\phi^m - f^2), \tag{1.199}$$

and the primary Hamiltonian is

$$H_P = H_C + \mu\Psi_1. \tag{1.200}$$

The derivation of all the secondary constraints is now a straightforward exercise in calculating Poisson brackets. We find three secondary constraints

$$\dot{\Psi}_1 = \{\Psi_1, H_P\} \approx f^2 - \phi^m\phi^m \equiv \Psi_2 \tag{1.201a}$$
$$\dot{\Psi}_2 = \{\Psi_2, H_P\} \approx -2\phi^m\pi_m \equiv \Psi_3 \tag{1.201b}$$
$$\dot{\Psi}_3 = \{\Psi_3, H_P\} \approx -2\pi_m\pi_m + 4\Lambda\phi^m\phi^m \equiv \Psi_4 \tag{1.201c}$$
$$\dot{\Psi}_4 = \{\Psi_4, H_P\} \approx 4\mu f^2 \tag{1.201d}$$

and the time-independence of the last one fixes $\mu = 0$. Thus, the Hamiltonian on the constrained submanifold is just

$$H = \frac{1}{2}\pi_a\pi_a. \tag{1.202}$$

The matrix of Poisson brackets of the constraints is

$$M = \begin{pmatrix} 0 & 0 & 0 & -4f^2 \\ 0 & 0 & 4f^2 & 0 \\ 0 & -4f^2 & 0 & 32f^2\Lambda \\ 4f^2 & 0 & -32f^2\Lambda & 0 \end{pmatrix}$$

and it is nondegenerate. From here one can calculate the Dirac brackets between the canonical variables

$$\{\phi^m, \phi^n\}_D = 0, \tag{1.203a}$$
$$\{\phi^m, \pi_n\}_D = \delta^m_n - \frac{1}{f^2}\phi^m\phi_n, \tag{1.203b}$$
$$\{\pi_m, \pi_n\}_D = -\frac{1}{f^2}(\phi_m\pi_n - \phi_n\pi_m). \tag{1.203c}$$

As in the Lagrangian treatment, this is a case where the constraints can be explicitly solved leaving a smooth phase space that is itself the cotangent bundle of a configuration space. For example if we choose polar coordinates as in (1.69), we define the momenta

$$\pi_1 = \frac{\pi_\Theta \cos\Theta\cos\Phi}{f} - \frac{\pi_\Phi \sin\Phi}{f\sin\Theta}, \tag{1.204a}$$
$$\pi_2 = \frac{\pi_\Theta \cos\Theta\sin\Phi}{f} + \frac{\pi_\Phi \cos\Phi}{f\sin\Theta}, \tag{1.204b}$$
$$\pi_3 = -\frac{\pi_\Theta \sin\Theta}{f}, \tag{1.204c}$$

where π_Θ and π_Φ are the momenta conjugate to Θ and Φ. These momenta satisfy the constraint $\Psi_3 = 0$ and the Hamiltonian (1.202) becomes

$$H = \frac{1}{2f^2}\left(\pi_\Theta^2 + \frac{1}{\sin^2\Theta}\pi_\Phi^2\right), \tag{1.205}$$

that is the correct Hamiltonian for the particle on the sphere in spherical coordinates.

It is left as an exercise to verify that the Dirac brackets (1.203) just reduce to the Poisson brackets, when the constraints are solved in terms of Θ, Φ, π_Θ, π_Φ (see Exercise 1.10).

1.7.4 Canonical treatment of Yang–Mills theory

Yang–Mills theory is the paradigm of a theory with gauge invariance, and in the canonical formulation this gauge invariance corresponds to the gauge transformations defined by first class constraints. In fact, as we shall see, this is an example of a theory where all constraints are first class.

Separating the space and time components of the field strength, the Yang Mills Lagrangian in d dimensions is

$$L_{YM} = \int d^d x \left(\frac{1}{2}E_i^a E_i^a - \frac{1}{4}F_{ij}^a F_{ij}^a\right), \tag{1.206}$$

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

$$D_i A_0^a = \partial_i A_0^a + e f_{abc} A_i^b A_0^c \tag{1.207}$$

for the 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 potentials A_0^a and A_i^a are

$$P_a^0 \equiv \frac{\partial L_{YM}}{\partial \dot{A}_0^a} = 0, \tag{1.208a}$$

$$P_a^i \equiv \frac{\partial L_{YM}}{\partial \dot{A}_i^a} = E_i^a. \tag{1.208b}$$

The second relation can be readily inverted, but the first defines the primary constraints

$$\Psi_a = P_a^0. \tag{1.209}$$

The canonical Hamiltonian can be written

$$H_C = \int d^d x \left[\frac{1}{2} P_i^a P_i^a + \frac{1}{4} F_{ij}^a F_{ij}^a + P_a^i D_i A_0^a \right], \tag{1.210}$$

and the last term can be integrated by parts to become $-A_0^a G_a$ where

$$G_a = D_i P_a^i = D_i E_i^a. \tag{1.211}$$

We have to impose that the primary constraints hold for all time. This means that $\{P_a^0(x), H\} = 0$, which results in the "secondary constraint"

$$G_a(x) = 0, \tag{1.212}$$

that is nothing but the Gauss law. We see that in the Hamiltonian formalism the fields A_0^a play the role of Lagrange multipliers enforcing the Gauss law constraint. It is useful to smear the constraint with a function ϵ^a having values in the Lie algebra of the gauge group:

$$G_\epsilon = \int d^d x \, \epsilon^a(x) G_a(x).$$

A short calculations shows that

$$\{G_\epsilon, H_C\} = 0, \tag{1.213}$$

so that no further constraints arise. The Gauss law commutes with P_0^a, and furthermore

$$\{G_a(x,t), G_b(y,t)\} = f_{ab}{}^c G_c(x,t) \delta^{(d)}(x-y), \tag{1.214}$$

or equivalently

$$\{G_{\epsilon_1}, G_{\epsilon_2}\} = G_{[\epsilon_1, \epsilon_2]}, \tag{1.215}$$

so that all the constraints commute weakly and we have a system of first class contraints only.

The primary constraints generate shifts of A_0^a, that is seen therefore to be a gauge variable. This is one example where we can see easily that also the secondary constraints generate gauge transformations. In fact

$$\{G_\epsilon, A_i^a\} = D_i \epsilon^a, \qquad \{G_\epsilon, P_a^i\} = -f_{abc} \epsilon^b P_c^i. \tag{1.216}$$

Remembering that ϵ depends on x but not on t, we conclude that the Gauss law generates time-independent gauge transformations.

When studying the canonical formulation of a YM theory it is most natural, and very convenient, to choose the gauge $A_0 = 0$.[19] In this way A_0^a and P_a^0 can be removed from the list of canonical variables, without changing the brackets between the other variables (their Dirac bracket is the Poisson bracket). However, this is only a partial gauge fixing: it leaves the freedom of performing time-independent gauge transformations. At an abstract level, the resulting phase space can be described as follows (we now assume $d = 3$). After fixing the gauge $A_0 = 0$, the configuration space is $\mathcal{C}$, the "space of connections" parametrized by the $3n$ fields A_i^a (where n is the dimension of the gauge group) and the phase space is its cotangent bundle $\mathcal{F}' = T^*\mathcal{C}$, parametrized by the $6n$ functions (A_i^a, P_a^i). The physical configuration space is the quotient $\mathcal{Q} = \mathcal{C}/\mathcal{G}$ representing the equivalence classes of gauge potentials modulo gauge transformations. Since the gauge group is parametrized by n functions, $\mathcal{Q}$ is parametrized by $2n$ functions, in agreement with the counting in perturbation theory. Following the Dirac–Bergmann procedure, we now define a space $\mathcal{F}''$ of fields satisfying the Gauss law, that is parametrized by $5n$ functions. Then we observe that if P satisfies the Gauss law with respect to A, $P + \delta P$ satisfies the Gauss law with respect to $A + \delta A$, where δA and δP are given in (1.216). Thus the action of the gauge group preserves the Gauss law and we define the physical phase space to be the quotient $\mathcal{F}''/\mathcal{G}$, that is parametrized by $4n$ functions. It consists of the gauge equivalence classes of pairs (A_i^a, P_i^a) satisfying the Gauss law, and is the cotangent bundle of $\mathcal{Q}$.

For the quantization of the theory there are also, in principle, various possibilities. In practice the procedure that is closest to the perturbative one is to gauge fix, compute the Dirac brackets of all variables and quantize by replacing them with quantum commutators. In principle the most elegant procedure would be to quantize only the degrees of freedom of the physical phase space. In this way the wave functions would have the form $\psi([A])$, where the square bracket denotes the gauge equivalence classes. In practice this is not possible, because the physical degrees of freedom are equivalence classes of gauge potentials and their momenta. An intermediate procedure is to quantize the variables in the intermediate phase space $\mathcal{F}'$, with wave functionals $\psi(A)$. One would then define the physical wave functionals to be those that are invariant under gauge transformations, i.e.

$$G_\epsilon \psi_{phys}(A) = 0. \tag{1.217}$$

[19] This can be done by performing the gauge transformation

$$g(x,t) = \mathrm{P}\exp\Big(-e\int^t dt'\, A_0(x,t')\Big),$$

where P stands for path ordering.

1.8 Exercises

Exercise 1.1: Fermionic Noether currents

Check that the currents (1.18) and (1.19) are Hermitian operators. Use the conventions for Fermi fields described in Appendix A.

Exercise 1.2: Noether currents of the $O(N)$ maodel

Write the Noether currents of the $O(N)$ transformations for the $O(N)$ model of Section 1.2.1 and verify their conservation. Do the same in the broken symmetry phase, reparametrizing the fields as $\phi^a = \pi^a$ for $a = 1, \dots, N-1$ and $\phi^N = f + \chi$. (This particular parametrization is used in Section 1.2.3.)

Exercise 1.3: Reductionism at work

The Ising model, in any dimension, can be easily solved in the mean field approximation, that consists in neglecting fluctuations. Define the average magnetization per site $m = \frac{1}{N}\sum_i S_i$ and expand $S_i = m + \delta S_i$. In the Hamiltonian (1.30) keep only terms linear in δS_i, then reexpress it in terms of S_i. At this point

$$\mathcal{H} = -J \sum_{i,j} (-m^2 + 2mS_i) - \mu H \sum_i S_i .$$

The double sum over S_i can be written as a single sum times z (the number of nearest neighbors) times one half (to compensate double counting). Discarding a constant, we have

$$\mathcal{H} = \frac{1}{2} JzNm^2 - \mu \bar{H} \sum_i S_i ,$$

where $N = \sum_i 1$ is the total number of spins and $\bar{H} = H + Jzm/\mu$. This is the Hamiltonian of N decoupled spins interacting with an effective magnetic field $\bar{H}$, which is the sum of the external field and a term representing the mean magnetic field generated by all the other spins.

The partition function of N decoupled spins is the partition function of a single spin raised to the power N:

$$Z = e^{-\beta JNzm^2/2} \left(2\cosh(\beta(\mu H + Jzm))\right)^N .$$

From here one gets the free energy density, or free energy per site

$$f = -\frac{\log Z}{\beta N} = \frac{Jz}{2}m^2 - \frac{1}{\beta}\log\left(2\cosh(\beta(\mu H + Jzm))\right).$$

The magnetization m must minimize the free energy. This leads to the equation

$$m = \tanh\left(\frac{Jz}{T}m\right).$$

This equation can be solved graphically. It has nonzero solutions for m provided $T \leq Jz$. We thus identify Jz with the critical temperature:

$$Jz = T_c.$$

Expanding the free energy in m, h and $T - T_c$ we obtain

$$f(m) = C - hm + \frac{1}{2}rm^2 + \frac{1}{3}sm^3 + \frac{1}{4}um^4 + \ldots$$

that we recognize as the Ginzburg–Landau free energy, with

$$h = \frac{T_c}{T}\mu H\,; \qquad r = T - T_c\,; \qquad s = \left(\frac{T_c}{T}\right)^3 \mu H\,; \qquad u = \frac{1}{3}\frac{T_c^4}{T^3}.$$

So this is a case in which the parameters of the low energy theory can be calculated, at least in some approximation, from those of the microscopic theory.

Exercise 1.4: Alternative chiral Lagrangian

In the linear chiral model, define the matrix-valued field

$$\Sigma = \sigma\mathbb{I} + i\pi^a\sigma_a = \begin{pmatrix} \sigma + i\pi^3 & \pi^2 + i\pi^1 \\ -\pi^2 + i\pi^1 & \sigma - i\pi^3 \end{pmatrix},$$

where σ_a are the Pauli matrices. Check that the Lagrangian

$$\begin{aligned}\mathcal{L}_{GML} = &-\frac{1}{4}\mathrm{tr}\partial_\mu\Sigma^\dagger\partial^\mu\Sigma - \frac{\lambda}{4}\left(\frac{1}{2}\mathrm{tr}\Sigma^\dagger\Sigma - f^2\right)^2 \\ &- \bar{N}_L\gamma^\mu\partial_\mu N_L - \bar{N}_R\gamma^\mu\partial_\mu N_R - 2g(\bar{N}_L\Sigma N_R + \bar{N}_R\Sigma^\dagger N_L).\end{aligned} \tag{1.218}$$

is equal to the Lagrangian (1.59). Verify that the transformations

$$\Sigma \longmapsto g_L^{-1} \Sigma g_R,$$

together with the fermion transformation (1.44), leave the action invariant, and are equivalent to the transformations (1.35), (1.34), (1.38). In this formulation chiral invariance is very easy to see.

This form of the Lagrangian is also useful to make contact with the nonlinear sigma model. In fact, when $\pi^a \pi^a + \sigma^2 = f^2$, we can write

$$\Sigma = fU$$

where $U \in SU(2)$. and then the Lagrangian (1.218) becomes identical to the sum of (1.80) and (1.86). However, the pion field in the two cases are not the same, because the matrix U defined above is linear in the pion field, whereas the matrix U in (1.83) is exponential in the pion field. The two pion fields agree at linear level, but differ by a field redefinition. Such redefinitions do not affect physical observables. See [DGH22] for more details.

Exercise 1.5: Coordinates on the sphere

The sphere S^{N-1} of radius r is the subset of $\mathbb{R}^N$ with

$$z_1^2 + ... + z_N^2 = r^2.$$

This embedding induces a metric on S^{N-1} with isometry group $O(N)$. Write the metric $h_{\alpha\beta}$ in the coordinate systems defined in (1.69), (1.71), (1.73) and confirm equations (1.70), (1.72), (1.74).

1. The first coordinate system are the familiar spherical coordinates colatitude and longitude. They are well defined everywhere except at the poles. One would need at least one other coordinate chart to cover the whole sphere (for example spherical coordinates where the x^N axis is rotated) but in practice this is never needed.

2. The second coordinate system is defined as follows: given a point P on the sphere project it vertically on the equatorial plane. Use the Cartesian coordinates of the projection $x_1,\ldots, x_{N-1}$, with $x_1^2 + ... + x_{N-1}^2 < r^2$ as independent coordinates. The map is

not invertible at the equator, so these coordinates cover only one hemisphere. There is a similar coordinate system for every hemisphere. How many such coordinate systems does one need to cover the whole sphere?

3. The third coordinate system are the stereographic coordinates, defined as follows: for any point P on S^{N-1} except the north pole, draw the line passing through that point and the north pole. The stereographic coordinates of P are the Cartesian coordinates of the intersection of that line with the tangent plane at the south pole. This coordinate system leaves out only the north pole, so to cover the whole sphere one typically uses another stereographic system based at the south pole. One could project the point on the equatorial plane instead of the plane tangent to the opposite pole, resulting in a simple rescaling of the coordinates by factors of two.

Exercise 1.6: Noether's theorems for Yang–Mills theory

For a YM theory minimally coupled to a fermion and a scalar field in arbitrary representations of the group, calculate the currents K^μ and k^μ defined in equations (1.126) and (1.131) and verify equations (1.133) and (1.134).

From Noether's first theorem, calculate the conserved current j^μ_ϵ that is associated to global gauge transformations. It is the same as the current K^μ_ϵ for constant gauge parameter. Compare to the Noether currents (1.12) and (1.18) in the absence of gauge fields, and to the covariantly conserved current J^μ_ϵ defined by (1.135). Show that the covariant conservation of J^μ_ϵ together with the YM equation of motion implies ordinary conservation of j^μ_ϵ.

Exercise 1.7: Covariant derivatives of nonlinear fields

Rewrite the second term in the Lagrangian (1.145) in spherical coordinates ρ, Θ, Φ. Identify the covariant derivatives $D_\mu\rho$, $D_\mu\Theta$, $D_\mu\Phi$ and show that they agree with the general formula (1.143), where the Killing vectors K^α_i are the usual generators of angular momentum.

Exercise 1.8: London penetration depth

Use the field equations of the action (1.146) for a superconducting half-space $x > 0$. Assume that outside the superconductor ($x < 0$) there is a constant magnetic field. Show that the orthogonal component of the magnetic field H_x must be zero for $x > 0$, whereas a tangential magnetic field will penetrate the superconductor. If the magnetic field is $H_x = H_y = 0, H_z = H$ for $x < 0$, in the interior it is He^{-x/λ_L} where $\lambda_L = 1/m_A$ is called the London penetration depth.

Exercise 1.9: Weakly vanishing functions

Let Σ be the submanifold of phase space defined by the primary constraints Φ_m. Show that if a function $F(q, p)$ on phase space is weakly zero, i.e. it is zero on the surface Σ, it satisfies (1.191).

Exercise 1.10: Dirac brackets

Using the spherical coordinates (1.69), verify that the Dirac brackets (1.203) are equivalent to the canonical Poisson brackets

$$\begin{aligned} \{\Theta, \Theta\} = \{\Theta, \Phi\} = \{\Phi, \Phi\} &= 0, \\ \{\pi_\Theta, \pi_\Theta\} = \{\pi_\Theta, \pi_\Phi\} = \{\pi_\Phi, \pi_\Phi\} &= 0, \\ \{\Theta, \pi_\Phi\} = \{\Phi, \pi_\Theta\} = 0, \qquad & \{\Theta, \pi_\Theta\} = \{\Phi, \pi_\Phi\} = 1. \end{aligned}$$

Chapter 2

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

A *soliton* is a classical solution of nonlinear field equations that

- is nonsingular,
- has finite energy and
- is localized in space.

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

In order to make this concept mathematically more precise, it is convenient to think of a field theory as a mechanical system with an infinite dimensional configuration space. Let us define the classical configuration space of the theory, $\mathcal{Q}$, to be the space of smooth, finite energy configurations of the field at some instant of time. Note that $\mathcal{Q}$ defines the kinematics of the theory, but also knows about the form of the energy. The theories that we will consider in this chapter will have the common characteristic that their configuration space is not connected. Instead, it will be the disjoint union of several connected components, indexed by a set $\pi_0(\mathcal{Q})$ (the reason for this notation is explained in Appendix E.1):

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

where $\mathcal{Q}_i$ are connected. Having determined the structure of the configuration space, the natural problem will be to find (if it exists) the absolute minimum of the static energy in each connected component. Such minima will automatically be solutions of the classical equations of motion. The minimum of the energy in some connected components will be the classical vacuum configuration, but in others it will correspond to nontrivial 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, unlike Noether currents, is conserved even without making use of the equations of motion. Associated to the topological current is the topological charge, which is a functional on $\mathcal{Q}$ that is locally constant. It is zero in the connected components containing the vacuum, and nonzero in those containing solitons.

When the system is quantized 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 that 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.

2.1 Scalar solitons in 1+1 dimensions

2.1.1 Classical kinks

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

$$S(\phi) = \int d^2x \left[-\frac{1}{2} \partial_\mu \phi \partial^\mu \phi - V(\phi) \right]. \tag{2.1}$$

The signature of the metric is such that $\partial_\mu \phi \partial^\mu \phi = -(\partial_0 \phi)^2 + (\partial_1 \phi)^2$. We demand that the potential V be bounded from below, and we assume without loss

of generality that the minimum value of V be zero. We call y_i, $i \in \mathcal{J}$, the minimum points. For definiteness one can think of the quartic potential

$$V = -\frac{1}{2}m^2\phi^2 + \frac{\lambda}{4}\phi^4 + \frac{m^4}{4\lambda} = \frac{\lambda}{4}\left(\phi^2 - f^2\right)^2, \tag{2.2}$$

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

With these assumptions, the energy:

$$E = \int_{-\infty}^{\infty} dx \left[\frac{1}{2}(\partial_0\phi)^2 + \frac{1}{2}(\partial_1\phi)^2 + V(\phi)\right] \tag{2.3}$$

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 (2.3) the first term represents the kinetic energy; the rest

$$E_S = \int_{-\infty}^{+\infty} dx \left[\frac{1}{2}(\partial_1\phi)^2 + V\right] \tag{2.4}$$

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 called *elastic energy*.

The field ϕ belongs to the space $\Gamma(\mathbb{R},\mathbb{R})$ of smooth real functions of one variable. (In general we will use the notation $\Gamma(X,Y)$ for the space of smooth maps from X to Y, where X and Y are manifolds. This space is itself an infinite dimensional smooth manifold, See Appendices G and G.1.) Finiteness of the energy demands that when $|x|$ tends to infinity, ϕ tends to one of the classical vacua, for otherwise E_S would diverge. We will call $\mathcal{Q}$ the subspace of $\Gamma(\mathbb{R},\mathbb{R})$ for which the static energy E_S is finite:

$$\mathcal{Q} = \{\phi \in \Gamma(\mathbb{R},\mathbb{R}) | E_S < \infty\}.$$

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

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

Every path in $\Gamma(\mathbb{R},\mathbb{R})$ joining $\mathcal{Q}_{ij}$ to $\mathcal{Q}_{i'j'}$ (with $ij \neq i'j'$) must necessarily pass through the complement of $\mathcal{Q}$. In fact, to change the asymptotic behaviour of ϕ one has to go through fields which do not tend to one of the minima at infinity, and these have infinite energy. So, the spaces $\mathcal{Q}_{ij}$ are separated by infinite energy barriers. For example in the case of the potential (2.2) there are

four connected components of $\mathcal{Q}$, labelled $\mathcal{Q}_{++}$, $\mathcal{Q}_{+-}$, $\mathcal{Q}_{-+}$, $\mathcal{Q}_{--}$. In general, the set $\pi_0(\mathcal{Q})$ of connected components of $\mathcal{Q}$ is the cartesian product of two copies of the set indexing the minima: $\pi_0(\mathcal{Q}) = \mathcal{J} \times \mathcal{J}$.

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

The natural problem is then to find the minimum of the energy in each connected component, if it exists. It is clear that in the "diagonal" connected components $\mathcal{Q}_{ii}$ the minima are the constant fields $\phi = y_i$. These are also the absolute minima of E on all $\mathcal{Q}$. The minima of the energy in the other sectors will be the sought-after solitons. They will interpolate between two different minima of the potential. It turns out that, of all the infinitely many degrees of freedom of the field, the crucial one is the "size of the soliton" (denoted ℓ), i.e. the length of the region where the field is significantly different from either minimum. The following argument then indicates that with the dynamics considered above there is an optimal size that minimizes the static energy. For definiteness let us consider the potential (2.2), The elastic energy is of order f^2/ℓ, and hence decreases with ℓ, while the potential energy is of order $\lambda f^4 \ell$, and hence increases with ℓ. The static energy will have a minimum at some finite value of order $\ell \approx 1/(\sqrt{\lambda} f)$. Inserting in the formula for the energy we also find that both elastic and potential energy of the soliton are of order $\sqrt{\lambda} f^3$. The soliton will therefore be the result of a balance between elastic and potential energy.

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

$$\frac{d^2\phi}{dx^2} = \frac{\partial V}{\partial \phi}, \tag{2.5}$$

with the appropriate boundary conditions. For the potential (2.2) the solutions of (2.5) in the sectors $\mathcal{Q}_{-+}$ and $\mathcal{Q}_{+-}$ are

$$\phi(x) = \pm \frac{m}{\sqrt{\lambda}} \tanh\left[\frac{m}{\sqrt{2}}(x - x_0)\right], \tag{2.6}$$

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 they

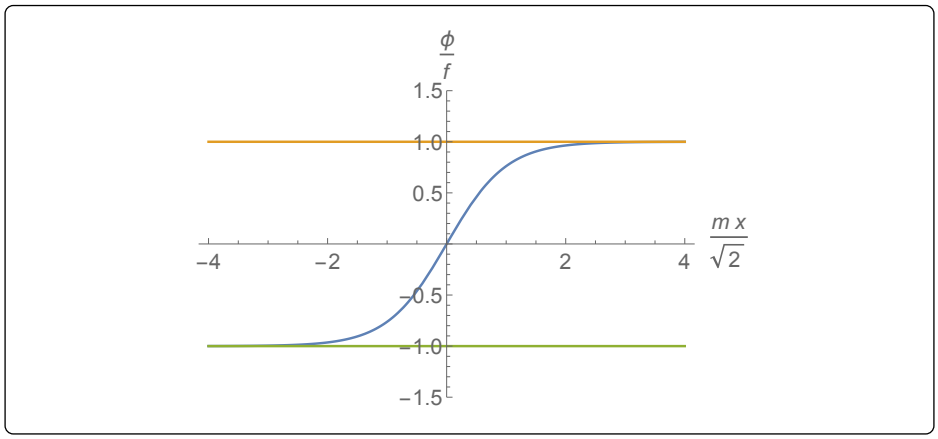


Figure 6. The kink of ϕ^4 theory.

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 6 shows a plot of ϕ/f as a function of $x\sqrt{2}/m$ for the kink at $x_0 = 0$. (The horizontal lines correspond to the minima of the potential.)

Inserting (2.6) in (2.4) we obtain

$$E_S = \frac{2\sqrt{2}m^3}{3\lambda} = \frac{2\sqrt{2}}{3} f^3 \sqrt{\lambda}. \tag{2.7}$$

It is useful to note that there is equipartition between elastic and potential energy (i.e. each of the two terms in (2.4) contributes exactly $E_S/2$). To see this, multiply both sides of the equation of motion (2.5) by $\frac{d\phi}{dx}$. The resulting equation can be written

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

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

In the theory with potential (2.2), consider the current

$$J^\mu_T = \frac{1}{2f}\varepsilon^{\mu\nu}\partial_\nu\phi\ ; \tag{2.8}$$

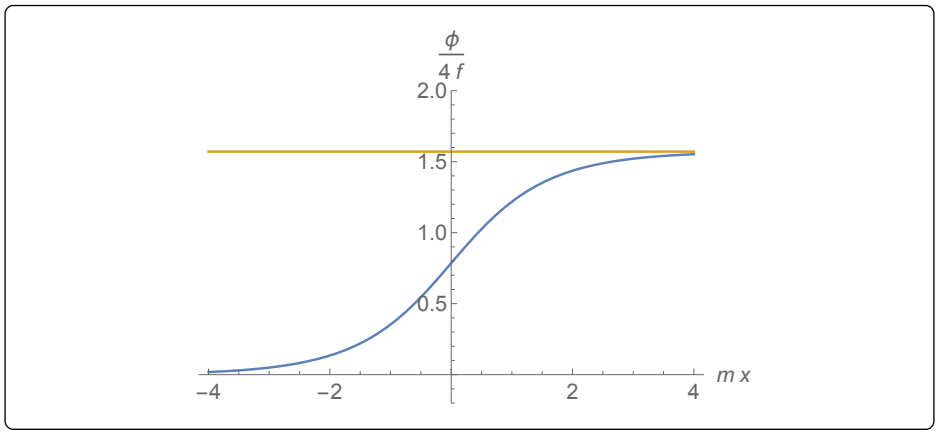


Figure 7. The kink of the sine–Gordon model.

clearly we have

$$\partial_\mu J_T^\mu = 0. \tag{2.9}$$

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

$$Q_T = \int_{-\infty}^{\infty} dx\, J_T^0 = \frac{1}{2f}\big[\phi(+\infty) - \phi(-\infty)\big] \tag{2.10}$$

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

Another interesting potential is

$$V(\phi) = \frac{m^4}{\lambda}\left[1 - \cos\left(\frac{\sqrt{\lambda}}{m}\phi\right)\right]. \tag{2.11}$$

This corresponds to the so called "sine–Gordon" (SG) model. The indexing set of minima is the set of the integers $\mathcal{J} = \mathbb{Z}$, 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 (2.8) and (2.10), 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 $\mathcal{Q}_{0\,1}$ and $\mathcal{Q}_{0\,-1}$

$$\phi(x) = \pm\frac{4m}{\sqrt{\lambda}} \arctan\{\exp[(x - x_0)m]\}. \tag{2.12}$$

This solution is plotted in Figure 7. Its static energy is

$$E_S = \frac{8m^3}{\lambda}. \tag{2.13}$$

Just adding $2nf$, we get the soliton and antisoliton, still with $Q_T = \pm 1$, which minimize the energy in $\mathcal{Q}_{n\,n+1}$ and $\mathcal{Q}_{n\,n-1}$. What about the sectors $\mathcal{Q}_{n\,n\pm k}$, with $k > 1$? If in the field equation (2.5) we reinterpret x as time and ϕ as the coordinate of a particle on a line, then we can regard it as Newton's equation of motion of the particle moving in a gravitational potential $-V$. Formula (2.12) represents a particle rolling from one maximum of the gravitational potential to the next. This motion takes infinite time, so the particle can never reach beyond that point. Using this analogy it becomes intuitively clear that there cannot be any static soliton of the SG model with $|Q_T| > 1$. Note that this reinterpretation links a field theory in $1+1$ dimensions to mechanics (a field theory in $0+1$ dimensions). In Chapter 3 we shall frequently use this trick of relating theories differing by one in dimension.

2.1.2 Quantum kinks

The kink is a solution of the equations of motion in a classical field theory. We now discuss the meaning of the kink in the quantum theory [DHN74a, DHN74b, GoJ74]. We start by recalling that when the potential is written in terms of the mass m and coupling constant λ, the nontrivial vacua occur at fields

$$\bar{\phi} = a\frac{m}{\sqrt{\lambda}}, \tag{2.14}$$

with $a = 1$ for the ϕ^4 theory and $a = 2\pi$ for the SG model. The occurrence of the inverse coupling is a sign that the nontrivial vacua, and the ensuing symmetry breaking, are of nonperturbative nature. These vacua are usually described in a semiclassical way by assuming that the constant classical field $\bar{\phi}$ is an approximation of the true vacuum expectation value of the quantum field, $\langle\phi\rangle$. Then one splits the full quantum field into a classical and quantum part

$$\phi(x) = \bar{\phi} + \varphi(x). \tag{2.15}$$

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

$$m_\pi = \sqrt{V''(\bar{\phi})} = b\,m. \tag{2.16}$$

with $b = \sqrt{2}$ for the ϕ^4 theory and $b = 1$ for the SG model. Note that ϕ is dimensionless and λ has dimensions of mass squared. In these theories weak coupling means $\lambda \ll m_\pi^2$.

This procedure deals with the "diagonal" classical sectors Q_{ii}, that contain the absolute minima of the energy. It can be extended also to the nontrivial soliton sectors, where the minimum energy is attained at the kinks. The properties of the kink are such that it is natural to interpret it as a particle. First of all, it is a localized excitation of the field with finite energy, and it is stable due to the conservation of the topological charge. Since the theory is Lorentz invariant, applying a boost to a static kink one obtains another solution describing a kink with finite momentum. Thus kinks can move in space, like any particle. If we think of the kink as a particle, its rest mass is the total energy of the solution. From (2.7), (2.13) and (2.16), the mass of the kink is

$$m_k = c\,\frac{m_\pi^3}{\lambda}. \tag{2.17}$$

where $c = 1/3$ for ϕ^4 and $c = 8$ for SG. Considering the ratio m_k/m_π, we see that the solitons are much heavier than the pions at weak coupling.

It is natural to interpret all this by saying that the quantum theory contains, in addition to the pions, also another type of much heavier particles that are not accessible in perturbation theory. Indeed, like the vacua that they interpolate, the kink solutions are themselves of order $m/\sqrt{\lambda}$, see (2.6), (2.12). We can use again the background field split

$$\phi(x) = \bar{\phi}(x) + \varphi(x), \tag{2.18}$$

and write a functional integral over the shifted field φ, that has trivial boundary conditions $\varphi \to 0$ for $x \to \pm\infty$. To this effect, we start by expanding the action around the background:

$$\begin{aligned} S(\phi) &= \int dt \left[\frac{1}{2}\int dx \left(\frac{d\phi}{dt}\right)^2 - E_S(\phi)\right] \\ &= S(\bar{\phi}) + \int dt\, dx \left[\frac{1}{2}\left(\frac{d\varphi}{dt}\right)^2 - \frac{1}{2}\varphi L \varphi - \lambda\left(\bar{\phi}\varphi^3 + \frac{1}{4}\varphi^4\right)\right] \end{aligned} \tag{2.19}$$

where

$$L = -\frac{d^2}{dx^2} + V''(\bar{\phi}). \tag{2.20}$$

The terms on the r.h.s. of (2.19) are ordered in powers of λ: the term $S(\bar{\phi}) = -m_k \int dt$ is of order λ^{-1} and hence non-perturbative; the first two

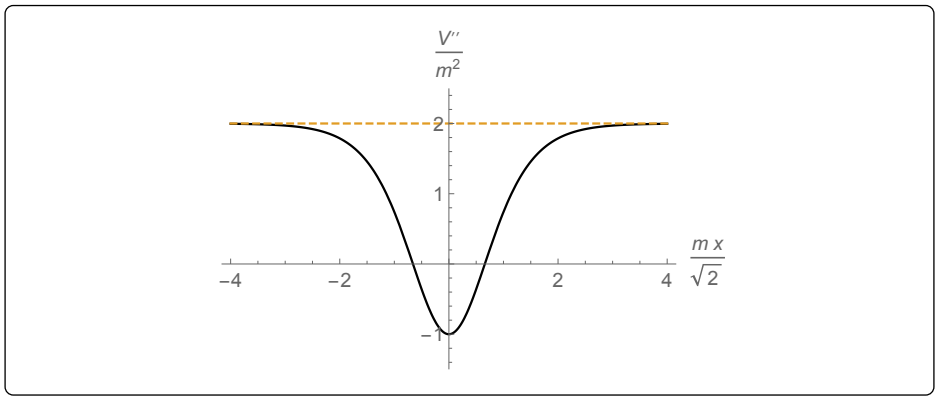


Figure 8. The potential in the operator L.

terms in the square bracket are of order λ^0, the term cubic in φ is of order $\sqrt{\lambda}$ ($\bar{\phi}$ contains a factor $\lambda^{-1/2}$) and the term quartic in φ is of order λ. We are going to evaluate quantum corrections at order λ^0, which is equivalent to a standard saddle point (one-loop) evaluation of the path integral.

We need to know the spectrum of the operator L:

$$L\eta_n = \omega_n^2 \eta_n. \tag{2.21}$$

(We use 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, as we shall see shortly.)

In the diagonal sectors Q_{ii}, $V''(\bar{\phi})$ is just a mass term, so the spectrum of L is $\omega_p^2 = m_\pi^2 + p^2$, with $-\infty < p < \infty$. The eigenfunctions are left- and right-moving plane waves that describe the motion of free particles. All the technical complications of the soliton problem derive from the fact that the "mass" term in the operator L is actually a function of x. In ϕ^4 theory

$$\begin{aligned} V''(\bar{\phi}) &= \lambda(3\bar{\phi}^2 - f^2) \\ &= \frac{1}{2}m_\pi^2\left(2 - \frac{3}{\cosh^2\left(\frac{m_\pi x}{2}\right)}\right). \end{aligned} \tag{2.22}$$

This function is shown in Figure 8. Away from the position of the kink it tends quickly to m_π^2, but near the kink it has a dip and becomes even negative. In the SG model we have $V''(\bar{\phi}) = m^2\cos[\arctan(e^{mx})]$, that can be rewritten in the form

$$V''(\bar{\phi}) = m^2\left(1 - \frac{2}{\cosh^2(mx)}\right), \tag{2.23}$$

and is qualitatively very similar to the one in Figure 8.

The presence of the kink deforms the spectrum but in a rather simple way. By simple rescalings, the eigenvalue problems of the operator L in the two theories are special cases of the equation

$$\left[-\frac{d^2}{dz^2} + \nu^2 - \frac{\nu(\nu+1)}{\cosh^2 z}\right]\eta_n = \omega_n^2 \eta_n, \tag{2.24}$$

with $\nu = 2$ for ϕ^4 and $\nu = 1$ for the SG model. This can be interpreted as the Schrödinger equation for a particle in the potential $U(x) = V''(\bar{\phi}(x))$ of the Pöschl–Teller form. This is a classic problem in quantum mechanics. The operator L is a self-adjoint, second order differential operator and therefore its eigenfunctions η_n form a basis for the space of square-integrable functions on the real line. The spectrum for the ϕ^4 theory consists of the following:

- an isolated eigenvalue $\omega_0^2 = 0$ with eigenfunction $\eta_0 = \frac{1}{\cosh^2\left(\frac{m_\pi x}{2}\right)}$;
- an isolated eigenvalue $\omega_1^2 = \frac{3}{2}m^2$ with eigenfunction $\eta_1 = \frac{\sinh\left(\frac{m_\pi x}{2}\right)}{\cosh^2\left(\frac{m_\pi x}{2}\right)}$,
- a continuous spectrum $\bar{\omega}_p^2 = m_\pi^2 + p^2$, with $-\infty < p < \infty$.

In the SG theory the spectrum is almost the same, except that the zero mode is $\eta_0 = \frac{1}{\cosh(mx)}$ and the isolated mode with eigenvalue ω_1^2 is absent.

The discrete mode η_1 has to be interpreted as an excited state of the kink. The continuous spectrum corresponds to pions propagating in the background of the kink and can be described as follows. A "right-moving" mode with momentum $p > 0$ is given for large negative x by $\bar{\eta}_p(x) \approx e^{ipx}$. Near the soliton the solution is more complicated, but it must have again a similar form for large positive x. It turns out that there is no reflected wave, and the transmitted wave, for large positive x is simply

$$\bar{\eta}_p(x) \approx e^{ipx + i\delta_p} \tag{2.25}$$

where the phase shift is given, for ϕ^4 theory, by

$$e^{i\delta_p} = \left(\frac{1 + ip/m_\pi}{1 - ip/m_\pi}\right)\left(\frac{1 + 2ip/m_\pi}{1 - 2ip/m_\pi}\right). \tag{2.26}$$

See Exercise 2.3. There is another eigenfunction with the same eigenvalue ω^2, which is given by the "left-mover" $\eta_p(-x)$. The general solution with given p is a linear combination of left- and right-moving waves:

$$A\bar{\eta}_p(x) + B\bar{\eta}_p(-x). \tag{2.27}$$

For later use, 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 (2.27) vanishes at $x = \pm L/2$ leads to $A = \pm B$ and $\bar{\eta}_p(L/2) = \pm\bar{\eta}_p(-L/2)$. Then, using the asymptotic behavior of the solutions, one obtains $\exp\left(ipL - i\delta_p\right) = \pm 1$, or

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

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

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

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

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

$$\varphi(t, x) = b_0(t)\eta_0(x) + b_1(t)\eta_1(x) + \sum_n a_n(t)\bar{\eta}_n(x), \tag{2.28}$$

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

$$\begin{aligned} H &= \int dx \left[\frac{1}{2}\dot{\varphi}^2 + \frac{1}{2}\varphi L \varphi\right] \\ &= \frac{1}{2}\dot{b}_0^2 + \frac{1}{2}\left(\dot{b}_1^2 + \omega_1^2 a_n^2\right) + \frac{1}{2}\sum_n \left(\dot{a}_n^2 + \bar{\omega}_n^2 a_n^2\right). \end{aligned} \tag{2.29}$$

By working in the basis of eigenfunctions of L we have managed to decompose the system into infinitely many decoupled degrees of freedom. Almost all of these are harmonic oscillators, but there is only one, namely the zero mode, which is not. Since the potential for this mode is zero, its wave function will not remain localized near the center of the soliton. This would invalidate our interpretation of the kink as a localized quantum state. Furthermore, it would give an infrared divergence in the evaluation of the effective action for the theory. The physical origin of the zero mode can be understood by noting that η_0 is proportional to the derivative of the classical solution. Among all possible deformations of the kink field, there is one that corresponds simply to an infinitesimal translation of the kink by δx:

$$\delta\phi(x) = \delta x \frac{d\bar{\phi}}{dx}. \tag{2.30}$$

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

This suggests that instead of quantizing the zero mode b_0, we quantize the position of the center of the kink. To do this, let us consider a slowly moving kink, which can be described by the solution (2.6), with x_0 replaced by $x_0(t)$: $\phi(x,t) = \bar{\phi}(x - x_0(t))$.[1] Inserting in the action we find

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

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

$$S = \int dt \left[\frac{1}{2} m_k \dot{x}_0^2 - m_k\right]. \tag{2.31}$$

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

$$H = m_k + \frac{p^2}{2m_k}. \tag{2.32}$$

This collective degree of freedom can be quantized simply imposing the standard commutation relation $[x_0, p] = i\hbar$.[2] When the motion of the kink is taken into account in this way, in order to avoid double counting, we must remove the zero mode from (2.29).

2.1.3 Renormalization of the kink mass

As a concrete calculation, we shall now investigate the quantum corrections to the kink mass. For definiteness we consider ϕ^4 theory. The theory is

[1] 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.

[2] 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.

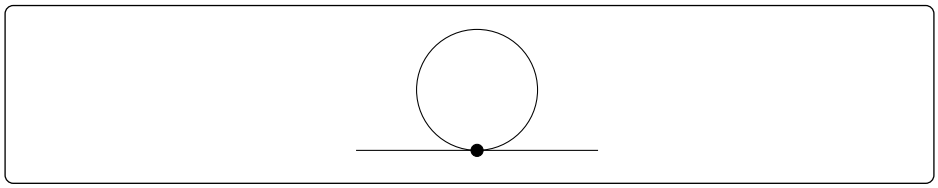


Figure 9. Renormalization of the pion mass.

superrenormalizable. The only divergence is logarithmic and renormalizes the pion mass, see Figure 9. Evaluation of this diagram gives for the renormalized mass

$$m_{\pi R}^2 = m_\pi^2 - \frac{3\lambda}{2\pi} \log\left(\frac{\Lambda^2}{m_\pi^2}\right), \tag{2.33}$$

where we employed a simple momentum cutoff Λ. Now recall that the kink mass is related to pion mass by

$$m_k = \frac{m_\pi^3}{3\lambda}. \tag{2.34}$$

We may wonder what happens to this relation when the pion mass gets renormalized.

From the discussion in the previous section, the energy of the quantum state describing a kink at rest, with the pion field in the Fock vacuum, is

$$H = m_k + \frac{\sqrt{3}}{4} m_\pi + \sum_n \frac{1}{2}\bar{\omega}_n, \tag{2.35}$$

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 energies of all the oscillators in the discretized continuous spectrum. For large n, $\bar{\omega}_n \sim n$, so the sum is quadratically divergent. This is the usual divergent contribution to the vacuum energy that one also encounters in any quantum field theory. It is also present in the vacuum sectors $\mathcal{Q}_{++}$ and $\mathcal{Q}_{--}$. We are thus led to define the renormalization of the kink mass as the difference between the sum of the vacuum energies of all the oscillators in the presence of the kink and the sum of the vacuum energies of all the oscillators in the absence of the kink. Both sums are quadratically divergent, and in the difference the quadratic divergences cancel. The renormalization of the kink mass is

therefore

$$\begin{aligned}\delta m_k &= \frac{\sqrt{3}}{4} m_\pi + \frac{1}{2}\sum_n (\bar{\omega}_n - \omega_n) \\ &= \frac{\sqrt{3}}{4} m_\pi + \frac{1}{2}\sum_n \frac{p_n \delta_p}{L\omega_n},\end{aligned} \tag{2.36}$$

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

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

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

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

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

Using the explicit form of the phase shift given in (2.26), we find

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

A direct calculation then yields for the renormalized kink mass, up to finite terms,

$$m_{kR} = m_k + \delta m_k = m_k - \frac{3}{4\pi} m_\pi \log\left(\frac{\Lambda^2}{m_\pi^2}\right). \tag{2.37}$$

For the unrenormalized mass on the r.h.s. we now use equation (2.34), which we can reexpress in terms of the renormalized pion mass, to first order in λ/m_π^2, as

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

We see that the logarithmic divergence cancels, so that the relation (2.34) is preserved under renormalization:

$$m_{kR} = \frac{m_{\pi R}^3}{3\lambda}. \tag{2.38}$$

2.1.4 Fractional charge

Peculiar phenomena happen when fermions propagate in the background of a kink. In this section we consider the scalar theory with potential (2.2) and couple it to a Dirac fermion, a complex two-component field ψ with Lagrangian

$$\mathcal{L}_F = -\bar{\psi}(\gamma^\mu \partial_\mu + g\phi)\psi \tag{2.39}$$

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

$$\psi \to e^{i\alpha}\psi\,; \qquad \bar{\psi} \to e^{-i\alpha}\bar{\psi} \tag{2.40}$$

as well as the discrete transformation

$$\phi \to -\phi\,; \qquad \psi \to \gamma_A \psi\,; \qquad \bar{\psi} \to -\bar{\psi}\gamma_A \tag{2.41}$$

where $\gamma_A = \gamma^0\gamma^1$ is the chirality operator. This $\mathbb{Z}_2$ symmetry is broken in the scalar vacuum $\phi = \pm f$, where the fermion acquires a mass $m_F = gf$.[3]

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

$$\psi = \int \frac{dp}{2\pi}\frac{1}{\sqrt{2E}}\left[b_p e^{-iEt}u_p(x) + d_p^\dagger e^{iEt}v_p(x)\right]. \tag{2.42}$$

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

$$u_p(x) = e^{ipx}\begin{pmatrix} \sqrt{E} \\ \frac{-p-im_F}{\sqrt{E}} \end{pmatrix}; \qquad v_p(x) = e^{-ipx}\begin{pmatrix} \sqrt{E} \\ \frac{-p+im_F}{\sqrt{E}} \end{pmatrix}. \tag{2.43}$$

The field is quantized by imposing the canonical anticommutation relations

$$\{b_p, b_{p'}^\dagger\} = 2\pi\delta(p-p')\,; \qquad \{d_p, d_{p'}^\dagger\} = 2\pi\delta(p-p').$$

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

For the fermion current it is best to use the definition

$$j^\mu = \frac{1}{2}\left(\bar{\psi}\gamma^\mu\psi - \bar{\psi}^c\gamma^\mu\psi^c\right), \tag{2.44}$$

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

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

$$Q = \int \frac{dp}{2\pi}\left(b_p^\dagger b_p - d_p^\dagger d_p\right), \tag{2.45}$$

whereas the Hamiltonian is given by

$$H = \int \frac{dp}{2\pi} E_p \left(b_p^\dagger b_p + d_p^\dagger d_p\right). \tag{2.46}$$

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

$$\begin{pmatrix} P^\dagger & \partial_t \\ -\partial_t & P \end{pmatrix} \qquad \text{where} \qquad P = \partial_x + g\bar{\phi},\ \ P^\dagger = -\partial_x + g\bar{\phi}. \tag{2.47}$$

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:

$$\begin{pmatrix} -P & \partial_t \\ -\partial_t & -P^\dagger \end{pmatrix}\begin{pmatrix} P^\dagger & \partial_t \\ -\partial_t & P \end{pmatrix} = \begin{pmatrix} -\partial_t^2 - PP^\dagger & 0 \\ 0 & -\partial_t^2 - P^\dagger P \end{pmatrix} \tag{2.48}$$

where

$$P^\dagger P = -\partial_x^2 + g^2\bar{\phi}^2 - g\partial_x\bar{\phi}, \qquad PP^\dagger = -\partial_x^2 + g^2\bar{\phi}^2 + g\partial_x\bar{\phi}. \tag{2.49}$$

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

$$L = \begin{pmatrix} PP^\dagger & 0 \\ 0 & P^\dagger P \end{pmatrix} \tag{2.50}$$

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

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

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

and demand that these functions are annihilated by $\partial_t^2 1 + L$. This implies that $\tilde{u}_1$ must be an eigenfunction of $PP^\dagger$ with eigenvalue E^2 and $\tilde{u}_2$ 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 $C_1 C_2 = 1/E^2$.[4]

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

$$\psi = b_0 \begin{pmatrix} 0 \\ u_0(x) \end{pmatrix} + \int \frac{dp}{2\pi} \frac{1}{\sqrt{2E}} \left[b_p e^{-iEt} \tilde{u}_p(x) + d_p^\dagger e^{iEt} \tilde{v}_p(x) \right]. \qquad (2.51)$$

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 (2.46), with the integral extending over all the non-zero modes. The zero mode is a discrete fermionic degree of freedom, so its creation and annihilation operators satisfy $\{b_0, b_0^\dagger\} = 1$. It can be in only 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 $|0'\rangle = b_0^\dagger |0\rangle$.

[4] In the case $\bar{\phi} = f$ these relations are satisfied by the solutions in (2.43), with $C_2 = -i/E$.

The surprise comes when we consider the charge of these states. When the decomposition (2.51) 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 (2.45). However, the zero mode does not have a partner and its contribution is different:

$$\frac{1}{2}\left(b_0^\dagger b_0 - b_0 b_0^\dagger\right) = b_0^\dagger b_0 - \frac{1}{2} \tag{2.52}$$

In the vacuum state where the zero mode is empty

$$Q|0\rangle = -\frac{1}{2}|0\rangle \tag{2.53}$$

while in the vacuum state where the zero mode is occupied

$$Q|0'\rangle = \frac{1}{2}|0'\rangle \tag{2.54}$$

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 can say that in the presence of the fermion field the kink itself carries a charge equal to $\pm 1/2$.

This phenomenon has analogues for the other, more complicated solitons that we will introduce in the following, but we shall not discuss it further.

2.2 Linear scalar fields in other dimensions

2.2.1 Domain walls

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

$$\sum_i \partial_i^2 \phi = V', \tag{2.55}$$

If we make an ansatz for the field

$$\phi(x_1, \dots, x_d) = \phi(x_1), \tag{2.56}$$

the equation of motion reduces to that of a scalar in one dimension. Thus, inserting in the ansatz the solutions we found in Section 2.1.1, we obtain 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 of codimension one where the scalar field vanishes. Domain walls are not solitons, because the energy of the solution is infinite:

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

represents a surface density of energy. For example, for the potential (2.2), one has from (2.7)[5]

$$\mathcal{E} = \frac{2\sqrt{2}}{3} f^3 \sqrt{\lambda}. \tag{2.57}$$

Domain walls can appear in cosmology when the universe undergoes a transition involving breaking of $\mathbb{Z}_2$ symmetry.

2.2.2 No go theorems

The existence of topological solitons requires that the configuration space has more than one connected component 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 $d > 1$ space 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 $r = |\vec{x}| \to \infty$, ϕ tends to one of the minima of V. Thus the configuration space $\mathcal{Q}$ will consist again of various connected components:

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

and $\mathcal{J}$ is the set of the minima of V. The absolute minimum of E_S in each $\mathcal{Q}_i$ is given by the constant $\phi = y_i$. These are just the classical vacua of the model. The essential difference with the case of the previous section is that in $d = 1$ the "sphere at infinity" S^0_∞ defined by the limit $r \to \infty$ consists of two

[5] One has to bear in mind that the dimension of f and λ is now different from Section 2.1.1, so that $\mathcal{E}$ has the correct dimension d in mass.

disconnected points, and the field can take different values at these two points, whereas in $d \geq 2$ the "sphere at infinity" S^{d-1}_∞ is connected. By continuity, the value of the field at infinity must be constant and there cannot be solutions with nontrivial boundary conditions.

Let us next consider the $O(N)$ model (1.21) with potential (1.25). The locus of the minima of the potential is a sphere S^{N-1}. The static energy is now

$$E_S = \int d^d x \left[\frac{1}{2} \partial_i \phi^a \partial_i \phi^a + V(|\phi|) \right]. \tag{2.58}$$

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

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

$$\tau = \frac{1}{2} + \frac{1}{\pi} \arctan t. \tag{2.59}$$

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

$$E_K = \frac{1}{2} \int d^d x \left(\frac{d\hat{\phi}}{dt} \right)^2$$

is the kinetic energy. Since $\frac{d\hat{\phi}}{dt}$ does not tend to zero as $r \to \infty$, it is clear that for finite t, E_K is divergent. We conclude that to go from ϕ to ϕ' one must go through configurations with infinite kinetic energy, so the boundary value of ϕ cannot change in the course of the time evolution. For this reason, we will always assume that the configuration space consists of fields with a fixed boundary condition at infinity. Using the $O(N)$ invariance of the theory, we can assume without loss of generality that the value of ϕ as $r \to \infty$ is $y_0 = (0, 0, \ldots, 0, f)$.

The limit $r \to \infty$ defines a "sphere at infinity" S^{d-1}_∞ and since the map ϕ must be constant on S^{d-1}_∞, all its points may be identified to a single point ∞.

Then ϕ may be regarded as a map from the one-point compactification

$$\mathbb{R}^d \cup \{\infty\} = S^d$$

into $\mathbb{R}^N$, mapping the "basepoint" ∞ of S^d to the "basepoint" y_0 of $\mathbb{R}^N$. Therefore $\mathcal{Q} = \Gamma_*(S^d, \mathbb{R}^N)$. All maps with these properties are homotopic to one another, so the space $\mathcal{Q}$ is connected.

Some remarks are in order at this point:

1. Fixing the boundary conditions is physically natural, and it is also natural from the point of view of homotopy theory, because maps preserving the basepoints have better properties, see Appendix E.1.

2. Having fixed the boundary conditions, the vacuum becomes unique. The only constant field in $\mathcal{Q}$ is the field that is everywhere equal to its boundary value.

3. The group $O(N)$ maps $\mathcal{Q}$ to a space $\mathcal{Q}'$ characterized by different boundary conditions. Fixing the boundary conditions breaks the symmetry group $O(N)$ to $O(N-1)$.

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*, implying that linear scalar field theories with action (1.21) 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 (2.58) as $E_S = E_1 + E_2$, where E_1 and E_2 are the elastic and potential energy, in the terminology introduced in the previous section. Let ϕ_λ be a one-parameter family of configurations defined by $\phi_\lambda(x) = \phi_1(\lambda x)$. We have

$$E_1(\phi_\lambda) = \lambda^{2-d} E_1(\phi_1), \qquad E_2(\phi_\lambda) = \lambda^{-d} E_2(\phi_1). \tag{2.60}$$

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

$$0 = \frac{d}{d\lambda} E_S(\phi_\lambda)\Big|_{\lambda=1} = (2-d)E_1(\phi_1) - dE_2(\phi_1). \tag{2.61}$$

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

For $d = 2$ we get $E_2(\phi_1) = 0$. This means that the field must be everywhere in the minimum of V, which implies that $\frac{\partial V}{\partial \phi^a} = 0$. Inserting in the equation of motion we obtain $\partial_\mu \partial^\mu \phi^a = 0$, which, together with the given boundary conditions, implies again $\phi =$ 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 Sections 2.6 and 2.7. Another way is to consider nonlinear scalar theories, and this is what we do next.

2.3 The $O(3)$ nonlinear scalar in $d = 2$

Derrick's theorem forbids the existence of solitons in linear scalar theories in $d > 1$. As discussed in Section 1.3.2, the $O(N)$ nonlinear sigma models can be viewed as low energy limits of the linear $O(N)$ models in the broken symmetry phase. Let us ask whether the nonlinear sigma models could have nontrivial solutions in $d > 1$. All solutions of the nonlinear sigma model have $E_2 = 0$, so (2.61) implies that if $d > 2$ the only static solution of the field equations is constant, but 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 $d = 2$.

2.3.1 Topology

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

$$E_S = \frac{f^2}{2} \int d^2x \, \partial_i \varphi^\alpha \partial_i \varphi^\beta h_{\alpha\beta}(\varphi) \tag{2.62}$$

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

$$\pi_0(\mathcal{Q}) = \pi_2(S^2) = \mathbb{Z}$$

(see Appendix E.3). So we can write

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

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

$$W(\varphi^\alpha) = \frac{1}{8\pi}\int d^2x\, \varepsilon^{ij}\partial_i\varphi^\alpha\partial_j\varphi^\beta\sqrt{\det h}\,\varepsilon_{\alpha\beta}. \tag{2.63}$$

For example, in the coordinate systems defined in (1.69), (1.71), (1.73), it has the expressions

$$W(\Theta,\Phi) = \frac{1}{4\pi}\int d^2x\, \sin\Theta\varepsilon^{ij}\partial_i\Theta\partial_j\Phi, \tag{2.64a}$$

$$W(\varphi_1,\varphi_2) = \frac{1}{4\pi}\int d^2x\, \frac{1}{\sqrt{1-\varphi_1^2-\varphi_2^2}}\varepsilon^{ij}\partial_i\varphi_1\partial_j\varphi_2, \tag{2.64b}$$

$$W(\omega_1,\omega_2) = \frac{1}{4\pi}\int d^2x\, \frac{4}{\left(\omega_1^2+\omega_2^2+1\right)^2}\varepsilon^{ij}\partial_i\omega_1\partial_j\omega_2, \tag{2.64c}$$

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

$$\sqrt{\det h(\varphi)}\epsilon_{\alpha\beta} = \omega_{\alpha\beta}(\varphi) \tag{2.65}$$

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

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

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

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

since the exterior derivative of the form ω vanishes. Thus W is a functional on $\mathcal{Q}$ that is constant on each connected component $\mathcal{Q}_n$. We shall encounter in the next section explicit solutions of the field equations for which one can check, by explicit calculation, that $W = n$ is an integer. Then, W is constant

and equal to n for all fields belonging to the same connected component $\mathcal{Q}_n$. A more general definition of the winding number and a theorem proving its integrality are discussed in Appendix E.2.

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

$$J_T^\lambda = \frac{1}{8\pi}\varepsilon^{\lambda\mu\nu}\partial_\mu\varphi^\alpha\partial_\nu\varphi^\beta\omega_{\alpha\beta}, \tag{2.66}$$

which is identically conserved:

$$\partial_\lambda J_T^\lambda = \frac{1}{8\pi}\varepsilon^{\lambda\mu\nu}\partial_\lambda\varphi^\gamma\partial_\mu\varphi^\alpha\partial_\nu\varphi^\beta\partial_\gamma\omega_{\alpha\beta} = 0, \tag{2.67}$$

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

$$Q_T = \int d^2x\, J_T^0 = \frac{1}{8\pi}\int d^2x\, \varepsilon^{ij}\partial_i\varphi^\alpha\partial_j\varphi^\beta\omega_{\alpha\beta} = W(\varphi). \tag{2.68}$$

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

2.3.2 Dynamics

Let us look at the absolute minimum of the static energy (2.62) in each topological sector $\mathcal{Q}_i$. Consider the following inequality [Pol75]

$$\begin{aligned} 0 &\le \int d^2x\, h_{\alpha\beta}\Big(\partial_i\varphi^\alpha \pm \varepsilon_{ik}\partial_k\varphi^\gamma\omega_\gamma{}^\alpha\Big)\Big(\partial_i\varphi^\beta \pm \varepsilon_{ij}\partial_j\varphi^\epsilon\omega_\epsilon{}^\beta\Big) \\ &= \int d^2x\Big[2h_{\alpha\beta}\partial_i\varphi^\alpha\partial_i\varphi^\beta \mp 2\varepsilon_{ij}\partial_i\varphi^\alpha\partial_j\varphi^\epsilon\omega_{\alpha\epsilon}\Big] \\ &= \frac{4}{f^2}E_S \mp 16\pi W, \end{aligned} \tag{2.69}$$

where in the product of the last two terms we used

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

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

$$E_S \ge 4\pi f^2|W|. \tag{2.70}$$

Furthermore, equality holds if and only if

$$\partial_i \varphi^\alpha = \mp \varepsilon_{ik} \partial_k \varphi^\gamma \varepsilon_{\gamma\delta} h^{\delta\alpha} \sqrt{\det h}. \tag{2.71}$$

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 (2.71) are first order equations, and therefore simpler than the Euler–Lagrange equations (1.78).

At this point it is convenient to specialize the discussion to stereographic coordinates ω_1 and ω_2. Equation (2.71) reduces to

$$\partial_i \omega_\alpha = \mp \varepsilon_{ik} \partial_k \omega_\gamma \varepsilon_{\gamma\alpha}, \tag{2.72}$$

and spelling these out

$$\begin{aligned} \partial_1 \omega_1 &= \pm \partial_2 \omega_2, \\ \partial_2 \omega_1 &= \mp \partial_1 \omega_2. \end{aligned} \tag{2.73}$$

If we define $\omega = \omega^1 + i\,\omega^2$ and $z = x^1 + i\,x^2$ we recognize (2.73) as the Cauchy–Riemann equations for the function $\omega = \omega(z)$. The solutions are the analytic or antianalytic functions, depending on the sign in (2.73). For example $\omega(z) = z^n$ and $\omega(z) = (z^*)^n$, with $n \geq 0$, are solutions of (2.73). Note that for large $|z|$, ω does not tend to an angle-independent limit, but since $|\omega| \to \infty$ it does not matter since all these points represent the north pole of S^2. These functions describe smooth maps $\varphi \in \Gamma_*(S^2, S^2)$ with winding number $W = n$ and $W = -n$ respectively. They are absolute minima of the static energy in the sectors $\mathcal{Q}_n$ and $\mathcal{Q}_{-n}$ respectively ($n \geq 0$).

The static energy (2.62) is invariant under space translations and rotations, dilatations and internal $O(3)$ rotations, so applying these transformations to the solutions we get other solutions. The collective coordinates parameterizing the physically distinct soliton solutions are also called the *moduli*. In principle there would seem to be seven moduli, corresponding to the number of generators of the symmetry group of the static energy. However, as already noticed in Section 2.2.2, fixing the boundary conditions of the fields reduces the internal symmetry group acting on $\mathcal{Q}$ to the unbroken group $O(2)$. Furthermore, combining such an internal $O(2)$ transformation with a space rotation by the same angle, will leave the solution invariant. This combined transformation is a symmetry of the solution and it does not give rise to a

modulus. Thus there can be only four moduli. The general solutions are

$$\omega(z) = \left(\frac{(z-z_0)e^{i\alpha}}{\lambda}\right)^n \tag{2.74}$$

$$\omega(z) = \left(\frac{(z-z_0)^* e^{-i\alpha}}{\lambda}\right)^n \tag{2.75}$$

where z_0, α and λ are the translational, rotational and dilatational moduli of the soliton.

2.3.3 No ferromagnetic transition in $d = 2$

As we saw in Section 1.3.2, The nonlinear $O(3)$ model can be regarded as the low energy limit of a Ginzburg–Landau $O(3)$ theory, which in turn is effective continuum description of a planar ferromagnetic lattice, with unit spins allowed to point in any direction in the 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 winding number $W = 0$. 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. The fields φ^α are the Goldstone bosons and their 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 in accordance with the Mermin–Wagner theorem [Mer66], stating that in two (or less) space dimensions at temperature $T > 0$, there cannot be a phase where a continuous symmetry is spontaneously broken. See also [Col73].

2.4 Skyrmions

2.4.1 Topology

Let us consider a general nonlinear sigma model with values in some target space N. The scaling argument rules out static solitons for the action (1.75)

in dimensions $d \neq 2$. Nevertheless let us ignore it for the moment and let us see for what choices of space dimension and target space the configuration space would have more than one connected component. Then we shall look for some alternative action functional that could have stationary points in the nontrivial topological sectors.

Following the same reasoning as in the case of the S^2 sigma model, the space of smooth finite energy configurations of the field is $\mathcal{Q} = \Gamma_*(S^d, N)$. Therefore, there is room for the existence of topological solitons whenever

$$\pi_0(\mathcal{Q}) = \pi_d(N) \neq 0.$$

One important case is when $N = G$, a Lie group. This is sometimes called a principal sigma model. If G is semisimple, one has $\pi_3(G) = \mathbb{Z}$, the fundamental class being realized by a homomorphism $SU(2) = S^3 \to G$. Phenomenologically, the most important cases are the chiral nonlinear sigma models in three dimensions, discussed in Section 1.3.3.

Let us focus on the chiral $SU(2)$ case, whose target space is diffeomorphic to $SU(2) = S^3$. The topological sectors in this case are classified by the winding number, which in terms of the fields U can be written (see Exercise 2.6):

$$W(U) = -\frac{1}{24\pi^2} \int d^3x\, \varepsilon^{\lambda\mu\nu} \mathrm{tr}\Big(U^{-1}\partial_\lambda U U^{-1}\partial_\mu U U^{-1}\partial_\nu U\Big). \tag{2.76}$$

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 the 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 (2.76), that (see Exercise 2.6)

$$W(U_1U_2) = W(U_1) + W(U_2); \qquad W(U^{-1}) = -W(U). \tag{2.77}$$

A field configuration with winding number one has the form

$$U(\vec{x}) = \exp\big[\hat{x}^a \tau_a g(r)\big], \tag{2.78}$$

where $\hat{x}^a = \frac{x^a}{r}$ and g is a function which is equal to -2π in the origin ($U = -\mathbb{I}$) and tends to zero as $r \to \infty$ ($U = \mathbb{I}$). Each 2-sphere in space, having radius r, is mapped in a one-to-one way to a 2-sphere in $SU(2)$. From (2.77), configurations with arbitrary winding numbers can be constructed simply taking powers of (2.78).

2.4.2 Dynamics

Unfortunately, it follows from the discussion in Section 2.3 that such fields cannot be solutions of the field equations obtained from the action (1.80). In fact, from (2.61) we get

$$\frac{dE_S(\phi_\lambda)}{d\lambda}\Big|_{\lambda=1} = -E_1(\phi_1) < 0,$$

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 [Sky61]. 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, as in (1.107). The total action considered by Skyrme was

$$S = \int d^4x \left[\frac{f^2}{4} \mathrm{tr}\left(U^{-1}\partial_\mu U U^{-1}\partial^\mu U\right) + \frac{1}{32e^2} \mathrm{tr}\left[U^{-1}\partial_\mu U, U^{-1}\partial_\nu U\right]\left[U^{-1}\partial^\mu U, U^{-1}\partial^\nu U\right] \right], \tag{2.79}$$

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

To see how the addition of the four-derivative terms circumvents the scaling argument of Derrick's theorem, suppose that the soliton has size ℓ, meaning that the function g in (2.78) goes from -2π to zero within a distance ℓ of the origin. Then, the static energy is of the order

$$E_S(\ell) \approx \ell^3 \left[\frac{f^2}{\ell^2} + \frac{1}{e^2\ell^4} \right]. \tag{2.80}$$

The first term, that comes from the standard two-derivative Lagrangian, is linear in ℓ. It means that one can gain energy by shrinking the profile of the field to zero. The second term has the opposite effect: it favors broad field profiles. In the presence of both terms, the energy has a minimum for some finite value of ℓ, suggesting that solitons can exist. In fact, as in Section 2.1.1, one can use this argument to derive some qualitative properties

of the solutions: the minimum of the energy occurs at $\ell \approx 1/fe$ and inserting this back in the formula for the energy, the mass of the soliton is of the order f/e. For weak coupling ($e \ll 1$), the soliton is much heavier than the pions.

In order to find the soliton with unit winding number, we have to insert the Ansatz (2.78) in the equations of motion that come from (2.79), and solve for the radial function g. Unfortunately the dynamics is sufficiently complicated to prevent an explicit solution. However, solutions can be found numerically.

As in previous examples, the global symmetries of the action imply that the solitons are not isolated solutions but come in families. Restricting our attention to static fields, the energy is invariant under translations, rotations and under $SU(2)_L \times SU(2)_R$, acting on the field as in (1.79). Each of these transformations could in principle give rise to moduli of solutions. However, we must restrict our attention to transformations that preserve the boundary condition $\lim_{r\to\infty} U = \mathbb{I}$ and this means that of $SU(2)_L \times SU(2)_R$ we must only consider the diagonal (isospin) subgroup $SU(2)_V$, acting as

$$U \longmapsto g^{-1}Ug.$$

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

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

When used in the ansatz (2.78), this action exponentiates to

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

We see that if we follow a rotation B by an isospin transformation $g = B^{-1}$ the solution is invariant. This is a symmetry of the skyrmion.[6] Thus the only moduli are given by translations and either rotations or isospin transformations.

These solitons are known as skyrmions. Skyrme also suggested that they could be interpreted as the baryons [Sky62]. In order to understand this claim, we have to study the quantum numbers of the skyrmions. This we shall do much later, in Sections 5.3 and 6.8.3.

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

2.5 Solitons in Yang–Mills theory

We now ask whether pure YM theory in d space dimensions can have solitons. In this section we shall use the unscaled form of the theory, with action (1.112), where $n = d + 1$. The canonical formulation of YM theory has been given in Section 1.7.4. Separating the space and time components of the curvature, the Lagrangian is

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

where $E_i^a = F_{0i}^a = \partial_0 A_i^a - D_i A_0^a$ is the nonabelian "electric" field. There are primary constraints $P_a^0 = 0$, where P_a^0 are the momenta conjugate to A_0^a, and secondary constraints $G_a = 0$, where $G_a = D_i P_a^i = D_i E_i^a$. The canonical Hamiltonian is

$$H_c = \int d^d x \left[\frac{1}{2}P_i^a P_i^a + \frac{1}{4}F_{ij}^a F_{ij}^a - A_0^a G_a\right], \tag{2.81}$$

where $P_a^i = E_a^i$ are the momenta conjugate to A_i^a. 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 $E_i^a = \dot{A}_i^a$, so the first term in (2.81) 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 YM theory can have static solitons. There is here a slight complication: if a gauge field configuration is time-independent, it can acquire a time dependence after a gauge transformation. In a gauge theory one calls a field "static" if there is a gauge in which A_μ is time-independent. This implies that all gauge invariant quantities constructed with the field (such as, for example, the energy density) are time-independent. Note that for a static configuration, the gauge $A_0 = 0$ may not be the gauge in which $\partial_0 A_\mu = 0$, so we do not make this gauge choice here. We shall now prove that pure YM theory does not admit static solitons if $d \neq 4$ (i.e. in five-dimensional spacetime) [Col77].

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

$$E_1 = \frac{1}{2}\int d^d x\, (D_i A_0)^2 > 0 \qquad \text{and} \qquad E_2 = \frac{1}{4}\int d^d x\, (F_{ij}^a)^2 > 0.$$

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

$$A_{(\sigma,\lambda)}{}_0^a(x) = \sigma\lambda A_0^a(\lambda x), \tag{2.82}$$
$$A_{(\sigma,\lambda)}{}_i^a(x) = \lambda A_i^a(\lambda x). \tag{2.83}$$

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

$$0 = \frac{d}{d\lambda}L\bigg|_{\lambda=\sigma=1} = (4-d)L\big(A_{(1,1)}\big), \tag{2.84}$$
$$0 = \frac{d}{d\sigma}L\bigg|_{\lambda=\sigma=1} = 2E_1\big(A_{(1,1)}\big), \tag{2.85}$$

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

This argument rules out nontrivial static solitons for pure YM theories, except in five spacetime dimensions. Since we are interested in topological solitons, it has to be complemented by an analysis of the topology of the configuration space. For this purpose it is very convenient to use the gauge $A_0 = 0$, in which case the configurations of the system are given by the equivalence classes of gauge potentials $A_i^a(x)$, with $i = 1, \ldots d$, modulo time-independent gauge transformations. We denote $\mathcal{C}$ the space of gauge potentials with finite static energy

$$E_S = \int d^dx F^a_{ij}F^a_{ij} \tag{2.86}$$

and $\mathcal{G}$ the group of time-independent gauge transformations that tend to the identity at infinity. Then we have

$$\mathcal{Q} = \mathcal{C}/\mathcal{G}.$$

The connected components of $\mathcal{Q}$ are in one-to-one correspondence with the connected components of $\mathcal{C}$. Thus we have to ask whether there exist gauge potentials that cannot be continuously deformed to a reference potential, say $A_i = 0$. We defer this analysis to Section 3.6.3, where we shall see that such potentials do exist in dimension $d = 4$. Thus, also the topological analysis points to five spacetime dimensions as the only interesting context for pure YM theory. We will see that topological 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 YM instantons.

2.6 Vortices

2.6.1 The Nielsen–Olesen vortex

We now consider scalar electrodynamics in two space dimensions [NiO73]. The dynamical variables are an abelian gauge field A_μ coupled to a complex scalar field ϕ, with action

$$S = \int d^3x \left[-\frac{1}{4} F_{\mu\nu} F^{\mu\nu} - \frac{1}{2} |D_\mu \phi|^2 - \frac{\lambda}{4} \left(|\phi|^2 - f^2 \right)^2 \right], \tag{2.87}$$

where $F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu$, $D_\mu \phi = \partial_\mu \phi - ieA_\mu \phi$. The theory is invariant under the local $U(1)$ gauge transformations $g(x) = e^{i\alpha(x)}$:

$$A_\mu \to A'_\mu = A_\mu - \frac{1}{e} \partial_\mu \alpha, \tag{2.88a}$$

$$\phi \to \phi' = e^{-i\alpha(x)} \phi. \tag{2.88b}$$

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

$$E_K = \int d^2x \left[\frac{1}{2} \dot{A}_i \dot{A}_i + \frac{1}{2} |\dot{\phi}|^2 \right]. \tag{2.89}$$

is the kinetic energy and E_S

$$E_S = \int d^2x \left[\frac{1}{2} B^2 + \frac{1}{2} |D_i \phi|^2 + \frac{\lambda}{4} \left(|\phi|^2 - f^2 \right)^2 \right], \tag{2.90}$$

with $B = F_{12}$, is the static energy. The absolute minimum of E_S, the classical vacuum, occurs for

$$B = 0, \qquad D_i \phi = 0, \qquad |\phi| = f. \tag{2.91}$$

A particular solution of these conditions is

$$A_i = 0, \qquad \phi = f. \tag{2.92}$$

This is the starting point for the usual perturbative discussion of the Higgs phenomenon, showing that the small fluctuations around this vacuum comprise a vector field with mass $m_A = ef$ and a scalar field with mass $m_S = \sqrt{2\lambda} f$. Any gauge transformation of (2.92) is obviously still a solution, but there are also other interesting states.

We will now look for static solitons, assuming that the gauge in which the field is time-independent is the gauge $A_0 = 0$. The classical configuration space of this theory consists of regular fields with finite static energy. Clearly (A, ϕ) will have finite energy only if the conditions (2.91) are satisfied asymptotically as $r \to \infty$. This requires that

$$\phi(r,\theta) \underset{r\to\infty}{\longrightarrow} \phi_\infty = f\, e^{-i\alpha_\infty}, \tag{2.93a}$$

$$A_i(r,\theta) \underset{r\to\infty}{\longrightarrow} -\frac{1}{e}\partial_i \alpha_\infty, \tag{2.93b}$$

where α_∞ depends only on the angular coordinate θ parameterizing the "circle at infinity" S^1_∞. We see that unlike the case of the sigma model, the condition $D_i\phi \to 0$ does not imply that ϕ tends to a constant at infinity: as long as $|\phi| \to f$, any dependence of ϕ on the angle θ is permitted, because one can always compensate for this dependence by choosing

$$A_i = \frac{1}{ie}\frac{\partial_i \phi}{\phi}.$$

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

$$W(\phi_\infty) = \frac{1}{2\pi}\int_0^{2\pi} d\theta\, \frac{d\alpha_\infty}{d\theta} = \frac{i}{2\pi}\int_0^{2\pi} d\theta\, \frac{1}{\phi_\infty}\frac{d\phi_\infty}{d\theta}. \tag{2.94}$$

The field ϕ has values in a linear space and therefore any field configuration can be smoothly deformed into any other. Figure 10 shows a homotopy between a field with $W = 1$ and the constant field $\phi = f$, that has $W = 0$. It is clear that in the intermediate steps of the deformation the modulus field $|\phi|$ does not tend to f as $r \to \infty$. Such fields have infinite static energy, so there is an infinite energy barrier between configurations with different winding numbers of ϕ_∞, or in other words the configuration space consists of infinitely many connected components, labelled by $W(\phi_\infty)$.

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

$$J^\lambda_T = \frac{1}{2\pi i}\varepsilon^{\lambda\mu\nu}\partial_\mu \hat\phi^* \partial_\nu \hat\phi, \tag{2.95}$$

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

$$Q_T = \int d^2x\, J^0_T = W(\phi_\infty). \tag{2.96}$$

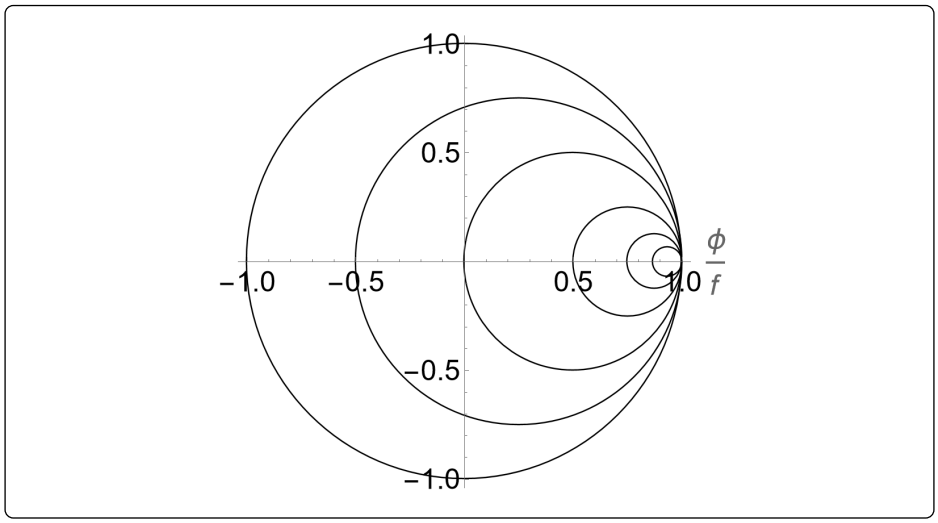


Figure 10. A homotopy in field space. The circles represent the images in field space of S^1_∞, for varying homotopy parameter. The circle of unit radius is the locus of the minima of the potential.

The physical meaning of the winding number can be understood by using (2.93) in (2.94) and then applying Stokes' theorem:

$$W(\phi_\infty) = \frac{e}{2\pi}\oint_{S^1_\infty} A_i dx^i = \frac{e}{2\pi}\int_{\mathbb{R}^2} d^2x\, B = \frac{e}{2\pi}\Phi. \tag{2.97}$$

If we think of two-dimensional space as a plane embedded in three-dimensional space, B is the magnetic field orthogonal to the plane and Φ is its magnetic flux. Since W is an integer, we get flux quantization:

$$\Phi = \frac{2\pi}{e}n. \tag{2.98}$$

We would like to find explicit *vortex* solutions in each topological sector. For the soliton with unit flux we make the ansatz

$$A_0 = 0, \tag{2.99a}$$
$$A_i = -\varepsilon_{ij}\hat{x}^j A(r), \tag{2.99b}$$
$$\phi = F(r)e^{i\varphi}, \tag{2.99c}$$

where A and F are functions of the radius such that

$$A(r) \to \frac{1}{er} \quad \text{and} \quad F(r) \to f + O(r^{-1})$$

when $r \to \infty$. Clearly the asymptotic conditions (2.93) are satisfied and $W(\phi_\infty) = 1$. However, it has so far proved impossible to solve explicitly the equations of motions and one has to resort to numerical calculations.

Vortices with $m_A = m_S$ are said to be *critical* and have special properties. One is that the equations of motion can be simplified by using a bound on the energy, see Exercise 2.4. Unfortunately, the ensuing reduction of the equations from second order to first order is still not enough to solve them analytically. A more physical consequence is that the force between two vortices is zero. Instead, two vortices with unit charge attract if $m_A > m_S$ (these are called theories of type I) and repel if $m_A < m_S$ (these are called theories of type II). See Exercise 2.5. One can heuristically understand this by noting that the interactions mediated by the vector field generate repulsion (one can think of two vortices as two charges with the same sign), while the interaction mediated by the scalar field generates attraction. In theories of type I the scalar is lighter than the vector and the attractive interaction prevails, whereas in theories of type II the vector is lighter and a net repulsion results. This also has the consequence that in type I theories, vortex solutions with $n > 1$ are stable, whereas in type II theories vortices with $n > 1$ tend to break up into n vortices of unit charge.

Finally we observe that, just as kinks can be reiterpreted as domain walls in $d > 1$, the two-dimensional Nielsen–Olesen vortex can be reinterpreted as a vortex line in three dimensions. If we assume that $A_3 = 0$ and that all the fields are independent of x_3, the equations of $d = 3$ scalar electrodynamics reduce to those of $d = 2$ scalar electrodynamics. Thus, the vortex soliton of $d = 2$ becomes 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 next.

2.6.2 Vortices in superconductors

As we discussed in Section 1.5.4, the Euclidean, three dimensional version of scalar electrodynamics is the Ginzburg–Landau theory of superconductivity. We focused on the properties of the superconducting phase where the Meissner effect means that a superconductor acts like a perfect diamagnet ($\chi = -1$). This phase can be described by the simple Lagrangian (1.149) where the modulus of the scalar is fixed ($|\phi| = f$) and only its phase, the Goldstone boson, is active. In order to understand also the other phases one needs the full Ginzburg–Landau Lagrangian (1.153).

Superconductors are classified into two types, labelled I and II. The phase diagram of a type I superconductor is shown on the left in Figure 11. There is a critical temperature T_c above which superconductivity is impossible. Below

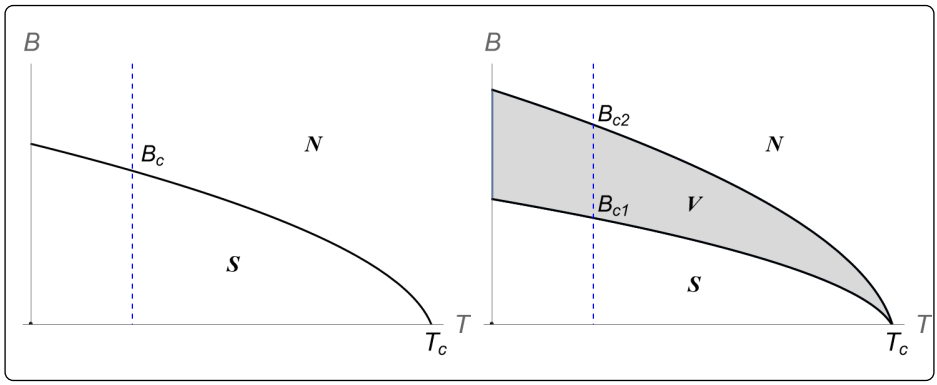


Figure 11. Phase diagram of a type I (left) and type II (right) superconductor. In type I superconductor, the material transitions homogeneously from the superconducting (S) to the normal phase (N) when the external magnetic field exceeds a critical value B_c. In the type II superconductor there is an intermediate phase V (shaded area) where the material is pierced by vortices.

T_c the material will be in the superconducting phase if the external magnetic field B is below a critical value B_c and in the normal phase above B_c.

Since superconductivity is related to a nonzero VEV of the scalar, the existence of a critical temperature can be understood exactly as in the $\mathbb{Z}_2$-symmetric model of Section 1.2.2: in the free energy, the mass squared is positive for $T > T_c$ and negative for $T < T_c$. The fact that here the group $U(1)$ is gauged does not play any role in this argument.

The existence of a critical magnetic field can be understood as follows. Consider a piece of material above the critical temperature, so with $|\phi| = 0$, in a magnetic field B, and lower the temperature below T_c, so that the minimum of the potential moves to $|\phi| = f$ and is lower by an amount

$$\Delta V = \frac{\lambda}{4} f^4 = \frac{m^4}{4\lambda}. \tag{2.100}$$

In order to transition to the superconducting state, the material must also expel the magnetic field by creating an opposite magnetic field in its interior. The energy density that has to be spent for this is $\frac{1}{2}B^2$. The difference in energy density between the superconducting and the normal state is

$$\mathcal{E}_S - \mathcal{E}_N = -\Delta V + \frac{1}{2}B^2. \tag{2.101}$$

Thus, the transition to the superconducting state will happen provided $\Delta V >$

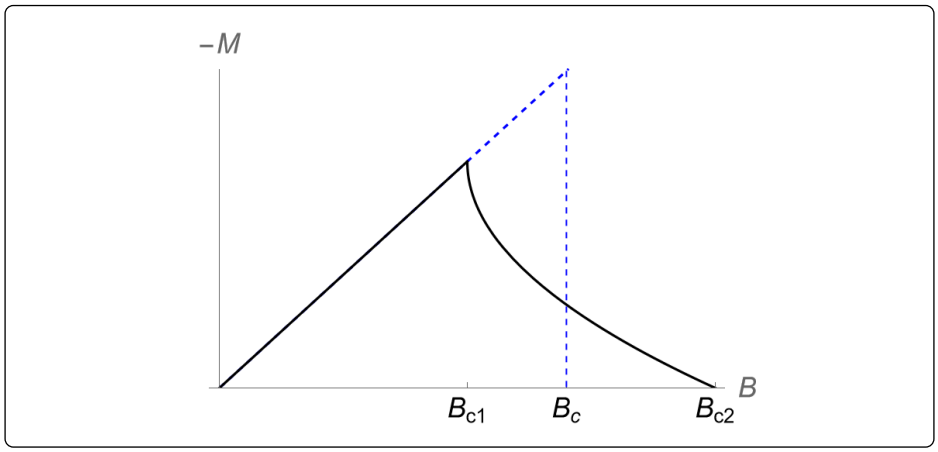


Figure 12. Induced magnetization as function of applied magnetic field for a type I superconductor (blue, dashed curve) and for a type II superconductor (black, continuous curve). For $B < B_{c_1}$ the system is a homogeneous superconductor. Here $M = -B$ and there is Meissner effect. For $B_{c_1} < B < B_{c_2}$ the system is in the mixed phase with vortices, for $B > B_{c_2}$ it is a normal conductor.

$\frac{1}{2}B^2$, i.e. if B is less than the critical value

$$B_c = \sqrt{2\Delta V}. \tag{2.102}$$

The phase diagram of a type II superconductor, shown on the right in Figure 11, is more complicated. When the magnetic field exceeds a lower critical value B_{c_1}, the magnetic field pierces the superconductor in the form of thin tubes. These tubes can be thought of as cylinders of normal phase embedded in the bulk of the superconductor. In the Landau–Ginzburg theory, the tubes carrying the magnetic field are nothing but the Nielsen–Olesen vortices of the preceding section. In the present context they are better known as Abrikosov vortices. Since the core of the vortex is not superconducting, the topology of a piece of superconductor that is pierced by a vortex line is the same as that of the thick torus discussed in Section 1.5.4. Therefore, the flux through the tube must be quantized as in (1.160):[7]

$$\Phi = \frac{\pi}{e} n. \tag{2.103}$$

[7]We now use natural units. Note the similarity and the difference between this quantum mechanical condition, that comes from BCS theory, and the classical quantization condition (2.98) of the effective Ginzburg–Landau theory: in the latter e is a classical parameter in the Lagrangian that could have any value, whereas here it is the electron charge (the factor two is due to the charge of the Cooper pairs). Yet we see that the topological information is present in both theories.

The density of flux tubes increases with B and for B greater than a higher critical value B_{c2}, the system returns to the normal, non-superconducting state. Thus, between the two critical magnetic fields, the thermodynamically favored state is inhomogeneous. Both critical magnetic fields depend on the temperature and go to zero when the temperature exceeds T_c. Figure 12 shows a plot of the induced magnetization as a function of the applied magnetic field for superconductors of type I and II.

This behavior, and also the difference between type I and type II superconductors, can be understood by a slightly more complicated version of the free energy argument given above.

The properties of the superconductors are determined to a large extent by two parameters: the London penetration depth λ_L, already defined in Section 1.5.4 and Exercise 1.8, and the *coherence length* ξ, which is the correlation length of the scalar field. In the Ginzburg–Landau theory, they are just the inverse of the photon mass and of the scalar mass:

$$\lambda_L = 1/m_A, \qquad \xi = 1/m_S,$$

Note that (2.100) can be written in terms of these measurable quantities as

$$\Delta V = \frac{1}{8e^2\lambda_L^2\xi^2}. \tag{2.104}$$

A vortex is a cylinder where the scalar is close to zero in an area of order $\pi\xi^2$ and the magnetic field is present in an area of order $\pi\lambda_L^2$. Thus, if we are in the homogeneous superconducting state, the formation of a vortex will increase the energy by an amount of order

$$\pi\xi^2\Delta V - \pi\lambda_L^2\frac{1}{2}B^2.$$

If there is an area density ρ of vortices, sufficiently low so that vortices do not overlap, the energy density difference between the vortex and homogeneous superconducting state is (up to numerical factors of order one)

$$\mathcal{E}_V - \mathcal{E}_S \approx \rho\pi\xi^2\Delta V - \rho\pi\lambda_L^2\frac{1}{2}B^2. \tag{2.105}$$

The formation of vortices will be energetically favored if

$$B > B_{c_1}, \quad \text{where} \quad B_{c_1} \approx \frac{\xi}{\lambda_L}\sqrt{2\Delta V} \approx \frac{1}{e\lambda_L^2} \tag{2.106}$$

is the lower critical field and in the last step we used (2.104). If $\xi < \lambda_L, B_{c_1} < B_c$ and therefore this can happen preserving superconductivity in the rest of the

bulk. This is what happens in type II superconductors. If $\xi > \lambda_L$, $B_{c_1} > B_c$, which means that the formation of vortices would only become energetically favored in a regime where superconductivity has already been lost. This is the situation in type I superconductors. We thus see that the difference between the two types of superconductors is dictated by the Ginzburg–Landau parameter

$$\kappa = \frac{\lambda_L}{\xi} = \frac{m_S}{m_A}. \tag{2.107}$$

Type I superconductors have $\kappa < 1$ and type II superconductors have $\kappa > 1$.

While the vortex solution exists in both situations, the profiles of the functions A and F in (2.99) have very different behavior: in a type I vortex, A decays to zero faster than F rises from zero to f, whereas in a type II vortex F grows to f faster than A goes to zero. This means that the type II vortices have a small superconducting core, immersed in a slowly decaying magnetic field.

We can now complete the analysis of the phase diagram of type II superconductors. Once it becomes energetically convenient for the magnetic field to funnel through the vortices, the whole magnetic field will do so. Each vortex carries a flux π/e, so the number density of vortices is:

$$\rho = \frac{e}{\pi} B. \tag{2.108}$$

Using (2.106) this implies

$$\rho \gtrsim \frac{1}{\pi \lambda_L^2}. \tag{2.109}$$

This means that, already just above the transition, the magnetic fields of the vortices overlap significantly and there is no Meissner effect anymore, while the non-superconducting cores can remain well separated. In the vortex phase the density of vortices grows linearly with the magnetic field, and it has an upper bound

$$\rho \lesssim \frac{1}{\pi \xi^2}, \tag{2.110}$$

where the normal metal cores begin to overlap significantly, which means that there is no superconductivity anymore. The two bounds on ρ are consistent, because $\lambda_L > \xi$. Using (2.108), this implies

$$B < B_{c_2}, \quad \text{where} \quad B_{c_2} \approx \frac{1}{e\xi^2} = \left(\frac{\lambda_L}{\xi}\right)^2 B_{c_1} = \frac{\lambda_L}{\xi}\sqrt{\Delta V}. \tag{2.111}$$

We can also see this by evaluating the energy density difference between the vortex and normal phase. When vortices overlap, $\rho\pi\lambda_L^2 \approx 1$ and (2.105) has to be replaced by $\mathcal{E}_V - \mathcal{E}_S \approx \rho\pi\xi^2\Delta V - \frac{1}{2}B^2$. Adding (2.101), and then using (2.108) and (2.104), we obtain

$$\mathcal{E}_V - \mathcal{E}_N \approx (\rho\pi\xi^2 - 1)\Delta V \approx \left(\frac{B}{B_{c_2}} - 1\right)\Delta V, \tag{2.112}$$

When $B > B_{c_2}$ the normal phase is energetically favored and superconductivity ceases to exist.

Since in type II superconductors the vortices repel, just under the upper critical field they tend to form a tightly packed triangular lattice.

2.7 Monopoles

2.7.1 Duality

Maxwell's equations can be written in the form

$$\partial_\mu F^{\mu\nu} = J_E^\nu, \tag{2.113a}$$
$$\partial_\mu {}^*F^{\mu\nu} = 0, \tag{2.113b}$$

where ${}^*F_{\mu\nu} = \frac{1}{2}g_{\mu\rho}g_{\nu\sigma}\varepsilon^{\rho\sigma\alpha\beta}F_{\alpha\beta}$ is the dual of the field strength (see Appendix A for the relevant definitions and conventions). In vacuum ($J_E^\nu = 0$) these equations are invariant under the duality transformation $F \to {}^*F$, ${}^*F \to {}^{**}F = -F$. Writing

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

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

$$F \to \cos\theta F + \sin\theta\, {}^*F \tag{2.114a}$$
$${}^*F \to -\sin\theta F + \cos\theta\, {}^*F. \tag{2.114b}$$

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 (2.113) is identically zero. They

could be made symmetric under duality transformations by introducing a magnetic current

$$\partial_\mu {}^*F^{\mu\nu} = J_M^\nu \tag{2.115}$$

and postulating the transformation

$$J_E \to \quad \cos\theta J_E + \sin\theta J_M, \tag{2.116a}$$
$$J_M \to -\sin\theta J_E + \cos\theta J_M. \tag{2.116b}$$

That J_M^ν is a magnetic current is seen by observing for example that the time component of (2.115) would read $\mathrm{div}B = J_M^0$. Thus J_M^0 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 is impossible to introduce a magnetic potential A_μ such that $F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu$. This complication does not arise if we limit ourselves to the study of pointlike magnetic sources. The Coulomb-like field

$$B_i = \frac{1}{4\pi}\frac{Q_M}{r^2}\hat{x}^i, \tag{2.117}$$

describing a static pointlike magnetic monopole in the origin, solves the equation $\mathrm{div}B = Q_M\delta(r)$. It is known as the Dirac monopole. Since the field is singular in the origin, one can remove this point from space and regard (2.117) as a smooth field on $\mathbb{R}^3\backslash\{0\}$. Since the field B given in (2.117) is divergence free on $\mathbb{R}^3\backslash\{0\}$, it is possible to introduce the magnetic potential there.

This solution of Maxwell's equations has interesting properties that we shall study in detail in Section 4.1. 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, whereas it certainly looks like a particle, it is a singular field and has infinite energy, so it does not satisfy the general requirements for being 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.

2.7.2 The 't Hooft–Polyakov monopole

We consider the Georgi–Glashow model, consisting of an $SO(3)$ gauge field $A_\mu = A_\mu^a T_a$ coupled to a Higgs field ϕ^a in the adjoint (triplet) representation.[8]

[8]This model is sometimes viewed as a $SU(2)$ gauge theory, but as long as there are no isospinor fields, the transformation $-\mathbb{I} \in SU(2)$ leaves all the fields invariant. The group that is represented faithfully on the fields is $SO(3)$.

Wenere use the unscaled gauge fields, with curvature (1.113) and action (1.112). The total Lagrangian density is

$$\mathcal{L} = -\frac{1}{4}F^a_{\mu\nu}F^{\mu\nu a} - \frac{1}{2}D_\mu\phi^a D^\mu\phi^a - \frac{\lambda}{4}\left(\phi^a\phi^a - f^2\right)^2 \tag{2.118}$$

where

$$D_\mu\phi^a = \partial_\mu\phi^a + e\varepsilon_{abc}A^b_\mu\phi^c. \tag{2.119}$$

It is invariant under the local gauge transformations (1.115). It is very convenient to choose the gauge $A^a_0 = 0$. Then

$$F^a_{0i} = \partial_0 A^a_i, \qquad D_0\phi^a = \partial_0\phi^a \tag{2.120}$$

and the static energy is simply

$$E_S = \int d^3x\left[\frac{1}{4}\left(F^a_{ij}\right)^2 + \frac{1}{2}\left(D_i\phi^a\right)^2 + \frac{\lambda}{4}\left(\phi^a\phi^a - f^2\right)^2\right]. \tag{2.121}$$

Its absolute minimum is obtained for

$$F_{ij} = 0, \tag{2.122a}$$

$$D_i\phi^a = 0, \tag{2.122b}$$

$$\phi^a\phi^a = f^2, \tag{2.122c}$$

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 $SO(3)$ transformations is not broken, however, and any gauge transform of this solution is also a solution.

Finiteness of E_S demands that the conditions (2.122) be satisfied asymptotically when $r \to \infty$. In particular for large r we must have $\phi^2 = f^2 + O(1/r^2)$, so the asymptotic behaviour of ϕ defines a map $\phi_\infty : S^2_\infty \to S^2_{\text{int}}$, where S^2_∞ denotes the "sphere at infinity" in $\mathbb{R}^3$ and S^2_{int} is the locus of the minima of the potential in 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 (2.122) does not restrict the map ϕ itself. The asymptotic field ϕ^a_∞ can depend on the angles in an arbitrary way and the condition $D_i\phi \to 0$ can then be solved by

$$A^a_i = \frac{1}{f^2 e}\varepsilon^{abc}\partial_i\phi^b\phi^c + \alpha_i\phi^a + O(1/r^2), \tag{2.123}$$

for arbitrary constants α_i.

The scalar fields ϕ fall into classes, labelled by the winding number of the map ϕ_∞. Fields with different winding numbers at infinity are separated by an infinite energy barrier. There follows that the configuration space of smooth finite energy configurations for this model consists of infinitely many connected components, labelled by the winding number of ϕ_∞. The configuration with $W = 0$ is the vacuum, the one with $W = 1$ is called a "hedgehog". The winding number cannot be altered in the course of the time evolution, so there must be a topological conservation law. Indeed, the topological current is

$$J_T^\mu = \frac{1}{8\pi}\varepsilon^{\mu\nu\rho\sigma}\varepsilon_{abc}\partial_\nu\hat{\phi}^a\partial_\rho\hat{\phi}^b\partial_\sigma\hat{\phi}^c, \tag{2.124}$$

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

$$\begin{aligned} Q_T &= \int d^3x J_T^0 = \frac{1}{8\pi}\int d^3x\varepsilon^{ijk}\varepsilon_{abc}\partial_i\hat{\phi}^a\partial_j\hat{\phi}^b\partial_k\hat{\phi}^c \\ &= \frac{1}{8\pi}\int\limits_{S^2_\infty} d^2x\varepsilon^{ij}\varepsilon_{abc}\hat{\phi}^a\partial_i\hat{\phi}^b\partial_j\hat{\phi}^c = W(\phi_\infty). \end{aligned} \tag{2.125}$$

The last equality can be proven by choosing a particular coordinate system on S^2, for example the spherical coordinates (1.69), and comparing with (2.64).

We are now in a position to explain why configurations with $W \neq 0$ can be interpreted as magnetic monopoles. When the Higgs phenomenon occurs, we can interpret the projection of the gauge field along the Higgs VEV as an abelian gauge field. If $\hat{\phi}^a = (0, 0, 1)$, the corresponding field strength is $\mathcal{F}_{\mu\nu} = \partial_\mu A^3_\nu - \partial_\nu A^3_\mu$. Following 't Hooft and Polyakov, we can generalize this to position-dependent Higgs fields [tHo74, Pol74]. Let $\mathcal{A}_\mu = A^a_\mu\hat{\phi}^a$. We define an abelian electromagnetic field $\mathcal{F}_{\mu\nu}$ by

$$\mathcal{F}_{\mu\nu} = \partial_\mu\mathcal{A}_\nu - \partial_\nu\mathcal{A}_\mu - \frac{1}{e}\varepsilon_{abc}\hat{\phi}^a\partial_\mu\hat{\phi}^b\partial_\nu\hat{\phi}^c. \tag{2.126}$$

The last term has been added to compensate the $SO(3)$ non-invariance of $\mathcal{A}$. In fact, this can also be written as

$$\mathcal{F}_{\mu\nu} = \hat{\phi}^a F^a_{\mu\nu} - \frac{1}{e}\varepsilon_{abc}\hat{\phi}^a D_\mu\hat{\phi}^b D_\nu\hat{\phi}^c, \tag{2.127}$$

which is manifestly invariant under $SO(3)$ gauge transformations (see Exercise 2.7). This tensor does not obey the Bianchi identities. Instead,

$$\partial_\nu {}^*\mathcal{F}^{\nu\mu} = \frac{4\pi}{e}J_T^\mu. \tag{2.128}$$

as one can check most easily using (2.126). Comparing with (2.115), we see that we can interpret $\frac{4\pi}{e}J_T^\mu$ as a magnetic current. The corresponding magnetic charge is

$$Q_M = \frac{4\pi}{e}Q_T = \frac{4\pi}{e}W. \tag{2.129}$$

Since W is an integer, we get a quantization condition for the magnetic charge, analogous to the flux quantization condition (2.98). We shall see in Section 3.1 that quantum mechanics requires the magnetic charge to be quantized in units of $\frac{2\pi\hbar c}{e}$, where e is the charge of the electron. The relation between these two conditions is closely analogous to that between (2.98) and (1.160).

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

$$\phi^a = \frac{x^a}{r}F(r), \tag{2.130a}$$

$$A_i^a = \varepsilon_{aij}\frac{x^j}{r}A(r), \tag{2.130b}$$

$$A_0^a = 0. \tag{2.130c}$$

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

$$D_i\phi^a = (\delta_{ia} - \hat{x}_i\hat{x}_a)\left(\frac{1}{r} - eA\right)F + \hat{x}_i\hat{x}_aF', \tag{2.131}$$

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

$$B_i^a = -(\delta_{ia} - \hat{x}_i\hat{x}_a)A' - \frac{1}{r}\delta_{ia}A + \hat{x}_i\hat{x}_a\left(eA^2 - \frac{1}{r}A\right). \tag{2.132}$$

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

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

$$\mathcal{B}_i = \frac{1}{2}\varepsilon_{ijk}\mathcal{F}_{jk} \to \hat{\phi}^a B_i^a = -\frac{1}{e}\frac{\hat{x}^i}{r^2} \tag{2.133}$$

while $\mathcal{E}_i = \mathcal{F}_{0i} = 0$. Therefore, for large r, the abelian field strength becomes identical to the one of the Dirac monopole with charge $Q_M = -\frac{4\pi}{e}$.

When the ansatz (2.130) is inserted into the Euler–Lagrange equations, these become coupled second order differential equations for the functions F and A. The exact solution to these equations has not been found; only numerical solutions have been given.

2.7.3 The Prasad–Sommerfield limit

There is one particular limit, known as the Prasad–Sommerfield limit, in which analytic expressions for the functions F and A are known: it is the limit in which λ and m^2 tend to zero with $f = \sqrt{m^2/\lambda}$ constant [PSo75]. This is in some sense the opposite of the limit we took in Section 1.3: the potential goes to zero, but we retain the boundary conditions of the field that came from minimizing it. Then one can derive a useful bound on the energy. We have

$$\begin{aligned} E &= \int d^3x \left[\frac{1}{4}F^a_{ij}F^a_{ij} + \frac{1}{2}D_i\phi^a D_i\phi^a\right] \\ &= \frac{1}{4}\int d^3x \left(F^a_{ij} \mp \varepsilon_{ijk}D_k\phi^a\right)^2 \pm \frac{1}{2}\int d^3x\, \varepsilon_{ijk}F^a_{ij}D_k\phi^a. \end{aligned} \tag{2.134}$$

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

$$\frac{1}{2}\int d^3x\, \partial_k\left(\varepsilon_{ijk}F^a_{ij}\phi^a\right) = f\int_{S^2_\infty} d\sigma^k \mathcal{B}_k = f\, Q_M = \frac{4\pi f}{e}\, W, \tag{2.135}$$

where we have used (2.133). Using this in (2.134) we get the so-called Bogomol'nyi bound [Bog75]

$$E \geq \frac{4\pi f}{e}|W|, \tag{2.136}$$

with equality holding if and only if

$$F^a_{ij} = \pm\varepsilon_{ijk}D_k\phi^a. \tag{2.137}$$

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 (2.137). In the Prasad–Sommerfield limit the explicit form of the functions appearing in (2.130), for the lower sign in (2.137), is

$$F(r) = \frac{f}{\tanh(efr)} - \frac{1}{er}, \tag{2.138a}$$

$$A(r) = \frac{1}{er} - \frac{f}{\sinh(efr)}. \tag{2.138b}$$

The profiles of these functions are shown in Figure 13.

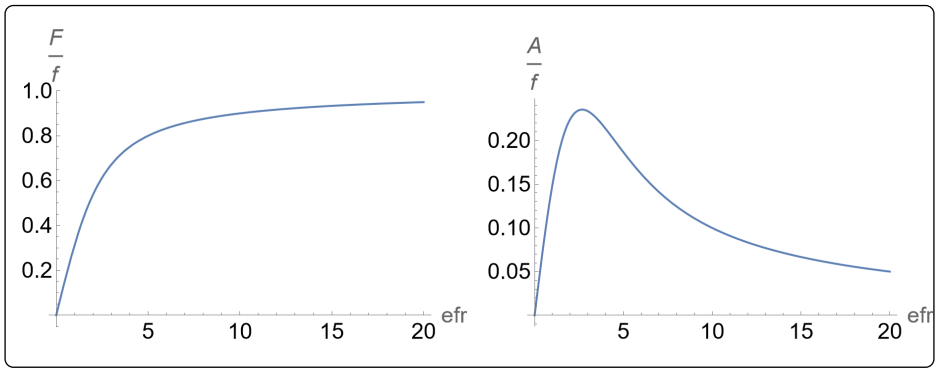


Figure 13. Monopole profiles in the Prasad–Sommerfield limit.

2.7.4 Symmetries and moduli

The symmetries of the Georgi–Glashow model are the Poincaré group and internal $SO(3)$ transformations with constant parameters (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 ϵ_I^a transforms

$$\delta_I \phi^a = \varepsilon_{abc}\epsilon_I^b \phi^c. \tag{2.139}$$

Under the rotation group a scalar transforms by

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

where

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

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

$$\delta_R \phi^a = \varepsilon_{abc}\epsilon_R^b \phi^c. \tag{2.140}$$

From (2.139) and (2.140) we see that ϕ is invariant under the combined transformation

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

where the infinitesimal parameters are the same in the two cases. The gauge field A_i^a is also invariant under the same transformations (see Exercise 3.7) so the monopole has a symmetry $SO(3)$ consisting of simultaneous internal and space rotations. This subgroup is unbroken and does not give rise to moduli.

There remains one $SO(3)$ subgroup that we can choose to correspond to the internal transformations and could give rise to moduli. However, recall that we work in a functional space with fixed boundary conditions. In this space the field ϕ at infinity is fixed and we do not allow transformations that change it. Since the field ϕ^a at infinity is direction-dependent, the transformations that leave it invariant must also be direction-dependent and are not strictly speaking a subgroup of the rigid internal $SO(3)$ rotations. They can be described in the following way. As is always the case, a field configuration in the Higgs phase can be brought to the unitary gauge, where the Higgs is aligned along the third direction. The transformation that does this for the monopole is [AFG74]

$$T = \begin{bmatrix} \frac{\hat{x}_2^2+\hat{x}_1^2\hat{x}_3}{1-\hat{x}_3^2} & -\frac{\hat{x}_1\hat{x}_2}{1+\hat{x}_3} & -\hat{x}_1 \\ -\frac{\hat{x}_1\hat{x}_2}{1+\hat{x}_3} & \frac{\hat{x}_1^2+\hat{x}_2^2\hat{x}_3}{1-\hat{x}_3^2} & -\hat{x}_2 \\ \hat{x}_1 & \hat{x}_2 & \hat{x}_3 \end{bmatrix}. \tag{2.141}$$

See Exercise 2.8. 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.

2.7.5 Monopoles in GUTs

The preceding discussion of the Georgi–Glashow model can be generalized to arbitrary groups G and H. The condition for the existence of monopoles is that the map ϕ_∞, mapping the sphere at infinity to the minima of the potential, has to be topologically nontrivial. Since the orbit where the potential is minimized

is diffeomorphic to the coset space G/H, the condition is that $\pi_2(G/H)$ be nontrivial.

The homotopy groups of this space are related to the homotopy groups of G and H by the so-called homotopy exact sequence. This is discussed in general in Appendix E.5. The part of the sequence that is relevant to us is

$$... \xrightarrow{\partial} \pi_2(H) \xrightarrow{\iota_*} \pi_2(G) \xrightarrow{\mu_*} \pi_2(G/H) \xrightarrow{\partial} \pi_1(H) \xrightarrow{\iota_*} \pi_1(G) \to ... \qquad (2.142)$$

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

We need the following properties of Lie groups. Of all the homotopy groups of $U(1)$ only the fundamental group π_1 is nontrivial and equal to the integers. If G is a compact, connected, simple Lie group G, $\pi_2(G) = 0$ and $\pi_3(G) = \mathbb{Z}$.

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

$$... \xrightarrow{\partial} 0 \xrightarrow{\iota_*} 0 \xrightarrow{\mu_*} \pi_2(G/H) \xrightarrow{\partial} \mathbb{Z} \xrightarrow{\iota_*} 0 \to ... \qquad (2.143)$$

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

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

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

and therefore $\iota_* : \pi_1(U(1)) \to \pi_1(G)$ is still an isomorphism. This implies that the image of ∂ is zero. On the other hand, since $\pi_2(SU(2)\times U(1)) = 0$, the map ∂ is still injective. Therefore we must have $\pi_2(SU(2) \times U(1)/U(1)) = 0$, and we conclude that the Standard Model does not admit monopole solutions.[9] One could have come to the same conclusion much more easily by noting that the orbit of the minima, defined by (1.168), is a three-sphere, and $\pi_2(S^3) = 0$.

Instead of appealing to the existence of the homotopy exact sequence, one can give an *ad hoc* proof of the above results, see Exercise 2.9.

[9] There actually exist monopole-like solutions in electroweak theory [ChM96], but they do not satisfy the regularity conditions we demand of solitons.

2.8 Exercises

Exercise 2.1: Bogomol'nyi bound for the kink

The kink is so simple that we could find the solution in closed form without great difficulty, but it is nevertheless instructive to have a bound on the static energy similar to the one of more complicated solitons.

Show that if we write the potential as

$$V(\phi) = \frac{1}{2}\left(\frac{dW}{d\phi}\right)^2,$$

for a suitable *prepotential* $W(\phi)$, the static energy can be written

$$E_S = \frac{1}{2}\int dx \left(\frac{d\phi}{dx} + W'\right)^2 - (W(\infty) - W(-\infty)).$$

Write the explicit prepotential for the potential (2.2) and show (a) that $W(\infty) - W(-\infty)$ is proportional to the topological charge and (b) that the kink (2.6) solves the first order equation

$$\frac{d\phi}{dx} + W' = 0.$$

Exercise 2.2: Interactions between kinks

Two widely separated static kinks can be treated as two particles exerting a mutual force. The force can be calculated as follows. Assume there is an antikink at $x = -a$ and a kink at $x = a$, with $a \gg 1/(\sqrt{\lambda} f)$ (the size of the kink). The configuration can be described as

$$\phi(x) = \phi_1(x) + \phi_2(x) + 1 \tag{2.144}$$

where $\phi_1(x)$ is the field of the static antikink and $\phi_2(x)$ is the field of the static kink. *Without using this ansatz*, let $-a \ll b \ll a$ and consider the momentum of the field in the half space left of b:

$$P = -\int_{-\infty}^{b} dx\, \dot{\phi}\phi'.$$

Derive a general formula for the force $F = \frac{dP}{dt}$. Since the integrand is a total derivative, the force can be expressed in terms of the field and its

derivatives at $-\infty$ and b. Now use the ansatz. For $x \leq b$, the quantity $\phi_2 + 1$ is exponentially small, so when (2.144) is used, one can treat it as a small perturbation and keep only linear terms. Also use the equation of motion, to obtain

$$F = -\phi_1'\phi_2' + (\phi_2 + 1)\phi_1''\Big|_{\infty}^{b}.$$

Use the explicit form of $\phi_{1,2}$ to calculate the force (use the approximation $\tanh(x) \approx 1 - 2e^{-2x}$ for $x \gg 1$, and the analogous one for $x \ll 1$).

Exercise 2.3: Renormalization of the kink mass

Solve the one-dimensional non-relativistic scattering problem in the Pöschl–Teller potential (2.22) and check the phase shifts (2.26).
See e.g. [MoF53], eq. (12.3.22) and following.

Fill in the details between (2.27) and (2.28).

Exercise 2.4: Critical vortices

The behavior of vortices depends qualitatively on the masses m_A and m_S. Show that when $m_A = m_S$ we can rewrite the energy as

$$E_S = \int d^2x \left\{ \frac{1}{2}\left[B + \frac{e}{2}\left(|\phi|^2 - f^2\right)\right]^2 + \frac{1}{2}|(D_1 + iD_2)\phi|^2 \right\} + \pi n f^2 \tag{2.145}$$

where n is the winding number. Thus we have the "Bogomol'nyi" bound

$$E_S \geq \pi n f^2 \tag{2.146}$$

and the inequality is saturated for

$$B + \frac{e}{2}\left(|\phi|^2 - f^2\right) = 0 \tag{2.147a}$$

$$(D_1 + iD_2)\phi = 0. \tag{2.147b}$$

Go to polar coordinates (r, θ). The Ansatz (2.99) means that the gauge field A only has component θ. Insert the Ansatz in the (second order) equations of motion and in the Bogomol'nyi equations (2.147). Check that the latter imply the former.

For large r the equations can be linearized. Deduce that the behavior of the fields for large r is given by

$$A_\theta \sim \frac{n}{er} - k_A K_1(m_A r) \tag{2.148a}$$

$$F(r) \sim f - k_S K_0(m_S r), \tag{2.148b}$$

where k_A, k_S are constants and K_0, K_1 are modified Bessel functions.

Exercise 2.5: Interaction of vortices

Let $(A^{(i)}, \phi^{(i)})$, $i = 1, 2$ be the fields corresponding to single vortices located at positions x_1 and x_2, with $R = |x_2 - x_1|$ much larger than the size of the vortex. Abrikosov made an Ansatz for the field of two vortices

$$\begin{aligned} \phi &= \phi^{(1)}\phi^{(2)} \\ A &= A^{(1)} + A^{(2)}. \end{aligned} \tag{2.149}$$

Far from the vortex cores $\phi^{(i)} = f(1-\sigma^{(i)})e^{i\chi^{(i)}}$ and $\sigma^{(i)}$ are exponentially small. Thus we can approximate:

$$\phi \sim f(1 - \sigma^{(1)} - \sigma^{(2)})e^{i(\chi^{(1)}+\chi^{(2)})}.$$

Furthermore

$$B = B^{(1)} + B^{(2)}$$

and we can approximate

$$D_i\phi = \phi^{(1)}D_i^{(2)}\phi^{(2)} + \phi^{(2)}D_i^{(1)}\phi^{(1)} \sim f(e^{i\chi^{(1)}}D_i^{(2)}\phi^{(2)} + e^{i\chi^{(2)}}D_i^{(1)}\phi^{(1)}).$$

The interaction energy of the vortices can be defined as the difference between the energy of this (approximate) solution and the sum of the energies of the single vortices. Inserting the Abrikosov ansatz (2.149) in the formula for the static energy and using (2.148) one can compute the interaction energy, which turns out to behave like

$$E_{\text{int}}(R) \sim C_A K_0(m_A R) - C_\phi K_0(m_S R), \tag{2.150}$$

where C_A, C_ϕ are (positive) constants. Thus the scalar contribution produces attraction and the vector contribution produces repulsion.

Prove (2.150) at criticality, where $C_A = C_\phi$. Since also $m_A = m_S$, vortices exert no mutual force in this case. Close to criticality one can assume that C_A and C_ϕ are almost the same. For $\kappa < 1$ ($m_S < m_A$) the interaction energy is negative and the vortices attract; for $\kappa > 1$ ($m_S > m_A$) the interaction energy is positive and the vortices repel. The former case corresponds to type-I superconductors, the latter to type-II superconductors.

Exercise 2.6: Formulae for Skyrmions

Using (1.81), check that equation (2.76) reproduces the general formulas for the winding number (E.2). Check the additivity property of the winding number, equation (2.77). Using the formula

$$\exp(g n^a \tau_a) = \cos\left(\frac{g}{2}\right) + 2 n^a \tau_a \sin\left(\frac{g}{2}\right)$$

show that the field given in (2.78) has winding number one.

Exercise 2.7: Formulae for the monopole

Prove that (2.127) is equivalent to (2.126).

Exercise 2.8: Monopole in unitary gauge

Show that the gauge transformation (2.141) brings the monopole field to unitary gauge.

Exercise 2.9: Direct calculation of $\pi_2(G/H)$

Without using the homotopy exact sequence, one can calculate $\pi_2(G/H)$ directly. What follows is taken form Coleman's 1975 Erice lectures on "Classical lumps and their quantum descendants", reprinted in [Col85].

We start from the map $\phi_\infty : S^2_\infty \to G/H$. As usual in homotopy, we think of it as a map $I \times I \to G/H$, such that $\phi_\infty(t_1, t_2) = eH$ whenever t_1 or t_2 is equal to 0 or 1. Using the gauge field A we construct a map

$g_\infty : S^2_\infty \to G$ as follows:

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

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

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

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

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

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

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

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

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

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

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

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

Chapter 3

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

We have seen in the previous chapter that when the configuration space of a theory is not connected, there is a conserved topological charge and, if the dynamics is properly chosen, topological solitons. In this chapter we will consider situations where the configuration space is connected but not simply connected. The paradigm of this phenomenon in finite dimensional quantum mechanics is the Aharonov–Bohm effect. In the first part of the chapter we give several examples of theories, both in quantum mechanics and quantum field theory, with multiply connected configuration spaces. They are all characterized by the existence of a topological term that does not change the equations of motion, and hence is immaterial in the classical theory, but affects the corresponding quantum theory: the Hilbert space is split again in superselection sectors, that in the most interesting cases are parametrized by an angle θ.

In the second part we introduce the instantons, solutions of the field equations representing the motion of the system in Euclidean time through a non-contractible loop in configuration space, and having a nontrivial topological invariant. Although physically very different, solitons and instantons are mathematically very similar objects: they are regular and localized solutions of Euclidean field equations minimizing some functional (the energy for solitons, the action for instantons). Insofar as these functionals are essentially the same, we shall see in Section 3.6 that all the solitons we encountered in Chapter 2 can be reinterpreted as instantons for the Euclidean version of the same theory in one less dimension. Finally in Section 3.7 we show that quantum fluctuations around instantons give nonperturbative contributions to the path integral that can be calculated by semiclassical methods in some simple cases.

3.1 Theta sectors

3.1.1 The Aharonov–Bohm effect

Consider an electron double slit interference experiment with a solenoid placed between the two slits, carrying a magnetic flux Φ. The electrons emerging from the two slits graze the solenoid and thereafter form an interference pattern on a screen. The interference pattern is observed to depend on the flux and to repeat itself when the flux changes by[1]

$$\Delta\Phi = \frac{2\pi\hbar c}{e},$$

where e is the charge of the electron [AhB59, AhB61].

We will now give an idealized theoretical interpretation of this phenomenon. Consider 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 is typically made of iron, and we neglect the probability of an electron tunnelling through it). 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, that we neglect). As a result of these approximations, the electrons move in a configuration space $\mathcal{Q}$ which is all of $\mathbb{R}^3$ with the solenoid removed and the magnetic field vanishes on $\mathcal{Q}$. The space $\mathcal{Q}$ is multiply connected, with $\pi_1(\mathcal{Q}) = \mathbb{Z}$. Consider the magnetic potential

$$\mathcal{A} = \theta\frac{\hbar c}{2\pi e}d\varphi, \tag{3.1}$$

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

The interference pattern on the screen arises from the phase difference between waves that travel above and below the solenoid. Consider first the case

[1] For quantum mechanical systems we use Heaviside–Lorentz units, see Appendix A.1.

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 on the flux. The Hamiltonian becomes

$$H = -\frac{\hbar^2}{2m}\mathcal{D}_i\mathcal{D}_i \tag{3.2}$$

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

$$\psi(q) = \psi_0(q)e^{\frac{ie}{\hbar c}\int^q \mathcal{A}}, \tag{3.3}$$

obeys the Schrödinger equation with Hamiltonian (3.2) and the same energy eigenvalue E. The phase difference between waves that travel above and below the solenoid in the presence of the magnetic flux is equal to the phase difference in the absence of magnetic flux, plus $\frac{e}{\hbar c}\oint \mathcal{A} = \theta$. This phase, and hence the interference pattern, varies linearly with flux. When θ changes by 2π, the phase remains the same. So the interference pattern has to be periodic in Φ with period $\frac{2\pi\hbar c}{e}$, as observed. This concludes the theoretical explanation of the Aharonov–Bohm effect.

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

$$\psi' = e^{-i\alpha}\psi, \qquad \mathcal{A}' = \mathcal{A} - \frac{\hbar c}{e}d\alpha, \tag{3.4}$$

where $g(x) = e^{i\alpha(x)}$ is a function from Q into $U(1)$. We assume that the wavefunctions ψ are periodic both before and after the gauge transformation, and therefore g has to be a well-defined, single valued function into $U(1)$. Two gauge potentials $\mathcal{A}$ and $\mathcal{A}'$ are $U(1)$-gauge related only if the function g in (3.4) is single valued.

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

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

and comparing with (3.4) we see that

$$\alpha(\varphi) = \frac{\theta - \theta'}{2\pi}\varphi.$$

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

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

3.1.2 Generalization

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

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

$$\chi(\ell) = e^{\frac{ie}{\hbar c}\oint_\ell \mathcal{A}}. \tag{3.5}$$

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: $\chi(\ell) = \chi([\ell])$ (we consider only loops starting and ending at a basepoint q_* in $\mathcal{Q}$). It is easy to see, using the definition of product of homotopy classes given in Appendix E.1, that

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

so χ defines a homomorphism from $\pi_1(\mathcal{Q})$ into $U(1)$. Conversely, given any character χ, it can be shown that there exists a flat connection $\mathcal{A}$ such that (3.5) holds.

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

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

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

In the following we shall encounter only two cases: $\pi_1(\mathcal{Q}) = \mathbb{Z}$ and $\pi_1(\mathcal{Q}) = \mathbb{Z}_2$. In the former case the characters are given by $\chi_\theta(n) = e^{i\theta n}$. Since θ and $\theta + 2\pi m$, with $\in \mathbb{Z}$ define the same character, we have

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

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

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

3.1.3 The topological term

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

$$L = \frac{1}{2} m g_{ij}(q) \dot{q}^i \dot{q}^j + \frac{e}{c} \mathcal{A}_i(q) \dot{q}^i - V(q). \tag{3.6}$$

The momentum conjugate to q^i is

$$p_i = m g_{ij}(q) \dot{q}^j + \frac{e}{c} \mathcal{A}_i(q), \tag{3.7}$$

and the canonical hamiltonian is

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

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

$$\widehat{p_i - e\mathcal{A}_i} = -i\hbar\left(\frac{\partial}{\partial q^i} - \frac{ie}{\hbar c}\mathcal{A}_i\right) = -i\hbar\mathcal{D}_i, \tag{3.8}$$

where $\mathcal{D}_i$ is the covariant derivative with respect to $\mathcal{A}_i$, acting now on wave-functions $\psi(q)$. The hamiltonian becomes the operator

$$\hat{H} = -\frac{\hbar^2}{2m}\frac{1}{\sqrt{g}}\mathcal{D}_i\sqrt{g}g^{ij}\mathcal{D}_j + V \tag{3.9}$$

where $g = \det(g_{ij})$.[2] Now let us consider the special case when the connection is flat: $\mathcal{F} = 0$. There exists at least locally a function Λ such that

$$\mathcal{A}_i = \frac{\hbar c}{e}\partial_i\Lambda \tag{3.10}$$

so the second term in (3.6) is a total derivative:

$$L_T = \frac{e}{c}\dot{q}^i\mathcal{A}_i(q) = \hbar\frac{d\Lambda}{dt}. \tag{3.11}$$

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

3.1.4 Multivalued wave functions

There is an alternative description of the θ sectors that does not rely on the existence of a topological term in the Lagrangian. To arrive at it, we observe that (3.10) can be interpreted by saying that $\mathcal{A}$ is locally a gauge transform of $\mathcal{A} = 0$, with gauge function $\alpha = -\Lambda$. The corresponding function in $U(1)$ is

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

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

We have

$$\mathcal{U}\,\mathcal{D}_i\,\mathcal{U}^{-1} = \partial_i. \tag{3.12}$$

[2] 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.

Therefore, acting on the Hamiltonian (3.9),

$$\hat{H}_\theta \longmapsto \mathcal{U}\hat{H}_\theta\mathcal{U}^{-1} = -\frac{\hbar^2}{2m}\frac{1}{\sqrt{g}}\partial_i\sqrt{g}g^{ij}\partial_j + V(\varphi) = \hat{H}_0. \tag{3.13}$$

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

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

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

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

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

and we have

$$\psi'(1) = e^{i\theta}\psi'(0). \tag{3.14}$$

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

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

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

3.2 Quantum mechanical examples

3.2.1 Spin and statistics

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

A classical model for a particle with spin is the rigid body. The configuration space of this system is $Q = \mathbb{R}^d \times SO(d)$, 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 $\mathbb{Z}_2$ in higher dimensions. This is indeed what happens. We have seen that the inequivalent quantizations can be described by choosing the periodicity conditions on the wave function:

$$\psi(\omega + 2\pi) = e^{i\theta}\psi(\omega), \tag{3.15}$$

where ω is a parameter along the loop. In the case of the rotation group, fixing the axis of rotation, the parameter ω is the angle of rotation and the fundamental noncontractible loop consists of rotating the body by 2π. Therefore (3.15) describes the behaviour of the wave function under a 2π rotation. It can be compared with the definition of spin in quantum mechanics. The wave function of a system with spin s acquires a phase $e^{2\pi i s}$ when the system is rotated by 2π. So we learn that θ is equal to $2\pi s \bmod 2\pi$.

In $d > 2$, s can be either integer or half-integer. Integer spin corresponds to $\theta = 0 \bmod 2\pi$, giving single-valued wave functions, whereas half integer spin corresponds to $\theta = \pi \bmod 2\pi$ giving wave functions that change sign under 2π rotations. In two dimensions the spin can take any real value and the corresponding particles are called *anyons*. Their wave functions change by a phase under a 2π rotation.

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

$$\psi(\dots, \vec{x}_i, \dots, \vec{x}_j, \dots) = e^{2\pi i\sigma}\psi(\dots, \vec{x}_j, \dots, \vec{x}_i, \dots). \tag{3.16}$$

The usual Bose–Einstein and Fermi–Dirac statistics correspond to σ integer and half-integer respectively. To see the connection between statistics and inequivalent quantizations, consider the classical configuration space of two identical particles in d dimensions. Let us also assume that the particles

cannot be at the same point in space.[3] The configuration space is then $\mathcal{Q} = (\mathbb{R}^{2d} \setminus \Delta)/S_2$, where Δ is the subset of $\mathbb{R}^{2d}$ for which the particle positions $\vec{x}_1$ and $\vec{x}_2$ coincide, and $S_2 = \mathbb{Z}_2$ is the permutation group of two objects. Passing from the coordinates $(\vec{x}_1, \vec{x}_2)$ to the center-of-mass coordinates $(x_{\mathrm{CM}}, \Delta\vec{x}) = \left(\frac{\vec{x}_1+\vec{x}_2}{2}, \frac{\vec{x}_2-\vec{x}_1}{2}\right)$ shows that the topology of the space $\mathbb{R}^{2d} \setminus \Delta$ is $\mathbb{R}^d \times \mathbb{R}^+ \times S^{d-1}$ (here $\mathbb{R}^d$ is parametrized by $\vec{x}_{\mathrm{CM}}$, $\mathbb{R}^+$ 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 $(x_{\mathrm{CM}}, \Delta\vec{x}) \to (x_{\mathrm{CM}}, -\Delta\vec{x})$ and therefore acts antipodally on S^{d-1}; the quotient has topology $\mathbb{R}^d \times \mathbb{R}^+ \times RP^{d-1}$ where $RP^{d-1} = S^{d-1}/\mathbb{Z}_2$ is a real projective space, whose fundamental group is $\mathbb{Z}_2$. The system of two particles can therefore be quantized in two inequivalent ways, corresponding to bosonic and fermionic statistics.

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

3.2.2 The pendulum

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

$$L_0 = \frac{1}{2} I \dot{\varphi}^2 - V(\varphi), \tag{3.17}$$

where $0 \le \varphi < 2\pi$ is the coordinate on S^1, I is the moment of inertia and $V(\varphi) = V_0(1 - \cos\varphi)$ 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 the θ-Heisenberg picture the Lagrangian contains in addition a total derivative term

$$L_T = \theta \frac{\hbar}{2\pi} \frac{d\varphi}{dt}, \tag{3.18}$$

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

$$S_T(\varphi) = \theta \frac{\hbar}{2\pi} \int_{-\infty}^{\infty} dt \frac{d\varphi}{dt} = \theta \hbar W(\varphi),$$

[3] This is necessary for $\mathcal{Q}$ to be a smooth manifold. Furthermore, equation (3.16) is compatible with $\vec{x}_1 = \vec{x}_2$ only for integer σ, so if we allowed this case, the statistics could only be bosonic.

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

From a physical point of view, the term L_T can be seen as the interaction of the particle (with charge $e = 1$) with the Aharonov–Bohm magnetic potential (3.1). In fact, here we have simply restricted the motion of the Aharonov–Bohm electrons by fixing the value of z (the axial coordinate along the solenoid) and r (the distance from the center of the solenoid). The potential is physically unimportant in the case of the actual Aharonov–Bohm experiment, but its presence will be necessary in Section 3.8 for the application of the WKB method.

The Hamiltonian operator in the θ-Heisenberg picture is

$$\hat{H}_{(H)} = -\frac{\hbar^2}{2I}\mathcal{D}_\varphi \mathcal{D}_\varphi + V(\varphi), \tag{3.19}$$

where $\mathcal{D}_\varphi = \frac{\partial}{\partial\varphi} - i\frac{\theta}{2\pi}$.[4] The Hilbert space is always $\mathcal{H} = L^2(S^1)$, the space of complex functions $\psi_{(H)}(\varphi)$ such that

$$\psi_{(H)}(\varphi + 2\pi) = \psi_{(H)}(\varphi) \tag{3.20}$$

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

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

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

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

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

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

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

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

In fact

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

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

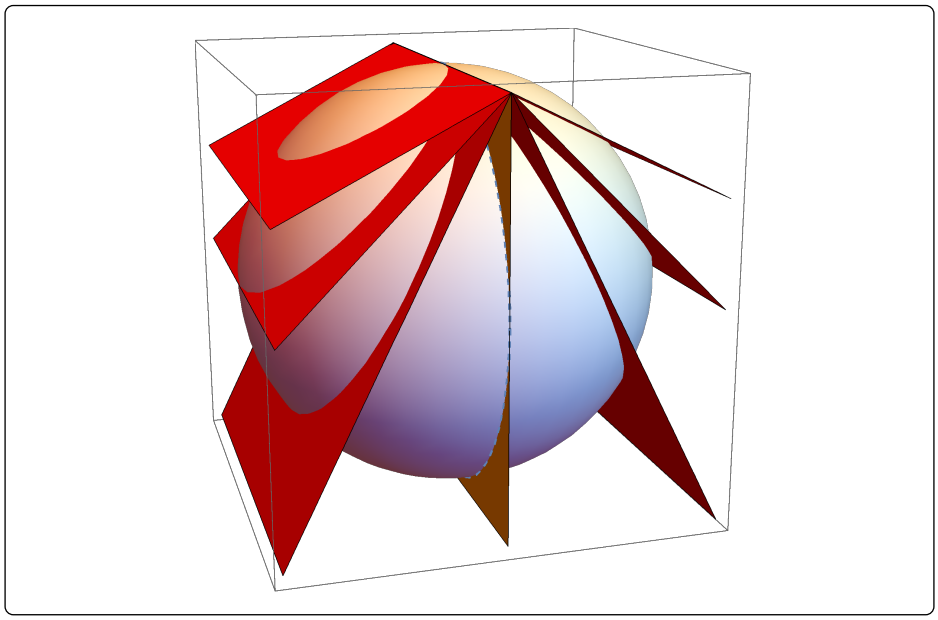


Figure 14. The intersections of the planes $z = ax + 1$ with the unit sphere, are the images of loops in the sphere based at the north pole. For $a = tan(\pi t)$ and $0 \leq t \leq 1$, this defines a noncontractible loop in the loop space of the sphere.

3.3 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 geometric formulation of Section 1.3.2, using of two fields φ^α that have the meaning of coordinates on S^2. The action is

$$S_0 = -\frac{f^2}{2}\int d^2x\, \partial_\mu \varphi^\alpha \partial^\mu \varphi^\beta h_{\alpha\beta}(\varphi), \tag{3.21}$$

where $h_{\alpha\beta}(\varphi)$ is the metric on the unit sphere. Using the same arguments of Section 2.3, but in one less dimension, the canonical configuration space of this model is $\mathcal{Q} = \Gamma_*(S^1, S^2)$, where the surfaces of constant time have been compactified to S^1 due to the requirement of finiteness of the energy. This space is called the loop space of S^2. Its fundamental group is $\pi_1(\mathcal{Q}) = \pi_2(S^2) = \mathbb{Z}$. So this theory will admit theta sectors, labelled by an angle $0 \leq \theta < 2\pi$.

The fundamental non-contractible loop in $\mathcal{Q}$ (i.e. the loop whose homotopy class generates $\pi_1(\mathcal{Q})$) can be described as follows. Points on $\mathcal{Q}$ are loops in S^2 beginning and ending at some basepoint y_0,i.e. maps $c : [0,1] \to S^2$ such that $c(0) = c(1) = y_0$. The basepoint of $\mathcal{Q}$ is the constant loop which maps all of $[0,1]$ into y_0. Consider the one-parameter family of loops c_t depicted in Figure 14. When $t = 0$ we have the constant loop. For growing t, the loops sweep out the whole sphere, and for $t \to 1$ they shrink back to the constant loop. Clearly c_t is a non-contractible loop of loops. More formally, the isomorphism between $\pi_0(\mathcal{Q})$ and $\pi_2(S^2)$ can be described as follows: if $c : I \to \mathcal{Q}$ is a loop in $\mathcal{Q}$ we define $\hat{c} : I \times I \to S^2$ by $\hat{c}(t,s) = \big(c(t)\big)(s)$, where $c(t)$, for fixed t, is regarded as a map $I \to S^2$. We have $\hat{c}(t,s) = y_0$ whenever t or s are equal to 0 or 1, so $\hat{c}$ defines a map $S^2 \to S^2$. Clearly homotopies of c correspond to homotopies of $\hat{c}$. So the desired isomorphism correspond to mapping $[c]$ to $[\hat{c}]$. (see Appendix G.2 for the general statement).

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

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

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

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

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

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

where

$$\mathcal{A}_\alpha(x) = \frac{\theta}{4\pi} \partial_x \varphi^\beta \sqrt{h} \, \varepsilon_{\alpha\beta}. \tag{3.22}$$

Comparing with equation (3.7) we see that $\mathcal{A}_\alpha$ can be regarded as a "functional magnetic potential" on $\mathcal{Q}$. In fact we can write the action $S = S_0 + S_T = \int dt (L_0 + L_T)$, with

$$L_0 = \frac{f^2}{2} h_{\alpha\beta} \dot{\varphi}^\alpha \dot{\varphi}^\beta - V(\varphi); \quad L_T = \mathcal{A}_\alpha \dot{\varphi}^\alpha.$$

This is an infinite dimensional version of the form (3.6), where we replaced the index i with the infinite indexing set (α, x). The potential is

$$V(\varphi)=\frac{f^2}{2}h_{\alpha\beta}\partial_x\varphi^\alpha\partial_x\varphi^\beta,$$

the magnetic potential is the one-form $\mathcal{A}=\int dx\mathcal{A}_\alpha(x)\delta\varphi^\alpha(x)$ and the riemannian metric is $g=\int dx h_{\alpha\beta}\delta\varphi^\alpha(x)\delta\varphi^\beta(x)$. In these formulae $\delta\varphi^\alpha(x)$ play the role of the differentials dq^i in the finite dimensional case. This terminology is further explained in Appendix G.

Since the topological term (i.e. the magnetic field) does not appear in the equation of motion, we expect that $\mathcal{F}=d\mathcal{A}=0$. This is indeed what one gets from a direct calculation based on formula (G.13), that in this case reduces to

$$d\mathcal{A}(v,w)=v\big(\mathcal{A}(w)\big)-w\big(\mathcal{A}(v)\big)-\mathcal{A}\big([v,w]\big). \tag{3.23}$$

See Exercise 3.1. There follows that $\mathcal{A}$ must be at least locally exact (as a one-form on $\mathcal{Q}$). In fact we have

$$\mathcal{A}=d\Lambda, \tag{3.24}$$

where

$$\Lambda=\theta\int dx\,\omega^0=\frac{\theta}{4\pi}\int dx\,\partial_x\varphi^\alpha\tau_\alpha. \tag{3.25}$$

If Λ was a smooth function on $\mathcal{Q}$, $\mathcal{A}$ would be a pure gauge potential in the strict sense. However, in general the function Λ is not single valued. The polidromy of Λ on the fundamental loop in $\mathcal{Q}$ is

$$\begin{aligned}\oint d\Lambda=\oint\mathcal{A}&=\int d\tau\Big[\frac{\theta}{4\pi}\int dx\partial_x\varphi^\alpha\frac{d\varphi^\beta}{d\tau}\sqrt{h}\,\varepsilon_{\alpha\beta}\Big] \\ &=\frac{\theta}{4\pi}\int d^2x\,\varepsilon^{\lambda\mu}\partial_\lambda\hat\varphi^\alpha\partial_\mu\hat\varphi^\beta\frac{1}{2}\sqrt{h}\,\varepsilon_{\alpha\beta}=\theta W(\hat\varphi)=\theta.\end{aligned} \tag{3.26}$$

In passing from the first to the second line we covariantized the expression and defined $\hat\varphi(t,x)=(\varphi(t))(x)$ (see (1.180)).

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

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

3.4 QED in 1+1 dimensions

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

$$S = S_M + S_T + S_m$$

where

$$S_M = -\frac{1}{4}\int d^2x\, F_{\mu\nu}F^{\mu\nu} \tag{3.27}$$

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

$$c_1 = \frac{1}{4\pi}\int d^2x\, \varepsilon^{\mu\nu}F_{\mu\nu} \tag{3.28}$$

is a "topological term". The topological significance of this term will be understood better in Section 3.9. For the time being we merely observe that

$$\frac{1}{4\pi}\varepsilon^{\mu\nu}F_{\mu\nu} = \partial_\mu C^\mu, \tag{3.29}$$

where

$$C^\mu = \frac{1}{2\pi}\varepsilon^{\mu\nu}A_\nu \tag{3.30}$$

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

$$P^1(x) = \frac{\partial\mathcal{L}}{\partial\partial_0 A_1(x)} = E_1(x) + \frac{\theta}{2\pi} \tag{3.31}$$

$$H = \int dx \left[\frac{1}{2}\left(P^1 - \frac{\theta}{2\pi}\right)^2 - A_0 G\right] \tag{3.32}$$

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

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

The configuration space $\mathcal{Q}$ of this theory consists of gauge and matter fields modulo gauge transformations. We denote $\mathcal{C} = \{(A_1, \Phi)\}$ the space of gauge and matter fields. and $\mathcal{G} = \Gamma_*(S^1, U(1))$ the gauge group, consisting of maps $g : \mathbb{R} \to U(1)$ such that $g \to 1$ for $|x| \to \infty$ (hence the possibility of compactifying $\mathbb{R}$ to S^1). So $\mathcal{Q} = \mathcal{C}/\mathcal{G}$. The action of $\mathcal{G}$ on $\mathcal{C}$ is free, i.e. it has no fixed points. Indeed, a gauge field A_1 that is a fixed point for a gauge transformation $g = e^{i\alpha}$ must satisfy

$$A_1 + \partial_x \alpha = A_1$$

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

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

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

The fact that $\pi_1(\mathcal{Q})$ and $\pi_0(\mathcal{G})$ are isomorphic can be proven using the homotopy exact sequence discussed in Appendix E.5. Here we describe the isomorphism. The gauge group $\mathcal{G}$ consists of infinitely many connected components $\mathcal{G}_n = \{g : S^1 \to U(1) \mid W(g) = n\}$. Now choose a basepoint $A_{(0)} = 0 \in \mathcal{C}$ (for definiteness we will take $A_{(0)} = 0$, but this is by no means necessary) and consider the orbit through $A_{(0)}$, i.e. the set of all connections of the form $A^g_{(0)} = g^{-1} dg$ for $g \in \mathcal{G}$. Since the action of $\mathcal{G}$ is free, there is a one-to-one correspondence between points of $\mathcal{G}$ and points of the orbit through $A_{(0)}$. This correspondence is also continuous in suitable topologies, so the topology of the orbit is the same as the topology of $\mathcal{G}$. There is a natural projection $p : \mathcal{C} \to \mathcal{Q}$, associating to A its gauge equivalence class $[A]$. Under this projection all points in the orbit through $A_{(0)}$ are mapped to the same point $[A_{(0)}]$ in $\mathcal{Q}$. It is natural to take $A_{(0)}$ as the basepoint in $\mathcal{C}$ and $[A_{(0)}]$ as a basepoint in $\mathcal{Q}$. Now consider a gauge transformation g with $W(g) = 1$. There is no continuous path in $\mathcal{G}$ joining g to the identity, and therefore there is also no path in the orbit through $A_{(0)}$ joining $A^g_{(0)} = g^{-1} dg$ to $A_{(0)}$. However, the space $\mathcal{C}$ is connected and so there is some path $\tilde{\ell}_t$ in $\mathcal{C}$, with $t \in [0, 1]$ such that $\tilde{\ell}_0 = A_{(0)}$

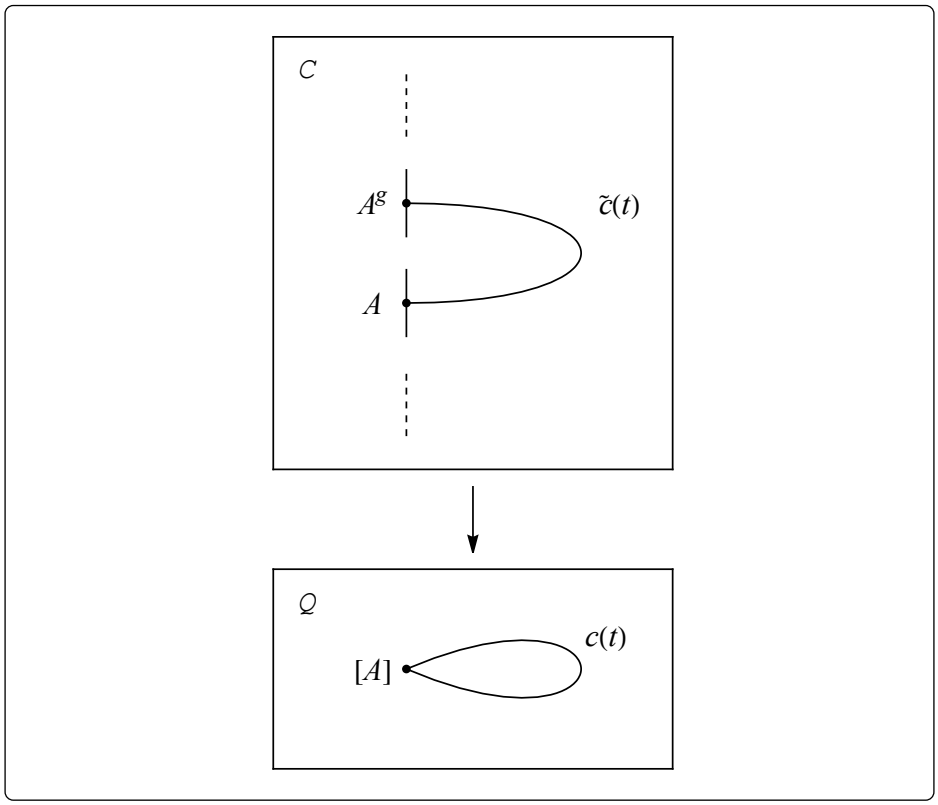


Figure 15. The space of connections $\mathcal{C}$ is a bundle over the physical configuration space $\mathcal{Q}$. The orbit through A has infinitely many connected components, depicted here as vertical segments. An open path $\tilde{c}$ joining A to A^g, with $W(g) = 1$, projects on a noncontractible loop c in $\mathcal{Q}$.

and $\tilde{\ell}_1 = A^g_{(0)}$. For instance one can take $c_t = t\, g^{-1}dg = t\, d\alpha$. The natural projection p maps this path in $\mathcal{C}$ to a path $\ell_t = [\tilde{\ell}_t]$ in $\mathcal{Q}$ beginning and ending at $[A_{(0)}]$. The desired isomorphism $\pi_1(\mathcal{Q}) \to \pi_0(\mathcal{G})$ maps the homotopy class of the loop ℓ_t in $\mathcal{Q}$ to the homotopy class of g in $\mathcal{G}$. See Figure 15.

Returning to equations (3.31) and (3.32) we see that the topological term θc_1 in the action can be written, in the gauge $A_0 = 0$, as $\int dt \int dx \dot{A}_1 \frac{\theta}{2\pi}$ and hence can be regarded as the interaction of a particle with unit charge and coordinate $A_1(x)$ with a magnetic potential (a one-form on $\mathcal{C}$)

$$\tilde{\mathcal{A}} = \int dx\, \frac{\theta}{2\pi} \delta A_1(x). \tag{3.33}$$

Since the components of the vector potential are constant, it is easy to verify

that the corresponding magnetic field $\tilde{\mathcal{F}} = d\tilde{\mathcal{A}} = 0$ (see Exercise 3.2). This is in accordance with the fact that the topological term does not contribute to the equation of motion: if it did, one could interpret the corresponding term in the equation of motion as a Lorentz force due to a nonzero $\tilde{\mathcal{F}}$. Since $d\tilde{\mathcal{A}}=0$, we can write at least locally $\tilde{\mathcal{A}}=d\tilde{\Lambda}$. The functional $\tilde{\Lambda}$ on $\mathcal{C}$ that has this property is

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

All this is on the contractible space $\mathcal{C}$. We would like now to see the corresponding steps being carried out in $\mathcal{Q}$. 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 $\mathcal{Q}$. This will be the case provided $\tilde{\Lambda}$ is constant on the orbits, i.e. if it is gauge invariant. Under a gauge transformation g,

$$\tilde{\Lambda}(A^g) - \tilde{\Lambda}(A) = \frac{\theta}{2\pi i}\int dx\, g^{-1}dg = \frac{\theta}{2\pi}\int dx\,\frac{d\alpha}{dx} = \theta W(g). \tag{3.34}$$

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

Similarly, we can ask if $\tilde{\mathcal{A}} = p^*\mathcal{A}$, i.e. if $\mathcal{A}$ is the pullback of a one-form $\mathcal{A}$ on $\mathcal{Q}$. This is true provided (see Lemma 1 on p. 294 of [KoN63], vol. II)

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

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

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

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

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

3.5 Nonabelian Yang–Mills theory in 3+1 dimensions

Except for algebraic complications, the discussion of a nonabelian YM 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 (1.120) and the gauge transformations act as in (1.122). We do not need to assume the existence of matter fields for the theory to be nontrivial, but their presence does not alter the discussion, as long as they carry linear representations of the gauge group. The total action is $S = S_{YM} + S_T$ where S_{YM} is given by (1.121), with $d = 3$ and $S_T = \theta c_2$, where

$$c_2 = \frac{1}{64\pi^2} \int d^4x \, \varepsilon^{\mu\nu\rho\sigma} F^a_{\mu\nu} F^a_{\rho\sigma} \tag{3.35}$$

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

$$\frac{1}{64\pi^2} \varepsilon^{\mu\nu\rho\sigma} F^a_{\mu\nu} F^a_{\rho\sigma} = \partial_\mu C^\mu, \tag{3.36}$$

where

$$C^\mu = \frac{1}{16\pi^2} \varepsilon^{\mu\nu\rho\sigma} \left(A^a_\nu \partial_\rho A^a_\sigma + \frac{1}{3} f_{abc} A^a_\nu A^b_\rho A^c_\sigma \right) \tag{3.37}$$

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

$$P^i_a = \frac{\partial L}{\partial \partial_0 A^a_i} = \frac{1}{e^2} E^a_i + \frac{\theta}{8\pi^2} B^a_i, \tag{3.38}$$

where $E_i^a = F_{0i}^a = \partial_0 A_i^a - D_i A_0^a$ and $B_i^a = \frac{1}{2}\varepsilon_{ijk}F_{jk}^a$. The canonical Hamiltonian is

$$H = \int d^3x \left[\frac{e^2}{2}\left(P_i^a - \theta\frac{e^2}{8\pi^2}B_i^a\right)^2 + \frac{1}{2e^2}\left(B_i^a\right)^2 - A_0^a G_a \right], \tag{3.39}$$

where $G_a = D_i E_a^i$. 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 $E_i^a = \partial_0 A_i^a$). Let $\mathcal{C}$ be the space of all gauge potentials A_i^a with finite static energy, i.e. such that $\int d^3x (B_i^a)^2$ is finite. Let $\mathcal{G}$ be the residual gauge group, consisting of time-independent gauge transformations such that $g(x) \to \mathbf{1}$ for $|\vec{x}| \to \infty$. With these boundary conditions, $\mathbb{R}^3$ can be compactified to S^3 and $\mathcal{G} = \Gamma_*(S^3, G)$. As in the previous section, $\mathcal{G}$ acts freely on $\mathcal{C}$. To see this note that if A is a fixed point for a gauge transformation g, we have

$$g^{-1}Ag + g^{-1}dg = A.$$

Thus g satisfies the equation $dg + [A, g] = 0$, which means that g is covariantly constant. If g is covariantly constant, its value at any point can be obtained from its value at another point by parallel transport. Since $g(\infty) = 1$ this implies $g = 1$ everywhere. Thus, the physical configuration space of the theory is the orbit space $\mathcal{Q} = \mathcal{C}/\mathcal{G}$, and the projection $p : \mathcal{C} \to \mathcal{Q}$ is a smooth infinite dimensional principal bundle [MiV81].

Since $\mathcal{C}$ is topologically trivial we have, following again the arguments of Appendix E.5,

$$\pi_1(\mathcal{Q}) = \pi_0(\mathcal{G}) = \pi_3(G) = \mathbb{Z}.$$

The isomorphism between $\pi_1(\mathcal{Q})$ and $\pi_0(\mathcal{G})$ is described again by Figure 15. The homotopy class $[g]$ of a gauge transformation corresponds to the homotopy class of the loop ℓ which is obtained by projecting to $\mathcal{Q}$ a curve $\tilde{\ell}$ joining $A = 0$ to $A^g = g^{-1}dg$. Comparing equations (3.38) and (3.39) with (3.9) and (3.8) we see that the topological term has given rise to a magnetic potential $\tilde{\mathcal{A}}$ on $\mathcal{C}$ defined by

$$\tilde{\mathcal{A}}(A) = \frac{\theta}{8\pi^2}\int d^3x\, B_i^a \delta A_i^a. \tag{3.40}$$

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

$$\tilde{\Lambda} = \theta\int d^3x\, C^0 = \frac{\theta}{16\pi^2}\int d^3x\, \varepsilon^{ijk}\left(A_i^a\partial_j A_k^a + \frac{1}{3}f_{abc}A_i^a A_j^b A_k^c\right). \tag{3.41}$$

(see Exercise 3.2). As in the previous section, one would like to describe the theory as a particle moving in $\mathcal{Q}$, rather than $\mathcal{C}$, so the question arises again whether the function $\tilde{\Lambda}$ and the form $\tilde{\mathcal{A}}$ can be projected onto a function Λ and a form $\mathcal{A}$ in $\mathcal{Q}$. Under a gauge transformation g, one finds

$$\tilde{\Lambda}(A^g) - \tilde{\Lambda}(A) = \theta W(g). \tag{3.42}$$

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

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

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

Then we have:

1. $\tilde{\mathcal{A}}$ is gauge invariant (B_i^a and δA_i^a both transform homogeneously);
2. $\tilde{\mathcal{A}}(v_\epsilon) = \frac{\theta}{8\pi^2}\int d^3x B_i^a D_i\epsilon^a = 0$ upon integrating by parts, using Bianchi's identity and the fact that $\epsilon \to 0$ for $|\vec{x}| \to \infty$.

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

Before closing this section we note the following geometrical interpretation of the Gauss law. Let $G_\epsilon = \int d^3x\, \epsilon^a G_a$. If the theory is quantized before eliminating all unphysical degrees of freedom, the wave functions are complex functionals on $\mathcal{C}$ and Gauss' law has to be imposed as a constraint on the physical states, as in (1.217). Upon using the quantization rule $\widehat{P}_a^i = -i\frac{\delta}{\delta A_i^a}$, we find

$$\begin{aligned} G_\epsilon\psi &= \int d^3x\, \epsilon^a D_i \left(P_i^a - \frac{\theta}{8\pi^2}B_i^a\right)\psi \\ &= i\int d^3x\, D_i\epsilon^a \left(\frac{\delta\psi}{\delta A_i^a} - i\frac{\theta}{8\pi^2}B_i^a\psi\right) \\ &= i\left(v_\epsilon\psi + i\tilde{\mathcal{A}}(v_\epsilon)\psi\right) = iv_\epsilon\psi. \end{aligned} \tag{3.43}$$

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.

3.6 Instantons

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

We call *instanton* a solution of nonlinear Euclidean field equations that

- is nonsingular,
- has finite action and
- 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 2. Indeed, we note first that a d-dimensional Euclidean spacetime is the same as "space" in a $d+1$-dimensional Lorentzian spacetime, and furthermore *the action of a d-dimensional Euclidean theory is identical to the static energy of the same theory in $d+1$-dimensional Minkowski space*. From a mathematical point of view, they are the same functional. Therefore, the static soliton solutions discussed in Chapter 2, can be recycled as instantons for the same theories in one less dimension.

As time evolves, the system traces out a path in configuration space. We will sometimes refer to such paths as *histories*. In Chapter 2 we have found it useful to visualize an instantaneous field configuration either as a function on space or as a point in the infinite-dimensional configuration space $\mathcal{Q}$. It will now be useful to visualize field histories alternatively as functions on spacetime or as paths in $\mathcal{Q}$. In particular, the histories beginning and ending at the vacuum are loops in $\mathcal{Q}$. A third point of view is to think of these histories as points in the loop space of $\mathcal{Q}$.

As with solitons, not all instantons have a topological meaning. Some non-topological instantons, and their applications, are discussed Section 3.11

and in Exercise 3.5. We will be mostly interested in topological instantons. 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 space infinity. Some of these histories can be continuously deformed into the vacuum. The topological instantons are histories that cannot. Let us now consider some examples.

3.6.1 The instanton of the pendulum and of the sigma model

The pendulum can be viewed as a field theory in zero space dimensions. The analog of the pendulum in one higher dimension is a scalar theory in 1+1 dimensions with a potential of the form $1-\cos\phi$. We have encountered this theory in Section 2.1.1: it was called the sine–Gordon (SG) model. The Euclidean action of the pendulum is

$$S_E(\varphi)=\int d\tau\left[\frac{1}{2}\left(\frac{d\varphi}{d\tau}\right)^2+\beta(1-\cos\varphi)\right]. \tag{3.44}$$

Apart from trivial changes of name of the variables, this is the same functional as the static energy (2.4) with potential (2.11). The soliton of the SG model, given in (2.12), is the instanton for the pendulum:

$$\varphi(\tau)=\pm 4\arctan\{\exp[\sqrt{\beta}(\tau-\tau_0)]\}. \tag{3.45}$$

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 (2.62). The instantons of this model in 1+1 Euclidean dimensions are the functions given in (2.75). 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 are not contractible and their homotopy class generates the fundamental group of $\mathcal{Q}$. Finally, viewed as points in the loop space of $\mathcal{Q}$, they belong to a connected component that does not contain the constant loop.

Quite generally, for any dimension of space d, the S^{d+1}-valued nonlinear sigma model has an instanton that maps compactified spacetime S^{d+1} to the target S^{d+1} with winding number one.

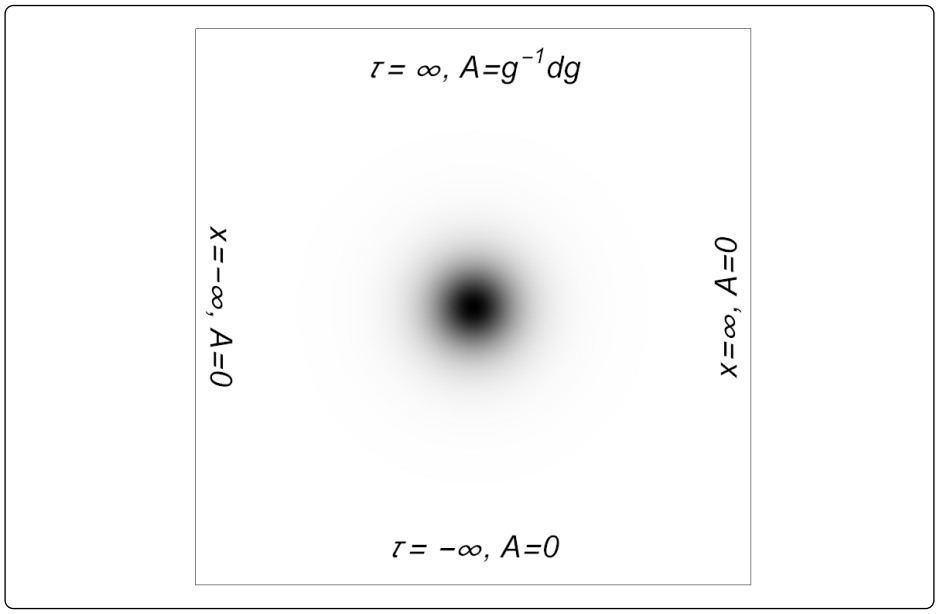


Figure 16. Density of action of the instanton. The field becomes pure gauge far away from the central core.

3.6.2 The instanton of scalar QED

The case of QED requires a little more discussion. As already anticipated in Section 3.4, in order to have a nontrivial theory we need some matter field. We consider the case of scalar QED with Euclidean action

$$S_{0E} = \int dx d\tau \left[\frac{1}{4} F_{\mu\nu} F_{\mu\nu} + \frac{1}{2} \left(D_\mu \phi \right)^* \left(D_\mu \phi \right) + \frac{\lambda}{4} \left(|\phi|^2 - f^2 \right)^2 \right], \qquad (3.46)$$

coinciding with the static energy of the same theory in 2+1 dimensions, in the gauge $A_0 = 0$, given in (2.90).

The unit 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 $\mathcal{Q}$. It follows from the discussion of Section 3.4 that this loop in $\mathcal{Q}$ is the projection of a path in $\mathcal{C}$ joining the classical vacuum $(A_{(0)}, \phi_{(0)}) = (0, f)$ to $\left(A^{g_1}_{(0)}, \phi^{g_1}_{(0)}\right) = \left(\frac{i}{e} g_1^{-1} dg_1, g_1^{-1} f\right)$, where $g_1 = e^{i\alpha}$ is a time-independent gauge transformation with winding number one, i.e. $\alpha(x \to -\infty) = 0$, $\alpha(x \to +\infty) = 2\pi$. We have found in Section 2.7 a stationary point of this functional with the boundary condition that when $r = \sqrt{x^2 + \tau^2} \to \infty$, $A_i \to \frac{i}{e} g_1^{-1} dg_1$ and $\phi \to f g_1$, where $g_1(\theta)$ is a map from S^1_∞ to $U(1)$ with winding number one: it was called the vortex. These are the

boundary conditions that we need, see Figure 16. However, in the explicit form of the solution given in (2.99), the density of winding number is uniform in all directions, whereas for our instanton discussion the density of winding number should be concentrated in the future directions. This difference is merely a matter of gauge choice. The vortex 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 3.6. Therefore, the vortex solution of the Abelian Higgs model in 2+1 dimensions with unit flux is the desired instanton solution of the Abelian Higgs model in 1+1 dimensions.

We can now understand better in what sense the integral c_1 in equation (3.28) is a "topological number". We restrict our attention to spacetime fields with finite euclidean action. This demands that when $r = \sqrt{x^2 + \tau^2} \to \infty$, $A_i \to i g_\infty^{-1} d g_\infty$ and $\phi \to g_\infty^{-1} f$, where $g_\infty(\theta)$ is a map from S^1_∞ to $U(1)$. Such maps are classified by their winding number, so the fields with finite action fall into disjoint classes, characterized by different asymptotic behaviour. These classes are usually called the topological sectors. We can now evaluate the quantity c_1 on such a field. Using (3.30) and the asymptotic form of A we get

$$c_1 = \frac{1}{4\pi} \int d^2x \, \varepsilon^{\mu\nu} F_{\mu\nu} = \frac{1}{2\pi} \int_{S^1_\infty} A = i \int_{S^1_\infty} g_\infty^{-1} d g_\infty = W(g_\infty). \qquad (3.47)$$

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

3.6.3 The BPST instanton

In our search for solitons in Chapter 2 we left out the case of pure YM theory, because we saw that the YM equations only have 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 YM theory.

We begin by giving a topological classification of four dimensional YM fields. For definiteness we consider the case $G = SU(2)$, but the generalization to other groups is quite straightforward. From the fact that the time evolution traces a continuous curve in $\mathcal{Q}$ and from the multiple connectedness of $\mathcal{Q}$, there follows that four-dimensional YM fields must fall into disjoint classes labelled by the integers. These classes can be described more explicitly as follows. We impose that $A^a_\mu(\vec{x}, \tau)$ has finite Euclidean action. This requires that at spacetime infinity, i.e. for $|x| = \sqrt{|\vec{x}|^2 + \tau^2} \to \infty$, $F^a_{\mu\nu} \to 0$. This in

turn implies

$$A_\mu \to g_\infty^{-1}\partial_\mu g_\infty, \tag{3.48}$$

where g_∞ is a function of the angles or equivalently a function from the sphere at infinity S^3_∞ to the gauge group $SU(2)$. Since $\pi_3\big(SU(2)\big) = \mathbb{Z}$, we find that the finite action gauge potentials A^a_μ fall into topologically distinct classes distinguished by their asymptotic behaviour. The topological invariant c_2 precisely measures these classes. In fact using (3.36) we can write

$$c_2 = \int_{\mathbb{R}^4} d^4x\, \partial_\mu C^\mu = \frac{1}{16\pi^2}\int_{S^3_\infty} d^3x\, \varepsilon^{ijk}\left(A^a_i\partial_j A^a_k + \frac{1}{3}f_{abc}A^a_i A^b_j A^c_k\right) = W(g_\infty). \tag{3.49}$$

The last equality is obtained by noting that on S^3_∞ we can replace A^a_i by its asymptotic form (3.48); the result then follows from Exercise 3.3. This calculation shows that for any YM field that has finite action, c_2 is an integer.[5]

The instanton has to represent the motion of the system through the fundamental noncontractible loop in $\mathcal{Q}$. We recall that in the gauge $A_0 = 0$ this means a path joining, for example $A_i = 0$ to $A_i = g^{-1}\partial_i g$, where g is a time-independent gauge transformation with winding number one. We also have $A_i = 0$ at spatial infinity. See again Figure 16. When viewed as a function on S^3_∞, g_∞ has $W = 1$, therefore such a field will have $c_2 = 1$.

To find the explicit form of the instanton, consider the inequality [BPST75]

$$\begin{aligned} 0 &\le \int d^4x\, \left(F^a_{\mu\nu} \pm {}^*F^a_{\mu\nu}\right)\left(F^{\mu\nu a} \pm {}^*F^{\nu a}\right) \\ &= 2\int d^4x\, F^a_{\mu\nu}F^{\mu\nu a} \pm 2\int d^4x\, F^a_{\mu\nu}{}^*F^{\mu\nu a}, \end{aligned} \tag{3.50}$$

which implies

$$S_E \ge \frac{8\pi^2}{e^2}|c_2|. \tag{3.51}$$

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

$$F^a_{\mu\nu} = \pm {}^*F^a_{\mu\nu}. \tag{3.52}$$

These fields are automatically solutions of the YM equations. So we have succeeded in replacing the second order YM equation by the simpler first order equations (3.52).

[5] A mathematically more sophisticated understanding of the topology of YM fields requires the language of fiber bundles and characteristic classes. See e.g. [Nak03].

Motivated by (3.48), for the instanton with $c_2 = 1$ we make an ansatz of the form

$$A_\mu(x) = f(r^2) g_1^{-1} \partial_\mu g_1, \tag{3.53}$$

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

$$g_1(x) = \begin{bmatrix} \hat{x}^4 + i\hat{x}^3 & \hat{x}^2 + i\hat{x}^1 \\ -\hat{x}^2 + i\hat{x}^1 & \hat{x}_4 - i\hat{x}^3 \end{bmatrix} = \hat{x}^\mu \alpha_\mu, \tag{3.54}$$

where $\alpha_k = i\sigma_k$ for $k = 1, 2, 3$ and $\alpha_4 = \mathbb{I}$. This function clearly has $W(g_1) = 1$. From here one finds

$$g_1^{-1}\partial_\mu g_1 = -2i\bar{\Sigma}_{\mu\rho}\frac{\hat{x}^\rho}{r}\,; \qquad g_1 \partial_\mu g_1^{-1} = -\partial_\mu g_1 g_1^{-1} = -2i\Sigma_{\mu\rho}\frac{\hat{x}^\rho}{r}, \tag{3.55}$$

where

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

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

The function f in (3.53) 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 (3.53):

$$\begin{aligned} F_{\mu\nu} &= \partial_\mu f(g^{-1}\partial_\nu g) - \partial_\nu f(g^{-1}\partial_\mu g) + f(f-1)[g^{-1}\partial_\mu g, g^{-1}\partial_\nu g] \\ &= 4i\left(\bar{\Sigma}_{\mu\rho}\hat{x}^\rho\hat{x}^\nu - \bar{\Sigma}_{\nu\rho}\hat{x}^\rho\hat{x}^\mu\right)\left(f' + \frac{1}{r^2}f(f-1)\right) - 4i\bar{\Sigma}_{\mu\nu}\frac{1}{r^2}f(f-1). \end{aligned} \tag{3.56}$$

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

$$\varepsilon_{\mu\nu\alpha\beta}\bar{\Sigma}_{\rho\beta} = -\delta_{\mu\rho}\bar{\Sigma}_{\nu\alpha} + \delta_{\nu\rho}\bar{\Sigma}_{\mu\alpha} - \delta_{\alpha\rho}\bar{\Sigma}_{\mu\nu} \tag{3.57}$$

and find

$${}^*F_{\mu\nu} = 4i\left(\bar{\Sigma}_{\mu\rho}\hat{x}^\rho\hat{x}^\nu - \bar{\Sigma}_{\nu\rho}\hat{x}^\rho\hat{x}^\mu\right)\left(f' + \frac{1}{r^2}f(f-1)\right) - 4i\bar{\Sigma}_{\mu\nu}f'. \tag{3.58}$$

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

$$f' + \frac{1}{r^2}f(f-1) = 0, \tag{3.59}$$

which is solved by

$$f(r^2) = \frac{r^2}{\lambda^2 + r^2} \tag{3.60}$$

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

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

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

where $\bar{\alpha}_k = -i\sigma_k$ for $k = 1, 2, 3$ and $\bar{\alpha}_4 = \mathbb{I}$. In this case we find

$$F_{\mu\nu} = 4i\left(\Sigma_{\mu\rho}\hat{x}^\rho\hat{x}^\nu - \Sigma_{\nu\rho}\hat{x}^\rho\hat{x}^\mu\right)\left(f' + \frac{1}{r^2}f(f-1)\right) - 4i\Sigma_{\mu\nu}\frac{1}{r^2}f(f-1) \tag{3.61}$$

and

$${}^*F_{\mu\nu} = -4i\left(\Sigma_{\mu\rho}\hat{x}^\rho\hat{x}^\nu - \Sigma_{\nu\rho}\hat{x}^\rho\hat{x}^\mu\right)\left(f' + \frac{1}{r^2}f(f-1)\right) + 4i\Sigma_{\mu\nu}f'. \tag{3.62}$$

Now solving the self-duality equation $0 = F_{\mu\nu} - {}^*F_{\mu\nu}$ leads again to equation (3.59), and hence to the same function f. Altogether the regular (antiself-dual) instanton and (self-dual) anti-instanton solutions can be written in the form

$$A_\mu = -2i\frac{\bar{\Sigma}_{\mu\nu}x^\nu}{\lambda^2 + r^2}\,; \qquad A_\mu = -2i\frac{\Sigma_{\mu\nu}x^\nu}{\lambda^2 + r^2}. \tag{3.63}$$

The respective field strengths are

$$F_{\mu\nu} = 4i\frac{\bar{\Sigma}_{\mu\nu}\lambda^2}{(\lambda^2 + r^2)^2}\,; \qquad F_{\mu\nu} = 4i\frac{\Sigma_{\mu\nu}\lambda^2}{(\lambda^2 + r^2)^2}. \tag{3.64}$$

What happens if we try to impose self-duality on the configuration $\hat{x}^\mu\alpha_\mu$ or antiself-duality on the configuration $\hat{x}^\mu\bar{\alpha}_\mu$? In both cases we arrive at the equation

$$f' - \frac{1}{r^2}f(f-1) = 0, \tag{3.65}$$

which is solved by

$$f(r^2) = \frac{\lambda^2}{\lambda^2 + r^2}. \tag{3.66}$$

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

$$A_\mu = -2i\lambda^2\frac{\Sigma_{\mu\nu}x^\nu}{r^2(\lambda^2 + r^2)}\,; \qquad A_\mu = -2i\lambda^2\frac{\bar{\Sigma}_{\mu\nu}x^\nu}{r^2(\lambda^2 + r^2)}. \tag{3.67}$$

These fields are singular in the origin. However, they are mere gauge transforms of the regular instanton and anti-instanton with a gauge transformation that is singular in the origin. In fact, let A be the antiself-dual instanton based on the ansatz (3.53) and consider the gauge transformation

$$\begin{aligned} A'_\mu &= g_1 A_\mu g_1^{-1} + g_1 \partial_\mu g_1^{-1} \\ &= (f-1)\partial_\mu g_1 g_1^{-1} \\ &= -2i\Sigma_{\mu\rho}\frac{x^\rho}{r^2}\frac{\lambda^2}{\lambda^2+r^2}, \end{aligned} \tag{3.68}$$

which coincides with the first field in (3.67). We observe that using (3.55) the same field can also be written

$$\frac{\lambda^2}{\lambda^2+r^2}(g_{-1})^{-1}\partial_\mu(g_{-1}),$$

so it becomes "pure gauge" at the origin, but with a gauge function that is the inverse of the one that describes its behavior at infinity in the regular gauge.

It is useful to observe that an instanton and an anti-instanton can be combined in a $SU(2)\times SU(2) = SO(4)$ gauge field. Then, the ansatz (3.55) is seen to be a special case of the construction discussed in Exercise 3.7.

As usual these instantons and anti-instantons are not isolated solutions but come in families parametrized by collective coordinates, or moduli. In order to discover these moduli we have to act on a solution with all the global symmetries of the theory and find when this gives rise to physically distinct solutions. The YM action is invariant under global $SU(2)$ gauge transformations and under the 15-dimensional conformal group, that consists of Poincaré transformations (10 parameters) the so-called special conformal transformations (4 parameters) and dilatations (1 parameter).[6]

The free parameter λ of the solution is clearly a modulus associated to the latter transformations. It can be shown that the special conformal transformations lead to gauge fields that are gauge equivalent to the original ones. Translations also generate four moduli: one just has to replace x^μ by $x^\mu - x_0^\mu$ in the solutions. There remains gauge transformations (3 parameters) and Euclidean rotations (6 parameters). We recall that the Euclidean rotation group $SO(4)$ is locally isomorphic to $SU(2)_L \times SU(2)_R$ (though mathematically identical, this should not be confused with the chiral rotation group of

[6]The sphere S^4 is conformally flat, so the flat space instanton can also be seen as a solution of the YM equations on the sphere. The instanton with λ equal to the radius of the sphere is invariant under $SO(5)$, see Exercise 3.8 and references [JaR76, Ore76].

the chiral models). The correspondence is as follows: if a rotation transforms x to x', then in (3.54)

$$g(x') = \hat{x}'^{\mu}\alpha_{\mu} = g_L g(x) g_R^{-1}. \tag{3.69}$$

Thus we see that the rotation transforms the gauge field (3.53) to

$$A'_{\mu} = g_R A_{\mu} g_R^{-1}. \tag{3.70}$$

The ansatz (3.53) is invariant under $SU(2)_L$ and is also invariant under the simultaneous action of $SU(2)_R$ and of the global gauge group $SU(2)$, with the same transformation parameters. Thus, of these nine transformation parameters, only three give rise to moduli, for example the global gauge transformations taken by themselves. Altogether the moduli spaces of the simple instantons and anti-instantons are eight-dimensional.

Much work has been done to find all self-dual and anti-self-dual solutions with $|c_2| > 1$. This line of research has led to important developments in mathematics, such as Donaldson theory. We shall see that such exact multi-instanton solutions are not of great practical use in the quantum theory, since the semiclassical evaluations of the path integral that we shall discuss below are based on more manageable approximate solutions.

3.7 Instantons and path integrals

3.7.1 Path integrals on multiply connected spaces

Recall that for a quantum mechanical 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

$$K(q_2, t_2 \mid q_1, t_1) = \int_{q_1,t_1}^{q_2,t_2} (dq) e^{\frac{i}{\hbar} S_0(q)}, \tag{3.71}$$

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

$$S_0 = \int dt \left[\frac{1}{2} m g_{ij}(q) \dot{q}^i \dot{q}^j - V(q) \right]. \tag{3.72}$$

We want to discuss the effects that arise when $\mathcal{Q}$ is multiply connected. We observe that the paths from q_1 to q_2 fall into homotopy classes. Clearly there are as many homotopy classes of paths from q_1 to q_2 as there are homotopy

classes of loops beginning and ending at the basepoint q_0, i.e. elements of $\pi_1(\mathcal{Q})$. However, the correspondence between homotopy classes of paths and elements of $\pi_1(\mathcal{Q})$ is not unique. To construct one such correspondence, choose two paths c_1 and c_2 joining q_0 to q_1 and q_2 respectively, see Figure 17. 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

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

where the subscript α in the measure means that the integral is performed over all paths such that $c_2^{-1} \cdot q \cdot c_1$ is in the class $\alpha \in \pi_1(\mathcal{Q})$. Since paths in different homotopy classes form disjoint sets, we can weigh differently the contribution of each homotopy class and write the total amplitude as

$$K(q_2, t_2 \mid q_1, t_1) = \sum_{\alpha \in \pi_1(\mathcal{Q})} \chi(\alpha) K_\alpha(q_2, t_2 \mid q_1, t_1). \tag{3.73}$$

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 of $\mathcal{Q}$ [LaD70]. Thus χ has to be a character of $\pi_1(\mathcal{Q})$. Each choice of χ defines an inequivalent quantum theory, so we have reached again the conclusion that inequivalent quantizations are labelled by $\mathrm{Hom}\big(\pi_1(\mathcal{Q}), U(1)\big)$.

This can be related to our preceding discussion as follows. As mentioned in Section 3.1.2 there is a one-to-one correspondence between the characters of $\pi_1(\mathcal{Q})$ and the gauge equivalence classes of flat $U(1)$ connections on $\mathcal{Q}$. If $\mathcal{A}$ is a flat connection, the corresponding character is given by

$$\chi(\alpha) = e^{\frac{ie}{\hbar c} \oint_\ell \mathcal{A}}, \tag{3.74}$$

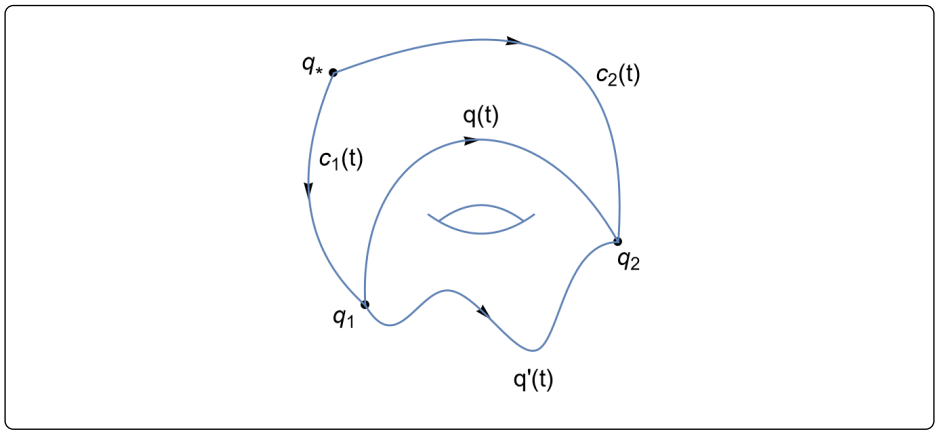


Figure 17. A multiply connected configuration space $\mathcal{Q}$ (hole in the middle) and two non-homotopic histories $q(t)$ and $q'(t)$ from q_1 to q_2. When composed with predetermined curves c_1 and c_2^{-1}, these define two loops based at q_*.

where ℓ is a loop in the homotopy class α (χ depends only on the homotopy class of ℓ since $\mathcal{A}$ is flat). There follows that if we define a "topological term"

$$S_T = \frac{e}{c}\int dt\, \dot{q}^i \mathcal{A}_i, \tag{3.75}$$

this term has the same value for all curves joining the point q_1 at the time t_1 to the point q_2 at the time t_2 and such that $c_2^{-1}\cdot q\cdot c_1$ is in a fixed homotopy class α. Thus we can absorb this term in the path integral and write:

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

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

3.7.2 Euclidean path integrals

To make the integral convergent we perform a Wick rotation to imaginary time $\tau = it$. The euclidean action is given by

$$S_{0E} = -iS_0(t=-i\tau) = \int d\tau \left[\frac{1}{2} m g_{ij}\left(\frac{dq^i}{d\tau}\right)\left(\frac{dq^j}{d\tau}\right) + V(q)\right]. \tag{3.76}$$

We will omit the subscript E from now on for notational simplicity. Putting aside the issues due to multiple connectedness for a moment, the euclidean amplitude is

$$K_E(q_2,\tau_2\,|\,q_1,\tau_1)=\int_{q_1,\tau_1}^{q_2,\tau_2}(dq)e^{-\frac{1}{\hbar}S_{0E}(q)}.$$

Let q_0 denote the vacuum (the state of lowest energy). To start with, we assume that it is unique. The vacuum-to-vacuum amplitude is also called the partition function:

$$Z_E(T)=K_E(q_0,T/2;q_0,-T/2). \tag{3.77}$$

We can extract the ground state energy of the system from the vacuum-to-vacuum amplitude, using the following trick. Denoting $\widehat{H}$ the Hamiltonian, the (Euclidean) evolution operator is $e^{-\frac{1}{\hbar}\widehat{H}\tau}$, and we have

$$Z_E(T)=\langle q_0|e^{-\frac{1}{\hbar}\widehat{H}T}|q_0\rangle=\sum_n|\langle q_0|E_n\rangle|^2e^{-\frac{1}{\hbar}E_nT},$$

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

$$\lim_{T\to\infty}Z_E(T)=\lim_{T\to\infty}|\langle q_0|E_0\rangle|^2e^{-\frac{1}{\hbar}E_0T}. \tag{3.78}$$

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 $V(q)=\frac{1}{2}\omega^2q^2$, the vacuum to vacuum amplitude turns out to be equal to (Exercise 3.4)

$$\left(\frac{\omega}{\pi\hbar}\right)^{1/2}e^{-\frac{\omega T}{2}}. \tag{3.79}$$

Comparing with (3.78) one finds the ground state energy $E_0=\frac{1}{2}\hbar\omega$.

Finally consider the case when $\mathcal{Q}$ is multiply connected and a topological term is present in the action. In general, it can be written as in (3.75). Since the topological Lagrangian only contains one time derivative, the Euclidean version of this action becomes imaginary:

$$S_{TE}=-i\frac{e}{c}\int d\tau\,\mathcal{A}_i(q)\dot{q}^i=-iS_T, \tag{3.80}$$

Thus in the path integral its contribution remains oscillatory:

$$K_E(q_2,\tau_2\,|\,q_1,\tau_1)=\int_{q_1,\tau_1}^{q_2,\tau_2}(dq)e^{-\frac{1}{\hbar}[S_{0E}(q)-iS_T(q)]}. \tag{3.81}$$

3.8 The path integral for the pendulum

The instantons give us a way of evaluating approximately the path integral on a multiply connected configuration space. The evaluation of path integrals involving instantons is often referred to as *instanton calculus*. The most important features of such calculations are already present in the simplest case of the pendulum. We will therefore begin by discussing in some detail this example.

As a preliminary, we observe that this problem is very similar to that of a particle in a periodic potential. This problem is well-known in solid-state physics. In a "zeroth-order" approximation one would expand the potential around a minimum $2\pi n$ and the lowest energy eigenfunction, with energy $E_0 = \frac{1}{2}\hbar\omega$, would be the one of the harmonic oscillator centered around $2\pi n$. There would be one such eigenfunction for each minimum, so the ground state would consist of infinitely many 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.

There is an important physical difference between this system and the pendulum: even though they have the same classical Lagrangian, in the case of the pendulum all points on the line are identified mod 2π, whereas for the particle in a periodic potential they are not. This leads to a different physical interpretation of the results.

3.8.1 The $n = \pm 1$ contributions

We are going to study the vacuum energy of the pendulum as a function of θ using the trick of Section 3.7.2. 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

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

where $n \in \mathbb{Z} = \pi_1(S^1)$ labels the homotopy classes of $\varphi(t)$ and we assume without loss of generality that $\varphi = 0$ for $T \to -\infty$. The partial amplitudes are computed here with the action S_0 corresponding to the Lagrangian (3.17) and we have introduced the characters of $\mathbb{Z}$ in the sum over the partial amplitudes, as discussed earlier. Since K_n is a path integral over loops in a fixed homotopy

class, we can bring the character inside the path integral and write

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

where we defined $\tilde{K}$ to be the amplitude in the presence of the topological term.

Next we perform the Wick rotation. As already mentioned, the euclidean topological term is imaginary: $S_{T,E} = -iS_T = -i\theta\hbar W$. Thus, the euclidean amplitude is

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

with

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

The partial amplitudes can be evaluated using the WKB, or saddle point approximation: we will now compute the contribution of fields which are near a stationary point of the Euclidean action.

Let us begin by evaluating certain contributions to $\tilde{K}_{E,1}(2\pi, T/2\,|\,0,-T/2)$, i.e. the sum over paths with winding number one. The action is minimized by the classical instanton solutions $\varphi_{\rm cl}$ given in (3.45), which are parametrized by the coordinate of the "center" τ_0, and the path integral will be dominated by configurations that are near one of these solutions. Thus we expand the action around $\varphi_{\rm cl}(\tau)$. We get

$$S_E(\varphi) = S_E(\varphi_{\rm cl}) + \frac{1}{2}\int d\tau d\tau' \eta(\tau)\mathcal{O}(\tau,\tau')\eta(\tau'),$$

where $\eta = \varphi - \varphi_{\rm cl}$ and

$$\mathcal{O}(\tau,\tau') = \frac{\delta^2 S_E}{\delta\varphi(\tau)\delta\varphi(\tau')}\bigg|_{\varphi_{\rm cl}} = \delta(\tau-\tau')\left(-\frac{d^2}{d\tau^2} + V''(\varphi_{\rm cl})\right). \tag{3.82}$$

In the WKB approximation

$$K_{E1}(2\pi, T/2\,|\,0,-T/2) = e^{-\frac{1}{\hbar}S_E(\varphi_{\rm cl})}\int (d\eta)\, e^{-\frac{1}{2}\int \eta\mathcal{O}\eta}$$
$$= e^{-\frac{1}{\hbar}S_E(\varphi_{\rm cl})}B(T)\,[\mathrm{Det}\mathcal{O}]^{-1/2} \tag{3.83}$$

where $B(T)$ is a measure factor.

The operator $-\frac{d^2}{dt^2}+V''(\varphi_{\rm cl})$ has a translational zero mode, corresponding to the fact that the position of the instanton is arbitrary. This zero mode makes the determinant zero, and the path integral ill-defined. As in the seminclassical quantization of solitons, the way out consists of replacing the zero mode by the corresponding collective coordinate τ_0. When the divergent integral over the zero mode is replaced by the integral over τ_0, it just yields a factor T. In this way we understand the meaning of the divergence of the path integral: it is automatically regulated by our device of putting the system in a "time box" and would reappear in the limit $T\to\infty$. The change in the integration variable produces a Jacobian J, whose evaluation we postpone to Section 3.8.3. The main result that is needed here is that J is independent of T for $T\to\infty$. So (3.83) can be rewritten

$$e^{-\frac{1}{\hbar}S_E(\varphi_{\rm cl})}B(T)JT\left[\mathrm{Det}'\mathcal{O}\right]^{-1/2}, \tag{3.84}$$

where Det′ is the product of the nonzero eigenvalues. The evaluation of the determinant is difficult because $\varphi_{\rm cl}$, which appears in the operator (3.82) depends explicitly on time. However, the size of the instanton was fixed by the form of the potential and is independent of T, so if we are only interested in the limit of large T, we see that "most of the time" $\varphi_{\rm cl}=0 \bmod 2\pi$ and therefore $V''(\varphi_{\rm cl})=\omega^2$. We can then write

$$\left[\mathrm{Det}'\mathcal{O}\right]^{-1/2}=K\left[\mathrm{Det}\left(-\frac{d^2}{dt^2}+\omega^2\right)\right]^{-1/2} \tag{3.85}$$

where K, the ratio of the determinants, becomes a T-independent constant 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 (3.79). We thus find

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

where we have written $S_E(\varphi_{\rm cl})=S_{0E}(\varphi_{\rm cl})-i\theta\hbar W(\varphi_{\rm cl})=S_{0E}-i\theta\hbar$. This is the contribution of the one-instanton sector to the total amplitude. By a similar argument, the one anti-instanton sector gives

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

3.8.2 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 Section 2.1.1 that there are no classical solutions to the equation $-\frac{d^2\varphi}{dt^2}+\frac{dV}{d\varphi}=0$ in the sectors $\mathcal{Q}_{0i}$ with $|i|>1$, i.e. solutions interpolating between nonadjacent minima. This means that there are no exact multi-instanton solutions around which to expand the action. Thus, we cannot directly apply the WKB method to compute the contribution of paths with winding number greater than one. In practice the calculation can still be done, but in a different way.

We observe that a configuration consisting of m_1 instantons and m_2 anti-instantons, all widely separated, will provide an approximate solution to the classical equation of motion with $W=m_1-m_2$. Such a configuration will contribute to the partial amplitude $\tilde{K}_E\big((m_1-m_2)2\pi, T/2\,|\,0,-T/2\big)$. The evaluation of the path integral for this case proceeds much as in the one-instanton case, with the following changes: every instanton gives a contribution to $S_E(\varphi_{\rm cl})$ equal to $S_{0E}-i\theta\hbar$ and each anti-instanton gives a contribution $S_{0E}+i\theta\hbar$; every instanton and anti-instanton has a translational zero mode contributing a factor TJ; as long as they are widely separated, every instanton and anti-instanton contributes a factor K when $V''(\varphi_{\rm cl})$ is replaced by ω^2 in the determinant. Altogether the contribution to the total amplitude due to configurations containing m_1 instantons and m_2 anti-instantons is

$$\frac{1}{m_1!m_2!}\exp\left[-\frac{1}{\hbar}(m_1+m_2)S_{0E}+i(m_1-m_2)\theta\right](KJT)^{m_1+m_2}\left(\frac{\omega}{\pi\hbar}\right)^{1/2}e^{-\frac{\omega T}{2}}. \tag{3.86}$$

The factor $\frac{1}{m_1!m_2!}$ is due to the indistinguishability of the instantons and anti-instantons (in the integral over the collective coordinates, the situation when instanton 1 is in position τ_1 and instanton 2 is in position τ_2 is physically the same as when instanton 1 is in position τ_2 and instanton 2 is in position τ_1). The total amplitude is obtained by summing over m_1 and m_2. This automatically includes a sum over winding numbers. The sums can be performed explicitly and we get

$$\begin{aligned} Z_\theta(T) &= \exp\left(KJTe^{-\frac{1}{\hbar}S_{0E}+i\theta}\right)\exp\left(KJTe^{-\frac{1}{\hbar}S_{0E}-i\theta}\right)\left(\frac{\omega}{\pi\hbar}\right)^{1/2}e^{-\frac{\omega T}{2}} \\ &= \left(\frac{\omega}{\pi\hbar}\right)^{1/2}\exp\left[-\frac{1}{\hbar}T\left(\frac{1}{2}\hbar\omega-2\hbar KJe^{-\frac{1}{\hbar}S_{0E}}\cos\theta\right)\right]. \end{aligned} \tag{3.87}$$

Comparing with (3.78) we find that the energy of the vacuum in the presence of the θ-term in the action is

$$E_\theta = \frac{1}{2}\hbar\omega - 2\hbar KJ e^{-\frac{1}{\hbar}S_{0E}} \cos\theta. \tag{3.88}$$

This way of computing the path integral for a theory with multiply connected $\mathcal{Q}$ is known as the *dilute instanton gas approximation*. We will see that it can be easily generalized to the case of fields theories.

A few remarks are in order here. First we observe that the non-analytic dependence on Planck's constant is a clear indication of the non-perturbative character of the result. Then, we see that instantons are related to tunnelling. If we unwrap S^1 by going to its covering space $\mathbb{R}$, we have a particle moving in a periodic potential. There are infinitely many classical vacua separated by energy barriers. Expanding the potential around one such vacuum, we get, for the low-lying states, an approximate harmonic oscillator spectrum. Thus there seem to be infinitely many degenerate vacua, each with vacuum energy $\frac{1}{2}\hbar\omega$. This picture is incorrect because it ignores tunnelling. In the WKB approximation the tunnelling amplitude is evaluated as a sum over trajectories that are near a classical solution of the equations of motion. No classical solutions exists in the real time, but as we have seen, solutions exist in the imaginary time. Thus it is the WKB approximation that requires performing the Wick rotation. In the end the amplitude can be analytically continued back to real time. The resulting θ-dependence of the energy is very similar to that of a particle in a periodic potential but with an important difference: in that case all states belong to the same Hilbert space and therefore transitions between states with different values of θ are permitted. In the case of the pendulum every value of θ defines a different theory and no transition between different θ-states can occur.

3.8.3 Evaluation of the Jacobian

The substitution of the integral over the zero mode by the integral over the collective coordinate can be justified by a procedure resembling the Faddeev–Popov method. We are going to constrain the projection of the quantum field φ onto the zero mode η_0 to be equal to the projection of the instanton on the zero mode. This requires a compensating factor that we shall evaluate.

We begin by recalling (from Section 2.1.1) that the action of the instanton comes in equal amounts from the kinetic and potential term. Thus the action of the instanton is

$$S_{cl} = \int dt \left[\frac{1}{2}\dot{\varphi}_{cl}^2 + V(\varphi_{cl})\right] = \int dt\, \dot{\varphi}_{cl}^2. \tag{3.89}$$

There follows that the normalized zero mode is

$$\eta_0(t-t_0) = \frac{1}{\sqrt{S_{cl}}} \frac{d\varphi_{cl}(t-t_0)}{dt}. \tag{3.90}$$

In what follows both the instanton φ_{cl} and the zero mode η_0 are located at a particular time t_0. Thus using (3.90) we find that the projection of the instanton on the zero mode is

$$\begin{aligned}(\varphi_{cl}, \eta_0) &= \int_{-\infty}^{\infty} dt\, \eta_0(t-t_0)\varphi_{cl}(t-t_0) \\ &= \frac{1}{2\sqrt{S_{cl}}} \int_{-\infty}^{\infty} dt \frac{d\varphi_{cl}(t-t_0)^2}{dt} \\ &= \frac{1}{2\sqrt{S_{cl}}} \varphi_{cl}^2\Big|_{-\infty}^{\infty} = \frac{2\pi^2}{\sqrt{S_{cl}}}\end{aligned} \tag{3.91}$$

and is independent of t_0. Then, we can write

$$\Delta[\varphi] \int dt_0\, \delta\left[(\varphi, \eta_0) - (\varphi_{cl}, \eta_0)\right] = 1 \tag{3.92}$$

where

$$(\varphi, \eta_0) = \int_{-\infty}^{\infty} dt\, \eta_0(t-t_0)\varphi(t)$$

will in general depend on t_0. The quantity Δ is

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

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

$$\begin{aligned}\Delta[\varphi] &= \int_{-\infty}^{\infty} dt\, \frac{d\eta_0(t-t_0)}{dt_0}\varphi(t) \\ &= \int_{-\infty}^{\infty} dt\, \eta_0(t-t_0)\frac{d\varphi(t)}{dt}.\end{aligned} \tag{3.93}$$

Now we insert the identity (3.92) in the path integral

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

and we expand the field around the instanton located at t_0. The argument of the delta function becomes (η, η_0), with η_0 centered again at t_0 and 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

$$Z = e^{-\frac{1}{\hbar}(S_{cl} + i\theta)} \sqrt{S_{cl}} \int dt_0 \int (d\eta) e^{-\frac{1}{2\hbar}(\eta, L\eta)} \delta[(\eta, \eta_0)]. \tag{3.94}$$

The path integral over η is now performed on fields that have no projection on the zero mode. It is therefore given by the primed determinant $(\mathrm{Det}\, L)^{-1/2}$. The integral over the zero mode has been replaced by the integral over the collective coordinate t_0, giving a factor T, and the Jacobian for the change of variable is

$$J = \sqrt{S_{cl}}, \tag{3.95}$$

which in the limit $T \to \infty$ is manifestly independent of t_0.

3.9 The abelian Higgs model

The perturbative spectrum of scalar QED depends on the sign of the mass term. For $f^2 < 0$, in dimension $d > 2$, it consists of a charged massive scalar, its antiparticle and a massless photon. In two dimensions there is no photon. Furthermore, the Coulomb potential grows linearly with distance and the force between two oppositely charged scalars is independent of distance. This means that the charged particles are confined in neutral bound states. When $f^2 > 0$, in any dimension including two, the theory is in the Higgs phase: there is only a neutral scalar (the radial mode) and a massive photon. Because of this, the force between charges falls off exponentially with distance. We shall see now that the effect of instantons changes the picture quite drastically: when $\theta \neq 0 \,\mathrm{mod} 2\pi$, there is no Higgs phase.

As already discussed in Section 3.6.2, the instanton of scalar QED_2 is the vortex of QED_3 with unit flux. We put the system in a large spacetime box of spatial extent L and time duration T. In the limit $T \to \infty$, the partition function

$$Z_\theta(T) = \int (dA\, d\phi\, d\phi^*)\, e^{-S_{0E} + i\theta c_1} \tag{3.96}$$

equals e^{-TE_θ}, and by evaluating Z_θ we shall obtain E_θ and other observables. Since the instantons have a fixed finite size that is negligible in the limit of large

T and L, we can evaluate Z_θ with a dilute instanton gas. The functional integral can be evaluated following the steps of the previous section. The main novelty is that now there are two translational zero modes for each instanton and anti-instanton, so the integration over the corresponding collective coordinates yields a factor LT for each instanton and anti-instanton. Thus we find

$$\lim_{T\to\infty} Z_\theta(T) = Ae^{-LT\left(C-e^{-S_{0E}}2B\cos\theta\right)} \tag{3.97}$$

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

$$\frac{E_\theta}{L} = C - e^{-S_{0E}}2B\cos\theta, \tag{3.98}$$

analogous to the result (3.88).

The physical meaning of the parameter θ can be further clarified by considering the vacuum expectation value of the electric field $\langle E_1\rangle_\theta = i\langle F_{01}\rangle_\theta$. Due to translation invariance

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

We have

$$\langle c_1\rangle_\theta = i\frac{d}{d\theta}\ln Z_\theta = -i\frac{d}{d\theta}(E_\theta T) = -iLTe^{-S_0}2B\sin\theta.$$

Therefore

$$\langle E_1(x,\tau)\rangle_\theta = 4\pi e^{-S_0}B\sin\theta. \tag{3.99}$$

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 $\widetilde{L}$. 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 charge-anticharge pair is created at some instant, brought to distance $\widetilde{L}$, then left there for a long time $\widetilde{T}$ and finally annihilated again. The classical contribution to the action due to the presence of the charges is

$$\int d^2x\, J^\mu A_\mu = q\oint A,$$

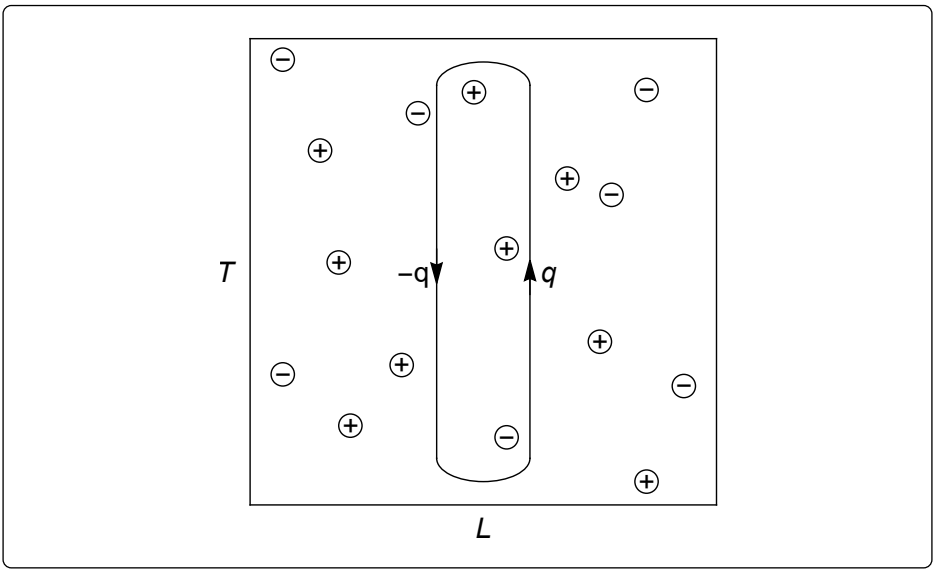


Figure 18. A sample configuration of the instanton gas with $n_+^{(\mathrm{in})} = 2$, $n_+^{(\mathrm{out})} = 7$. $n_-^{(\mathrm{in})} = 1$ and $n_-^{(\mathrm{out})} = 6$, thus total instanton charge $n = 2$.

where $J^\mu(x) = q\delta^{(2)}(x - x(t))\frac{dx^\mu}{dt}$ is the current generated by the charges. The quantity $W = e^{iq\oint A}$ is called the *Wilson loop*. As before, we enclose the system in a spacetime volume of sides $L \gg \widetilde{L}$ and $T \gg \widetilde{T}$. In the limit $T \to \infty$ the Euclidean functional integral gives the exponential of the energy in the presence of the charges:

$$\int (dAd\phi d\phi^*) e^{-S_{0E} + i\theta c_1} W = \exp\bigl(-TE_\theta - \widetilde{T}\Delta E_\theta(\widetilde{L})\bigr). \tag{3.100}$$

We have then

$$\lim_{\widetilde{T}\to\infty} \langle W\rangle = \frac{1}{Z_\theta(T)} \int (dAd\phi d\phi^*) e^{-S_{0E} + i\theta c_1 - iq\oint A} = e^{-\widetilde{T}\Delta E_\theta(\widetilde{L})}. \tag{3.101}$$

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

$$\Delta E_\theta(\widetilde{L}) = -\lim_{\widetilde{T}\to\infty} \frac{1}{\widetilde{T}} \ln\langle W\rangle_\theta. \tag{3.102}$$

We use again the dilute instanton gas approximation. We divide $n_\pm = n_\pm^{(\mathrm{in})} + n_\pm^{(\mathrm{out})}$, counting separately instantons and anti-instantons that lie inside or

outside the spacetime loop traced by the charges. See Figure 18. The reason for this is that the Wilson loop can be rewritten:

$$W = e^{iq\oint A} = e^{\frac{iq}{2}\int_U d^2x\epsilon^{\mu\nu}F_{\mu\nu}} = e^{2\pi iq(n_+^{(\text{in})}-n_-^{(\text{in})})}$$

where U is the region enclosed by the loop. Then, the functional integral (3.100) can be evaluated as follows:

$$\begin{aligned} &Ae^{-LTC}\sum_{n_+^{(\text{in})},n_-^{(\text{in})},n_+^{(\text{out})},n_-^{(\text{out})}}\frac{1}{n_+^{(\text{in})}!n_-^{(\text{in})}!n_+^{(\text{out})}!n_-^{(\text{out})}!} \\ &\times\exp\Big[-(n_+^{(\text{in})}+n_-^{(\text{in})}+n_+^{(\text{out})}+n_-^{(\text{out})})S_{0E} \\ &\qquad +i\theta(n_+^{(\text{in})}-n_-^{(\text{in})}+n_+^{(\text{out})}-n_-^{(\text{out})})\Big] \\ &\times\left[B(LT-\widetilde{L}\widetilde{T})\right]^{n_+^{(\text{out})}+n_-^{(\text{out})}}(B\widetilde{L}\widetilde{T})^{n_+^{(\text{in})}+n_-^{(\text{in})}}\exp\left[2\pi iq(n_+^{(\text{in})}-n_-^{(\text{in})})\right] \\ &\quad = A\exp\left\{-LTC+2Be^{-S_{0E}}\left[\widetilde{L}\widetilde{T}\cos(\theta+2\pi q)+(LT-\widetilde{L}\widetilde{T})\cos\theta\right]\right\}. \end{aligned} \tag{3.103}$$

Using (3.97) and (3.103) in (3.102) we get

$$\Delta E_\theta(\widetilde{L}) = 2Be^{-S_{0E}}\widetilde{L}\left[\cos\theta-\cos(\theta+2\pi q)\right]. \tag{3.104}$$

From this formula we see that the potential grows with distance, leading again to confinement of the charges. Thus, the physical picture is the same for $f^2>0$ as for $f^2<0$. From the factor $e^{-S_{0E}}$ we see, however, that the force is strictly nonperturbative (the numerator contains a hidden factor $1/\hbar$) and that it vanishes exponentially in the classical limit.

To get a physical intuition for the θ- and q-dependence, we can expand for small θ and q and find

$$\langle E_1\rangle_\theta = 4\pi Be^{-S_{0E}}\theta \tag{3.105a}$$

$$E_\theta = BLe^{-S_{0E}}\theta^2 \tag{3.105b}$$

$$\Delta E_\theta(\widetilde{L}) = B\tilde{L}e^{-S_{0E}}\left[(\theta+2\pi q)^2-\theta^2\right]. \tag{3.105c}$$

In the θ vacuum there is a constant electric field and an energy density proportional to the square of this electric field. External charges in one dimension act as the plates of a capacitor and produce an additional constant electric field in the space between them. The shift in energy due to the charges is the distance between the charges, times the difference in the energy density in the presence and in the absence of the charges.

Returning to the full result (3.104), we see that if the charges are integer, the force vanishes. 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.

3.10 Vacuum tunnelling in Yang–Mills theory

The results of the previous section raised hopes that instantons may provide an understanding of confinement also in four dimensions. This did not work, for various reasons. One is that two dimensions are special because a loop in the plane (or on a sphere) divides the space in two disjoint regions. Thus, there is a clear meaning to the statement that an instanton is inside or outside the Wilson loop. This does not happen in higher dimensions. Another problem is that the dilute gas approximation is questionable: the YM action is scale invariant, and, at the same cost in action, instantons can have arbitrary size. Thus, at fixed instanton density, there will always be large instantons that overlap. If one nevertheless tries to repeat the instanton gas calculation, a more technical issue appears: the integral over quantum fluctuations has both ultraviolet and infared divergences.

Let us give here the main steps of such a calculation. We use the background field method and split

$$A^a_\mu = \bar{A}^a_\mu + a^a_\mu,$$

where $\bar{A}$ is the BPST instanton solution. Denoting $\bar{F}^a_{\mu\nu}$ the curvature of the background field, we have

$$F^a_{\mu\nu} = \bar{F}^a_{\mu\nu} + \bar{D}_\mu a^a_\nu - \bar{D}_\nu a^a_\mu + f^a{}_{bc} a^b_\mu a^c_\nu,$$

where

$$\bar{D}_\mu a^a_\nu = \partial_\mu a^a_\nu + \bar{A}^b_\mu f^a{}_{bc} a^c_\nu$$

is the covariant derivative with respect to the background field. We see that the Yang–Mills action is quartic in the quantum field a^a_μ. For our purposes it is enough to keep terms up to second order in a^a_μ:

$$S_{YM}(A) = A_{YM}(\bar{A}) + \frac{1}{2g^2}\int d^x\, a^a_\mu \left(-\bar{D}^2\delta^{\mu\nu} a^a_\nu - \bar{D}^\nu \bar{D}^\mu a^a_\nu - [\bar{F}^{\mu\nu}, a_\nu]^a\right) + O(a^3). \tag{3.106}$$

The terms linear in a vanish because the background satisfies the YM equations.

Now one faces the usual problem that the operator appearing in this expression has an infinite dimensional kernel consisting of infinitesimal gauge transformations applied to the background. An infinitesimal gauge transformation with parameter ϵ^a

$$\delta_\epsilon A^a_\mu = D_\mu \epsilon^a = \partial_\mu \epsilon^a + f^a{}_{bc} A^b_\mu \epsilon^c$$

can be split in different ways between background and fluctuation. One is to keep the background fixed and attribute all the variation to the quantum field:

$$\begin{aligned} \delta^{(Q)}_\epsilon \bar{A}^a_\mu &= 0, \\ \delta^{(Q)}_\epsilon a^a_\mu &= D_\mu \epsilon^a. \end{aligned} \tag{3.107}$$

These are called "quantum gauge transformations". The other is to split the transformation evenly so that the background transforms as a connection and the quantum field as a matter field in the adjoint representation:

$$\begin{aligned} \delta^{(B)}_\epsilon \bar{A}^a_\mu &= \bar{D}_\mu \epsilon^a = \partial_\mu \epsilon^a + f^a{}_{bc} \bar{A}^b_\mu \epsilon^c, \\ \delta^{(B)}_\epsilon a^a_\mu &= f^a{}_{bc} a^b_\mu \epsilon^c. \end{aligned} \tag{3.108}$$

These are called "background gauge transformations". The Yang–Mills action is obviously invariant under both quantum and background transformations. The gauge fixing term is meant to break the quantum gauge transformations but it is possible, and in fact extremely advantageous, to choose it in such a way as to preserve the background gauge invariance. We choose the covariant gauge condition $\bar{D}_\mu a^{\mu a} = 0$, which is implemented in the functional integral by adding to the action the gauge-fixing term

$$S_{GF}(a; \bar{A}) = \frac{1}{2g^2\alpha} \int d^4x (\bar{D}_\mu a^{\mu a})^2,$$

where α is a gauge parameter. The corresponding ghost operator is obtained by varying the gauge condition under a quantum gauge transformation:

$$\delta^{(Q)}_\epsilon \bar{D}_\mu a^{\mu a} = \Delta_{gh} \epsilon^a,$$

which yields $\Delta_{gh} = \bar{D}_\mu D^\mu$. We will only consider the case when the expectation value of the quantum field a_μ is zero, so that the ghost operator is just the covariant Laplacian acting on $\mathfrak{su}(2)$-valued functions

$$\Delta^{(0)} = -\bar{D}^2.$$

Thus we have to add to the action the ghost term

$$S_{gh} = \int d^4x\, \bar{c}_a \Delta^{(0)ab} c_b. \tag{3.109}$$

It is convenient to choose the Feynman gauge $\alpha = 1$. In this case a straightforward calculation shows that non-minimal terms of the form $a_\nu \bar{D}_\mu \bar{D}^\nu a^\mu$ in (3.106) are removed, and the remaining quadratic part of the action is

$$S^{(2)} = S^{(2)}_{YM} + S_{GF} = \frac{1}{2g^2} \int d^4x\, a^a_\mu \Delta^{(1)\mu\nu}_{ab} a^b_\nu, \tag{3.110}$$

where

$$\Delta^{(1)\mu\nu}_{ab} = -g^{\mu\nu} \bar{D}^2_{ab} + E^{\mu\nu}_{ab}\,; \qquad E^{\mu\nu}_{ab} = -2 f_{acb} \bar{F}^{\mu\nu c}. \tag{3.111}$$

With less index clutter, one can write

$$\Delta^{(1)} a_\mu = -D^2 a_\mu - 2[F_{\mu\nu}, a^\nu].$$

The one-loop contribution to the partition function of the one-instanton sector is the Gaussian integral

$$\begin{aligned} Z_1(\bar{A}) &= \frac{\int (da\, d\bar{c}\, dc) e^{-S^{(2)}(\bar{A},a) + S_{gh}(\bar{A},\bar{c},c))}}{\int (da\, d\bar{c}\, dc) e^{-S^{(2)}(0,a) + S_{gh}(0,\bar{c},c))}} \\ &= e^{-S_{YM}(\bar{A})} \frac{\sqrt{\det \Delta^{(1)}(0)}}{\sqrt{\det \Delta^{(1)}(\bar{A})}} \frac{\det \Delta^{(0)}(\bar{A})}{\det \Delta^{(0)}(0)}. \end{aligned} \tag{3.112}$$

We have normalized Z_1 dividing by the functional integral in the absence of the instanton.[7]

The calculation of this amplitude is rather lengthy and has been performed first by 't Hooft in [tHo76]. We follow a somewhat simpler method put forward in [BeP77, CDDN77], that uses the $SO(5)$-invariance of the instanton on a sphere, discussed in Exercise 3.8. The main steps of the evaluation of the amplitude are given in Exercise 3.9, with the final result

$$Z_1 = \text{constant} \times \left(\frac{8\pi^2}{g_B^2}\right)^4 \int d^4x_0 \int \frac{d\lambda}{\lambda^5} e^{-\frac{8\pi^2}{g_B^2} + \frac{11}{3} \log(\lambda^2 \Lambda^2_{UV})}, \tag{3.113}$$

[7] This is analogous to what we did in Section 2.1.3 when we calculated the renormalization of the kink mass, and is useful to remove certain divergences.

where Λ_{UV} is an ultraviolet cutoff. We have written the YM coupling as g_B to emphasize that it is the bare coupling. If in the exponential we replace it by the renormalized coupling, defined by

$$\frac{1}{g_B^2} = \frac{1}{g_R^2(\mu)} + \frac{1}{8\pi^2}\frac{11}{3}\log\frac{\Lambda_{UV}^2}{\mu^2}. \tag{3.114}$$

we find that the ultraviolet divergence is removed. The exponent becomes

$$-\frac{8\pi^2}{g_R^2(\mu)} + \frac{11}{3}\log\lambda^2\mu^2 + \text{constant},$$

and given that the renormalized coupling satisfies, in perturbation theory, the renormalization group equation

$$\mu\frac{dg_R}{d\mu} = -\frac{1}{(4\pi)^2}\frac{22}{3}g_R^3, \tag{3.115}$$

it is actually independent of the arbitrary renormalization scale μ. In fact, it is natural to choose $\mu = 1/\lambda$, in which case the exponent becomes

$$-\frac{8\pi^2}{g_R^2(1/\lambda)} + \text{constant}.$$

Altogether the one-instanton contribution to the tunnelling amplitude is given by

$$Z_1 = \text{constant} \times \left(\frac{8\pi^2}{g_B^2}\right)^4 \int d^4x_0 \int \frac{d\lambda}{\lambda^5} e^{-\frac{8\pi^2}{g_R^2}}. \tag{3.116}$$

Solving the renormalization group equation

$$\frac{8\pi^2}{g_R^2(1/\lambda)} = \frac{8\pi^2}{g_0^2} - \frac{22}{3}\log\left(\frac{\lambda}{\lambda_0}\right).$$

and inserting back in the amplitude the integral over instanton size is

$$\int \frac{d\lambda}{\lambda^5}\left(\frac{\lambda}{\lambda_0}\right)^{22/3}.$$

We see that the integration is convergent for small λ but divergent for large λ. Unlike the divergent integral over the position of the instanton, that can be understood as the spacetime volume where the instanton can be, this divergence does not have a satisfactory interpretation. It occurs in the infrared regime where YM theory is strongly coupled, and the present one loop calculation is not reliable. Thus, it is a signal that a different, nonperturbative estimate of the effect is needed. We will not pursue this topic further.

3.11 False vacuum decay

We discuss here an application of instantons that is not related to a multiply connected configuration space, namely the decay of a metastable vacuum. This issue arises in first order phase transitions. As discussed in Section 1.2.2, a typical example in statistical physics would be a ferromagnet below the Curie temperature. The main difference is that in our treatment we do not consider thermal fluctuations and the transition is driven by quantum fluctuations. To be specific we will assume that the system is described by a scalar theory with quartic potential, tilted by the addition of an infinitesimal linear term:

$$V(\phi) = \frac{\lambda}{4}\left(\phi^2 - f^2\right)^2 + \frac{\epsilon}{2f}\phi, \tag{3.117}$$

where $\epsilon > 0$. To first order in ϵ, the two minima and the maximum are located at

$$\phi_\pm = \pm f - \frac{\epsilon}{4\lambda f^3}, \qquad \phi_{\max} = \frac{\epsilon}{2\lambda f^3} > 0$$

and the respective potentials are

$$V(\phi_+) = \epsilon/2, \qquad V(\phi_-) = -\epsilon/2, \qquad V(\phi_{\max}) = \lambda f^4/4 + O(\epsilon^2).$$

The difference in energy density of the two minima is $\mathcal{E} = \epsilon$. Recall from Section 2.2.1 that in $d > 1$ the kink can be reinterpreted as the profile of the walls separating domains of different vacua. The thickness of the wall is $\ell \sim 1/(\sqrt{\lambda} f)$ and the "surface" energy density of the wall is $\mathcal{T} \sim \sqrt{\lambda} f^3$. These formulas were derived for planar domain walls, but they will still be approximately correct if the radius of curvature of the surface is $R \gg \ell$. This is called the *thin wall approximation*.

The other simplifying assumption is spherical symmetry. Thus consider a d-dimensional system in the homogeneous metastable state ϕ_+ and suppose that a spherical bubble of true vacuum forms inside the metastable vacuum. In the thin wall approximation, this is an easily understandable one-dimensional problem, whose only variable is the radius of the bubble. If we shift the potential in such a way that the energy of the metastable vacuum is zero, the energy of the bubble has a negative term proportional the bulk volume and a positive one proportional to the surface:

$$E(r) = -\mathcal{E} V_{B^d}(r) + \mathcal{T} V_{S^{d-1}}(r). \tag{3.118}$$

The volume of the $(d-1)$-dimensional sphere of radius r is

$$V_{S^{d-1}}(r) = (4\pi)^{(d-1)/2} r^{d-1} \Gamma((d-1)/2)/\Gamma(d-1)$$

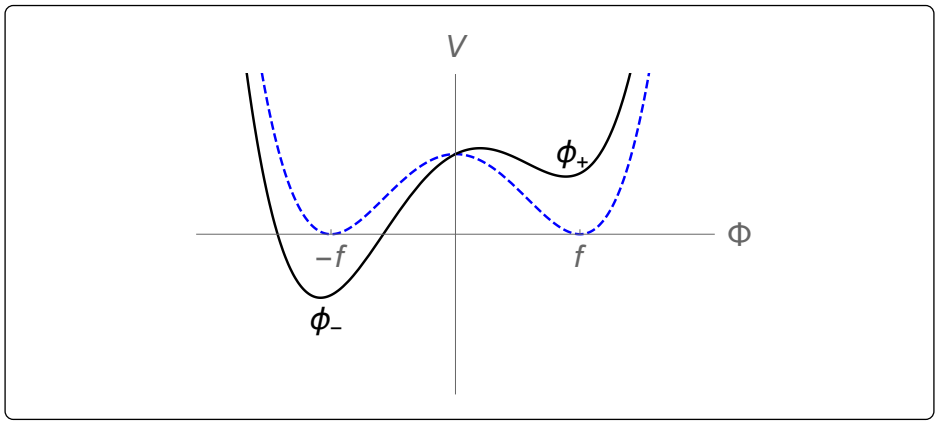


Figure 19. Quartic potential with linear tilt (black curve) vs. the original potential (dashed). The two minima have moved to the left and the maximum has moved to the right.

and the volume of the d-dimensional ball of radius r is

$$V_{B^d}(r) = \pi^{d/2} r^d / \Gamma((d+2)/2).$$

The energy difference is plotted for various dimensions in Figure 20.

The bubble will be in equilibrium if

$$\frac{dE(r)}{dr} = 0$$

and this happens at

$$r_e = (d-1)\frac{\mathcal{T}}{\mathcal{E}}. \tag{3.119}$$

(Note that there can be no equilibrium in $d = 1$.) If the bubble has radius $r < r_e$ the surface tension will dominate and make the bubble shrink and disappear. If it has radius $r > r_e$ the bulk energy will dominate causing the bubble to expand to infinity.

It is also of interest to consider bubbles that involve neither a gain nor a loss of energy. Such bubbles can form spontaneously by quantum tunnelling, as we shall discuss below. Imposing that $E = 0$ we see that the radius of such a bubble is

$$r_0 = d\frac{\mathcal{T}}{\mathcal{E}} > r_e \tag{3.120}$$

and therefore such zero-energy bubbles will expand.

Let us see under what circumstances the results of this simplified one-dimensional picture are reliable. The thin wall approximation is justified if

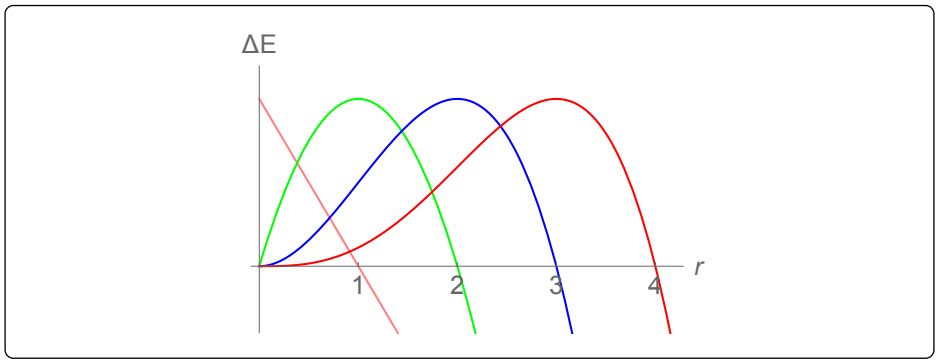


Figure 20. Energy of a vacuum bubble with $\mathcal{T} = 1$ and $\mathcal{E} = 1$ in $d = 1$ (straight pink line), $d = 2$ (green), $d = 3$ (blue) and $d = 4$ (red), all rescaled so as to have the same maximum. Note that the zero-energy radius in d dimensions is the same as the equilibrium radius in $d + 1$ dimensions.

$r_e \gg \ell$. Neglecting numerical factors of order one, this means

$$1 \ll \frac{r_e}{\ell} \approx \frac{\mathcal{T}}{\mathcal{E}\ell} \approx \lambda \frac{f^4}{\epsilon},$$

which can always be satisfied if ϵ is sufficiently small. In the following we will always assume that this condition is satisfied and that the radius of the bubble is not much smaller than r_e.

So far we have limited ourselves to discussing equilibrium conditions. Let us now study the bubble dynamics in the physically most interesting case $d = 3$. Again, the assumption of spherical symmetry reduces this to a one-dimensional problem. The bubble acts like a particle at position r with a position-dependent mass $M = 4\pi r^2 \mathcal{T}$, so the Lagrangian is given by a kinetic term, minus the "potential" energy (3.118):

$$L = \frac{1}{2} M \dot{r}^2 + \frac{4}{3}\pi r^3 \mathcal{E} - M. \tag{3.121}$$

The first and third terms can be seen as coming from the expansion of the square root of the following relativistic Lagrangian:

$$L = -M\sqrt{1 - \dot{r}^2} + \frac{4}{3}\pi r^3 \mathcal{E}. \tag{3.122}$$

From here we now derive the relativistic equations for the system. The momentum conjugate to r is

$$p = \frac{M\dot{r}}{\sqrt{1 - \dot{r}^2}}, \tag{3.123}$$

which can be inverted to

$$\dot{r} = \frac{p}{\sqrt{p^2 + M^2}}. \tag{3.124}$$

The Hamiltonian, written as a function of the velocity, is

$$H = \frac{M}{\sqrt{1 - \dot{r}^2}} - \frac{4}{3}\pi r^3 \mathcal{E}. \tag{3.125}$$

From here, using (3.123), (3.124), we obtain the relation

$$\left(H + \frac{4}{3}\pi r^3 \mathcal{E}\right)^2 = \left(\frac{M}{\sqrt{1 - \dot{r}^2}}\right)^2 = \left(\frac{p}{\dot{r}}\right)^2 = p^2 + M^2. \tag{3.126}$$

Thus for a bubble nucleated from vacuum, with $H = 0$, we have

$$p^2 = \left(\frac{4}{3}\pi r^3 \mathcal{E}\right)^2 - M^2 = M^2 \left(\frac{r^2}{r_0^2} - 1\right), \tag{3.127}$$

and inserting this in (3.124) we obtain

$$\dot{r} = \sqrt{1 - \frac{r_0^2}{r^2}}. \tag{3.128}$$

This is real only if $r \geq r_0$. Let us begin with a bubble of radius $r = r_0$ and $\dot{r} = 0$. Solving the equation of motion (3.128) one finds

$$r(t) = \sqrt{r_0^2 + t^2}. \tag{3.129}$$

This is a relativistic uniformly accelerated motion, starting at rest and approaching asymptotically a light cone.

We now wish to discuss the process of bubble nucleation. This will be done using instanton methods. If the system is at rest in the state ϕ_+, there is no classical solution of the equations of motion that will generate a spherical bubble of true vacuum; it does not have the energy that is necessary to overcome the potential barrier. However, there are solutions of the Euclidean equations that can do this. The simplest way to see this is to observe that, as in all cases considered before, the Euclidean action of the theory in d dimensions is the same as the static energy of the same theory in $d + 1$ dimensions. Thus, for a spherically symmetric Euclidean bubble in the thin wall approximation, the action is equal to (3.118), with $d = 4$. We know already that this action

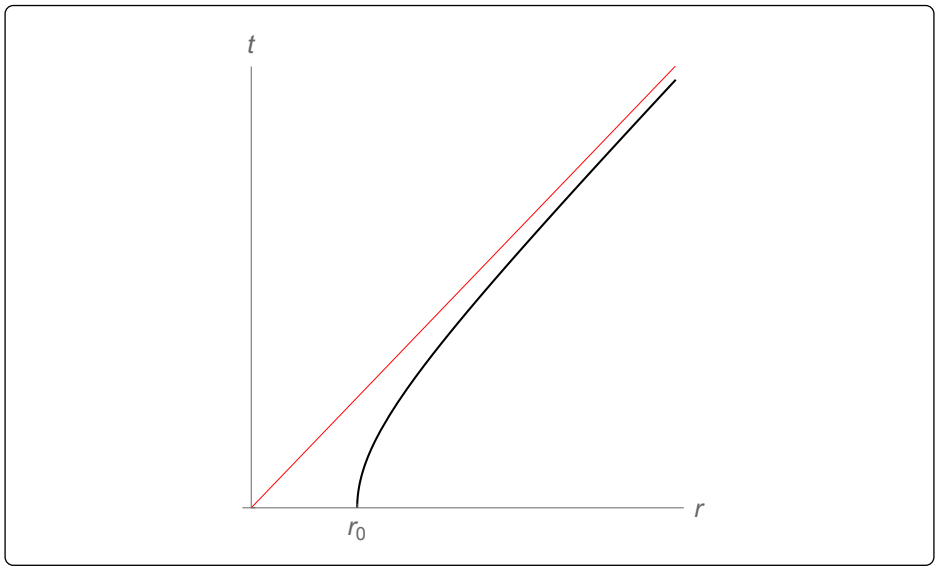


Figure 21. Expansion of a vacuum bubble starting at rest.

has a stationary point: it is the maximum of the red curve in Figure 20. Therefore this represents the instanton in the simplified model with one degree of freedom.

For a proper evaluation of the nucleation rate we need to know the instanton of the original scalar theory from which (3.118) was derived, and in this case it is a bit more complicated to see that an instanton exists. The instanton will be spherically symmetric in the four-dimensional sense[8] and thus will be given by a function $\phi(r)$ (where $r = \sqrt{x_1^1 + x_2^2 + x_3^2 + \tau^2}$) with the boundary conditions that $\phi \to \phi_+$ for $r \to \infty$, and $\phi'(0) = 0$ (as required by spherical symmetry). The Euclidean field equation for this function is

$$\frac{d^2\phi}{dr^2} + \frac{3}{r}\frac{d\phi}{dr} = V'(\phi). \tag{3.130}$$

By the usual device of reinterpreting ϕ as position and x as time, this can be seen as Newton's equation for a particle in the potential $-V$, in the presence of a time-dependent friction term. Following Coleman, the existence of a solution with the desired boundary conditions can be argued as follows. Let us shift the potential such that $V(\phi_+) = 0$ and let ϕ_0 be the zero of V that lies nearest to ϕ_+, see Figure 22. The boundary condition $\phi'(0) = 0$ means that

[8] It can be shown that an $SO(4)$-symmetric instanton has lower action than any non-symmetric instanton, see [CGM77].

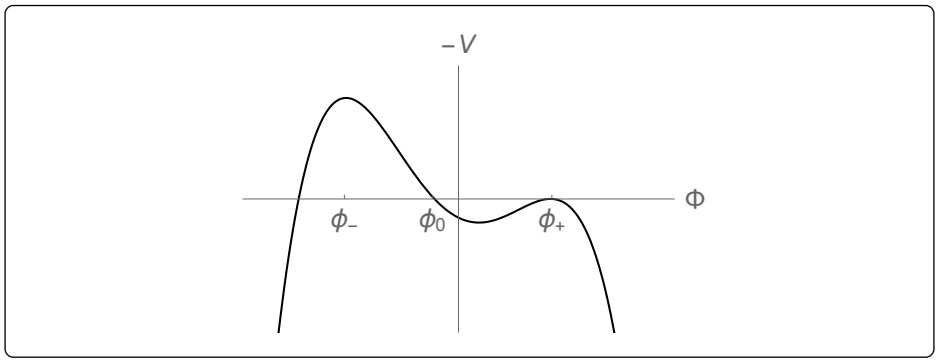


Figure 22. For the solution of (3.119).

the particle is released at time zero at rest. If $\phi(0)$ is on the right of ϕ_0, the particle does not have enough energy to reach ϕ_+. If $\phi(0)$ is on the left of ϕ_0 and very close to ϕ_-, the particle will spend a long time near ϕ_-. During this time the friction term can become arbitrarily small. When the particle finally rolls down the potential, the friction term can be neglected and the particle overshoots. Thus there must exist an initial position ϕ_* between ϕ_- and ϕ_0 such that if $\phi(0) = \phi_*$, the particle will just reach ϕ_+ and come to rest there. When we replace the fictitious time by radius, the resulting function $\phi(r)$ is a solution of (3.119) with the desired boundary conditions. The profile of the solution is not known analytically, but in the thin wall limit $\epsilon \to 0$, $\phi_* \to \phi_-$ and the profile becomes that of a kink at radius r_e. In this limit the action of the instanton is

$$S_{cl} = -\mathcal{E}V_{B^4}(r_e) + \mathcal{T}V_{S^3}(r_e) = \frac{27}{2}\pi^2\frac{\mathcal{T}^4}{\mathcal{E}^3} \tag{3.131}$$

and the tunnelling amplitude in the classical approximation is proportional to $e^{-S_{cl}}$. We note that this conclusion is in agreement with a standard one-dimensional quantum mechanical treatment. Indeed, in the WKB approximation, the tunnelling amplitude is proportional to

$$\exp\left(-2\int_0^{r_0} dr|p(r)|\right)=\exp\left(-2\int_0^{r_0} dr 4\pi r^2\mathcal{T}\sqrt{1-\frac{r^2}{r_0^2}}\right)=\exp\left(-\frac{\pi^2}{2}r_0^3\mathcal{T}\right),$$

and we see that the exponent agrees with (3.131) for $d = 3$.

Returning to the field-theoretic picture, we could proceed as with the other instanton calculations, and try to interpret the partition function as e^{-ET}, where E is the energy of the ground state. Unlike the other calculations,

however, the system does not have a ground state because the energy is unbounded from below. The instability of the state manifests itself in the energy having an imaginary part, as we shall see next.

Let ϕ_{cl} be the instanton solution described above and let $\phi_\lambda(x) = \phi_{cl}(\lambda x)$. We have

$$S(\phi_\lambda) = \frac{1}{2}\lambda^{-2}\int d^4x(\partial\phi_{cl})^2 + \lambda^{-4}\int d^4xV(\phi_{cl}). \tag{3.132}$$

Deriving with respect to λ and putting $\lambda = 1$ we find that

$$0 = \frac{dS(\phi_\lambda)}{d\lambda}\Big|_{\lambda=1} = -\int d^4x(\partial\phi_{cl})^2 - 4\int d^4xV(\phi_{cl}). \tag{3.133}$$

This implies that $S(\phi_{cl}) = \frac{1}{4}\int d^4x(\partial\phi_{cl})^2 > 0$. Deriving a second time

$$0 = \frac{d^2S(\phi_\lambda)}{d\lambda^2}\Big|_{\lambda=1} = 3\int d^4x(\partial\phi_{cl})^2 + 20\int d^4xV(\phi_{cl}) = -2\int d^4x(\partial\phi_{cl})^2 < 0. \tag{3.134}$$

This shows that the kinetic operator has one negative mode. It corresponds to the radius in the simplified model with one degree of freedom, and we know already that the solution is a maximum for the radius. It can be shown that there are no other negative modes. Therefore the determinant $\left(\det\left(\frac{\delta^2S}{\delta\phi\delta\phi}\right)\right)^{-1/2}$ is purely imaginary. This fact is important for the following reason. We can evaluate the functional integral in the dilute gas approximation leading to the expression

$$Ae^{-VT(C-J^4Ke^{-S_{cl}})},$$

where A is the determinant of the harmonic oscillator states, J is the Jacobian associated to each zero mode, calculated in (3.95), and K is the ratio of determinants defined as in (3.85), adapted to the present problem. J appears to the fourth power because the instanton has four translational zero modes that get converted to the spacetime volume VT. It follows from the previous remark that K is purely imaginary. When we read off the energy from the previous expression we find that it has an imaginary part. However, this was to be expected, because the system does not have a stable ground state. Instead, the imaginary part of the energy is just the decay probability. The decay probability per unit time and unit volume is equal to

$$\Gamma/V = J^4|K|e^{-S_{cl}}. \tag{3.135}$$

Finally we see from equation (3.119) and (3.120) that the radius r_e of the instanton (in $d = 4$) is equal to the radius r_0 of a bubble of zero energy

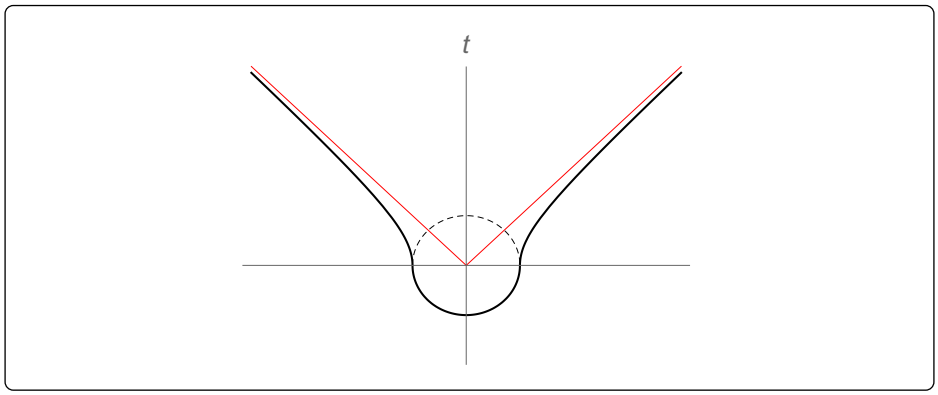


Figure 23. Euclidean instanton ($t < 0$) matching the expanding bubble in Minkowski space ($t > 0$).

(in $d = 3$). Thus, the instanton can be seen as interpolating continuously between the initial false vacuum state and a state where there is a bubble of true vacuum at rest and zero total energy (this state being represented by the evaluation of the instanton at the time $\tau = 0$). Actually, since the Minkowskian field equations are the analytic continuation of the Euclidean ones, and since at time $\tau = 0$ the four-dimensional radius is the same as the three-dimensional radius, the same profile $\phi(r)$ that gives the $SO(4)$-invariant Euclidean solutions, when evaluated at $\tau = 0$, also solves the static field equations for an $SO(3)$-invariant bubble of zero energy. Thus at the midpoint of its time evolution, the instanton is exactly a bubble of radius r_0. One can then smoothly match the Euclidean solution to the Minkowskian solution representing the expanding bubble, as shown in Figure 23.

3.12 Exercises

Exercise 3.1: Functional gauge potential for the nonlinear sigma model

Check that the functional gauge potential $\mathcal{A}$ given in (3.22) satisfies $d\mathcal{A} = 0$.

Exercise 3.2: Functional gauge potential for gauge theories

Check that the functional gauge potentials (3.33) and (3.40) for 2-dimensional QED and 4-dimensional YM theory satisfy $d\tilde{\mathcal{A}} = 0$.

Exercise 3.3: Chern–Simons form

The Chern–Simons 3-form, introduced in (3.37), can be written

$$C^0(A) = \frac{1}{16\pi^2}\varepsilon^{ijk}\mathrm{tr}\left(A_i\partial_j A_k + \frac{2}{3}A_iA_jA_k\right).$$

Use this form to check the gauge transformation property

$$\int d^3xC^0(A^g) - \int d^3xC^0(A) = W(g),$$

that is used in Equation (3.42).

Exercise 3.4: Path integral of the harmonic oscillator

Calculate the Euclidean partition function of the harmonic oscillator $Z_T = \langle 0|e^{-\frac{1}{\hbar}\hat{H}T}|0\rangle$ and derive Equation (3.79).

Exercise 3.5: Instantons for the double well potential

This is an example of a non-topological instanton. For a one-dimensional quantum mechanical system with Lagrangian

$$L = \frac{1}{2}\dot{q}^2 + \frac{\lambda}{4}(q^2 - f^2)^2$$

find the instanton that interpolates between the two degenerate minima of the potential. In the WKB approximation, use the instanton to calculate the splitting of the energy between the (classically degenerate) vacua.

Exercise 3.6: The vortex as instanton

Rewrite the vortex ansatz (2.99) in the gauge $A_1 = 0$, as required by the discussion in Section 3.6.2. Remember that the direction 1 now corresponds to Euclidean time.

Exercise 3.7: Symmetric gauge fields

Let $\Xi_{ab} = -iM_{ab}$ be generators for the Lie algebra $\mathfrak{so}(N)$. The commutators (B.12) become

$$[\Xi_{ab}, \Xi_{cd}] = -i\delta_{ac}\Xi_{bd} + i\delta_{ad}\Xi_{bc} + i\delta_{bc}\Xi_{ad} - i\delta_{bd}\Xi_{ac}. \tag{3.136}$$

Consider an $SO(N)$ YM field in $\mathbb{R}^N$ of the form

$$A_a = \Xi_{ab}x^b f(r), \tag{3.137}$$

where $r = \sqrt{x_1^1 + \ldots x_N^2}$. Show that this field is invariant under a simultaneous $SO(N)$ rotation and $SO(N)$ gauge transformation, with the same transformation parameters.

The monopole and the instanton are the $N = 3$ and $N = 4$ cases of such fields.

Exercise 3.8: The BPST instanton on the sphere

The Yang–Mills equatons are invariant under conformal transformations and the sphere S^4 is conformally flat. (Here we are not taking the sphere just as a topological device, but we give it its natural $SO(5)$-invariant metric.) Thus every solution of the YM equations in flat space is also a solution on the sphere. Write the instanton of size λ on a sphere of radius λ and show that it is invariant under $SO(5)$.

Hint: instead of pulling back the flat space instanton to S^4, it is easier to start from the SO(5)-invariant gauge field on $\mathbb{R}^5$ given in the previous exercise, show that it is tangential to $S^4 \subset \mathbb{R}^5$ and that the 5-th component can be eliminated by a gauge choice. When mapped to flat space, the remaining $SO(4)$ gauge field is recognized to be the combination of an instanton and an anti-instanton, as discussed in the previous exercise.

Exercise 3.9: Quantum fluctuations around the YM instanton

Following the logic that has been explained in detail for the pendulum and for 2-dimensional QED, compute the one loop amplitude (3.112). Instead of performing the calculation in flat space, where the one-point compactification to S^4 is just a device that keeps track of the boundary conditions on the fields, it is convenient to work on an actual metric 4-sphere, where the instanton exhibits $SO(5)$-symmetry, as discussed

in Exercise 3.8. The calculation then proceeds as follows:

1. Calculate the spectra of the operators on the sphere.
2. Convert the integrals over zero modes to integrals over the collective coordinates.
3. Using a suitable regulator, evaluate the determinants.

The final result is formula (3.113) for the amplitude, that exhibits a logarithmic divergence.

This exercise requires techniques that have not been explained in this book.

Chapter 4

$\pi_2(\mathcal{Q})$ and the quantization of parameters

Let us consider again the motion of a charged particle on a manifold $\mathcal{Q}$, in a background magnetic field $\mathcal{F}$, with a potential $\mathcal{A}$, and Lagrangian (3.6):

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

In Chapter 3 we considered the case in which $\mathcal{Q}$ is multiply connected and $\mathcal{F} = 0$. We will now consider situations in which $\mathcal{Q}$ is simply connected and $\mathcal{F} \neq 0$. As discussed in Section 3.1.1, quantization demands that $\mathcal{A}$ be a $U(1)$-connection, rather than an $\mathbb{R}$-connection. How can we tell whether this is the case? A necessary and sufficient condition for $\mathcal{F}$ to be the field strength of a $U(1)$ connection is that

$$\int_m \mathcal{F} = \frac{2\pi\hbar c}{e} n \qquad\qquad n \in \mathbb{Z}, \tag{4.1}$$

for any two-dimensional submanifold m of $\mathcal{Q}$ without boundary. In mathematical terms, $\mathcal{F}$ has to define an integral cohomology class in $H^2(\mathcal{Q})$ (see Appendix F for a brief survey).

If $\mathcal{Q}$ is topologically trivial, every two-dimensional submanifold without boundary is itself the boundary of a three-dimensional submanifold. Then, the condition (4.1) is trivially satisfied, because the integral is always zero. In order to have a nontrivial condition, there must be some two-dimensional submanifold without boundary that is not the boundary of a three-dimensional submanifold. In typical examples, there will be a 2-sphere that cannot be shrunk continuously to a point. In fact, while the proper topological setting for

these phenomena is cohomology, if $\mathcal{Q}$ is simply connected, Hurewicz' theorem (Appendix F) states that $H^2(\mathcal{Q},\mathbb{Z}) = \pi_2(\mathcal{Q})$, so in these cases one could loosely say that these phenomena are related to a nontrivial second homotopy group.

We will prove (4.1) in the simplest case $\mathcal{Q} = S^2$, where it is equivalent 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 in the equations of motion of a charged particle. Nevertheless, because of their topological origin, we will still call them "topological terms".

4.1 The Dirac quantization condition

Let us consider the Coulomb-like magnetic field (in Heaviside units)

$$B_i = \frac{1}{4\pi}\frac{Q_M}{r^2}\hat{x}^i. \tag{4.2}$$

We will regard it as a fixed background, and seek consistency conditions for the quantization of a charged particle moving in this background. The following description is due to Wu and Yang [WuY75] and was one of the earliest applications of fiber bundles to physics.

As discussed in Section 2.7.1, the field (4.2) is singular in the origin and can be regarded as a vacuum solution of Maxwell's equations on $\mathcal{Q}=\mathbb{R}^3\setminus\{0\}$. On this manifold, $d\mathcal{F}=0$ and so one expects that there exists a magnetic potential $\mathcal{A}$. It turns out that there is no magnetic potential for the monopole which is regular everywhere on $\mathbb{R}^3\setminus\{0\}$. To see this, suppose a magnetic potential $\mathcal{A}$ is given and consider the line integral $\Phi(\theta)=\oint_{\ell(\theta)}\mathcal{A}$, where $\ell(\theta)$ is a parallel at colatitude θ on a sphere of radius r, see Figure 24. 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

$$\mathcal{F} = \frac{Q_M}{4\pi}\sin\theta\, d\theta\wedge d\varphi. \tag{4.3}$$

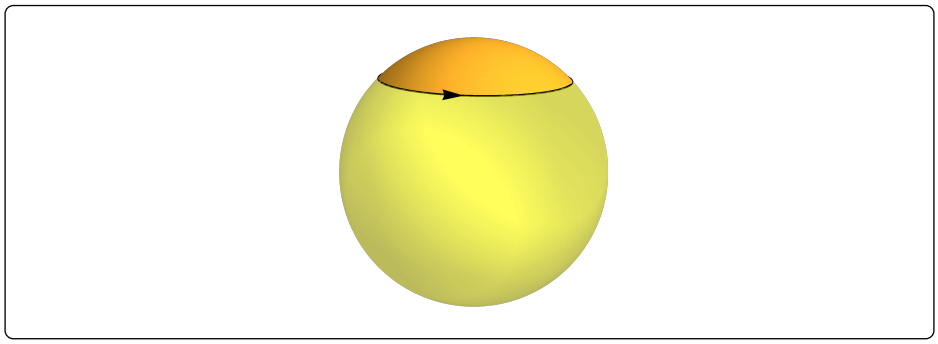


Figure 24. A parallel at colatitude θ and the cap it bounds.

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

$$\mathcal{A} = \mathcal{A}^{(+)} = \frac{Q_M}{4\pi}(1 - \cos\theta)d\varphi. \tag{4.4}$$

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, and it can be moved by gauge transformations. For example, another choice of magnetic potential is

$$\mathcal{A}^{(-)} = \frac{Q_M}{4\pi}(-1 - \cos\theta)d\varphi, \tag{4.5}$$

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

$$\mathcal{A}^+ - \mathcal{A}^- = \frac{Q_M}{2\pi}d\varphi. \tag{4.6}$$

Now we consider a particle moving in the monopole background. Its classical configuration space is $\mathcal{Q}$. To set up the quantum theory we introduce a wave function on $\mathcal{Q}$. Actually, we need two wavefunctions $\psi^\pm$, each of which need only be well-defined on $U^\pm$ respectively, related by a gauge transformation $g = e^{i\alpha}$ on the intersection:

$$\psi^+(\theta,\varphi) = g(\theta,\varphi)^{-1}\psi^-(\theta,\varphi) \qquad \text{on } U^+ \cap U^-. \tag{4.7}$$

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

$$\mathcal{A}^+ = \mathcal{A}^- - \frac{\hbar c}{ie} g^{-1} dg = \mathcal{A}^- - \frac{\hbar c}{e} d\alpha, \tag{4.8}$$

so comparing with (4.6) we see that the appropriate gauge transformation is

$$g(\theta, \varphi) = e^{-i\frac{eQ_M}{2\pi\hbar c}\varphi}. \tag{4.9}$$

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

$$Q_M = \frac{2\pi\hbar c}{e} n. \tag{4.10}$$

Noting that in Heaviside units $Q_M = \int_{S_2} \mathcal{F}$, this is precisely the same as (4.1).

There is a path integral argument leading to the same conclusion [Wit83b]. The action of a particle moving in a background magnetic field is

$$S = \int_{-\infty}^{\infty} dt \left[\frac{m}{2} \left(\frac{dq^i}{dt} \right)^2 + \frac{e}{c} \frac{dq^i}{dt} \mathcal{A}_i \right]. \tag{4.11}$$

This action suffers from two related problems. First, in the case of the monopole, $\mathcal{A}$ has singularities, as we have seen. This form of the action is therefore only appropriate for those histories of the particle that do not cross the Dirac string. On the other hand we know that the Dirac string is not a physical singularity, so this must be a shortcoming of our description of the system, not of the system itself. Second, the action is not gauge invariant. Under the gauge transformation $g = e^{i\alpha}$, $S' = S - \hbar(\alpha(\infty) - \alpha(-\infty))$.

To avoid these problems we consider a closed orbit[1] with $q^i(\infty) = q^i(-\infty)$ and we apply Stokes' theorem to write

$$\frac{e}{c} \oint_c \mathcal{A} = \frac{e}{c} \int_U \mathcal{F}, \tag{4.12}$$

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

[1] For the action considered here this may seem arbitrary, but in the presence of a potential it is motivated by the requirement that the system be in the lowest energy state in the far past and future. See e.g. the choice $\varphi(\infty) = \varphi(-\infty)$ in the discussion of the pendulum in Section 2.2.

trajectory of the particle. Since $d\mathcal{F} = 0$, the integral (4.12) 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. For example, they could correspond to the orange and yellow regions in Figure 24. The difference ΔS in the actions $(e/c)\int_{U_1}\mathcal{F}$ and $(e/c)\int_{U_2}\mathcal{F}$ is equal to the integral of $\mathcal{F}$ on the closed surface formed by joining U_1 and U_2 along the boundary. Since this surface contains the origin, the integral is equal to $(e/c)Q_M$. This arbitrariness in the action will not affect the functional integral if

$$e^{\frac{i}{\hbar}\Delta S} = 1, \tag{4.13}$$

which implies (4.10).

Finally we observe that the Dirac quantization condition can also be seen as an application of the old Bohr–Sommerfeld quantization conditions $\oint pdq = 2\pi\hbar n$. From (3.7) we have

$$\oint pdq = \int \left(mg_{ij}\dot{q}^i\dot{q}^j + \frac{e}{c}\dot{q}^i\mathcal{A}_i\right)dt. \tag{4.14}$$

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 $(e/c)\oint\mathcal{A} = (e/c)\int_{S^2}\mathcal{F}$, having applied Stokes' theorem to a surface bounded by the loop and not containing the string. The Bohr–Sommerfeld rule then gives again (4.10).

4.2 Wess–Zumino–Witten terms

4.2.1 Two dimensions

Consider a nonlinear sigma model with values in $SU(2)\equiv S^3$, in $d=1$ space dimensions. The configuration space is $\mathcal{Q} = \Gamma_*(S^1, S^3)$. Using the, by now, familiar technique of Appendix G.2 we find that

$$\begin{aligned}\pi_0(\mathcal{Q}) &= \pi_1\big(SU(2)\big) = 0,\\ \pi_1(\mathcal{Q}) &= \pi_2\big(SU(2)\big) = 0,\\ \pi_2(\mathcal{Q}) &= \pi_3\big(SU(2)\big) = \mathbb{Z}.\end{aligned}$$

The generator of $\pi_2(\mathcal{Q})$ is a map $m : S^2 \to \mathcal{Q}$ which is defined by $\big(m(t_1,t_2)\big)(t_3) = \hat{m}(t_1,t_2,t_3)$, where $\hat{m}$ is a map of S^3 (a cube $I\times I\times I$ with the boundary identified to a point) to $SU(2)$, sending $\partial(I\times I\times I)$ into the identity element,

and with winding number $W(\hat{m}) = 1$. If the map c in Section 2.3 could be referred to as a "loop of loops", the map m defined here could be called a "sphere of loops". By Hurewicz' theorem, one concludes that $H^0(\mathcal{Q}, \mathbb{Z}) = 0$, $H^1(\mathcal{Q}, \mathbb{Z}) = 0$ and $H^2(\mathcal{Q}, \mathbb{Z}) = \mathbb{Z}$.

The low homotopy and cohomology groups of $\mathcal{Q}$ are the same as in the previous section, so one may expect to find some analogue of the Dirac quantization condition in this theory. This is indeed the case. As with theta sectors, in order to reveal the occurrence of topological phenomena, it is necessary to add an appropriate term to the action, that is called the *Wess–Zumino–Witten term*, henceforth abridged WZW term.

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 a fictitious three dimensional ball B^3 and regard the fields φ as boundary values of some field $\bar{\varphi}$ defined on B^3. This is always possible because $\pi_2(SU(2)) = 0$, so all fields φ are homotopically trivial and have a continuation in the interior of B^3. The topological term we are after is just the topological term of $\bar{\varphi}$, which now depends only on φ and not on the value of the fields in the interior of the ball. We therefore have

$$\begin{aligned} S_{WZW} &= c \int_{B^3} \bar{\varphi}^* \omega = -\frac{c}{24\pi^2} \int d^3x \, \mathrm{tr}\, (\bar{U}^{-1} d\bar{U})^3 \\ &= c \int_{S^2} \varphi^* \tau = \frac{c}{2} \int d^2x \, \varepsilon^{\mu\nu} \partial_\mu \varphi^\alpha \partial_\nu \varphi^\beta \tau_{\alpha\beta}, \end{aligned} \tag{4.15}$$

where $\bar{U}(x)$ is the matrix representative of $\bar{\varphi}(x)$, ω is the volume form on $SU(2)$, normalized so that $\int_{SU(2)} \omega = 1$ and $\omega = d\tau$ or, in components,

$$\omega_{\alpha\beta\gamma} = 3(\partial_\alpha \tau_{\beta\gamma} + \partial_\beta \tau_{\gamma\alpha} + \partial_\gamma \tau_{\alpha\beta}). \tag{4.16}$$

We have renamed c the constant that previously was called θ. For example, suppose we choose on $SU(2)$ a coordinate system given by the Euler angles (Θ, Φ, Ψ) (see Appendix D). The volume form is given by

$$\omega = \frac{1}{16\pi^2} \sin\Theta \, d\Theta \wedge d\Phi \wedge d\Psi \tag{4.17}$$

and a choice of τ is

$$\tau = -\frac{1}{16\pi^2} \cos\Theta \, d\Phi \wedge d\Psi. \tag{4.18}$$

Since $d\tau \neq 0$, the WZW term (4.15) 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 (4.16) is that τ is not uniquely defined: if τ satisfies (4.16), also $\tau' = \tau + d\beta$ does. This amounts to adding a total derivative to the action (4.15). Another fact of the greatest importance is that τ is not globally defined. If it was, ω would define a trivial cohomology class. But we know from Appendix F that the volume-form on a compact manifold always defines a non-trivial cohomology class. This means that the form τ is singular somewhere on $SU(2)$. For example the form τ defined in (4.18) is singular for $\Theta = 0$ or π.

Now consider a field $\varphi(x,t)$; we regard it as a map from S^2 (the one-point compactification of spacetime) into $SU(2)$. Since the image of φ has, generically, dimensions 2 and the singular set of any form τ has dimension zero, generically, φ will not meet the singular points of τ. Thus there will be an open subset $\mathcal{U}$ of $\Gamma_*\big(S^2, SU(2)\big)$ where $\mathrm{Im}\varphi \cap \{\text{singular set}\} = \emptyset$, and the WZW action (4.15) will be well defined on $\mathcal{U}$. However, there are also maps φ whose image intersects the singular set of τ. For such maps (4.15) is not well defined. We can use the freedom of adding a total derivative term to the action to move the singularity elsewhere. In this way one can cover $\Gamma_*\big(S^2, SU(2)\big)$ with open sets, such that on each set there is a well defined function $c\int \varphi^*\tau$, and on the intersection of two sets these functions differ by a total derivative term $c\int d(\varphi^*\beta)$. The collection of these locally defined functions is the WZW term.

Let us consider the nonlinear sigma model with the action given by $S = S_0 + S_{WZW}$, where

$$S_0 = -\frac{1}{2}\int d^2x\, \partial_\mu\varphi^\alpha \partial^\mu\varphi^\beta h_{\alpha\beta}. \tag{4.19}$$

The equation of motion reads

$$h_{\alpha\beta}\partial_\mu\partial^\mu\varphi^\beta + \Gamma_{\alpha,\beta\gamma}\partial_\mu\varphi^\beta\partial^\mu\varphi^\gamma + \frac{c}{2!}\varepsilon^{\mu\nu}\partial_\mu\varphi^\beta\partial_\nu\varphi^\gamma\omega_{\alpha\beta\gamma} = 0 \tag{4.20}$$

where $\Gamma_{\alpha,\beta\gamma} = \frac{1}{2}\left(\partial_\beta h_{\alpha\gamma} + \partial_\gamma h_{\alpha\beta} - \partial_\alpha h_{\beta\gamma}\right)$ are the Christoffel symbols of the metric $h_{\alpha\beta}$ on $SU(2)$. The last term is the contribution of the WZW term. It can be interpreted as follows. The WZW term is linear in the time derivative and therefore can be written as $\int dt\, \dot\varphi^\alpha \mathcal{A}_\alpha(\varphi)$, where

$$\mathcal{A} = -c\int dx\, \partial_1\varphi^\alpha \tau_{\alpha\beta}\delta\varphi^\beta \tag{4.21}$$

is a one-form on $\mathcal{Q}$. When we think of the sigma model as a particle moving on $\mathcal{Q}$, $\mathcal{A}$ can be interpreted as a "functional vector potential". Unlike the cases discussed in Chapter 3, the corresponding "functional magnetic field" is now non-vanishing. A direct calculation using the methods of Appendix G yields

$$\mathcal{F} = d\mathcal{A} = \frac{c}{2}\int dx\, \partial_1\varphi^\alpha \omega_{\alpha\beta\gamma}\delta\varphi^\beta\delta\varphi^\gamma. \tag{4.22}$$

To confirm our interpretation of $\mathcal{A}$ and $\mathcal{F}$ note that the last term in (4.20) can be written $c\dot{\varphi}^\beta\mathcal{F}_{\alpha\beta}$ and therefore can be interpreted as the Lorentz force due to $\mathcal{F}$. The one-form $\mathcal{A}$ is only well defined on a subset $\mathcal{V}$ of $\mathcal{Q}$ such that the image of $\varphi(x)$ (a loop in $SU(2)$) does not intersect the singular set of τ. By contrast, $\mathcal{F}$ is well defined everywhere on $\mathcal{Q}$. We are therefore in the same situation as in the previous section, with a magnetic field $\mathcal{F}$ that cannot be derived from a globally defined vector potential $\mathcal{A}$. We can apply directly the general result (4.1). Let us therefore compute the integral of $\mathcal{F}$ on the "sphere of loops" m described above. We have

$$\begin{aligned}\int_m \mathcal{F} &= \int_0^1 dt_1 \int_0^1 dt_2 \left\{ c\int dx\, \partial_1\varphi^\alpha \omega_{\alpha\beta\gamma}\frac{\partial\varphi^\beta}{\partial t_1}\frac{\partial\varphi^\gamma}{\partial t_2}\right\} \\ &= \frac{c}{3!}\int d^3x\, \varepsilon^{\lambda\mu\nu}\frac{\partial\bar{\varphi}^\alpha}{\partial x^\lambda}\frac{\partial\bar{\varphi}^\beta}{\partial x^\mu}\frac{\partial\bar{\varphi}^\gamma}{\partial x^\nu}\omega_{\alpha\beta\gamma} \\ &= cW(\bar{\varphi}) = c. \end{aligned} \tag{4.23}$$

Using (4.1) we find that the theory can be quantized only for

$$c = 2\pi n. \tag{4.24}$$

This is the analogue of the Dirac quantization condition.

The quantization of the parameter c can also be proven in the functional integral formalism by means of the following argument [Wit83b]. The extension $\bar{\varphi}$ of the map φ to the interior of B^3 is not unique. Consider two extensions $\bar{\varphi}_1 : B_1^3 \to SU(2)$ and $\bar{\varphi}_2 : B_2^3 \to SU(2)$, with $\bar{\varphi}_1\big|_{S^2} = \bar{\varphi}_2\big|_{S^2} = \varphi$. Since they coincide on $\partial B_1 = \partial B_2 = S^2$, we can think of them as a single map $\bar{\varphi} : S^3 \to SU(2)$, where S^3 is obtained by glueing the two B^3 along their boundaries S^2 (in this picture the two balls are the hemispheres of S^3, and S^2 is the equator of S^3). The difference of the two WZW actions is therefore equal to

$$\Delta S_{WZW} = S_{WZW}(\bar{\varphi}_2) - S_{WZW}(\bar{\varphi}_1) = cW(\bar{\varphi}).$$

This arbitrariness will not affect the functional integral if $e^{i\Delta S_{WZW}} = 1$, which again implies (4.24).

4.2.2 Four dimensions

Can one write a WZW term for a sigma model in 3+1 dimensions? To answer this question, let us review what we have done in $2+1$ dimensions. We started from a closed three-form ω representing a nontrivial cohomology class of the target space. This form could be written *locally* as the exterior differential of a two-form τ. The WZW action was the integral of the pullback of τ. In 3+1 dimensions ω would have to be a closed five form and τ a four-form. There are no five-forms on $SU(2)$, but there are nontrivial five-forms on $SU(N)$ for $N \geq 3$. In fact, $H^5(SU(N), \mathbb{R}) = \mathbb{R}$, and the generator of this group is the left- and right-invariant five-form[2]

$$\omega = -\frac{i}{480\pi^3}\mathrm{tr}R^5 = -\frac{i}{480\pi^3}\mathrm{tr}L^5 = d\tau,$$

where $R = dUU^{-1}$ and $L = U^{-1}dU$. Therefore, the WZW term can be written in either one of the following two forms:

$$\begin{aligned} S_{WZW} &= c\int_{S^4} d^4x\, \varepsilon^{\mu\nu\rho\sigma}\partial_\mu\varphi^\alpha\partial_\nu\varphi^\beta\partial_\rho\varphi^\gamma\partial_\sigma\varphi^\delta\tau_{\alpha\beta\gamma\delta} \\ &= -\frac{ic}{480\pi^3}\int_{B^5} d^5x\, \varepsilon^{\lambda\mu\nu\rho\sigma}\mathrm{tr}(R_\lambda R_\mu R_\nu R_\rho R_\sigma), \end{aligned} \tag{4.25}$$

where spacetime has been compactified to a four-sphere and in the last line the integral is over a five-ball having spacetime as a boundary. Note that if spacetime was a compact connected five-dimensional manifold without boundary, this integral would be a topological invariant. Its integral formula is very similar to the one for the winding number in three dimensions.

The WZW term gives rise to the magnetic potential

$$\mathcal{A} = -c\int d^3x\, \varepsilon^{ijk}\partial_i\varphi^\alpha\partial_j\varphi^\beta\partial_k\varphi^\gamma\tau_{\alpha\beta\gamma\delta}\delta\varphi^\delta \tag{4.26}$$

on $\mathcal{Q}$. The corresponding field strength is

$$\mathcal{F} = d\mathcal{A} = \frac{c}{2}\int d^3x\, \varepsilon^{ijk}\partial_i\varphi^\alpha\partial_j\varphi^\beta\partial_k\varphi^\gamma\omega_{\alpha\beta\gamma\delta\eta}\delta\varphi^\delta\delta\varphi^\eta. \tag{4.27}$$

One can now repeat the arguments given for the two-dimensional case, leading again to the quantization of the parameter c as in (4.24). See [Ram84] for

[2]The normalization has been chosen so that it gives one when integrated on the S^5 submanifold that is the image of the fundamental generator of $\pi_5(SU(3)) = \mathbb{Z}$. This differs by a factor 2π from the normalization used in [Wit83b]. As a consequence also the functional Γ used there is 2π times the functional S_{WZW} defined below.

a discussion in the canonical context. Alternatively, the same result can be obtained from the functional integral formalism, as in the original paper of Witten [Wit83b].

Finally we observe that the relation between ω and $\mathcal{F}$ is a special example of a general construction that relates cohomology classes of N to cohomology classes of $\Gamma(M,N)$. This is discussed in Appendix G.3.

4.3 Chern–Simons terms

Next we consider an $SU(2)$ gauge theory in $2+1$ dimensions. We use the rescaled, geometrical gauge fields, with curvature defined by (1.120) and gauge transformations (1.122). Instead of writing explicitly the Lie algebra indices, we use matrix notation and write $A_\mu = A_\mu^a T_a$, where T_a are matrices representing the generators of the algebra (see Appendix B). In this notation the YM action (1.121) reads

$$S_{YM} = \frac{1}{2e^2}\int d^3x \,\mathrm{tr}\, F_{\mu\nu}F^{\mu\nu}. \tag{4.28}$$

As in Sections 3.4 and 3.5, we choose the gauge $A_0 = 0$; then the static energy reads

$$E_S = -\frac{1}{e^2}\int d^2x \,\mathrm{tr}\, B^2, \tag{4.29}$$

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

This configuration space is connected and furthermore has

$$\pi_1(\mathcal{Q}) = \pi_0(\mathcal{G}) = \pi_2(SU(2)) = 0$$
$$\pi_2(\mathcal{Q}) = \pi_1(\mathcal{G}) = \pi_3(SU(2)) = \mathbb{Z}.$$

In each line, the first equality comes from the homotopy exact sequence, following the same arguments of Sections 3.4 and 3.5. The generator of the group $\pi_2(\mathcal{Q})$ can be described as follows. The gauge group $\mathcal{G}$ is connected but not simply connected. Let $\ell(t)$ be a loop whose homotopy class generates $\pi_1(\mathcal{G})$. Fix a reference point $A_{(0)}$ in $\mathcal{C}$ and consider the loop in the orbit through $A_{(0)}$ given by $A_{(0)}^{\ell(t)}$. This loop cannot be shrunk to a point within the orbit but it can be shrunk to a point in $\mathcal{C}$. Thus there is a map $\tilde{m}$ from a two dimensional ball B^2 to $\mathcal{C}$ which is equal to $A_{(0)}^{\ell(t)}$ on the boundary. Now compose this map

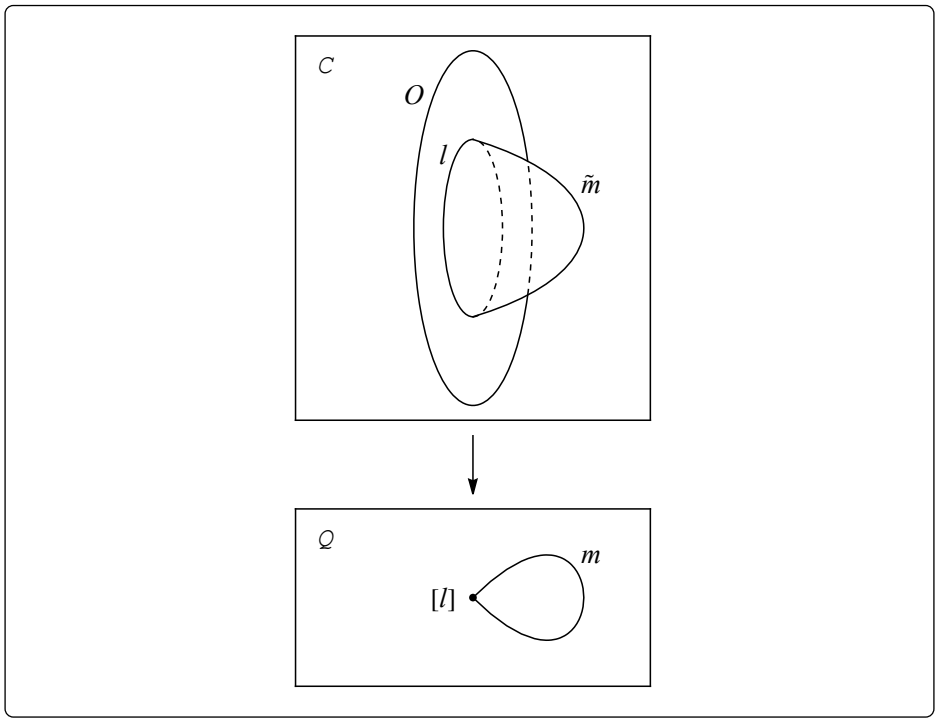


Figure 25. The space of connections $\mathcal{C}$ is a bundle over the physical configuration space $\mathcal{Q}$. The orbit $\mathcal{O}$ contains a noncontractible loop l that projects to a point in $\mathcal{Q}$. A disk $\tilde{m}$ with boundary l projects to a noncontractible sphere m.

with the projection $\mathcal{C} \to \mathcal{Q}$. Since all points on the boundary of the disk are mapped to the same point, we get a map m from S^2 to $\mathcal{Q}$ which is not homotopic to a constant (see Figure 25). The isomorphism between $\pi_1(\mathcal{G}) = \mathbb{Z}$ and $\pi_2(\mathcal{Q})$ is the map that sends the homotopy class of ℓ to the homotopy class of m.

Once again we have exactly the same homotopy groups as in the case of a particle in the field of a monopole, 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 sphere S^3, and regard this sphere as the boundary of a four dimensional ball B^4. Gauge fields A_μ on a three sphere are topologically trivial and can be thought of as boundary values of gauge fields $\bar{A}_\mu$ defined on B^4. The $SU(2)$ gauge theory in 3+1 dimensions was discussed in Section 3.5, where we added to the action a topological term $S_T = \theta c_2$. With the boundary conditions of Section 3.5, c_2 was an integer, but with the boundary conditions used here, the integral $c_2(A)$ becomes a functional of the boundary values $\bar{A}$. Using that the integrand

of c_2 is the exterior differential of the Chern–Simons three form (3.37), the appropriate topological term to be added to S_{YM} in three dimensions is the *Chern–Simons term*

$$S_{CS}(A) = \mu \frac{8\pi^2}{e^2} \int d^3x \Omega, \tag{4.31}$$

where

$$\Omega = -\frac{1}{8\pi^2} \varepsilon^{\lambda\mu\nu} \mathrm{tr} \left(A_\lambda \partial_\mu A_\nu + \frac{2}{3} A_\lambda A_\mu A_\nu \right) \tag{4.32}$$

is the (dual of the) Chern–Simons form. Note that apart from the overall coefficient, S_{CS} is identical to the functional $\tilde{\Lambda}$ defined in (3.41). The constant μ has dimension of mass. In fact simple manipulations on the equations of motion show that this theory describes spin one particles with mass $|\mu|$ (Exercise 4.1). For this reason it is called a "topologically massive gauge theory" [DJT81].

From our previous discussion of the WZW action, we are led to expect that the coefficient of the CS action, the mass μ, has to be quantized in certain units. This is indeed what happens. The proof of this fact turns out to be rather involved at the canonical level, (see [AMi85]) so we will depart from our standard procedure and only give a proof at the level of functional integrals.

For field configurations with finite S_{YM}, the group of gauge transformations is $\mathcal{G} = \Gamma_*(S^3, SU(2))$, and it consists of infinitely many connected components, labelled by their winding number. The dual of the Chern–Simons form transforms as follows

$$\Omega(A^g) = \Omega(A) - \frac{1}{8\pi^2} \varepsilon^{\lambda\mu\nu} \mathrm{tr}\, \partial_\lambda \left(\partial_\mu g g^{-1} A_\nu \right) + \frac{1}{24\pi^2} \varepsilon^{\lambda\mu\nu} \mathrm{tr} \left(g^{-1} \partial_\lambda g^{-1} \partial_\mu g g^{-1} \partial_\nu g \right), \tag{4.33}$$

and since we assume g to tend to the identity at infinity, upon integration we find

$$S_{CS}(A^g) = S_{CS}(A) + \mu \frac{8\pi^2}{e^2} W(g). \tag{4.34}$$

See Exercise 3.3. Thus, the Chern–Simons action is gauge invariant under gauge transformations that are homotopic to the identity (in particular, it is invariant under infinitesimal gauge transformations), but not under "large" gauge transformations. We demand that the functional integral be insensitive to this ambiguity. This requires that $e^{i\Delta S_{CS}} = 1$, or

$$\mu = \frac{e^2}{4\pi} n, \qquad n \in \mathbb{Z}. \tag{4.35}$$

Note that in the Euclidean functional integral one would demand $e^{-\Delta S_{CS,E}} = 1$, where $S_{CS,E}$ is the Euclidean Chern–Simons action. Since S_{CS} is linear in the time derivative, $S_{CS,E} = iS_{CS}$, so we are led again to (4.35).

To see what would go wrong if we did not impose the quantization condition (4.35), consider the formal Faddeev–Popov procedure to eliminate 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 (4.34) we find

$$Z = V_0 \sum_n e^{-i\mu \frac{8\pi^2}{e^2} n} \int (dA) \Delta_{FP}(A) \delta(f(A)) e^{iS(A)},$$

where V_0 is the volume of one connected component of the gauge group. The sum in front of the integral gives zero unless μ satisfies the quantization condition (4.35). Thus if (4.35) is not satisfied, the functional integral, and similarly the expectation value of any gauge invariant observable, is ill-defined.

4.4 Exercise

Exercise 4.1: Topologically massive gauge theory

Derive the Euler–Lagrange equation from the action $S_{\rm YM} + S_{\rm CS}$ and show that it describes the propagation of a particle with mass μ.

Chapter 5

The spin of solitons

In previous chapters we have considered examples of field theories that have either solitons or theta sectors or quantized parameters. For clarity we have considered these phenomena in isolation, but there are interesting cases when they occur together. In this chapter we shall discuss three examples where solitons, theta vacua and/or quantized parameters are simultaneously present. We will see that the topological terms (whether total derivatives or not) have an effect on the physical properties of the solitons. In particular, there are several cases where they determine the spin of the soliton. There are various ways of seeing this, but one way that works in all cases is the following.

Let $|i\rangle$ be a quantum state describing the soliton, whose spin we want to compute. Imagine a process whereby the soliton is rotated by 2π adiabatically slowly. The final state is $|f\rangle = e^{2\pi i R}|i\rangle$, where R is the generator of the rotation. As mentioned in Section 3.2.1, the final state is identical to $|i\rangle$, up to a phase that is related to the spin:

$$\langle f|i\rangle = e^{2\pi i s}.$$

This phase can be calculated as a functional integral for the process of a soliton that is rotated by 2π in a time T:

$$\langle f|i\rangle = \frac{1}{Z_0}\int (d\varphi) e^{iS(\varphi)},$$

where Z_0 is the functional integral for a soliton that is not rotated. In the leading order of the semiclassical approximation, the integrals are given by $e^{iS(\varphi_{\rm cl})}$, where $\varphi_{\rm cl}$ is the classical action. An important fact about the topological terms S_T is that they are linear in time derivatives. In the adiabatic limit $T \to \infty$ such terms give a finite contribution. By contrast, the kinetic terms (containing two derivatives) give a contribution of order $1/T$ and the static

energy contribution $-MT$, is canceled by Z_0. We conclude that the spin is given just by the topological term, evaluated on the classical field:

$$2\pi s = S_T(\varphi_{cl}). \tag{5.1}$$

The identification of the appropriate topological term, its parameter and the classical field have to be made case by case.

5.1 Sigma model anyons

5.1.1 The Hopf invariant

In Section 2.4 we discussed the S^2 nonlinear sigma model in 2+1 dimensions and showed that its configuration space consists of infinitely many connected components $\mathcal{Q}_n$ labelled by the winding number. In each component we were able to find the absolute minimum of the static energy, and these minima were the Belavin–Polyakov solitons. That these solitons could have fractional spin and statistics was recognized in [WiZ83, WuZ84]. In order to arrive at this result we will begin by examining the topology of the configuration space, and more precisely its fundamental group.

In general, when a space has many connected components, one has to choose a basepoint in each of them and compute the fundamental group separately for each component. In principle, these groups could all be different. Let us therefore start by considering the component consisting of homotopically trivial maps, $\mathcal{Q}_0 = \Gamma_*(S^2, S^2)_0$. Using the familiar rule we have

$$\pi_1(\Gamma_*(S^2, S^2)_0) = \pi_3(S^2)\,.$$

The fundamental generator of this group is the (homotopy class of the) Hopf map $h : S^3 \to S^2$, which is just the projection of the Hopf bundle $SU(2) \to SU(2)/U(1)$ (see Appendix D). The homotopy groups of S^2 are related to the homotopy groups of S^3 by the homotopy exact sequence of the Hopf bundle. In particular, we have

$$\ldots \to \pi_3(S^1) \to \pi_3(S^3) \xrightarrow{h_*} \pi_3(S^2) \to \pi_2(S^1) \to \ldots$$

Since the first and the last group in this sequence are trivial, the map h_* is an isomorphism. Since $\pi_3(S^3) = \mathbb{Z}$, we have proven that also $\pi_3(S^2) = \mathbb{Z}$. The integer that labels the homotopy classes of maps from S^3 to S^2 is called the Hopf invariant. It can be defined as follows [BoT82]. Let Ω be the invariant volume form on the sphere S^2 with unit radius. As discussed in Section 4.1, it

is closed but not exact, so it defines a nontrivial cohomology class in $H^2(S^2)$ (see also Appendix F). Given a map $f : S^3 \to S^2$, the pullback $f^*\Omega$ is a two form on S^3, and since $H^2(S^3) = 0$, it must be the differential of a globally well-defined one-form α: $\Omega = d\alpha$. The Hopf invariant of f is then

$$H(f) = \int_{S^3} \alpha \wedge f^*\Omega \,. \tag{5.2}$$

In order to calculate the Hopf invariant of the Hopf map, it is convenient to use spherical coordinates on S^2, so that $\omega = \frac{1}{4\pi} \sin\theta d\theta \wedge d\phi$, and Euler coordinates θ, ϕ, ψ on S^3, so that $h(\theta, \phi, \psi) = (\theta, \phi)$ and $h^*\Omega = \frac{1}{4\pi} \sin\theta d\theta \wedge d\phi$. On S^3, $h^*\Omega = d\alpha$ with $\alpha = \frac{1}{4\pi} L^3$, where L^3 is the left-invariant form (D.7c). Thus

$$H(h) = \frac{1}{16\pi^2} \int_0^\pi d\theta \cos\theta \int_0^{2\pi} d\phi \int_0^{4\pi} d\psi = 1 \,. \tag{5.3}$$

In summary, the Hopf invariant classifies the homotopy classes of loops in $\mathcal{Q}_0$, and since the configuration space is multiply connected, there will be theta-vacua. We discuss this first in the sector $\mathcal{Q}_0$ (no solitons), because we have all the tools already at hand, and return later to the question of the topology of the sectors $\mathcal{Q}_n$, $n \neq 0$, and its effect on the spin of the solitons.

As usual, the appearance of the theta-sectors is related to the existence of a topological term, which in this case is $S_T = \theta H$. So, we need an integral representation for the Hopf invariant. If we use the minimal geometric formulation of the nonlinear sigma model, it is impossible to write a local expression for H, because the form α is defined by solving the differential equation $d\alpha = \omega$ (we shall return to this point below). The same problem is present if we think of the nonlinear sigma model as a constrained linear $O(3)$ model. It is therefore advantageous to use the formulation of the nonlinear sigma model with gauge invariance, that was discussed in Section 1.3.4. Let φ be a field with values in S^2 on three-dimensional Euclidean spacetime. Finiteness of the Euclidean action imposes that φ goes to a constant at infinity, so we can compactify spacetime to S^3. Every map $\varphi : S^3 \to S^2$ has a *lift* to S^3, i.e. a map $\tilde{\varphi} : S^3 \to S^3$ such that $h \circ \tilde{\varphi} = \varphi$. It follows from the previous arguments that the Hopf invariant of the map φ is the winding number of its lift $\tilde{\varphi}$. We can therefore use the integral representation of the winding number, provided we work not with the original field φ but rather with its lift $\tilde{\varphi}$. Let us see how the Lagrangian looks like in various formulations of the model.

Lifted fields

In the notation of Section 1.3.4, we take $P = S^3$ and $N = S^2$, and we have the natural connection form ω which is just the $U(1)$-component of the left-invariant Maurer–Cartan form. With this connection one can construct a composite $U(1)$ gauge field $\mathcal{B}_\mu = \partial_\mu \tilde{\varphi}^{\tilde{\alpha}} \omega_{\tilde{\alpha}}$, and the covariant derivatives

$$D_\mu \tilde{\varphi}^{\tilde{\alpha}} = \partial_\mu \tilde{\varphi}^{\tilde{\alpha}} - \mathcal{B}_\mu F^{\tilde{\alpha}} \,, \tag{5.4}$$

where F is the fundamental vectorfield generating right $U(1)$ multiplication. In terms of the variables $\tilde{\varphi}$, the action (1.75) becomes

$$S_0(\tilde{\varphi}) = -\frac{f^2}{2} \int d^3x \; \tilde{h}_{\tilde{\alpha}\tilde{\beta}}(\tilde{\varphi}) D^\mu \tilde{\varphi}^{\tilde{\alpha}} D_\mu \tilde{\varphi}^{\tilde{\beta}} \,, \tag{5.5}$$

where $\tilde{h}_{\tilde{\alpha}\tilde{\beta}}$ is an invariant metric on S^3, and the Hopf term is

$$H(\varphi) = W(\tilde{\varphi}) = \frac{1}{16\pi^2} \int d^3x \, \varepsilon^{\lambda\mu\nu} \partial_\lambda \tilde{\varphi}^{\tilde{\alpha}} \partial_\mu \tilde{\varphi}^{\tilde{\beta}} \partial_\nu \tilde{\varphi}^{\tilde{\gamma}} \frac{1}{3!} \sqrt{\det \tilde{h}} \varepsilon_{\tilde{\alpha}\tilde{\beta}\tilde{\gamma}} \,. \tag{5.6}$$

The total action is $S(\tilde{\varphi}) = S_0(\tilde{\varphi}) + \theta W(\tilde{\varphi})$.

As we saw above, it is convenient to use the Euler angles as coordinates on S^3, because they are adapted to the Hopf fibration, in the sense that Ψ is the coordinate in the fibers S^1 while Θ and Φ are coordinates in the base space S^2, and are constant on the fibers. The invariant connection form ω is given in Euler coordinates in (D.7c), and the composite gauge field is

$$\mathcal{B}_\mu = \partial_\mu \tilde{\varphi}^{\tilde{\alpha}} \omega_{\tilde{\alpha}} = -\partial_\mu \Psi - \cos\Theta \partial_\mu \Phi \,. \tag{5.7}$$

The fundamental vectorfield is $F = L_3 = -\frac{\partial}{\partial \Psi}$, as given in (D.11c). The covariant derivatives of the fields are therefore

$$D_\mu \Theta = \partial_\mu \Theta \,, \tag{5.8}$$
$$D_\mu \Phi = \partial_\mu \Phi \,, \tag{5.9}$$
$$D_\mu \Psi = -\cos\Theta \partial_\mu \Phi \,. \tag{5.10}$$

The field Ψ disappears from S_0, which reduces to (1.70), and

$$W(\Theta, \Phi, \Psi) = \frac{1}{16\pi^2} \int d^3x \, \varepsilon^{\lambda\mu\nu} \sin\Theta \partial_\lambda \Theta \partial_\mu \Phi \partial_\nu \Psi \,. \tag{5.11}$$

Thus, the solitons discussed in Section 1.4 are still present, and their form is unaffected. This form of the theory has been discussed in [PaP90].

CP^1 formulation

In the CP^1 form of the theory, the variables are two complex fields z_1 and z_2 forming a spinor of $SU(2)$

$$z = \begin{pmatrix} z_1 \\ z_2 \end{pmatrix}$$

with $z^\dagger z = 1$. The action is $S_0 + \theta H$, where S_0 is given by (1.96) and the Hopf invariant is just the Chern–Simons term for the composite abelian gauge field:

$$S_0(z) = -\frac{f^2}{2} \int d^n x \, D_\mu z^\dagger D^\mu z \, . \tag{5.12a}$$

$$H(z) = -\frac{1}{4\pi^2} \int d^3 x \, \epsilon^{\lambda\mu\nu} B_\lambda \partial_\mu B_\nu \, , \tag{5.12b}$$

with $D_\mu z = \partial_\mu z + i B_\mu z$ and $B_\mu = i z^\dagger \partial_\mu z$. Using this, the Hopf invariant takes the form

$$H(z) = -\frac{1}{4\pi^2} \int d^3 x \, \epsilon^{\lambda\mu\nu} (z^\dagger \partial_\lambda z)(\partial_\mu z^\dagger \partial_\nu z) \, . \tag{5.13}$$

The proof that this expression is a total derivative can be found in [WuZ84, DiZ84].

The relation to the lifted fields is as follows. From Appendix D.4, the spinor z is related to the Euler angles by

$$z_1 = \cos\frac{\Theta}{2} \exp\left(-\frac{i}{2}(\Phi + \Psi)\right) , \tag{5.14a}$$

$$z_2 = -i \sin\frac{\Theta}{2} \exp\left(\frac{i}{2}(\Phi - \Psi)\right) , \tag{5.14b}$$

Then we find that

$$B_\mu = \frac{1}{2}(\partial_\mu \Psi + \cos\Theta \partial_\mu \Phi) = -\frac{1}{2}\mathcal{B}_\mu \tag{5.15}$$

and

$$\epsilon^{\lambda\mu\nu} \partial_\mu B_\nu = -\frac{1}{2} \epsilon^{\lambda\mu\nu} \sin\Theta \partial_\mu \Theta \partial_\nu \Phi \, . \tag{5.16}$$

Using these, we confirm that (5.12) is equal to (5.11).

Minimal formulation

If one sticks to the minimal formulation of the model in terms of only two fields (coordinates on S^2), it is impossible to write a local expression for the Hopf invariant. In fact, let us start from the observation that since the topological current (2.67) is conserved, we can define a potential α_μ such that

$$J_T^\lambda = \epsilon^{\lambda\mu\nu}\partial_\mu\alpha_\nu \,. \tag{5.17}$$

It is defined up to a gauge transformation $\alpha_\mu \to \alpha_\mu + \partial_\mu\Lambda$. If we choose, for example, the gauge $\partial_\mu\alpha^\mu = 0$, we can write

$$\alpha_\mu = -\frac{1}{\partial^2}\epsilon_{\mu\rho\sigma}\partial^\rho J_T^\sigma \,, \tag{5.18}$$

showing explicitly that it is nonlocal. Comparing (5.17) with (5.16) and with the topological current (2.67), that in spherical coordinates reads

$$J_T^\lambda = \frac{1}{4\pi}\epsilon^{\lambda\mu\nu}\sin\Theta\partial_\mu\Theta\partial_\nu\Phi \,, \tag{5.19}$$

we see that we can identify the nonlocal potential α_μ with the composite abelian gauge field of the lifted or CP^1 formulation:

$$\alpha_\mu = -\frac{1}{2\pi}B_\mu = -\frac{1}{4\pi}\mathcal{B}_\mu \,. \tag{5.20}$$

5.1.2 Theta vacua

The vacuum (no solitons) connected component of configuration space $Q_0 = \Gamma_*(S^2, S^2)_0$ can be seen as the quotient $\mathcal{P}/\mathcal{G}$, where $\mathcal{P} = \Gamma_*(S^2, S^3)$ is the configuration space of the lifted fields and $\mathcal{G} = \Gamma_*(S^2, S^1)$ is the gauge group. All these spaces are connected and $\mathcal{P}$ is a principal bundle over $\mathcal{Q}_0$. The homotopy exact sequence of this bundle gives

$$... \to \pi_1(\mathcal{G}) \to \pi_1(\mathcal{P}) \to \pi_1(\mathcal{Q}_0) \to \pi_0(\mathcal{G}) \to ...$$

We have $\pi_1(\mathcal{G}) = \pi_3(S^1) = 0$ and $\pi_0(\mathcal{G}) = \pi_2(S^1) = 0$, so the map in the middle is an isomorphism.[1] This means that the non-contractible loops in $\mathcal{P}$ are lifts of non-contractible loops in $\mathcal{Q}_0$.

[1] Note the difference with the YM theories discussed in Sections 3.4 and 3.5, where the total space was contractible: here it is the fiber that is contractible.

As in the examples that we considered in Chapter 3, the topological term gives rise to a magnetic potential $\tilde{\mathcal{A}}$ on the configuration space $\mathcal{P}$ of the lifted fields.[2] Using Euler coordinates for convenience,

$$\tilde{\mathcal{A}} = \frac{\theta}{16\pi^2}\int d^2x\ \sin\Theta\varepsilon^{ij}\big(\partial_i\Phi\partial_j\Psi\delta\Theta + \partial_i\Psi\partial_j\Theta\delta\Phi + \partial_i\Theta\partial_j\Phi\delta\Psi\big)\,. \quad (5.21)$$

The corresponding magnetic field strength is $\tilde{\mathcal{F}} = d\tilde{\mathcal{A}} = 0$, so $\tilde{\mathcal{A}}$ is a flat connection. In fact, we can write, locally $\tilde{\mathcal{A}} = d\tilde{\Lambda}$, where

$$\tilde{\Lambda} = \theta\int d^2x\,\omega^0 = -\frac{\theta}{16\pi^2}\int d^2x\,\varepsilon^{ij}\cos\Theta\partial_i\Phi\partial_j\Psi\,. \quad (5.22)$$

Still, $\tilde{\mathcal{A}}$ is not a trivial connection, because $\mathcal{P}$ is multiply connected and Λ is not single-valued.

The generator of $\pi_1(\mathcal{P})$ is the class of the loop $(\Theta_\tau, \Phi_\tau, \Psi_\tau)$ defined by $\Theta_\tau(\vec{x}) = \hat{\Theta}(\vec{x},\tau)$, $\Phi_\tau(\vec{x}) = \hat{\Phi}(\vec{x},\tau)$, $\Psi_\tau(\vec{x}) = \hat{\Psi}(\vec{x},\tau)$, where $(\hat{\Theta},\hat{\Phi},\hat{\Psi})$ is a map $S^3 \to S^3$ with winding number one (the instanton of the model). The polydromy of $\tilde{\Lambda}$ on this loop is given by

$$\begin{aligned}
&\tilde{\Lambda}(1) - \tilde{\Lambda}(0) = \oint\tilde{\mathcal{A}}\\
&= \int_0^1 d\tau\Big[\frac{\theta}{16\pi^2}\int d^2x\ \sin\Theta\varepsilon^{ij}\Big(\partial_i\Phi\partial_j\Psi\frac{d\Theta}{d\tau} + \partial_i\Psi\partial_j\Theta\frac{d\Phi}{d\tau} + \partial_i\Theta\partial_j\Phi\frac{d\Psi}{d\tau}\Big)\Big]\\
&= \frac{\theta}{16\pi^2}\int d^3x\ \sin\hat{\Theta}\varepsilon^{\lambda\mu\nu}\partial_\lambda\hat{\Theta}\partial_\mu\hat{\Phi}\partial_\nu\hat{\Psi} = \theta W(\hat{\Theta},\hat{\Phi},\hat{\Psi})\,. \quad (5.23)
\end{aligned}$$

Thus, $\tilde{\mathcal{A}}$ is a $U(1)$ pure gauge field if $\theta = 2\pi m$, for $m \in \mathbb{Z}$. The gauge inequivalent flat connections on $\mathcal{P}$, and hence the inequivalent quantizations, are parametrized by $0 < \theta \leq 2\pi$.

5.1.3 Anyons

Let us now consider the soliton sectors. Writing $z = re^{i\varphi}$ and using spherical coordinates in the target space, the solution is given by

$$\Theta(r,\varphi) = g(r) \equiv 2\,\mathrm{arccot}(r/2)\,, \qquad \Phi(r,\varphi) = \varphi\,.$$

Composing with the map s_- defined in (D.19) one obtains a lifted field, described by the same functions Θ and Φ and in addition $\Psi(r,\varphi) = \varphi$. It gives rise to the composite magnetic potential $\mathcal{B}_\mu = \frac{1}{4\pi}(-1-\cos\Theta)\partial_\mu\Phi$, which is

[2] Unlike in the examples of Chapter 3, here $\tilde{\mathcal{A}}$ is not the pullback of a one-form on $\mathcal{Q}_0$.

just the potential (4.5) for a magnetic monopole of unit charge. This lift is regular in the origin, but not at infinity. In fact, at infinity we have $\Theta = 0$, so that different values of Φ correspond to the same point on S^2, namely the North pole, but different values of Ψ parameterize a whole circle (the fiber of the Hopf bundle over the North pole). Thus, the lift is incompatible with the compactification of $\mathbb{R}^2$ to S^2. This is true for all maps in $\mathcal{Q}_n$ with $n \neq 0$. Still, one can define a space $\mathcal{P}_1$ of lifted fields by acting with $\Gamma_*(S^2, S^3)$ on the unit soliton described above. Since the transformations in $\Gamma_*(S^2, S^3)$ tend to the identity at infinity, all the fields in $\mathcal{P}_1$ have the same boundary conditions and form a bundle over $\mathcal{Q}_1$ with fiber $\Gamma_*(S^2, S^1)$. This defines a one-to-one correspondence between fields in $\mathcal{Q}_1$ and $\mathcal{Q}_0$, in particular one has $\pi_1(\mathcal{Q}_1) = \pi_1(\mathcal{Q}_0) = \mathbb{Z}$.

The fundamental noncontractible loop in $\mathcal{Q}_1$ can be realized as a loop in the moduli space $\mathcal{M}_1$. In fact, recall that $\mathcal{M}_1 = \mathbb{R}^2 \times \mathbb{R}^+ \times S^1$. Apart from the scale, which is topologically irrelevant anyway, this is the same as the configuration space of a two dimensional rigid body, which has been discussed in Section 2.2. The space $\mathcal{M}$ can be embedded in $\mathcal{Q}_1$ by simply associating to a certain set of collective coordinates the corresponding soliton field. We then see that the generator of $\pi_1(\mathcal{Q}_1) = \mathbb{Z}$ consists of a soliton being rotated by 2π. Now looking at Figure 30 in Appendix D.4, if we identify the poloidal coordinate with the angular coordinate in two-dimensional space, and the toroidal coordinate with (periodic) time, we recognize that the Hopf map precisely describes such a process. Therefore, its Hopf invariant is 1 and $S_T = \theta$. The calculation of the topological term for such a field can also be done directly, either using the formulation of the sigma model as a triplet of fields with a constraint $\phi_1^2 + \phi_2^2 + \phi_3^2 = f^2$ [WiZ83] or in the $\mathbb{CP}^1$ formulation [WuZ84, DiZ84], with the same result. Then from (5.1) one finds that the spin of the soliton is

$$s = \frac{\theta}{2\pi} \,. \tag{5.24}$$

There is an alternative way of reaching the same conclusion, that does not require consideration of $\mathcal{Q}_1$. Imagine a process whereby a soliton-antisoliton pair is created at time $-T/2$, the soliton undergoes an adiabatic rotation by 2π and at time $T/2$ the soliton and antisoliton annihilate. See Figure 26. Since the field tends to the identity at spatial infinity, for all time, and everywhere for $t \to \pm\infty$, this process describes a loop in $\mathcal{Q}_0$. We can continuously deform the process, so that it becomes again identical to the process shown in Figure 30, but now with time running from left to right. The topological action of this process is again θ.

Yet another way of arriving at the same result consists of calculating the angular momentum of the field [BKW86]. In this way one finds a more general

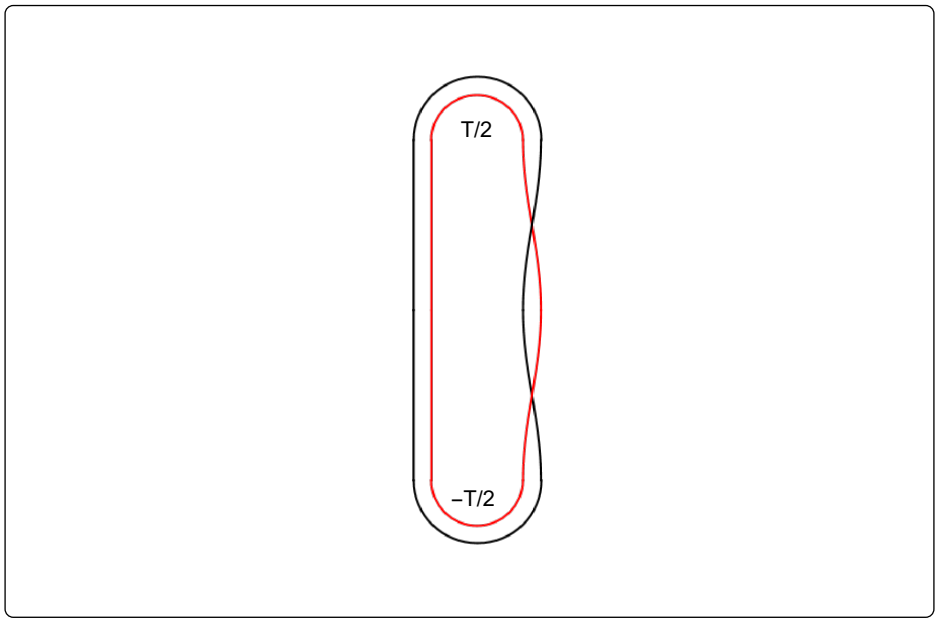


Figure 26. A soliton-antisoliton pair is created at time $-T/2$, the soliton is rotated adiabatically by 2π in a time T and then the pair annihilates. The black and red line represent two antipodal points on the soliton.

result, namely for a soliton of charge n, the spin is

$$s = \frac{\theta}{2\pi} n^2 \,. \tag{5.25}$$

5.2 Dyons

The Georgi–Glashow model is a non-abelian gauge theory with gauge group $SO(3)$. Even though this group is topologically different from $SU(2)$, it is again true that $\pi_3(SO(3)) = \mathbb{Z}$, so also this theory has θ-sectors. We shall now discuss the effect of the θ-parameter on the monopoles.

As we saw in Section 2.7.4, magnetic monopoles come in four-parameter families, depending on the coordinates of the center of mass, $\vec{x}_0$ and an internal angular parameter α. When we try to quantize the system semiclassically, along the lines of the quantization of the kink in Section 2.1.2, the zero modes of the operator appearing in the expansion of the action have to be traded for these moduli, that can then be treated as ordinary quantum mechanical degrees of freedom.

In the case of the translational zero mode, this corresponds to the quantization of a free particle with mass M, equal to the energy of the classical solution. (If we consider the Prasad–Sommerfield limit, $M = 4\pi f/e$.) Its Lagrangian is simply

$$\frac{1}{2}M\dot{\vec{x}}_0^2. \tag{5.26}$$

The quantization of the fourth modulus is more interesting. By definition, moduli are flat direction of the static energy, so their Lagrangian cannot contain a potential term. In order to write the Lagrangian for $\alpha(t)$ we have to first understand how the classical fields depend on $\dot{\alpha}$. We recall that α parametrizes a "global" $SO(2)$ rotation around the direction of the Higgs field. Here "global" means that the parameter of the tranformation is constant, but since the direction of the Higgs field depends on position, this looks formally like a gauge transformation. Under an infinitesimal variation of the parameter α we must therefore have, in the gauge $A_0 = 0$,[3]

$$\delta\phi^a = 0, \qquad \delta A_i^a = \frac{1}{ef}\delta\alpha D_i\phi^a.$$

Thus, the Higgs part of the classical Lagrangian does not contribute anything. In the YM Lagrangian there will be a contribution coming from the electric components of the field strength.

$$E_i^a = \dot{A}_i^a = \frac{1}{ef}\dot{\alpha}D_i\phi^a \tag{5.27}$$

Inserting in the Lagrangian we find

$$L = \frac{1}{2}\int d^3x E_i^a E_i^a = \frac{\dot{\alpha}^2}{2f^2e^2}\int d^3x (D_i\phi^a)^2. \tag{5.28}$$

Using the BPS condition (2.137), $\int d^3x (D_i\phi^a)^2 = M$, so we can rewrite the Lagrangian as

$$L = \frac{1}{2}\frac{M}{m_A^2}\dot{\alpha}^2, \tag{5.29}$$

where $m_A = fe$. Since $0 \le \alpha \le 2\pi$ is an angular variable, this can be interpreted as the Lagrangian of a rotator with moment of inertia $I = M/m_A^2$.

[3] Note that the angle of rotation depends on r, being α at infinity and zero in the origin. This is necessary to have a smooth gauge transformation, but ultimately only the transformation at infinity matters.

Now we observe that motion in the α-direction corresponds to the presence of an electric field:

$$\mathcal{E}_i = E_i^a \hat{\phi}^a = \dot{A}_i^a \hat{\phi}^a = \frac{\dot{\alpha}}{fe} \hat{\phi}^a D_i \phi^a = \frac{\dot{\alpha}}{fe} \hat{\phi}^a B_i^a = \frac{\dot{\alpha}}{fe} \mathcal{B}_i, \tag{5.30}$$

where in the second last step we used the BPS condition. Integrating over a two-dimensional sphere at $r \to \infty$ we obtain

$$Q_E = \frac{\dot{\alpha}}{fe} Q_M. \tag{5.31}$$

We see that a monopole moving the direction of the angular modulus has an electric charge proportional to the velocity. Electrically charged monopoles are called *dyons*.

The quantum mechanical wave functions must be periodic in α, hence they must be of the form $\psi = \exp(ik\alpha)$ with $k \in \mathbb{Z}$. The momentum conjugate to α therefore has integer eigenvalues

$$\pi_\alpha \psi \equiv -i \frac{\partial \psi}{\partial \alpha} = k\psi$$

and the velocity has eigenvalues

$$\dot{\alpha} = \frac{m_A^2}{M} k.$$

Using this in (5.31) we see that the electric charge of the dyon is quantized:

$$Q_E = ek, \qquad k \in \mathbb{Z}. \tag{5.32}$$

Let us now add to the Lagrangian of the theory the topological term θc_2, with c_2 given by (3.35), that can be rewritten as

$$S_T = \frac{\theta}{8\pi^2} \int d^4x E_i^a B_i^a. \tag{5.33}$$

This contributes to the Lagrangian of the modulus a term

$$L_\theta = \frac{\theta}{8\pi^2} \frac{\dot{\alpha}}{ef} \int d^3x B_i^a D_i \phi^a. \tag{5.34}$$

Integrating by parts, only the surface term survives, and it gives

$$L_\theta = \frac{\theta}{8\pi^2} \frac{\dot{\alpha}}{e} \int_{S^2_\infty} \mathcal{B} = \frac{\theta}{8\pi^2} \frac{\dot{\alpha}}{e} Q_M = \frac{\theta}{2\pi e^2} \dot{\alpha}. \tag{5.35}$$

Now the Lagrangian of the modulus reads

$$L = \frac{1}{2}\frac{M}{m_A^2}\dot{\alpha}^2 + \frac{\theta}{2\pi e^2}\dot{\alpha}. \tag{5.36}$$

The presence of the term linear in the time derivative changes the relation between velocity and momentum. The total Hamiltonian of the moduli is then

$$H = M + \frac{\vec{p}^2}{2M} + \frac{1}{2}\frac{m_A^2}{M}\left(\pi_\alpha - \frac{\theta}{2\pi e^2}\right). \tag{5.37}$$

Repeating the preceding steps, we find

$$\dot{\alpha} = \frac{m_A^2}{M}\left(k - \frac{\theta}{2\pi e^2}\right) \tag{5.38}$$

and therefore

$$Q_E = e\left(k - \frac{\theta}{2\pi e^2}\right). \tag{5.39}$$

In the presence of the θ angle, the quantized charge is shifted by a constant amount. This is known as the Witten effect [Wit79].

5.3 The spin of the Skyrmion

As discussed in Sections 1.3.3, 1.4.2, the nonlinear sigma model is quite successful in describing the low energy scattering of mesons. It is natural to ask, how should baryons be described in this theory? Skyrme suggested that the nucleons be described by the solitons of the same theory, that were discussed in Section 2.4.2. This may seem impossible at first, since the nonlinear sigma model is a purely bosonic theory and the baryons are spin 1/2 fermions.

However, it was also observed that in principle the skyrmions could be quantized as spin 1/2 fermions [FiR68]. To see this one has to look at the configuration space of the $SU(N)$ sigma model in three space dimensions, which is $\mathcal{Q} = \Gamma_*(S^3, SU(N))$. We have $\pi_0(\mathcal{Q}) = \pi_3(SU(N)) = \mathbb{Z}$, so $\mathcal{Q} = \cup_{n\in\mathbb{Z}}\mathcal{Q}_n$ is the disjoint union of infinitely many connected components characterized by the winding number. $\mathcal{Q}$ is a group under pointwise multiplication of maps; since the group action maps any connected component into any other, and the group action is given by diffeomorphisms, there follows that all connected components $\mathcal{Q}_n$ are diffeomorphic, and in particular have the same fundamental group. The fundamental group of the $n = 0$ component can be computed using the familiar rule $\pi_1(\Gamma_*(S^3, SU(N))_0) = \pi_4(SU(N))$. The discussion now follows different paths in the cases $N = 2$ and $N > 2$.

For $N = 2$, $\pi_4(SU(2)) = \mathbb{Z}_2$. As with the sigma model anyons of Section 5.1, the nontrivial topology of $\mathcal{Q}_1$ is given by the moduli space $\mathcal{M}$. According to the discussion in Section 2.4.2, this moduli space contains a factor $SO(3)$, and indeed $\pi_1(SO(3)) = \mathbb{Z}_2$. The nontrivial loop in $\mathcal{Q}_1$, whose homotopy class generates the first homotopy group, consist of a rotation of the soliton by 2π. Then, treating the soliton as a rigid body, it follows from the reasoning of Section 3.2.1 that we have the choice of quantizing it either as a boson or as a fermion.

Let us see here how this result can obtained in the functional integral. In this case one works with space-time dependent fields. Imposing finiteness of the action, Euclidean spacetime can be compactified to S^4, so that the space that one is formally integrating over is $\Gamma_*(S^4, SU(2))$. This function space consists of exactly two connected components, corresponding to the two homotopy classes in $\pi_4(SU(2))$. One of them consists of homotopically trivial maps. The other contains a map that describes the following process: a skyrmion-antiskyrmion pair is created in the far past, the skyrmion is rotated by 2π and finally the pair annihilates again. To understand that this is the right map, note that this map tends to the identity at infinity in all directions, as is required for elements of $\Gamma_*(S^4, SU(2))$, and it describes the non-contractible loop in $\mathcal{Q}_0$. This can be visualized again as in Figure 26, the only difference being that now space and spacetime have one dimension more. As discussed in Section 3.7.1, in the functional integral, homotopically distinct classes of paths can be summed with arbitrary weights given by characters of $\pi_1(\mathcal{Q})$. In this case we have

$$Z = \int (dU)_0 e^{-S} \pm \int (dU)_1 e^{-S}, \tag{5.40}$$

where the two terms correspond to the two homotopy classes of paths and the sign of the second term is a character of $\mathbb{Z}_2$. Choosing the lower sign corresponds to quantizing the skyrmion as a fermion.

For $N > 2$, $\pi_4(SU(N)) = 0$ and the previous arguments do not apply. However, in this case $\pi_2(\Gamma_*(S^3, SU(N)) = \pi_5(SU(N)) = \mathbb{Z}$ and one can add to the action a new topological term, the Wess–Zumino–Witten term defined in Section 4.2 [Wit83b]. In fact, this term is necessary to describe processes such as

$$K^+K^- \to \pi^+\pi^0\pi^-$$

that are allowed in QCD and occur in nature, but cannot be accounted for by the standard chiral Lagrangian. If we expand the WZW term by inserting (1.83) in (4.25) we obtain a total derivative that can be converted to a normal four-

dimensional integral. The leading term is

$$c\frac{2}{15\pi^2 f^2}\int d^4x\,\varepsilon^{\mu\nu\rho\sigma}\pi^a\partial_\mu\pi^b\partial_\nu\pi^c\partial_\rho\pi^d\partial_\sigma\pi^e B_{abcde} + \text{higher order terms}$$

where $B_{abcde} = \mathrm{tr}(T_a T_b T_c T_d T_e)$. This term can indeed be used to describe the process of two kaons going to three pions.

Now we can address the question of the spin of the skyrmion in the $N = 3$ chiral model. Using the general argument (5.1), the spin of the soliton is given by the WZW action evaluated for a slowly rotating skyrmion. A direct calculation shows that $S_{WZW} = c/2$. Recalling that $c = 2\pi n$, we have integer spin when n is even and half-integer spin when n is odd. As in the $N = 2$ chiral model, this seems to leave us with the freedom to quantize the skyrmions either way. As a final step in this chain of reasoning, we shall see in Section 6.8.3 that in the real world this freedom is fixed by QCD, and that the soliton must be a fermion.

Chapter 6

Anomalies

It is sometimes impossible to quantize a system preserving all the invariances of the classical theory. One then says that there is an anomaly. There are various types of anomalies, both from a mathematical and from a physical point of view. One can distinguish between anomalies for discrete groups of tranformations, for finite dimensional continuous groups (Lie groups) and for infinite dimensional groups. Another distinction of a more physical nature is whether the invariance that cannot be preserved is a genuine symmetry of the system (meaning that it consists of transformations that can be physically observed) or a gauge invariance (in which case the transformed object is physically indistinguishable from the original one).

As we saw in Sections 1.1 and 1.6, invariance under continuous transformation groups give rise to conserved currents (covariantly conserved, in the case of infinite dimensional invariance groups) and the anomaly manifests itself in the violation of these conservation law. The physical implications of the anomaly are then very different in the two cases. In the case of a genuine continuous symmetry (typically a symmetry with constant transformation parameters) the current of interest is the Noether current. The anomaly appears as a nonzero divergence of this current and does not have harmful consequences. The standard example is the Adler–Bell–Jackiw (ABJ) anomaly of the axial current [Adl69, BeJ69]. On the other hand in the case of a current coupled to gauge fields (when the transformation parameters are functions on spacetime), failure of current conservation jeopardizes the consistency of the theory. We will generally refer to such anomalies as *gauge anomalies*.[1]

[1]The term "local anomaly" is also used, referring to the fact that such an anomaly affects local gauge transformations. By the same token, then, the anomalies for finite dimensional symmetry groups could be called *global anomalies*. Unfortunately, the same terms is also

Gauge invariance removes certain states from the physical spectrum, and if it is violated by an anomaly it means that the quantum theory does not describe the same degrees of freedom as the initial classical theory. In these cases the anomaly gives rise to pathologies and requiring the absence of anomalies becomes a powerful tool to select physically viable theories.

Anomalies are a vast subject of which we will discuss only certain aspects. Our goal will be limited to highlighting the connections between the anomalies and the topological effects discussed in the preceding chapters. We shall therefore restrict ourselves to anomalies for internal transformations, either (finite dimensional) global symmetry groups or (infinite dimensional) YM invariance groups or certain discrete gauge transformations. We will completely omit a discussion of anomalies for spacetime transformations, such as the trace anomaly, related to the breaking of scale invariance, or gravitational anomalies. We will not calculate triangle diagrams, that are the main source of anomalies in perturbation theory, nor give Fujikawa's derivation of anomalies from the functional integral.

From the mathematical point of view the existence of anomalies is related to a very rich vein of results in algebraic topology and geometry. In particular, axial anomalies are intimately related to the index theorem for the Dirac operator, while the existence of gauge anomalies can be proven using a generalization of the index theorem involving two-parameter families of Dirac operators. The whole subject can also be recast in cohomological language. There are few other fields where progress of physics and mathematics has been so close.

Since the anomalies that we are interested in occur in even dimensional spacetime, in this chapter we adopt the convention of calling the dimension of spacetime $2n$, instead of n.

6.1 The axial anomaly

Here we consider the ABJ anomaly, which was historically one of the earliest examples of anomaly. It appears in the case of a single massless complex Dirac field coupled to electromagnetism. One finds that it is impossible to satisfy simultaneously the conservation of the vector and of the axial symmetry. Therefore this theory is anomalous. Depending on the regularization we choose, we can decide which symmetry is actually realized in the quantum theory: since the vector symmetry is in some sense more important than

used for certain anomalies that have to do with transformations that are not homotopic to the identity.

the axial symmetry, one usually preferes to give up the latter. Once this is understood, one then says that the axial symmetry is anomalous. In the next section we shall generalize the results to the case of a multiplet of fermions fields, carrying a representation of some global symmetry group.

We begin by setting up some notation. The action for a fermion in $2n$ spacetime dimensions, coupled to an external electromagnetic potential A_μ is

$$S_F(\psi, \bar{\psi}, A) = -\int d^{2n}x\, \bar{\psi}\left(\gamma^\mu D_\mu + m\right)\psi. \tag{6.1}$$

It is invariant under the (global) vector transformations

$$\psi' = e^{-i\alpha}\psi; \qquad \bar{\psi}' = \bar{\psi}e^{i\alpha} \tag{6.2}$$

with associated vector current

$$j_V^\mu = \bar{\psi}\gamma^\mu\psi. \tag{6.3}$$

We are also interested in the axial transformations

$$\psi' = e^{i\beta\gamma^A}\psi; \qquad \bar{\psi}' = \bar{\psi}e^{i\beta\gamma^A}, \tag{6.4}$$

where γ^A, anticommutes with the gamma matrices and squares to one. Under an infinitesimal axial transformation the variation of the action (6.1) is

$$\delta S_F = -2i\beta m \int d^{2n}x\, \bar{\psi}\gamma^A\psi, \tag{6.5}$$

showing that (6.4) is a symmetry of the Dirac action only if the mass vanishes. The corresponding Noether current is

$$j_A^\mu = \bar{\psi}\gamma^A\gamma^\mu\psi. \tag{6.6}$$

In general, the divergence of this current is

$$\partial_\mu j_A^\mu = -2m\bar{\psi}\gamma^A\psi, \tag{6.7}$$

so the axial current is conserved only if the mass is zero: in the following we will consider the massless case.

6.1.1 Point splitting

Let us now quantize the theory and ask whether $\partial_\mu \langle j_A^\mu \rangle = 0$ in the massless case (it is understood that j_A^μ now denotes the quantum operator corresponding to the axial current (6.6) and the brackets its vacuum expectation value in the A_μ background). The formal manipulations leading to the result (6.7) cannot be trusted because the operator j_A^μ is the product of two fields at the same spacetime point and is therefore singular: in other words the naive definition of composite operator in quantum field theory leads to divergent result. One has to resort to some kind of regularization. Physically, the most transparent regularization for problems of this type is point splitting: the axial current operator is defined to be the $\epsilon \to 0$ limit of the following expression:

$$j_A^\mu(x,\epsilon) = \bar\psi\left(x+\frac{\epsilon}{2}\right)\gamma^A\gamma^\mu \exp\left(ie\int_{x-\frac{\epsilon}{2}}^{x+\frac{\epsilon}{2}} A\right)\psi\left(x-\frac{\epsilon}{2}\right). \tag{6.8}$$

The regulator ϵ is a vector representing an infinitesimal displacement in spacetime; in order not to break Lorentz invariance it will be necessary, in taking the limit $\epsilon \to 0$, to average over all directions.

Under the local gauge transformation

$$\psi'(x) = e^{-i\alpha(x)}\psi(x); \qquad \bar\psi'(x) = e^{i\alpha(x)}\bar\psi(x); \qquad A'_\mu = A_\mu - \frac{1}{e}\partial_\mu\alpha$$

we have

$$\exp\left(ie\int_{x-\frac{\epsilon}{2}}^{x+\frac{\epsilon}{2}} A'\right) = \exp\left(-i\alpha\left(x+\frac{\epsilon}{2}\right)\right)\exp\left(ie\int_{x-\frac{\epsilon}{2}}^{x+\frac{\epsilon}{2}} A\right)\exp\left(i\alpha\left(x-\frac{\epsilon}{2}\right)\right).$$

The two outer exponentials cancel the transformation of the fermions and the regulated current (6.8) is gauge invariant. We have made this choice to ensure that the vector current, that couples to the gauge field, remains conserved. Different regularizations would spoil this desired property.

To compute the divergence of the current in the quantum theory, the prescription is to take first the divergence and then the limit $\epsilon \to 0$. Using the equations of motion

$$\begin{aligned} \gamma^\mu\partial_\mu\psi &= ie\gamma^\mu A_\mu\psi - m\psi, \\ \partial_\mu\bar\psi\gamma^\mu &= -ie\bar\psi\gamma^\mu A_\mu + m\bar\psi, \end{aligned} \tag{6.9}$$

one finds that

$$\partial_\mu j^\mu_A(x,\epsilon) = ie j^\mu_A(x,\epsilon)\left[A_\mu\left(x-\frac{\epsilon}{2}\right) - A_\mu\left(x+\frac{\epsilon}{2}\right) + \partial_\mu \int_{x+\frac{\epsilon}{2}}^{x+\frac{\epsilon}{2}} A\right], \tag{6.10}$$

plus the classical term given in (6.7), that we shall disregard from now on. For small ϵ the square bracket can be expanded as

$$\epsilon^\alpha\left(\partial_\alpha A_\mu - \partial_\mu A_\alpha\right) + O(\epsilon^2) = \epsilon^\alpha F_{\alpha\mu} + O(\epsilon^2).$$

Note that the classical result is recovered if one takes the limit $\epsilon \to 0$ naively. We have already said that this incorrect, since $j^\mu_A(x,\epsilon)$ is singular in the limit. To see this concretely, let us take the vacuum expectation value of both sides of (6.10). We find

$$\langle \partial_\mu j^\mu_A(x,\epsilon)\rangle = -ie\langle j^\mu_A(x,\epsilon)\rangle \epsilon^\alpha F_{\alpha\mu} + O(\epsilon^2). \tag{6.11}$$

6.1.2 Calculation of the anomaly

We will now show that the coefficient of ϵ is divergent. The VEV on the r.h.s. can be rewritten, for $\epsilon^0 > 0$,

$$\begin{aligned}\langle j^\mu_A\rangle &= -\mathrm{Tr}\gamma^A\gamma^\mu\left\langle T\psi\left(x-\frac{\epsilon}{2}\right)\bar\psi\left(x+\frac{\epsilon}{2}\right)\right\rangle e^{ie\int A} \\ &= -\mathrm{Tr}\gamma^A\gamma^\mu G\left(x-\frac{\epsilon}{2}, x+\frac{\epsilon}{2}\right)e^{ie\int A},\end{aligned} \tag{6.12}$$

where the trace is over Dirac indices, T denotes time ordering and $G(x,y)$ denotes the Dirac propagator in an external electromagnetic field, defined by

$$\gamma^\mu\left(\partial_\mu - ieA_\mu(x)\right)G(x,y) = \delta(x-y). \tag{6.13}$$

Note that due to the presence of the external field, G is not simply a function of the difference $x-y$. Let $S(x-y)$ denote the free Dirac propagator, which is defined by equation (6.13) with A_μ set equal to zero. Multiplying (6.13) by S on the left (in the sense of kernel composition) and using

$$-\frac{\partial}{\partial y^\mu}S(x-y)\gamma^\mu = \delta(x-y)$$

one finds the equation

$$G(x,y) = S(x-y) + ie\int d^{2n}z\, S(x-z)\gamma^\mu A_\mu(z)G(z,y). \tag{6.14}$$

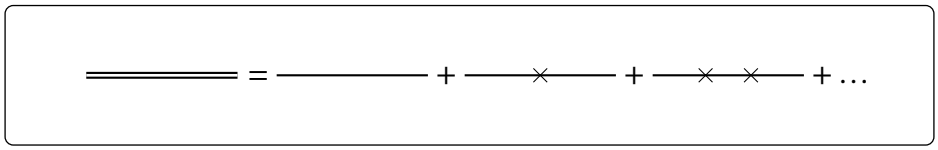


Figure 27. The full propagator G (double line) expanded as the free propagator S (single line) with insertions of the external field A (crosses).

This equation can be solved by iteration:

$$\begin{aligned} &G\left(x-\frac{\epsilon}{2},x+\frac{\epsilon}{2}\right) \\ &=S(-\epsilon)+ie\int d^{2n}yS\left(x-\frac{\epsilon}{2}-y\right)\gamma^{\mu}A_{\mu}(y)S\left(y-x-\frac{\epsilon}{2}\right) \\ &\quad -e^2\int d^{2n}yd^{2n}z\,S\left(x-\frac{\epsilon}{2}-y\right)\gamma^{\mu}A_{\mu}(y)S(y-z)\gamma^{\nu}A_{\nu}(z)S\left(z-x-\frac{\epsilon}{2}\right)+... \end{aligned} \tag{6.15}$$

This is represented graphically in Figure 27.

At this point the analysis begins to depend upon the dimension of spacetime. The free propagator can be written in Fourier space

$$S(x)=\int\frac{d^{2n}p}{(2\pi)^{2n}}e^{-ip\cdot x}\frac{\gamma^{\rho}p_{\rho}}{p^2}=\gamma^{\rho}S_{\rho}(x) \tag{6.16}$$

and inserting in (6.15) we see that the first term diverges for $\epsilon\to 0$ like $\epsilon^{-(n-1)}$, the second like $\epsilon^{-(n-2)}$ and so on, until the n-th term, which diverges logarithmically. All subsequent terms are convergent.

The expression (6.15) contains an odd number of gamma matrices. When inserted in (6.12) we have a trace of γ^A times an even number of gamma matrices. The first nonzero result occurs when the number of gamma matrices is equal to the dimension of spacetime. This leading term is proportional to the totally antisymmetric Levi–Civita symbol. It is always linearly divergent and when inserted in (6.12) it gives a finite contribution. The subsequent logarithmically divergent and finite terms of G are irrelevant in the limit $\epsilon\to 0$.

Let us discuss first the two-dimensional case ($n=1$). We use the gamma matrices (A.10). We have to take into account only the first term of the expansion (6.15) so the right hand side of (6.11) becomes:

$$ie\mathrm{Tr}[\gamma^A\gamma^{\mu}\gamma^{\nu}]S_{\nu}(-\epsilon)\epsilon^{\alpha}F_{\alpha\mu}+O(\epsilon).$$

The trace gives $\mathrm{Tr}\left[\gamma^A\gamma^{\mu}\gamma^{\nu}\right]=2\varepsilon^{\mu\nu}$, while for the fermionic propagator we obtain:

$$S_{\nu}(-\epsilon)=-\frac{i}{2\pi}\frac{\epsilon_{\nu}}{\epsilon^2}. \tag{6.17}$$

Taking the limit in ϵ and averaging over all the directions gives

$$\lim_{\epsilon\to 0}\frac{\epsilon_\alpha\epsilon_\nu}{\epsilon^2}=\frac{1}{2}\eta_{\alpha\nu}. \tag{6.18}$$

The factor $1/2$ comes from imposing that both sides of the equation have the same trace. In this way we arrive at the following expression for the anomaly:

$$\langle\partial_\mu j_A^\mu\rangle=\frac{e}{2\pi}\varepsilon^{\mu\nu}F_{\mu\nu}. \tag{6.19}$$

We note that this is twice the integrand of the topological invariant c_1 defined in (3.28).

Things are a bit more complicated in four dimensions ($n=2$), where the relevant term in (6.15) is the second one. Then, the right hand side of (6.11) becomes:

$$-e^2\epsilon^\alpha F_{\alpha\mu}\mathrm{Tr}\left[\gamma^A\gamma^\mu\gamma^\nu\gamma^\rho\gamma^\lambda\right]\int d^4y\, S_\nu\left(x-\frac{\epsilon}{2}-y\right)A_\rho(y)S_\lambda\left(y-x-\frac{\epsilon}{2}\right). \tag{6.20}$$

The calculation of the relevant Fourier transforms and integrals is left to Exercise 6.1. The final result is

$$\langle\partial_\mu j_A^\mu\rangle=\frac{e^2}{16\pi^2}\varepsilon^{\mu\nu\rho\lambda}F_{\mu\nu}F_{\rho\lambda}. \tag{6.21}$$

We notice in these computations that the anomaly originates from a *classical* zero that multiplies a *quantum* infinity, giving rise to a finite term in the conservation laws. Moreover if we restore the Planck constant $\hbar$ we find that the coefficient of the anomaly depends linearly on it, manifesting the quantum mechanical nature of the phenomenon.

What is the fate of the vector current (6.3) and its conservation law? In the point splitting we have been careful to preserve invariance under vector transformations. It is easy to understand that no anomaly arises using this regularization procedure. For the explicit calculation in two dimensions see Exercise 6.7.

6.1.3 Other axial anomalies

In the preceding calculation we have considered a fermion coupled to the electromagnetic field. This calculation can be generalized to the case of a multiplet of fermions ψ^A, carrying a representation of a global symmetry group G (in realistic applications, G is $SU(N)$, and is called the flavor group). The

matrices representing the generators in the Lie algebra of G wil l be denoted T_a. They are assumed to be antihermitian and to satisfy

$$T_a^\dagger = -T_a, \quad [T_a, T_b] = f_{ab}{}^c T_c, \quad f^*_{abc} = f_{abc}, \quad \mathrm{tr}\, T_a T_b = -\frac{1}{2}\delta_{ab}. \quad (6.22)$$

For example for $SU(2)$, $T_a = \frac{i}{2}\sigma_a$ whereas for $SU(3)$, $T_a = \frac{i}{2}\lambda_a$, where λ_a are the Gell–Mann matrices. We will not write explicitly the indices of the fermions, neither the spinor indices nor the indices pertaining to the representation of G. Thus ψ will now denote a column vector on which the group acts by left multiplication and the Dirac conjugate $\bar\psi$ is a row vector on which the group acts by right multiplication. The action can be written again as in (6.1), with our new interpretation of symbols. The field $A_\mu = A^a_\mu T_a$ is an external (non-dynamical) non-abelian gauge field. This action has a global symmetry $U(1)_V \times G_V$, where $U(1)_V$ is defined by (6.2), with all components of ψ transforming by the same phase, and G_V is defined by

$$\psi' = e^{-\alpha^a T_a}\psi; \qquad \bar\psi' = \bar\psi e^{\alpha^a T_a}, \quad (6.23)$$

where α^a are the transformation parameters (defining a function with values in the Lie algebra). The Noether current associated to $U(1)_V$ is (6.3), and the Noether current associated to G_V is

$$j_{Va}^{\ \mu} = \bar\psi \gamma^\mu T_a \psi. \quad (6.24)$$

It transforms according to the adjoint representation. In the massless case, the action is also invariant under a group $U(1)_A \times G_A$, where $U(1)_A$ is defined by (6.4), and G_A is given by the transformations

$$\psi' = e^{\alpha^a T_a \gamma^A}\psi; \qquad \bar\psi' = \bar\psi e^{\alpha^a T_a \gamma^A}. \quad (6.25)$$

The Noether current corresponding to $U(1)_A$ is (6.6), and the current associated to G_A is

$$j_{Aa}^{\ \mu} = \bar\psi \gamma^A \gamma^\mu T_a \psi. \quad (6.26)$$

As in the abelian case, the vector and axial currents cannot be simultaneously conserved. The anomaly can be computed using the method described above (the only difference consists in replacing the exponential in (6.8) by a path ordered exponential) and one finds[2]

$$\partial_\mu j^\mu_A = \frac{i}{2\pi}\epsilon^{\mu\nu}\mathrm{tr} F_{\mu\nu} \qquad \text{for} \quad n = 1, \quad (6.27a)$$

$$\partial_\mu j^\mu_A = \frac{1}{16\pi^2}\varepsilon^{\mu\nu\rho\sigma}\mathrm{tr} F_{\mu\nu}F_{\rho\sigma} \qquad \text{for} \quad n = 2, \quad (6.27b)$$

[2]Here and in the rest of this section we use rescaled gauge fields, so that the explicit factors of e in (6.19) and (6.21) are no longer present.

$F_{\mu\nu}$ now being the non-abelian field strenght. Note that the r.h.s. of (6.27a) is zero for the group $SU(N)$, and the r.h.s. of (6.27b) is twice the topological invariant c_2.

The argument given above can be generalized straightforwardly also to the calculation of the (covariant) divergence of the non-singlet current (6.26). The result is

$$(D_\mu j^\mu_A)_a = \frac{1}{2\pi}\varepsilon^{\mu\nu}\operatorname{tr} T_a F_{\mu\nu}, \qquad \text{for } n = 1 \tag{6.28a}$$

$$(D_\mu j^\mu_A)_a = \frac{i}{16\pi^2}\varepsilon^{\mu\nu\rho\sigma}\operatorname{tr} T_a F_{\mu\nu}F_{\rho\sigma}, \qquad \text{for } n = 2 \tag{6.28b}$$

We note that the factors of i in the formulas (6.27a) and (6.28b) are needed for these traces to be real, since they contain an odd number of antihermitian matrices. We end the section stressing again that the axial symmetry we have discussed is a global symmetry. We shall see that the quantum mechanical breaking of this symmetry has interesting physical implications.

6.2 The index theorem

We have observed that the anomaly of the axial current in two and four dimensions is twice the integrand of the topological invariants c_1 and c_2, that we introduced in Sections 3.4 and 3.5. This is no coincidence, and in fact the axial anomaly can be shown to affect the phenomenon of theta vacua in gauge theories in the presence of fermion fields. We begin by observing that since the fermionic configuration space is linear, the general topological arguments for the existence of theta sectors given in Sections 3.4 and 3.5 continue to hold also in the presence of fermions. However, *massless* fermions have a dramatic influence on the dynamics of such theories: it turns out that the VEVs of gauge invariant operators become independent of θ. The proof of this statement relies on a profound mathematical result, known as the Atiyah–Singer index theorem, that encodes the topological meaning of the axial anomaly.

6.2.1 Statement of the theorem

In order to state the theorem, we go to Euclidean signature and assume that spacetime is even dimensional, compact and without boundary. As usual, this can be achieved by imposing suitable boundary conditions so that spacetime can be compactified to a sphere. Then, the Dirac operator

$$\mathcal{D} = \gamma^\mu \left(\partial_\mu + A_\mu\right)$$

acting on the fermionic space V, is self-adjoint and has a discrete spectrum. Since $(\gamma^A)^2 = 1$, we can split

$$V = V_+ \oplus V_-$$

where $V_\pm$ are eigenspaces of γ^A with eigenvalues ± 1 respectively. Now let $\{\psi_n\}$ be a complete set of orthonormal eigenfunctions of $\mathcal{D}$:

$$\mathcal{D}\psi_n = \lambda_n \psi_n; \qquad \int d^{2n}x\, \bar{\psi}_m \psi_n = \delta_{mn}.$$

Since γ^A anticommutes with γ^μ, if ψ is in V_+, $D\psi$ is in V_-, and vice-versa. So, if $\lambda_n \neq 0$, ψ_n cannot be an eigenfunction of γ^A. However, the eigenfunctions with zero eigenvalue (the zero modes) can be chosen to belong either to V_+ or V_-. Let n_+ and n_- be the numbers of linearly independent zero modes of $\mathcal{D}$ with positive and negative chirality respectively. The index theorem states that

$$n_+ - n_- = c_n, \tag{6.29}$$

where c_n is the topological invariant defined by (3.28) and (3.35) for $n = 1$ and 2 respectively. We now give a "physicist's proof" of this result that highlights its connection to the axial anomaly.

6.2.2 Derivation from the anomaly

Let us start with a massive fermion interacting with an external gauge field via the vector current (6.24) in four dimensions. Here the mass plays the role of a regulator and will be sent to zero in the end. The divergence of the axial current is

$$\partial_\mu \langle j^\mu_A \rangle = -2m\langle \bar{\psi}\gamma^A \psi\rangle + \frac{i}{8\pi} \mathrm{Tr}\, F_{\mu\nu}{}^* F^{\mu\nu} \tag{6.30}$$

(the factor i appears because we are now in Euclidean signature). We integrate both sides and take the expectation value in the fermionic vacuum. The l.h.s. is zero because we are on a manifold without boundary.[3] From the r.h.s. we obtain

$$2m \int d^{2n}x\, \langle \bar{\psi}\gamma^A \psi\rangle = -2ic_2. \tag{6.31}$$

Now we want to evaluate the VEV on the left:

$$\langle \bar{\psi}\gamma^A \psi\rangle = \frac{\int (d\psi d\bar{\psi}) e^{-S_F} \left(\int d^4x\, \bar{\psi}\gamma^A \psi\right)}{\int (d\psi d\bar{\psi}) e^{-S_F}}. \tag{6.32}$$

[3] In the physically more realistic case of a non-compact spacetime, $\int d^4x \partial_\mu j^\mu_A = \int d\Sigma_\mu j^\mu_A = 0$ because the fermion field is massive.

The eigenfunctions ψ_n of $\mathcal{D}$ are also eigenfunctions of $\mathcal{D}-im$ with eigenvalues $\lambda_n - im$. Thus we can decompose

$$\psi(x) = \sum_n a_n \psi_n(x); \qquad \bar{\psi}(x) = \sum_n \bar{a}_n \bar{\psi}_n(x),$$
$$S_F(\psi, \bar{\psi}, A) = \sum_n \bar{a}_n a_n (\lambda_n - im); \qquad (d\psi d\bar{\psi}) = \prod_n da_n d\bar{a}_n$$

The functional integrals in (6.32) can be performed using Berezin's rules for the integration over fermion fields:

$$\int da_n a_m = \delta_{nm}; \qquad \int d\bar{a}_n \bar{a}_m = \delta_{nm} \tag{6.33}$$

The denominator of (6.32) is

$$\begin{aligned} \int (d\psi d\bar{\psi}) e^{-S_F} &= \prod_n \int d\bar{a}_n da_n (1 - (\lambda_n - im)\bar{a}_n a_n) \\ &= \prod_n (\lambda_n - im) = \det(\mathcal{D} - im). \end{aligned} \tag{6.34}$$

The numerator is a bit more complicated:

$$\prod_n \int da_n d\bar{a}_n \, (1 - \bar{a}_n a_n (\lambda_n - im)) \left[\sum_{rs} \bar{a}_r a_s \int d^4y \, \bar{\psi}_r(y) \gamma^A \psi_s(y) \right].$$

We consider separately each term in the sums. If $r \neq s$, all the factors with n not equal to either r or s are zero, so the whole term vanishes. If $r = s$, the only nonvanishing contribution comes from picking the first term in the round bracket for $n = r = s$ and the second term for all the other values of n. This gives

$$\sum_r \int d^{2n}y \bar{\psi}_r(y) \gamma^A \psi_r(y) \prod_{n \neq r} (\lambda_n - im). \tag{6.35}$$

In the above formulas the formal determinant of the Dirac operator appears. It can be given a meaning by choosing a specific regularization procedure. In any case, for the calculation we are interested in, this is not necessary. In fact, we have

$$\int d^{2n}y \, \langle \bar{\psi}(y) \gamma^A \psi(y) \rangle = \sum_r \frac{\int d^{2n}y \, \bar{\psi}_r(y) \gamma^A \psi_r(y)}{\lambda_r - im}. \tag{6.36}$$

Since $\mathcal{D}$ anticommutes with γ^A, if ψ_n is an eigenfunction with eigenvalue λ_n, $\gamma^A\psi_n$ is an eigenfunction with eigenvalue $-\lambda_n$. Therefore, using orthogonality we find that if $\lambda_s \neq 0$, $\int d^{2n}y\, \bar{\psi}_s(y)\gamma^A\psi_s(y) = 0$. On the other hand, if ψ_n is a zero mode, also $\gamma^A\psi_n$ is a zero mode, and we have chosen the zero modes to have definite chirality. Therefore, if $\lambda_n = 0$, $\int d^{2n}y\, \bar{\psi}_s(y)\gamma^A\psi_s(y) = \pm 1$, depending on the chirality. So we find that

$$\int d^{2n}y\, \langle\bar{\psi}(y)\gamma^A\psi(y)\rangle = -\frac{1}{im}(n_+ - n_-), \tag{6.37}$$

and inserting back in (6.31) we obtain the index theorem (6.29). We have shown in this way that with certain assumptions about the boundary conditions, the index theorem follows from the existence of the axial anomaly. Conversely we can understand the appearence of the axial anomaly as a consequence of the dependence of the spectrum of the Dirac operator on the topology of the gauge field.

6.3 Consequences of the anomaly

We now discuss two physical consequences of the ABJ anomaly and of the index theorem.

6.3.1 Neutral pion decay

The dominant decay mode of the neutral pion is into two photons. This process is described by the triangle diagram shown in Figure 28 (and a similar one with crossed photons), where the particles circulating in the loop are fermions. This diagram had been calculated as early as 1949 by Steinberger with the nucleon in the loop, and gave a good agreement with experiments. Yet, almost two decades later, this decay still posed a puzzle. The dominant theoretical framework at the time was current algebra (see Section 1.2.3) and there was an argument of Sutherland and Veltman implying that in the chiral limit the neutral pion cannot decay into photons if the axial current is conserved, as was then assumed in applications of PCAC. The small mass of the pion could not account for the discrepancy. It was eventually understood by ABJ that the triangle graph implies a choice between vector and axial current conservation. Given that the former is essential for gauge invariance in QED, it makes more sense to sacrifice the latter. In this way, the triangle graph invalidates one of the essential assumptions of the Sutherland–Veltman argument, and current algebra could be reconciled with the experimental facts.

We give here a very quick summary of the calculation in the context of the chiral model. The pion-nucleon vertex has been given in Equations (1.86)

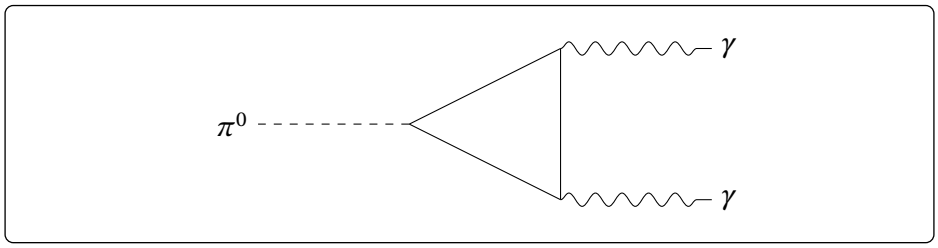


Figure 28. The triangle diagram.

and (1.87). Since the neutron does not couple to photons, we are only interested in the pion-proton vertex, and since we are interested in the neutral pion, we choose $a = 3$. Then we find a coupling of the neutral pion to the proton axial current:

$$ig_{\pi NN}\pi^0 \bar{p}\gamma^A p,$$

with $g_{\pi NN} = m_N/F_\pi$ (recall the Goldberger–Treiman relation (1.54)).
We refer e.g. to [Sch14], Section 30.1 for the calculation of the triangle diagram, leading to the amplitude

$$\mathcal{M} = g_{\pi NN}\frac{e^2}{4\pi^2 m_N}\varepsilon^{\mu\nu\rho\sigma}\epsilon^1_\mu\epsilon^2_\nu q^1_\rho q^2_\sigma, \tag{6.38}$$

where ϵ^i and q^i, with $i = 1, 2$ are the polarizations and momenta of the two photons. From here one derives the decay rate

$$\Gamma(\pi^0 \to 2\gamma) = \frac{\alpha^2}{64\pi^2}\frac{m_\pi^3}{F_\pi^2} \approx 7.77\text{eV},$$

that is quite close to the experimentally measured value 7.73eV. Considering the primitive state of QFT at the time when this calculation was first performed, and even today, considering that the chiral model is only a non-renormalizable low energy effective field theory, this agreement is quite surprising.

One important feature of this result is that the mass of the nucleon, that is present in $g_{\pi NN}$, cancels with the factor of m_N in (6.38), so that the final result does not depend on the mass of the fermions in the loop. In fact, a more modern version of the same calculation consists in using the two lightest quarks in the loop. They carry the same isospin representation as the nucleons, but different electric charges: in the case of the quarks, both isospin states contribute to the process. Whereas the charges of the nucleons give a factor

$$(1^2 - 0^2) = 1,$$

the quarks give

$$\left(\left(\frac{2}{3}\right)^2 - \left(-\frac{1}{3}\right)^2\right) N_c = \frac{N_c}{3},$$

where N_c is the number of colors. The two results only agree for $N_c = 3$, and this was one of the early ways of measuring the number of colors. Once again, the success of this calculation is surprising, this time because the pions are nonperturbative states in QCD. The reason why it works so well is that the one loop calculation is exact. It was shown by Adler and Bardeen [ABa69] that there are no corrections in perturbation theory, and there are topological arguments suggesting that this should be true also in a nonperturbative sense.

The independence of the triangle diagram on the mass of the fermions in the loop suggests that it should also hold in the limit when the fermions are massless, which is the context in which the anomaly arises. One can therefore say that the decay of the neutral pion is made possible by the anomaly. In fact, the amplitude (6.38) can be interpreted as an effective pion-photon interaction in the low-energy effective action (see e.g. [DGH22])

$$\frac{e^2 N_c}{96\pi^2 F_\pi} \int d^4x\, \varepsilon^{\mu\nu\rho\sigma} \pi^0 F_{\mu\nu} F_{\rho\sigma}, \tag{6.39}$$

where we see that the pion couples to the anomalous divergence of the axial current. We shall need this result in Section 6.8.3 when we address again the spin of the skyrmion.

6.3.2 Consequences for the theta sectors

In the calculation of the anomaly, the gauge field was treated as a background field. Let us now see the implications of these results for the quantization of the full theory, with dynamical gauge field, in four dimensions. We are specifically interested in the tunnelling amplitude through the noncontractible path in $\mathcal{Q}$, since this amplitude was responsible for the θ-dependence of the vacuum energy. It is given by the Euclidean functional integral

$$\int_{c_2 \neq 0} (dA d\bar{\psi} d\psi) e^{-S(\psi,\bar{\psi},A)} = \int_{c_2 \neq 0} (dA) e^{-S_{\rm YM}(A)} \det(\mathcal{D}), \tag{6.40}$$

where we have used (6.34), with $m = 0$. The integral is restricted to those gauge field configurations that have nonvanishing topological invariant. Now the index theorem (6.29) implies that for these fields there must be at least one zero mode, and therefore the determinant of the Dirac operator is identically zero, for all A in that sector. This implies that the tunnelling amplitude is zero,

and therefore the theta vacua are all degenerate, in sharp contrast to what happened without fermions, or with massive fermions. Note that in deriving this result we did not have to make use of the WKB approximation.

There is also a more formal argument that directly relates the anomaly to the degeneracy of the theta vacua. Consider a gauge- and chiral-invariant operator $\mathcal{O}$. The VEV of $\mathcal{O}$ in the vacuum specified by a value θ can be computed as

$$\langle \mathcal{O} \rangle_\theta = \frac{\delta \log Z_\theta(J)}{\delta J} \tag{6.41}$$

where

$$Z_\theta(J) = \int (dA d\bar{\psi} d\psi) e^{-S_{YM}(A) + i\theta c_2(A) - S_F(\psi, \bar{\psi}, A) + \int d^4x J\mathcal{O}}. \tag{6.42}$$

To simplify the notation we do not write explicitly the gauge fixing and ghost terms, since they are irrelevant for what follows. A priori, the VEV of $\mathcal{O}$ seems to depend upon θ. Let us now examine how Z_θ behaves under the axial transformations (6.4). Since the action is manifestly invariant, if the measure was also invariant, the whole functional integral would be invariant. This is incompatible with the existence of the anomaly, so the measure cannot be invariant [Fuj80]. One can compute the variation of the measure directly. However, knowing the form of the anomaly, one can deduce that under an infinitesimal axial transformation $\delta\psi = \delta\beta\gamma^A\psi$ the measure must change by $(d\bar{\psi} d\psi) = (d\bar{\psi}' d\psi') e^{\delta S}$. This is equivalent to a transformation of the action, that can be inferred from Noether's theorem (Equation (1.8)) to be

$$\delta S = \int d^4x \, \delta\mathcal{L} = \int d^4x \, \partial_\mu j^\mu_A = 2i\delta\beta c_2(A). \tag{6.43}$$

Therefore the effect of an axial transformation on the fermion fields is equivalent to a shift of θ by $2\delta\beta$:

$$Z_\theta(J) = \int (dA d\bar{\psi}' d\psi') e^{-S_{YM}(A) + i(\theta+2\beta) c_2(A) - S_F(\psi', \bar{\psi}', A) + \int d^4x J\mathcal{O}} = Z_{\theta+2\beta}(J). \tag{6.44}$$

In the last step we have replaced ψ' and $\bar{\psi}'$ by ψ and $\bar{\psi}$, since they are integration variables. The conclusion is that the value of θ is irrelevant: the expectation value of every gauge and chiral invariant observable is independent of θ.

We emphasize once again that this does not mean that there are no theta sectors anymore. The topological arguments remain valid. One can also argue that the theta sectors have to be still distinct in order for the cluster property be satisfied. All that has happened is that the theta sectors are now completely degenerate.

6.4 Gauge anomalies

Next we consider anomalies in a current that couples to gauge fields. We will call these *gauge anomalies*. Let us consider quite generally a fermionic current J_a^μ coupled to a gauge field A_μ^a via an interaction term $\mathcal{L}_I = J_a^\mu A_\mu^a$. To begin with, we do not specify whether J is a vector, axial or other current. All we assume is that the classical action S is gauge invariant. The classical current can be defined as

$$J_a^\mu = \frac{\delta S}{\delta A_\mu^a}. \tag{6.45}$$

Functional integration over the fermions yields a contribution to the action for the gauge fields:

$$W(A) = -i \ln \int (d\psi d\bar{\psi}) e^{iS_F(\psi,\bar{\psi},A)}. \tag{6.46}$$

We will refer to W as the fermionic effective action. The expectation value of the current in the fermionic vacuum is given by

$$\langle J_\mu^a \rangle = \frac{\delta W}{\delta A_\mu^a}. \tag{6.47}$$

We have shown in (1.136), (1.137) how gauge invariance of the classical action implies covariant conservation of the gauge current. We shall now repeat that discussion for the effective action.

For an infinitesimal gauge transformation parameter $\epsilon = \epsilon^a T^a$, define the operator

$$\delta_\epsilon = \int d^{2n}x \, D_\mu \epsilon^a(x) \frac{\delta}{\delta A_\mu^a(x)}. \tag{6.48}$$

It can be thought of as a vector tangent to the gauge orbit through A in the space $\mathcal{C}$ of all gauge fields. The derivative of W in the direction of this vector is

$$\begin{aligned} \delta_\epsilon W(A) &= \int d^{2n}x \, D_\mu \epsilon^a(x) \frac{\delta W}{\delta A_\mu^a(x)} \\ &= \int d^{2n}x \, D_\mu \epsilon^a(x) \langle J_a^\mu \rangle \\ &= -\int d^{2n}x \, \epsilon^a(x) \langle D_\mu J_a^\mu(x) \rangle. \end{aligned} \tag{6.49}$$

Since the classical action was gauge invariant, the r.h.s. can be interpreted as the anomaly, so we define the integrated anomaly

$$\mathcal{A}(\epsilon, A) \equiv \int d^{2n}x \, \epsilon^a(x) \langle D_\mu J_a^\mu(x) \rangle = -\delta_\epsilon W(A). \tag{6.50}$$

Covariant conservation of the gauge current is a very important property in the full quantum gauge theory: in perturbation theory, it ensures unitarity and renormalizability. Its violation is therefore to be avoided.

6.4.1 Chiral gauge theories

In Section 6.1 we discussed the case when a gauge field is coupled to a fermionic vector current. We proved that there exists a quantization scheme that preserves the conservation of this current, violating the conservation of the axial current. Since the axial current was not coupled to gauge fields, no problem arose in that case.

The situation is different if the coupling is not purely vectorial. The most general situation is to have the vector and axial currents coupled to two different gauge fields

$$\mathcal{L}_I = J_{Va}^{\ \mu} A_{V\mu}^{\ a} + J_{Aa}^{\ \mu} A_{A\mu}^{\ a}. \tag{6.51}$$

In order to ensure the invariance of the action under the vector and axial gauge transformations (6.23) and (6.25) the gauge fields have to transform as follows:

$$\begin{aligned} &\delta_{V\epsilon} A_{V\mu} = D_\mu \epsilon; && \delta_{V\epsilon} A_{A\mu} = [A_{A\mu}, \epsilon]; \\ &\delta_{A\epsilon} A_{V\mu} = [A_{A\mu}, \epsilon]; && \delta_{A\epsilon} A_{A\mu} = D_\mu \epsilon. \end{aligned} \tag{6.52}$$

These transformations obey the following algebra:

$$[\delta_{V\epsilon_1}, \delta_{V\epsilon_2}] = \delta_{V[\epsilon_1,\epsilon_2]}\,; \tag{6.53a}$$
$$[\delta_{V\epsilon_1}, \delta_{A\epsilon_2}] = \delta_{A[\epsilon_1,\epsilon_2]}\,; \tag{6.53b}$$
$$[\delta_{A\epsilon_1}, \delta_{A\epsilon_2}] = \delta_{V[\epsilon_1,\epsilon_2]}. \tag{6.53c}$$

The vector and axial transformations are deeply entangled. As in (1.42), it is convenient to define left and right currents

$$J_{La}^{\ \mu} = \frac{J_{Va}^{\ \mu} - J_{Aa}^{\ \mu}}{2} = \bar{\psi} T_a \gamma^\mu P_- \psi \tag{6.54a}$$
$$J_{Ra}^{\ \mu} = \frac{j_{Va}^{\ \mu} + J_{Aa}^{\ \mu}}{2} = \bar{\psi} T_a \gamma^\mu P_+ \psi \tag{6.54b}$$

and left and right gauge fields

$$A_{L\mu}^{\ a} = A_{V\mu}^{\ a} - A_{A\mu}^{\ a} \tag{6.55a}$$
$$A_{R\mu}^{\ a} = A_{V\mu}^{\ a} + A_{A\mu}^{\ a} \tag{6.55b}$$

In term of these new variables the interaction reads

$$\mathcal{L}_I = J_{La}^{\ \mu} A_{L\mu}^{\ a} + J_{Ra}^{\ \mu} A_{R\mu}^{\ a}, \tag{6.56}$$

and defining $\delta_L = \delta_V - \delta_A$ and $\delta_R = \delta_V + \delta_A$ the algebra becomes

$$[\delta_{L\epsilon_1}, \delta_{L\epsilon_2}] = \delta_{L[\epsilon_1,\epsilon_2]}\,; \tag{6.57a}$$
$$[\delta_{L\epsilon_1}, \delta_{R\epsilon_2}] = 0\,; \tag{6.57b}$$
$$[\delta_{R\epsilon_1}, \delta_{R\epsilon_2}] = \delta_{R[\epsilon_1,\epsilon_2]}. \tag{6.57c}$$

In terms of these variables the left and right gauge transformations are completely decoupled. The left and right gauge fields transform in the usual way under the left and right gauge transformations and are coupled to the left and right currents respectively. In discussing the possible anomalies of this theory it is therefore more convenient to use the left–right decomposition than the vector–axial decomposition.

Since the left and right sectors of the theory are classically decoupled, it will be enough to study only one of them. From now on we will assume that there is only one gauge field A coupled to only one of the chiral components of the fermion. We are now going to consider anomalies for the local gauge transformations in this chirally coupled theory. The action is a chirally modified non-abelian version of (6.1):

$$S_F^{L/R}(\psi, \bar\psi, A) = -\int d^{2n}x\, \bar\psi \gamma^\mu D_\mu^{L/R} \psi, \tag{6.58}$$

where the new operator is defined as:

$$D_\mu^{L/R} = \partial_\mu - ieP_\mp A_\mu. \tag{6.59}$$

This action S^L has a local symmetry G_L,

$$\psi_L' = g^{-1}\psi_L; \qquad \bar\psi_L' = \bar\psi_L g, \tag{6.60a}$$
$$\psi_R' = \psi_R; \qquad \bar\psi_R' = \bar\psi_R, \tag{6.60b}$$
$$A_\mu' = g^{-1}A_\mu g + g^{-1}\partial_\mu g \tag{6.60c}$$

where $\psi_L = P_-\psi$, $\psi_R = P_+\psi$ and $g = e^{\alpha^a T_a}$. Interactions of this type occur in the Standard Model. Similar formulas hold for the right-chiral models.

We do not give here the calculation of the gauge anomaly.[4] The result is [GJa72, Bar69]

$$D_\mu \langle J_L^{\ \mu}\rangle_a = \pm\frac{1}{4\pi}\varepsilon^{\mu\nu}\operatorname{tr} T_a \partial_\mu A_\nu \qquad \text{for } n=1, \tag{6.61a}$$
$$D_\mu \langle J_L^{\ \mu}\rangle_a = \pm\frac{i}{24\pi^2}\varepsilon^{\mu\nu\lambda\rho}\operatorname{tr} T_a \partial_\mu \left(A_\nu \partial_\lambda A_\rho + \frac{1}{2}A_\nu A_\lambda A_\rho\right) \qquad \text{for } n=2, \tag{6.61b}$$

where the overall sign depends on the chirality of the fermions.

[4] The result for an abelian theory in two dimensions is derived in Exercise 6.7.

The form (6.61) of the anomaly is not unique: it depends on the chosen regularization of the fermionic determinant. Another regularization could result in another form of the effective action, differing by a local functional of the gauge field. To understand the extent of this ambiguity, recall that the fermionic determinant is given by a sum of one loop graphs with any number of insertions of the external field A. The first term contains one power of A and diverges like Λ^{n-1}, where Λ is some ultraviolet cutoff; the second contains two powers of A and diverges like Λ^{n-2} and so on. The n-th term contains A^n and is logarithmically divergent. All subsequent terms are finite. Divergent terms give rise to ambiguities in the effective action. One is free to change the renormalization conditions so as to add to the effective action finite terms proportional to the coefficients of these divergences. Therefore, one is free to modify the effective action by adding a polynomial in A of order n (and containing terms of dimension n). If the expression (6.61) was itself the variation of such a polynomial, then by a different choice of renormalization one could remove the anomaly. We shall now see that this question can be cast as a problem of cohomology.

6.4.2 The Wess–Zumino consistency condition

The operators δ_ϵ defined in (6.48) form a representation of the gauge algebra:

$$[\delta_{\epsilon_1}, \delta_{\epsilon_2}] = \delta_{[\epsilon_1,\epsilon_2]}. \tag{6.62}$$

If we now apply the above operatorial relation to the vacuum functional $W(A)$ we get an equation for the integrated anomaly $\mathcal{A}(\epsilon, A)$:

$$\delta_{\epsilon_1}\mathcal{A}(\epsilon_2, A) - \delta_{\epsilon_2}\mathcal{A}(\epsilon_1, A) = \mathcal{A}([\epsilon_1, \epsilon_2], A). \tag{6.63}$$

This is called the *Wess–Zumino (WZ) consistency condition*. If the anomaly is defined as gauge variation of the effective action W, as in (6.50), then it must satisfy the above constraint. Such anomalies are called *consistent* anomalies. If on the other hand one defines the anomaly as $\langle D_\mu J^\mu \rangle$, then the result of a calculation may or may not satisfy this condition, depending on the regularization procedure.

Multiplying (6.61) by an infinitesimal gauge parameter ϵ^a and integrating over spacetime we obtain the expressions[5]

$$\mathcal{A}(A,\epsilon) = \mp\frac{1}{4\pi}\int d^2x\,\varepsilon^{\mu\nu}\,\mathrm{tr}\,\partial_\mu\epsilon\,A_\nu \qquad \text{for } n=1; \tag{6.64a}$$

$$\mathcal{A}(A,\epsilon) = \mp\frac{i}{24\pi^2}\int d^4x\,\varepsilon^{\mu\nu\lambda\rho}\,\mathrm{tr}\,\partial_\mu\epsilon\left(A_\nu\partial_\lambda A_\rho + \frac{1}{2}A_\nu A_\lambda A_\rho\right) \qquad \text{for } n=2. \tag{6.64b}$$

One can verify by explicit calculation that they satisfy the WZ consistency condition, see Exercise 6.2. In fact, the WZ consistency condition determines the anomaly completely, up to an overall normalization.

The definitions in this section have a clear geometrical meaning. Let $\mathcal{C}$ be the space of connections A, and $\mathcal{G}$ the gauge group.[6] For a fixed infinitesimal gauge transformation ϵ, δ_ϵ is a first order (functional) differential operator corresponding to the directional derivative along a vector field tangent to the orbits of the gauge group in $\mathcal{C}$. We can think of it as a vertical vectorfield on $\mathcal{C}$, i.e. a vectorfield that is in the kernel of the projection $\mathcal{C}\to\mathcal{C}/\mathcal{G}$. Fix a reference gauge field A and consider its gauge orbit $\mathcal{O}_A$. It is diffeomorphic to the gauge group $\mathcal{G}$. The anomaly $\mathcal{A}$ is a linear functional that maps vectors on $\mathcal{O}_A$ to real numbers. Thus, we can think of it as a one-form on $\mathcal{O}_A$. Equation (6.63) is the statement that $\mathcal{A}$ is a closed form (a one-cocycle). If W was a globally well-defined functional on $\mathcal{O}_A$, equation (6.50) would mean that $\mathcal{A}$ is an exact form, and $\mathcal{A}$ would be in the trivial cohomology class in $H^1(\mathcal{O}_A)$, or equivalently of $H^1(\mathcal{G})$. However, at this stage we do not really know W well enough. Equation (6.50) must be interpreted as saying that $\mathcal{A}$ is *locally* exact, i.e. the differential the locally-defined functional $-W$. In Section 6.5 we will consider the global properties of W and we shall see that it is not globally well-defined. This will be a non-perturbative proof that the anomaly is a genuine physical phenomenon.

6.4.3 The covariant anomaly

One peculiar aspect of the expressions (6.61) is that they are not gauge covariant. One can see this as a further indication that gauge invariance is broken in these theories.

[5] For the present purposes it proves convenient to perform an integration by parts so that one derivative acts on the gauge parameter. This is legitimate, since ϵ vanishes at infinity.

[6] Unlike earlier sections, here we consider connections and gauge transformations on spacetime, not just space.

Define an arbitrary infinitesimal variation of the connection to be

$$\delta_B A^a_\mu = B^a_\mu, \tag{6.65}$$

in such a way that

$$\delta_B W = \int d^{2n}x \frac{\delta W}{\delta A^a_\mu} B^a_\mu = \int d^{2n}x \langle J_{La}{}^\mu \rangle B^a_\mu = -2 \int d^{2n}x \, \mathrm{tr} \langle J^\mu_L \rangle B_\mu. \tag{6.66}$$

Following a similar logic as in the derivation of the WZ consistency condition, we see that[7]

$$[\delta_\epsilon, \delta_B] = \delta_{[B,\epsilon]}. \tag{6.67}$$

We must therefore have

$$\delta_B \delta_\epsilon W - \delta_\epsilon \delta_B W = \delta_{[B,\epsilon]} W. \tag{6.68}$$

The l.h.s. is

$$-\delta_B \mathcal{A}(A,\epsilon) - \int d^{2n}x \, \delta_\epsilon \langle J_{La}{}^\mu \rangle B^a_\mu$$

and the r.h.s. is

$$\int d^{2n}x \langle J_{La}{}^\mu \rangle [B_\mu, \epsilon]^a = -\int d^{2n}x \, B^a_\mu [\langle J^\mu_L \rangle, \epsilon]_a,$$

where we used total antisymmetry of the structure constants. Putting these pieces together, and keeping in mind that B is arbitrary, the gauge variation of the current is determined by the following condition:

$$\int d^{2n}x \, \delta_\epsilon \langle J_{La}{}^\mu \rangle B^a_\mu = \int d^{2n}x \, [\langle J^\mu_L \rangle, \epsilon]_a B^a_\mu - \delta_B \mathcal{A}(A,\epsilon). \tag{6.69}$$

The first piece on the r.h.s. is the usual covariant transformation of the current. This relation tells us that the current is covariant only if the anomaly is zero.

We may then ask whether we can define a new current[8]

$$J'^\mu_{La} = J_{La}{}^\mu + X^\mu_a, \tag{6.70}$$

[7] Since B is itself a transformation parameter, we assume that it does not to vary under gauge transformations.

[8] Note that the freedom of adding a piece to the current is more general than the freedom of adding a local counterterm ΔW to the effective action. In the latter case one would have $X^\mu_a = \frac{\delta \Delta W}{\delta A^a_\mu}$, which is not true in general, and in particular not for the X's given below.

transforming covariantly. For this we need that the transformation property of X_a^μ be given by

$$\int d^{2n}x\,(\delta_\epsilon X)_a^\mu B_\mu^a = \int d^{2n}x\,[X^\mu,\epsilon]_a B_\mu^a + \delta_B \mathcal{A}(A,\epsilon). \tag{6.71}$$

In this case the anomalous terms cancel and

$$\int d^{2n}x\,\delta_\epsilon \langle J'^\mu_{La}\rangle B_\mu^a = \int d^{2n}x\,[\langle J'^\mu_L\rangle,\epsilon]_a B_\mu^a. \tag{6.72}$$

or, due to the arbitrariness of B_μ^a,

$$\delta_\epsilon \langle J'^\mu_L\rangle = [\langle J'^\mu_L\rangle,\epsilon]. \tag{6.73}$$

It is not a priori clear that a solution of (6.71) exist. Since we insist on locality, X^μ must be of dimension one for $n=1$ and three for $n=2$. Thus in two spacetime dimension X^μ must be proportional to $\epsilon^{\mu\nu}A_\nu$ and in four dimensions it must be a linear combination of terms of the form AdA or A^3. One then finds [BaZ84]

$$X_a^\mu = \mp\frac{1}{8\pi}\varepsilon^{\mu\nu}A_\nu \qquad \text{for } n=1, \tag{6.74a}$$

$$X_a^\mu = \pm\frac{i}{24\pi^2}\varepsilon^{\mu\nu\rho\sigma}\operatorname{tr}\left(A_\nu\partial_\rho A_\sigma + \partial_\nu A_\rho A_\sigma + \frac{3}{2}A_\nu A_\rho A_\sigma\right) \qquad \text{for } n=2. \tag{6.74b}$$

The polynomial X_a^μ is called the *Bardeen–Zumino counterterm* and the covariant divergence of $\langle J'^\mu_{La}\rangle$ is known as the *covariant anomaly*:

$$\langle D_\mu J'^\mu_L\rangle_a = \pm\frac{1}{2\pi}\varepsilon^{\mu\nu}\operatorname{tr}T_a F_{\mu\nu} \qquad \text{for } n=1 \tag{6.75a}$$

$$\langle D_\mu J'^\mu_L\rangle_a = \pm\frac{i}{16\pi^2}\varepsilon^{\mu\nu\lambda\rho}\operatorname{tr}T^a F_{\mu\nu}F_{\lambda\rho} \qquad \text{for } n=2. \tag{6.75b}$$

One can then check that the integrated covariant anomalies

$$\mathcal{A}'(A,\epsilon) = \pm\frac{1}{2\pi}\int d^2x\,\varepsilon^{\mu\nu}\operatorname{tr}\epsilon F_{\mu\nu} \qquad \text{for } n=1 \tag{6.76a}$$

$$\mathcal{A}'(A,\epsilon) = \pm\frac{i}{16\pi^2}\int d^4x\,\varepsilon^{\mu\nu\lambda\rho}\operatorname{tr}\epsilon F_{\mu\nu}F_{\lambda\rho} \qquad \text{for } n=2 \tag{6.76b}$$

do not satisfy the WZ consistency condition, see Exercise 6.2. Therefore, these anomalies are not the variation of a functional $W(A)$, not even locally. They can still be used as diagnostics for the existence of the anomaly, as we shall see in Section 6.8.1.

6.4.4 Commutator anomalies

Recall that in the canonical formulation of YM theory, the generators of time-independent gauge transformations G_ϵ, that we refer to as the Gauss law generators, satisfy the Poisson brackets (1.215). In the quantum theory, we should therefore have

$$[G_{\epsilon_1}, G_{\epsilon_2}] = iG_{[\epsilon_1,\epsilon_2]}. \tag{6.77}$$

In the presence of chirally coupled fermions, the Gauss law contains also the charge density $J_{La}^{\ 0}$ or $J_{Ra}^{\ 0}$. When integrated, the corresponding classical charges satisfy the algebra (1.17). Thus the full Gauss law generators

$$G_\epsilon = \int d^3x\, \epsilon^a (D_i E_{ia} + J_{L/Ra}^{\ 0}) \tag{6.78}$$

are still expected to satisfy (6.77). Instead, in an anomalous gauge theory, the algebra is modified by the appearance of a *Schwinger term*

$$[G_{\epsilon_1}, G_{\epsilon_2}] = iG_{[\epsilon_1,\epsilon_2]} + \Omega(A, \epsilon_1, \epsilon_2). \tag{6.79}$$

This term has the same origin as the anomalous divergence of the current: the fermionic charge density

$$J_{La}^{\ 0}(x) = \psi^\dagger(x) T_a P_- \psi(x), \quad J_{Ra}^{\ 0}(x) = \psi^\dagger(x) T_a P_+ \psi(x),$$

being the product of two fermion fields evaluated at the same point, is an ill-defined operator. In two dimensions, the current can be defined by normal ordering. The calculation of the commutator of two fermionic charge densities is left as Exercise 6.3. Let us call G' the Gauss law operator with this definition of the current. Then one finds

$$[G'_a(x), G'_b(y)] = if_{ab}{}^c G'_c(x)\delta(x-y) \pm \frac{i}{2\pi}\delta_{ab}\frac{d}{dx}\delta(x-y), \tag{6.80}$$

with the sign depending on the chirality. Thus the Schwinger term is

$$\Omega(\epsilon_1, \epsilon_2) = \pm\frac{1}{2\pi}\int dx\, \mathrm{tr}\, \epsilon_1 \partial_x \epsilon_2. \tag{6.81}$$

In this particular case the Schwinger term originates entirely from the fermionic term. Also note that the Schwinger term is independent of A: it defines merely a central extension of the algebra.

A gauge invariant regularization of the current akin to (6.8) leads instead to an extension of the algebra depending explicitly on A:

$$[G_a(x), G_b(y)] = if_{ab}{}^c G_c(x)\delta(x-y) \mp \frac{1}{8\pi} f_{abc} A_1^c \delta(x-y), \tag{6.82}$$

which means that

$$\Omega(\epsilon_1, \epsilon_2) = \mp\frac{1}{4\pi}\int dx\, \mathrm{tr}\,[\epsilon_1, \epsilon_2]A_1. \tag{6.83}$$

One easily sees that these two results differ by a redefinition

$$G'_a(x) = G_a(x) \pm \frac{1}{8\pi}A^a_1(x). \tag{6.84}$$

Notice that the additional term is not invariant under infinitesimal gauge transformations, so in (6.82) the Schwinger term has contributions from both terms of the Gauss law. The additional term proportional to A is recognized as the time component of the Bardeen–Zumino form (6.74a). Thus we may say that (6.81) is the Schwinger term for the covariant current, whereas (6.83) is the Schwinger term for the consistent current.

Similar results hold in four dimensions [Jo851, Jo852]. The Gauss law defined with the consistent current satisfies

$$\begin{aligned}[G_a(x), G_b(y)] = {} & if_{ab}{}^c G_c(x)\delta(x-y) \\ & \mp \frac{i}{48\pi^2}\varepsilon_{ijk}\mathrm{tr}\big[[T_a, T_b](\partial_i A_j A_k + A_i \partial_j A_k + A_i A_j A_k) \\ & + \partial_i(T_a A_j T_b A_k)\big]\delta(x-y),\end{aligned} \tag{6.85}$$

corresponding to the Schwinger term

$$\begin{aligned}\Omega(\epsilon_1, \epsilon_2) = {} & \mp\frac{i}{48\pi^2}\int d^3x\, \varepsilon_{ijk}\mathrm{tr}\big[[\epsilon_1, \epsilon_2](\partial_i A_j A_k + A_i \partial_j A_k + A_i A_j A_k) \\ & + \epsilon_1 \partial_i A_j \epsilon_2 A_k - \epsilon_1 A_i \epsilon_2 \partial_j A_k\big].\end{aligned} \tag{6.86}$$

The redefinition

$$G'_a(x) = G_a(x) + X^0_a(x), \tag{6.87}$$

with X^0_a given by (6.74b), leads to the algebra

$$\begin{aligned}[G'_a(x), G'_b(y)] = {} & if_{ab}{}^c G'_c(x)\delta(x-y) \\ & \mp \frac{i}{24\pi^2}\varepsilon_{ijk}\mathrm{tr}\big[\{T_a, T_b\}\,\partial_i A_j\big]\partial_k\delta(x-y),\end{aligned} \tag{6.88}$$

that corresponds to the Schwinger term

$$\Omega'(\epsilon_1, \epsilon_2) = \mp\frac{i}{24\pi^2}\int d^3x\, \varepsilon_{ijk}\mathrm{tr}\big[\{\partial_i\epsilon_1, \partial_j\epsilon_2\}A_k\big]\,. \tag{6.89}$$

Also in this case, the Schwinger terms coming from the covariant current contain fewer powers of A, and derivatives of the transformation parameters.

The anomalies in commutators that we have just encountered satisfy a consistency condition of cohomological nature, that is analogous to the WZ condition. Quite generally, let us assume that the algebra of generators of time-independent gauge transformations has an abelian extension (6.79). In order that the generators G_ϵ be representable as linear operators on a Hilbert space, they must satisfy the Jacobi identity:

$$[G_{\epsilon_1}, [G_{\epsilon_2}, G_{\epsilon_3}]] + [G_{\epsilon_2}, [G_{\epsilon_3}, G_{\epsilon_1}]] + [G_{\epsilon_3}, [G_{\epsilon_1}, G_{\epsilon_2}]] = 0. \tag{6.90}$$

Therefore we get the following consistency condition:

$$\begin{aligned} &\delta_{\epsilon_1}\Omega(A;\epsilon_2,\epsilon_3) + \delta_{\epsilon_2}\Omega(A;\epsilon_3,\epsilon_1) + \delta_{\epsilon_3}\Omega(A;\epsilon_1,\epsilon_2) \\ &+\Omega(A;\epsilon_1,[\epsilon_2,\epsilon_3]) + \Omega(A;\epsilon_2,[\epsilon_3,\epsilon_1]) + \Omega(A;\epsilon_3,[\epsilon_1,\epsilon_2]) = 0. \end{aligned} \tag{6.91}$$

This means that Ω, regarded as a differential 2-form on $\mathcal{G}$, must be closed, or equivalently that Ω has to be a two-cocycle for the action of the gauge group. A redefinition of the Gauss law

$$G'_\epsilon = G_\epsilon + X(A;\epsilon), \tag{6.92}$$

changes the Schwinger term by the coboundary of X:

$$\Omega'(A;\epsilon_1,\epsilon_2) = \Omega(A;\epsilon_1,\epsilon_2) + \delta_{\epsilon_1}X(A;\epsilon_2) - \delta_{\epsilon_2}X(A;\epsilon_1) - X(A,[\epsilon_1,\epsilon_2]). \tag{6.93}$$

If $X(A;\epsilon) = \delta_\epsilon F(A)$, i.e. if X is the coboundary of a zero-cochain F, equation (6.62) implies that Ω is unchanged.

6.5 The Wess–Zumino functional

In the preceding section we have considered the effect of infinitesimal gauge transformations on the fermionic determinant. Let us now consider the effect of finite gauge transformations. Define the *Wess–Zumino (WZ) functional* $\Gamma_{WZ}(A,g)$ to be (minus) the change in the fermionic effective action under a gauge transformation:

$$W(A^g) - W(A) = -\Gamma_{WZ}(A,g). \tag{6.94}$$

When g differs infinitesimally from the identity, Γ_{WZ} becomes the anomaly:

$$\Gamma_{WZ}(A,1+\epsilon) = \mathcal{A}(A,\epsilon). \tag{6.95}$$

From the definition one finds that

$$\Gamma_{WZ}(A^{g_1}, g_2) - \Gamma_{WZ}(A, g_1 g_2) + \Gamma_{WZ}(A, g_1) = 0. \tag{6.96}$$

This condition has a cohomological significance and is the analogue of the WZ consistency condition for finite transformations. A functional satisfying it is said to be a one-cocycle for the action of the gauge group with coefficients in the smooth functionals of A_μ.

From the way it was derived, Γ_{WZ} is seen to depend on a connection A and a gauge transformation g. However, g is just a map from spacetime to the group G and we can also think of it as a configuration for a chiral model. In this case we denote it as U and we can think as $\Gamma_{WZ}(A, U)$ as a possible term in the action for a chiral model coupled to gauge fields. In this case (6.96) can be rewritten in the suggestive form

$$\Gamma_{WZ}(A^g, U^g) - \Gamma_{WZ}(A, U) = -\Gamma_{WZ}(A, g). \tag{6.97}$$

where $U^g = g^{-1}U$ can be thought of as the gauge transform of U by g. Comparing with (6.94), this formula shows that the WZ functional has the same anomalous transformation property as the fermionic determinant. The important difference, that we shall now see, is that whereas W is a non-local functional, Γ_{WZ} is a local functional.[9]

We can compute Γ_{WZ} explicitly by integrating the anomaly. We begin by fixing a reference gauge field A_μ. Then we can identify the orbit through A_μ with $\mathcal{G}$ (by mapping g to A_μ^g) and we can regard W as a function on $\mathcal{G}$. As above, we think of the anomaly $\mathcal{A}$ as a one-form on $\mathcal{G}$. Since $\mathcal{A}$ is closed, its integral along a curve does not change under continuous deformations of the curve, as long as the endpoints remain fixed. It can only change in a discontinuous way if we change the homotopy class of the curve. Let $g(r)$ be a one-parameter family of gauge transformations interpolating between g and the identity:

$$\bar{g}(r) = e^{r\epsilon^a T_a}\,; \qquad \bar{g}(0) = e\,; \qquad \bar{g}(1) = g. \tag{6.98}$$

and let

$$\bar{A}_\mu(r) = A^{\bar{g}(r)} = \bar{g}^{-1} A_\mu \bar{g} + \bar{g}^{-1}\partial_\mu \bar{g} \tag{6.99}$$

be the gauge transform of A_μ at the point r along the path. Then the WZ functional can be written

$$\Gamma_{WZ}(A, g) = \int_0^1 dr\, \mathcal{A}(\bar{A}, \bar{g}^{-1}\partial\bar{g}). \tag{6.100}$$

[9] Here local means that it depends only on the fields and finitely many derivatives of the fields. It is also local in the sense that it is not smooth on the whole field space, as we shall discuss later.

Let us perform the integral explicitly in two dimensions ($n = 1$). Using the anomaly (6.64) we have to compute

$$\frac{1}{4\pi}\int_0^1 dr \int d^2x\,\varepsilon^{\mu\nu}\mathrm{tr}\,\bar{g}^{-1}\partial_r\bar{g}\partial_\mu(\bar{g}^{-1}A_\nu\bar{g}+\bar{g}^{-1}\partial_\nu\bar{g})$$
$$=\frac{1}{4\pi}\int_0^1 dr \int d^2x\,\varepsilon^{\mu\nu}\mathrm{tr}\,\partial_r\bar{g}\bar{g}^{-1}\left[\partial_\mu A_\nu-\partial_\mu\bar{g}\bar{g}^{-1}A_\nu+A_\nu\partial_\mu\bar{g}\bar{g}^{-1}-\partial_\mu\bar{g}\bar{g}^{-1}\partial_\nu\bar{g}\bar{g}^{-1}\right].$$

Now we rewrite this in a covariant form in the three coordinates r, x_1, x_2, which parametrize a three-dimensional ball with boundary S^2 (it is assumed that the r-component of A_μ is zero):

$$\frac{1}{4\pi}\int d^3x\,\varepsilon^{\lambda\mu\nu}\mathrm{tr}\left[\partial_\lambda\bar{g}\bar{g}^{-1}\partial_\mu A_\nu-\partial_\lambda\bar{g}\bar{g}^{-1}\partial_\mu\bar{g}\bar{g}^{-1}A_\nu-\frac{1}{3}\partial_\lambda\bar{g}\bar{g}^{-1}\partial_\mu\bar{g}\bar{g}^{-1}\partial_\nu\bar{g}\bar{g}^{-1}\right].$$

The first two terms are a total derivative, and can be rewritten as an integral on the boundary S^2, so we obtain

$$\Gamma_{WZ}(A,g)=-\frac{1}{4\pi}\int d^2x\,\varepsilon^{\mu\nu}\mathrm{tr}\,R_\mu A_\nu-\frac{1}{12\pi}\int d^3x\,\varepsilon^{\lambda\mu\nu}\mathrm{tr}\,\bar{R}_\lambda\bar{R}_\mu\bar{R}_\nu, \quad (6.101)$$

where $R_\mu=\partial_\mu g g^{-1}$ and $\bar{R}_\mu=\partial_\mu\bar{g}\bar{g}^{-1}$. We note that if $g=1+\epsilon$, $R_\mu=\partial_\mu\epsilon$, so the first term gives back $\mathcal{A}(A,\epsilon)$, as expected. Also, we recognize that the second term is the WZW action S_{WZW}, defined in (4.15), with the correct normalization of the coefficient, $c=2\pi$.

Recall that, even though the integrand of $S_{WZW}(\bar{g})$ is the same as that of the winding number, $S_{WZW}(\bar{g})$ is not a topological invariant, since it depends on the boundary values of $\bar{g}$. The WZW action that appears in the WZ functional corresponds to the choice $n=1$, or $c=2\pi$, for the coefficient. Thus, the WZ functional can be viewed as a left-gauged extension of the WZW functional.

One can proceed in the same way in higher dimensions. Integrating the anomaly (6.64) one arrives at the following expression for the Wess–Zumino functional in four dimensions

$$\begin{aligned}\Gamma_{WZ}(A_\mu,g)=&-\frac{i}{48\pi^2}\int d^4x\,\varepsilon^{\mu\nu\rho\sigma}\mathrm{tr}\Big[\big(A_\mu\partial_\nu A_\rho+\partial_\mu A_\nu A_\rho+A_\mu A_\nu A_\rho\big)R_\sigma\\&-\frac{1}{2}A_\mu R_\nu A_\rho R_\sigma-A_\mu R_\nu R_\rho R_\sigma\Big]\\&-\frac{i}{240\pi^2}\int_B d^5x\,\varepsilon^{\lambda\mu\nu\rho\sigma}\mathrm{tr}\,\bar{R}_\lambda\bar{R}_\mu\bar{R}_\nu\bar{R}_\rho\bar{R}_\sigma. \quad (6.102)\end{aligned}$$

Once again we recognize that the last term is the WZW functional with the correctly normalized coefficient $c=2\pi$.

Finally, let us return to the question whether $W(A)$ is a globally well-defined functional on the orbits of the gauge group. We will discuss this in the two-dimensional case, for an $SU(2)$ YM theory. Assuming that spacetime has been compactified to S^2, the gauge group is $\mathcal{G} = \Gamma_*(S^2, SU(2))$ and

$$\pi_1(\mathcal{G}) = \pi_3(SU(3)) = \mathbb{Z}.$$

Thus we need to ask whether W is single-valued along a non-contractible path in the orbit. As observed above, the gauge variation of W is the same as the gauge variation of the WZ functional. Therefore, up to an additive constant, these two functionals are the same, when restricted to a gauge orbit. We therefore ask whether Γ_{WZ} is single-valued along a non-contractible path in the orbit. To answer this question one just has to integrate the anomaly along a closed loop, in which case in (6.98) we have to set $\bar{g}(1) = e$. Then, of the whole WZ action, only the WZW term remains, and is equal to 2π. Thus, W is not single-valued, but e^{iW} is.

Now consider a $SU(N)$ YM theory in four dimensions, with $N > 2$. Assuming that spacetime has been compactified to S^4, the gauge group is $\mathcal{G} = \Gamma_*(S^4, SU(N))$ and

$$\pi_1(\mathcal{G}) = \pi_5(SU(N)) = \mathbb{Z}.$$

One can repeat the argument given above for the two-dimensional case, and we conclude that W changes by 2π when one follows a noncontractible path in the orbit.

6.6 The descent equations

The axial anomaly, the gauge anomaly and the Schwinger terms of gauge theories in different dimensions, are strictly related. In fact, we will see that one can obtain the solution of the WZ consistency condition, i.e. the consistent anomaly, including the correct normalization, by a series of manipulations, starting from the axial anomaly in a space of two more dimensions. We will define the cocycles ω_r^k, for $k = 0, 1, 2$, where $r = 2n - k - 1$ is the degree of ω as a form in spacetime (and hence also the dimension of the space over which ω has to be integrated), and k is its degree as a form in the space of connections (more precisely, in an orbit of the gauge group in the space of connections). This means that, given k infinitesimal gauge transformation parameters $\epsilon_1, \dots, \epsilon_k$, and r vectorfields $v_1, \dots, v_r$, $\omega_r^k(\epsilon_1, \dots, \epsilon_k, v_1, \dots, v_r)$ is a real number. The fact that these are cocycles means that they are closed, both

as r-forms on space(time) and as k-forms on the orbit of the gauge group. The relations between all these forms constitute the so-called "descent equations".

In order to minimize the index clutter it is convenient to use the algebra of differential forms. Hence we write $A = A^a_\nu T_a dx^\nu$, $F = \frac{1}{2} F^a_{\mu\nu} T_a dx^\mu \wedge dx^\nu$. The exterior derivative acting on a p-form ω can be defined in a coordinate-independent way by specifying the result of acting with $d\omega$ on $p+1$ vector-fields:

$$\begin{aligned} d\omega(v_1, \dots, v_{p+1}) = & \sum_{1<i<p+1} (-1)^{i+1} v_i(\omega(v_1, \dots, \hat{v}_i, \dots, v_{p+1})) \\ & + \sum_{1<i<j<p+1} (-1)^{i+j} \omega([v_i, v_j], v_1, \dots, \hat{v}_i, \dots, \hat{v}_j, \dots, v_{p+1}), \end{aligned} \tag{6.103}$$

where a hat over a vector means that it is missing. If the components of ω are defined by

$$\omega = \frac{1}{p!} \omega_{\mu_1,\dots,\mu_p} dx^{\mu_1} \wedge \dots \wedge dx^{\mu_p} \tag{6.104}$$

we can also write

$$d\omega = \frac{1}{p!} d\omega_{\mu_1,\dots,\mu_p} \wedge dx^{\mu_1} \wedge \dots \wedge dx^{\mu_p}, \tag{6.105}$$

where the differential acts only on the components. Thus we can write $F = dA + A \wedge A$, and to further condense the notation also wedge products will not be written explicitly, so $F = dA + A^2$.

One begins from the expression for the Chern class, a $2n$-form in $2n$ dimensions:

$$c_n = k_n \operatorname{tr} F^n, \tag{6.106}$$

where

$$k_n = \frac{1}{n!} \left(\frac{i}{2\pi} \right)^n \tag{6.107}$$

is a normalization constant, such that the integral of c_n is an integer. Note that since in our conventions the Lie algebra generators T_a (and hence also F) are antihermitian, the power of i is needed to make the expression real. In particular, in the following we need

$$k_2 = -\frac{1}{8\pi^2}, \quad k_3 = -\frac{i}{48\pi^3}, \quad k_4 = \frac{1}{384\pi^4}. \tag{6.108}$$

The Chern class is gauge invariant

$$\delta_\epsilon c_n = 0 \tag{6.109}$$

and closed

$$dc_n = 0. \tag{6.110}$$

Thus we can write c_n, at least locally, as the exterior differential of an $2n-1$ form ω^0_{2n-1}, called the Chern–Simons form

$$c_n = d\omega^0_{2n-1}. \tag{6.111}$$

A general formula can be given in any dimension, but we limit ourselves to the cases $n = 2, 3, 4$, where we have

$$\omega^0_3(A) = k_2 \mathrm{tr}\Big(FA - \frac{1}{3}A^3\Big), \tag{6.112a}$$

$$\omega^0_5(A) = k_3 \mathrm{tr}\Big(F^2A - \frac{1}{2}FA^3 + \frac{1}{10}A^5\Big), \tag{6.112b}$$

$$\omega^0_7(A) = k_4 \mathrm{tr}\Big(F^3A - \frac{2}{5}F^2A^3 - \frac{1}{5}FAFA^2 + \frac{1}{5}FA^5 - \frac{1}{35}A^7\Big). \tag{6.112c}$$

The gauge variation of the Chern–Simons form is closed, because

$$d\delta_\epsilon \omega^0_{2n-1} = \delta_\epsilon d\omega^0_{2n-1} = \delta_\epsilon c_n = 0,$$

therefore it is locally the differential of a $(2n-2)$-form:

$$\delta_\epsilon \omega^0_{2n-1}(A) = d\omega^1_{2n-2}(A, \epsilon). \tag{6.113}$$

This form can be written as

$$\omega^1_{2n-2}(A, \epsilon) = \mathrm{tr}\, d\epsilon\, \phi_{2n-3}(A), \tag{6.114}$$

where the $(2n-3)$-form $\phi_{2n-3} = \phi^a_{2n-3} T_a$ is a polynomial in A and F. For $n = 2, 3, 4$ this polynomial is given by

$$\phi_1 = -k_2 A, \tag{6.115a}$$

$$\phi_3 = -\frac{k_3}{2}(FA + AF - A^3), \tag{6.115b}$$

$$\begin{aligned}\phi_5 = -\frac{k_4}{3}\Big[&(F^2A + FAF + AF^2) \\ &- \frac{4}{5}(A^3F + FA^3) - \frac{2}{5}(A^2FA + AFA^2) + \frac{3}{5}A^5\Big].\end{aligned} \tag{6.115c}$$

From these formulae we recognize that

$$2\pi \int \omega^1_{2n-2}(A, \epsilon) = \mathcal{A}(A, \epsilon) \tag{6.116}$$

is the consistent anomaly in dimension $2n-2$. The origin of the factor 2π will be explained below.

The coboundary of ω^1_{2n-2} (in the sense of Lie algebra cohomology) is a closed $(2n-2)$-form, thus locally it is the differential of a $(2n-3)$-form

$$\delta_{\epsilon_1}\omega^1_{2n-2}(A,\epsilon_2)-\delta_{\epsilon_2}\omega^1_{2n-2}(A,\epsilon_1)-\omega^1_{2n-2}(A,[\epsilon_1,\epsilon_2])=d\omega^2_{2n-3}(A,\epsilon_1,\epsilon_2). \tag{6.117}$$

When integrated on a closed manifold without boundary, this is just the WZ consistency condition (6.63). For $n=2,3,4$ we find

$$\omega^2_1(A,\epsilon_1,\epsilon_2) = 2k_2\mathrm{tr}\,\epsilon_1 d\epsilon_2, \tag{6.118a}$$

$$\omega^2_3(A,\epsilon_1,\epsilon_2) = k_3\mathrm{tr}\,\{d\epsilon_1,d\epsilon_2\}A, \tag{6.118b}$$

$$\omega^2_5(A,\epsilon_1,\epsilon_2) = \frac{k_4}{15}\mathrm{tr}\,(5F-3A^2)\,[2A\{d\epsilon_1,d\epsilon_2\}-d\epsilon_1 A d\epsilon_2+d\epsilon_2 A d\epsilon_1]\,. \tag{6.118c}$$

The 2-cocycles in (6.118a) and (6.118b) are just the Schwinger terms (6.81) and (6.89):

$$\int d^{2n}x\,\omega^2_{2n-3}(\epsilon_1,\epsilon_2)=\Omega(A,\epsilon_1,\epsilon_2). \tag{6.119}$$

The factor 2π has the same origin as in (6.116).

It is clear from (6.113) and (6.117) that ω^1_{2n-2} and ω^2_{2n-3} are only defined up to a closed form. In particular one could add to ω^1_{2n-2} the closed form $-d(\mathrm{tr}\epsilon\phi(A))$ and get

$$\hat\omega^1_{2n-2}(A,\epsilon)=-\mathrm{tr}\,\epsilon d\phi_{2n-3}, \tag{6.120}$$

which is another form of the consistent anomaly. Applying the coboundary to $\hat\omega^1_{2n-2}$ defines a different 2-cocycle $\hat\omega^2_p$:

$$\delta_{\epsilon_1}\hat\omega^1_{2n-2}(A,\epsilon_2)-\delta_{\epsilon_2}\hat\omega^1_{2n-2}(A,\epsilon_1)-\hat\omega^1_{2n-2}(A,[\epsilon_1,\epsilon_2])=d\hat\omega^2_{2n-3}(A,\epsilon_1,\epsilon_2). \tag{6.121}$$

For $n = 2, 3, 4$

$$\hat{\omega}_1^2(A, \epsilon_1, \epsilon_2) = k_2 \mathrm{tr}\,[\epsilon_1, \epsilon_2]\, A, \tag{6.122a}$$

$$\begin{aligned}\hat{\omega}_3^2(A, \epsilon_1, \epsilon_2) = \frac{1}{2} k_3 \mathrm{tr}\,\big[&[\epsilon_1, \epsilon_2](FA + AF - A^3) \\ &+ \epsilon_1 dA \epsilon_2 A - \epsilon_1 A \epsilon_2 dA\big],\end{aligned} \tag{6.122b}$$

$$\begin{aligned}\hat{\omega}_5^2(A, \epsilon_1, \epsilon_2) = \frac{1}{3} k_4 \mathrm{tr}\big\{&[\epsilon_1, \epsilon_2]\big[(F^2 A + FAF + AF^2) \\ &- \frac{4}{5}\{A^3, F\} - \frac{2}{5}\{A, AFA\} + \frac{3}{5}A^5\big] \\ &- \frac{1}{5}[\epsilon_1, d\epsilon_2][F, A^2] - \frac{3}{5}(d\epsilon_1 A \epsilon_2 + \epsilon_2 A d\epsilon_1)(FA + AF - A^3) \\ &+ \frac{1}{5}[\epsilon_2, d\epsilon_1][F, A^2] - \frac{3}{5}(d\epsilon_2 A \epsilon_1 + \epsilon_1 A d\epsilon_2)(FA + AF - A^3)\big\}.\end{aligned} \tag{6.122c}$$

The cocycles (6.122a) and (6.122b) are the Schwinger terms (6.83) and (6.86).

Finally we discuss the normalization of the gauge anomaly, in particular the origin of the factor 2π in equation (6.116) and (6.119). Consider the integral of $\omega^0_{2n-1}(g^{-1}dg)$ on a $2n-1$-dimensional sphere. Since the gauge field is pure gauge, we can put $F = 0$ and so in Equations (6.112) only the last term survives. One can actually show that the general formula for this term is

$$(-1)^{n-1} \frac{n!(n-1)!}{(2n-1)!} \mathrm{tr} A^{2n-1}. \tag{6.123}$$

Now consider the formula for c_n integrated not on a compact manifold without boundary, but on $\mathbb{R}^n$. Further assume that $A \to g^{-1}dg$ on the sphere at infinity. We then find that

$$c_n = (-1)^{n-1} \left(\frac{i}{2\pi}\right)^n \frac{(n-1)!}{(2n-1)!} \int_{S^{2n-1}} \mathrm{tr}(g^{-1}dg)^{2n-1}. \tag{6.124}$$

Since c_n is an integer, so is the integral on the right. Indeed for $n = 2$ the integral on the right is just the winding number.[10] When the integral is extended only over one hemisphere this is the two-dimensional WZW action, and the analogous integral with $n = 2$ is precisely the four-dimensional WZW action (4.25). Now we recall that the coefficient c of the WZW action, in any dimension, is quantized in integral multiples of 2π, and also that the WZ action is just a gauged version of the WZW action, keeping the same normalization. Since the gauge anomaly is the variation of the WZ action, it has an additional factor 2π compared to the result of the dimensional descent.

[10] In fact the argument given here is just the reverse of the argument given in Section 3.6.3 to show that c_2 is an integer.

6.7 A global gauge anomaly

In studying gauge anomalies we have discussed the effect of infinitesimal gauge transformations on the effective action of chiral fermions. If a theory is invariant under infinitesimal gauge transformations, it may still fail to be invariant under gauge transformations that are not homotopic to the identity. The classic example is the $SU(2)$ gauge theory in four dimensions coupled to Weyl fermions [Wit82].

The discussion will be in the context of the Euclidean functional integral. We start by noticing that the usual non-abelian anomaly vanishes in the pure $SU(2)$ case, as one can easily check by the explicit computation of the trace in (6.61). However, the fourth homotopy group of $SU(2)$ is non-trivial

$$\pi_4(SU(2)) = \mathbb{Z}_2. \tag{6.125}$$

This means that when we compactify the four-dimensional spacetime to S^4, there are gauge transformations $U(x)$, with $U(x) \to 1$ as $|x| \to \infty$, that cannot be continuosly deformed to the identity. The fact that the homotopy group is $\mathbb{Z}_2$ means that if U is such a transformation, U^2 is homotopic to the identity. We would like to determine the transformation properties of the fermionic effective action under a homotopically nontrivial transformation $U(x)$.

Let $\mathcal{D}[A]$ be the Dirac operator acting on an $SU(2)$ doublet of Dirac fermions. The effective action $W(A)$ for Dirac fermions is defined by

$$e^{-W_D(A)} = \int (d\psi d\bar{\psi})_D e^{-\int \bar{\psi}\mathcal{D}[A]\psi} = \det\mathcal{D}[A]$$

and can be regularized in a way that is invariant under all gauge transformations. The effective action for two $SU(2)$ doublets of Weyl fermions is the same as the effective action for one $SU(2)$ doublet of Dirac fermions. Thus the determinant of the Dirac operator acting on one Weyl doublet must be the square root of the Dirac operator acting on one Dirac doublet:

$$e^{-W_W(A)} = \int (D\bar{\psi}D\psi)_W e^{-\int \bar{\psi}\mathcal{D}[A]\psi} = (\det\mathcal{D}[A])^{1/2}. \tag{6.126}$$

Now there is an ambiguity in extracting the square root, because we have to choose the sign, and when we choose it for a particular A, we fix this ambiguity everywhere, because the determinant is a continuous functional on the space of the gauge connections. Consider a path $A(t)$ joining A to A^U, where U is homotopically nontrivial. For example, we may choose

$$\tilde{A}(t) = (1-t)A + tA^U. \tag{6.127}$$

The spectrum of A^U is exactly the same as the spectrum of A, but there may be some rearrangement of the eigenvalues along the way. If one eigevalue (or an odd number of eigenvalues) changes sign along the path, then

$$(\det\mathcal{D}[A^U])^{1/2} = -(\det\mathcal{D}[A])^{1/2}, \tag{6.128}$$

or equivalently,

$$W(A^U) = W(A) + i\pi. \tag{6.129}$$

Witten proved, using a mod2 version of the Atiyah–Singer index theorem for the Dirac operator in five dimensions, that there has to be an odd number of eigenvalues crossing zero along the path, so that indeed (6.128) must hold. Since for every gauge configuration there is another one that is gauge equivalent and for which $(\det\mathcal{D}[A])^{1/2}$ has opposite sign, this has the consequence that the functional integral over the gauge fields must be exactly zero. This makes the theory ill defined.[11]

There is an alternative way of reaching the same conclusion that does not rely on the index theorem [Kli90]. First we observe that, by continuity, if (6.128) holds for a particular A, it holds for all A. It is therefore enough to prove (6.128) for $A = 0$. Let us embed $SU(2)$ into $SU(3)$, and represent the homotopically nontrivial map $U_1 : S^4 \to SU(2)$ by $\widetilde{U}_1 : S^4 \to SU(3)$:

$$\widetilde{U}_1 = \begin{bmatrix} U_1 & 0 \\ 0 & 1 \end{bmatrix}$$

Since $\pi_4(SU(3)) = 0$, it is possible to find a continuous path $\widetilde{U}(t)$ in $SU(3)$ that joins $\widetilde{U}_1$ to the identity (this path will obviously not lie entirely in the subgroup $SU(2)$):

$$\widetilde{U}(1) = \widetilde{U}_1, \quad \widetilde{U}(0) = \mathbb{I}. \tag{6.130}$$

Unlike (6.127), this path in the space of gauge fields consisting entirely of pure gauge fields, albeit in a larger gauge group:

$$\tilde{A}(t) = \widetilde{U}(t)^{-1} d\widetilde{U}(t), \tag{6.131}$$

with $\tilde{A}(0) = 0$. Choosing $W(0) = 0$, the difference

$$W(\tilde{A}(1)) - W(\tilde{A}(0)) = W(\widetilde{U}_1^{-1} d\widetilde{U}_1) \tag{6.132}$$

[11] This pathology is similar to the one encountered in Section 4.3 when the quantization condition is not satisfied.

is given by the integral of the $SU(3)$ (infinitesimal) gauge anomaly along the path, i.e. by the WZ functional, and since this WZ functional is evaluated for $\tilde{A} = 0$, it reduces just to the WZW term[12]

$$S_{WZW}(\widetilde{U}_1) = \frac{i}{240\pi^2}\int_{B^5} \mathrm{tr}(\widetilde{U}_1^{-1} d\widetilde{U}_1)^5 = i\pi. \tag{6.133}$$

Finally, since the endpoints of the path lie in the subgroup $SU(2)$ we arrive again at (6.129).

The lesson we learn from this theory is that the local gauge anomaly does not exhaust the information about the anomalous behavior of a gauge theory, and the global structure of the gauge group has to be taken into account to discuss the consistency at the quantum level.

6.8 Some applications

6.8.1 Anomaly cancellation

The presence of a gauge anomaly, whether local or global, renders a theory inconsistent. In particular, in perturbation theory maintaining local gauge invariance is necessary for renormalizability [GJa72, BIM72]. Therefore, the most important role of gauge anomalies is as a criterion to select viable theories. If we make the Lie algebra indices explicit, the four-dimensional gauge anomaly (6.61) reads

$$\left[D_\mu \langle J^\mu_{L/R}\rangle\right]^a = \pm\frac{i}{24\pi^2}\varepsilon^{\mu\nu\lambda\rho} d_{abc}\partial_\mu\left(A^b_\nu \partial_\lambda A^c_\rho + \frac{1}{2} f_{cde} A^b_\nu A^d_\lambda A^e_\rho\right), \tag{6.134}$$

where

$$d_{abc} = \frac{1}{2}\mathrm{tr}\, T_a\{T_b, T_c\}. \tag{6.135}$$

Note that the covariant form of the anomaly (6.75b) is also proportional to the same tensor, so in both cases the anomaly cancellation criterion is $d_{abc} = 0$.

In the canonical formalism, the manifestation of the anomaly is a Schwinger term in the commutators of gauge generators, see (6.79). For example, in four dimensions, the Schwinger term (6.118) is given more explicitly by

$$-\frac{i}{12\pi^2}\int d^3x\, \varepsilon^{ijk} d_{abc}\partial_i \epsilon_1^a \partial_j \epsilon_2^b A_k^c, \tag{6.136}$$

[12]This is the same integral that was used in Section 5.3 to evaluate the spin of the $N = 3$ skyrmion. We observe that the integral over S^5 would give 2π, and we obtain half that because the integral is over one hemisphere only.

showing that the vanishing of d_{abc} is necessary for the closure of the Gauss law algebra. Not unexpectedly, the condition for the absence of anomalies is the same in the covariant and in the canonical approach.

The presence of anomalies thus depends on the tensor d_{abc}. There are some groups all whose representations have $d_{abc} = 0$. This is the case of $SU(2)$ (but no other special unitary group), all orthogonal groups except $SO(6) \approx SU(4)$ and all symplectic groups. These are called *safe* groups. For the other groups, d_{abc} may or may not be zero, depending on the representation. If in a given representation the generators T_a are antihermitian, the generators in the complex conjugate representation are $T_a^* = -T_a^T$. If the representation is real or pseudoreal, it is equivalent to its complex conjugate

$$T_a = -UT_a^T U^{-1} \tag{6.137}$$

for some unitary matrix U. But then, it follows that the tensor d_{abc} is automatically zero:

$$Tr[T_a\{T_b, T_c\}] = -Tr[T_a^T\{T_b^T, T_c^T\}] = -Tr[T_a\{T_b, T_c\}]. \tag{6.138}$$

Thus, anomalies can only arise when a group has complex representations. The only such groups are $SU(N)$ with $N > 2$, $SO(6) \approx SU(4)$ and E_6.

So far we have considered the case of a gauge field coupled to only one chiral component of the fermions. If it couples to both, but they are in different representations or have different charges, the anomaly cancellation criterion becomes

$$d_{abc}(L) - d_{abc}(R) = 0, \tag{6.139}$$

where $d_{abc}(L/R)$ are the d-tensors of the representations carried by the left- and right-handed fermions. This is the case of the Standard Model. We leave it to Exercise 6.6 to check the anomaly cancellation in this important case.

6.8.2 Anomaly matching

Another important aspect of anomalies in general, and gauge anomalies in particular, is that their presence or absence in a certain theory should be reflected in any effective theory that approximates it at low energy. This is known as *anomaly matching* [tHo79] and it implies, among other things, that the anomalies must be the same above and below a phase transition. It gives useful restrictions for model building.

Here we mention only one particular case, that contains some of the essential features of the Standard Model. Suppose we have a fermion that is

coupled chirally to a gauge field and also has Yukawa couplings to a Higgs field:

$$\mathcal{L}_F = -\bar{\psi}_L \gamma^\mu D_\mu \psi_L - \bar{\psi}_R \gamma^\mu \partial_\mu \psi_R - g\bar{\psi}_L \Phi \psi_R - g\bar{\psi}_R \Phi^\dagger \psi_L. \tag{6.140}$$

We do not need to write the gauge and scalar parts of the action. When the theory is in the symmetric phase, both the Higgs and the fermions are massless, and the gauge current has an anomaly given by (6.61). Now consider what happens in the Higgs phase: the Higgs field acquires a VEV $\langle\Phi\rangle = \overline{\Phi}$ with $|\overline{\Phi}| = v$, and the Yukawa coupling gives rise to a fermion mass term with $m_F = gv$. When one looks at the theory at energy scales below this mass, one would normally assume that, due to the Appelquist–Carazzone decoupling theorem [ApC74], all effects due to the fermion vanish. This could have very unpleasant effects. For example, suppose the full theory contains other fermions in such representations that the total gauge anomaly vanishes. If such fermions are massless, or have a mass that is much smaller than m_F, then removing the massive fermion would ruin the anomaly cancellation and the low energy theory would seem to be pathological.

In fact, in this case the decoupling theorem does not work in the usual way. Let us write the Higgs field as $\Phi = \rho\widehat{\Phi}U$, where $\widehat{\Phi}$ is the unit vector pointing in the direction of the VEV $\overline{\Phi}$, ρ is the massive radial mode, and the field U describes the massless Goldstone bosons. We assume that we are looking at the theory at an energy much below the mass of the ρ field, so that only the Goldstone degrees of freedom are active. If we take the limit $g \to \infty$, the fermion disappears, but it leaves behind a WZ action $S_{WZ}(A, U)$ as given in (6.102) [DHF84a, DHF84b]. The gauge variation of this functional is equal to the anomaly of the original theory, so if the fermion was involved in an anomaly cancellation mechanism, this role of the fermion is taken over by the Goldstone bosons and the anomaly cancels also in the low energy effective theory. Gauge invariance is maintained at the level of the effective action. One can verify that this cancellation works also in the canonical formalism: the Gauss law algebra in the presence of the WZ term has exactly the same Schwinger term as the Gauss law algebra of the original anomalous fermionic theory, so if the total Schwinger term is zero in the fundamental theory, it is also zero in the low energy effective theory where one fermion has decoupled [PeR88]. The difference with the original fermionic theory is that whereas the latter may be renormalizable (if the total anomaly cancels, and depending on the form of the Lagrangian), the low energy effective theory is a nonlinear sigma model and hence definitely non-renormalizable. This is not a problem, though, because it was meant from the outset to be only a description of the low energy world.

6.8.3 Skyrmions as baryons, the final word

In Section 5.3 we saw that the skyrmions of the $N = 3$ chiral model can be quantized either as bosons or as fermions, depending whether the coefficient of the WZW term is an even or odd multiple of 2π. We now have all the ingredients that are needed to fix this ambiguity,

We have seen that the decay of the neutral pion can be represented by a term of the form (6.39) in the low energy effective field theory of QCD. That expression came from a one-loop calculation in QCD, but we can try to derive it directly by coupling the $N = 3$ chiral model to the electromagnetic field. The chiral action, including the WZW term, is invariant under the vector transformations $U \to g^{-1}Ug$. Consider in particular the case $g = \exp\{i\alpha Q\}$, or infinitesimally

$$\delta U = -i\alpha[Q, U], \tag{6.141}$$

where Q is the electric charge matrix of the three lightest quarks

$$Q = \begin{pmatrix} 2/3 & 0 & 0 \\ 0 & -1/3 & 0 \\ 0 & 0 & -1/3 \end{pmatrix}.$$

The terms discussed in Section 2.4 are invariant under this global $U(1)$ transformation, but not under its local counterpart with parameter $\alpha(x)$. They can be made invariant by the usual procedure of replacing partial derivatives with the covariant derivatives

$$D_\mu U = \partial_\mu U + iA_\mu[Q, U]. \tag{6.142}$$

This procedure does not work for the WZW term (there is no extension of the gauge potential in the interior of the 5-dimensional space). Nevertheless, the covariantization of the WZW action can be achieved by an ad hoc procedure [Wit83a]. One can check that the following action is a $U(1)$-gauge invariant generalization of S_{WZW}:

$$\begin{aligned} S(A, U) = S_{WZW}(U) &- \frac{ne}{48\pi^2}\int d^4x\, \varepsilon^{\mu\nu\rho\sigma} A_\mu \mathrm{tr} Q(R_\nu R_\rho R_\sigma + L_\nu L_\rho L_\sigma) \\ &+ \frac{ine^2}{24\pi^2}\int d^4x\, \varepsilon^{\mu\nu\rho\sigma}\partial_\mu A_\nu A_\rho \mathrm{tr}\Big[Q^2(R_\sigma + L_\sigma) \\ &+ \frac{1}{2}(QUQU^{-1}L_\sigma + QU^{-1}QUR_\sigma)\Big]. \end{aligned} \tag{6.143}$$

Here $n = c/2\pi$ is the integer appearing in the coefficient of the WZW action, see (4.24). Since

$$U = e^{i\pi^a \lambda_a / F_\pi},$$

where λ_a are the (hermitian) Gell–Mann matrices, we have

$$R_\mu = \frac{i}{F_\pi}\partial_\mu \pi^a \lambda_a + \dots, \quad L_\mu = \frac{i}{F_\pi}\partial_\mu \pi^a \lambda_a + \dots.$$

Then, the terms quadratic in A_μ give

$$\frac{ine^2}{24\pi^2}\int d^4x\, \varepsilon^{\mu\nu\rho\sigma}\partial_\mu A_\nu A_\rho \mathrm{tr}\left[3Q^2 \frac{i}{F_\pi}\partial_\sigma \pi^a \lambda_a\right]$$

The term that matters for the neutral pion decay has $a = 3$, and since $\mathrm{tr}Q^2\lambda_3 = 1/3$, we find

$$\frac{ne^2}{96\pi^2 F_\pi}\int d^4x\, \varepsilon^{\mu\nu\rho\sigma}\pi^0 F_{\mu\nu}F_{\rho\sigma}.$$

Comparing with the QCD result (6.39) we find that

$$n = N_c. \tag{6.144}$$

Since in the real world there are three colors, we conclude that the skyrmion *must* be quantized as a fermion.

6.9 Exercises

Exercise 6.1: The ABJ anomaly in $d = 4$

Complete the calculation of the ABJ anomaly in four dimensions. Write the Fourier transform of $S_\nu A_\rho(y) S_\lambda$ in (6.20) and perform the integrals leading to (6.21).

Exercise 6.2: The WZ consistency conditions

Check that the consistent anomalies (6.61) satisfy the WZ consistency condition (6.63), while the covariant anomalies (6.75b) do not.

Exercise 6.3: Anomalies in commutators

Evaluate the commutator of the normal ordered charge density for a multiplet of free chiral fermions in two dimensions.

Exercise 6.4: The two-dimensional WZ functional

Check that the two-dimensional WZ functional (6.101) satisfies the condition (6.97).

Exercise 6.5: $U(1)$ gauged WZW action

Check that the action (6.143) is $U(1)$-invariant.

Exercise 6.6: Anomalies in the Standard Model

Verify that the gauge anomalies cancel for the group $SU(2) \times U(1)$ when the fermions have the quantum numbers of the Standard Model.

Exercise 6.7: The Schwinger model

The Schwinger model is 2-dimensional QED with vectorial fermion coupling. The fermionic action (6.1) is

$$\begin{aligned} S_F(\psi, \bar{\psi}) &= -\int d^2x\, \bar{\psi}\gamma^\mu D_\mu \psi \\ &= \int d^2x\, \left[-\bar{\psi}\gamma^\mu \partial_\mu \psi + ieA_\mu j_V^\mu\right]. \end{aligned} \tag{6.145}$$

The gamma matrices are given by (A.10).

1. Compute the expectation value of the vector current and show that it is conserved.
2. use the identity
$$\gamma^A \gamma^\mu = \varepsilon^\mu{}_\nu \gamma^\nu \tag{6.146}$$
to reobtain the anomaly of the axial current (6.19).

3. Integrate the equation

$$\langle J_V^\mu \rangle = \frac{1}{e} \frac{\delta W(A)}{\delta A_\mu}. \tag{6.147}$$

to give a closed form expression for W.

4. adding now the Maxwell action to the fermionic effective action, derive the quantum equation of motion and show that it describes a massive state with mass $m^2 = e^2/\pi$.

5. using only (6.146), from the result of 3. derive the effective action for the chiral Schwinger model, where j_V^μ is replaced by j_L^μ in the action. Compute its gauge variation and compare with the consistent anomaly (6.64a).

Appendix A

Notations and conventions

A.1 Units

In most of the text natural units are used, where c and $\hbar$ are taken as the units of velocity and angular momentum (or action). In some sections dealing with finite dimensional quantum mechanics, the Heaviside–Lorentz system of units is used, that is the rationalized version of Gauss units. In this system the electric field generated by a charge Q is

$$E_i = \frac{1}{4\pi}\frac{Q}{r^2}\hat{x}_i, \tag{A.1}$$

and the charge contained in a sphere is

$$Q = \int_{S^2} d\sigma_i E_i \tag{A.2}$$

without factors 4π. Maxwell's equations are

$$\partial_\mu F^{\mu\nu} = J^\nu \tag{A.3}$$

and the electromagnetic energy density is

$$\mathcal{E} = \frac{1}{2}(\vec{E}^2 + \vec{H}^2), \tag{A.4}$$

both without factors 4π. In these units charges and fluxes have dimensions $M^{1/2}L^{3/2}T^{-1}$, the electromagnetic field $F_{\mu\nu}$ has dimensions $M^{1/2}L^{-1/2}T^{-1}$ and the potential A_μ has dimensions $M^{1/2}L^{1/2}T^{-1}$. The fine structure constant is

$$\alpha = \frac{e^2}{4\pi\hbar c}. \tag{A.5}$$

A.2 Tensors and spinors

The dimension of space is d and the dimension of spacetime is $n = d+1$ in most of the book, except for Chapter 6, where $d + 1 = 2n$. Latin indices i, j, k ... are spatial indices, greek indices μ, ν, ρ ... are spacetime indices. Internal indices may be taken from various parts of the latin or greek alphabet,

The Minkowski metric is

$$\eta_{\mu\nu} = \begin{pmatrix} -1 & & & \\ & 1 & & \\ & & 1 & \\ & & & \ldots \end{pmatrix} \tag{A.6}$$

One advantage of this signature is that the Wick rotation (the replacement $t \longmapsto \tau \equiv -it$) directly leads to a Euclidean metric. With this signature the Lagrangian for a scalar is

$$-\frac{1}{2}\partial_\mu\phi\partial^\mu\phi - \frac{1}{2}m^2\phi^2 \tag{A.7}$$

and the Lagrangian for a free spinor, following the conventions of [Wei95], is

$$-\bar{\psi}(\gamma^\mu\partial_\mu + m)\psi \tag{A.8}$$

where

$$\bar{\psi} = i\psi^\dagger\gamma^0.$$

The gamma matrices are defined by

$$\{\gamma^\mu, \gamma^\nu\} = 2\eta^{\mu\nu}. \tag{A.9}$$

In order to guarantee that the action be real we further demand that γ^0 be antihermitian and γ^i be hermitian. The chirality operator will be called γ^A. Specific choices in two dimensions are

$$\gamma^0 = i\sigma_2, \quad \gamma^1 = \sigma_1, \quad \gamma^A = \gamma^0\gamma^1 = \sigma_3 \tag{A.10}$$

and in four dimensions

$$\gamma^0 = \begin{pmatrix} 0 & -i\mathbb{I} \\ -i\mathbb{I} & 0 \end{pmatrix}, \quad \gamma^k = \begin{pmatrix} 0 & -i\sigma_k \\ i\sigma_k & 0 \end{pmatrix}, \quad \gamma^A = -i\gamma^0\gamma^1\gamma^2\gamma^3 = \begin{pmatrix} \mathbb{I} & 0 \\ 0 & -\mathbb{I} \end{pmatrix}, \tag{A.11}$$

where σ_k are the Pauli matrices (B.21).

The Euclidean action is defined as $S_E = -iS|_{t\to -it}$. For bosonic fields this amounts to replacing the Minkowski metric $\eta_{\mu\nu}$ with the Euclidean metric $\delta_{\mu\nu}$ and changing the sign of the action.

The totally antisymmetric Levi–Civita symbol is defined by $\varepsilon^{0123} = 1$. It is used to define the dual of a two-form by

$$({}^*\omega)_{\mu\nu} = \frac{1}{2}\eta_{\mu\alpha}\eta_{\nu\beta}\varepsilon^{\alpha\beta\gamma\delta}\omega_{\gamma\delta}. \tag{A.12}$$

In Minkowski space ${}^{**}\omega = -\omega$, whereas in Euclidean space one has ${}^{**}\omega = \omega$.

A.3 List of symbols

j^μ	Noether current
J^μ	current coupled to gauge fields
J^μ_T	topological current
$\mathcal{Q}$	configuration space
$\mathcal{A}$	abelian gauge field on $\mathcal{Q}$ (Chapter 3) or anomaly (Chapter 6)
$\mathcal{C}$	space of connections
$\mathcal{G}$	gauge group
ϕ	linear scalar field (possibly with constraints)
φ	coordinates
$\Gamma(M,N)$	space of maps from M to N
n	dimension of spacetime (Chapters 1–5)
d	dimension of space
E_S	static energy
F_π	pion decay constant
W	winding number

Appendix B

Lie groups and Lie algebras

Let G be a Lie group, $\mathfrak{g}$ its Lie algebra and $\{e_a\}$ a basis in $\mathfrak{g}$. Let V be a vector space carrying a representation ρ of G. Then the basis elements have explicit representations as matrices $T_a = \rho(e_a)$ acting on V. They are assumed to be antihermitian, satisfy the commutation relations

$$[T_a, T_b] = f_{ab}{}^c T_c. \tag{B.1}$$

and are normalized so that

$$\mathrm{tr} T_a T_b = -\frac{1}{2}\delta_{ab}. \tag{B.2}$$

The structure constants are real for a real Lie algebra, and in general antisymmetric only in the first two indices.

The adjoint representation Ad is a representation of G on $\mathfrak{g}$. For a matrix group it can be defined simply by

$$Ad(g)Y = gYg^{-1}, \tag{B.3}$$

for any $g \in G$ and $X \in \mathfrak{g}$. The corresponding representation of the Lie algebra is called ad:

$$ad(X)Y = [X, Y]. \tag{B.4}$$

More explicitly, the matrix representation of $ad(e_a)$ is given by the structure constants:

$$ad(e_a)_{bc} = -f_{abc}. \tag{B.5}$$

and the algebra

$$[ad(e_a), ad(e_b)] = f_{ab}{}^c \, ad(e_c) \tag{B.6}$$

follows from the Jacobi identity. The Killing form is the Ad-invariant quadratic form

$$B(X,Y) = \operatorname{tr} ad(X)ad(Y), \tag{B.7}$$

or explicitly in components

$$B_{ab} = \operatorname{tr} ad(e_a)ad(e_b) = f_{ac}{}^d f_{bd}{}^c. \tag{B.8}$$

For compact simple Lie groups the Killing form is negative definite, so one can use $-B$ as a positive definite inner product in $\mathfrak{g}$. One can then lower the third index of the structure constants and the resulting tensor f_{abc} is antisymmetric in all three indices.

When a field ϕ carries a representation ρ of a group G, the finite transformation is conventionally

$$\phi' = \rho(g^{-1})\phi \tag{B.9}$$

and for an infinitesimal transformation $g = 1 + \epsilon$, the variation of the field is

$$\delta_\epsilon \phi = -\rho(\epsilon)\phi. \tag{B.10}$$

When there are no ambiguities, the symbol ρ indicating the representation is omitted.

Let us now focus on the orthogonal groups. The generators of $\mathfrak{so}(n)$ are the vectorfields in $\mathbb{R}^n$

$$M_{ab} = x_a\partial_b - x_b\partial_a. \tag{B.11}$$

They are orthogonal to the radial vector $x_c\partial_c$ and therefore are tangent to the sphere $S^{n-1} \subset \mathbb{R}^n$. They satisfy the algebra (Lie brackets)

$$[M_{ab}, M_{cd}] = -\delta_{ac}M_{bd} + \delta_{ad}M_{bc} + \delta_{bc}M_{ad} - \delta_{bd}M_{ac}. \tag{B.12}$$

The vectorfield M_{ab} generates counterclockwise rotations in the plane (a,b).

For $\mathfrak{so}(3)$ we define

$$K_a = \frac{1}{2}\epsilon_{abc}M_{bc} \tag{B.13}$$

or

$$M_{ab} = \epsilon_{abc}K_c, \tag{B.14}$$

so K_a generates counterclockwise rotations around the axis a. In spherical coordinates these vectorfields read

$$K_1 = -\sin\Phi\frac{\partial}{\partial\Theta} - \cot\Theta\cos\Phi\frac{\partial}{\partial\Phi}, \tag{B.15a}$$

$$K_2 = \cos\Phi\frac{\partial}{\partial\Theta} - \cot\Theta\sin\Phi\frac{\partial}{\partial\Phi}, \tag{B.15b}$$

$$K_3 = \frac{\partial}{\partial\Phi}. \tag{B.15c}$$

Their algebra is

$$[K_a, K_b] = -\epsilon_{abc} K_c. \tag{B.16}$$

We take this as the basic algebra of $\mathfrak{so}(3)$. From (B.5), the corresponding adjoint representation matrices are $(t_a)_{bc} = ad(K_a)_{bc} = \epsilon_{abc}$. Explicitly

$$t_1 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & -1 & 0 \end{pmatrix}, \quad t_2 = \begin{pmatrix} 0 & 0 & -1 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}, \quad t_3 = \begin{pmatrix} 0 & 1 & 0 \\ -1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}. \tag{B.17}$$

The fundamental representation of $SO(3)$ coincides with the adjoint.

For the group $SO(4)$ we define

$$K_a^{\pm} = \frac{1}{2}(K_a \pm M_{a4}) \tag{B.18}$$

and it is easy to check that

$$[K_a^+, K_b^+] = -\epsilon_{abc} K_c^+ \tag{B.19a}$$
$$[K_a^-, K_b^-] = -\epsilon_{abc} K_c^- \tag{B.19b}$$
$$[K_a^+, K_b^-] = 0. \tag{B.19c}$$

proving that $\mathfrak{so}(4) = \mathfrak{so}(3) \oplus \mathfrak{so}(3)$.

The group $SU(2)$ is a double covering of $SO(3)$, so they have the same Lie algebra. In the fundamental representation of $SU(2)$ the generators K_a are represented by the matrices

$$\tau_a = \frac{i}{2}\sigma_a, \tag{B.20}$$

where σ_a are the Pauli matrices

$$\sigma_1 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad \sigma_2 = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}, \quad \sigma_3 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}. \tag{B.21}$$

They satisfy an algebra isomorphic to (B.16)

$$[\tau_a, \tau_b] = -\epsilon_{abc} \tau_c \tag{B.22}$$

and the trace normalization (B.2).

The generators of $SU(3)$ in the fundamental representation are $\frac{i}{2}\lambda_a$, where λ_a are the Gell–Mann matrices

$$\begin{aligned}
\lambda_1 &= \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \quad \lambda_2 = \begin{pmatrix} 0 & -i & 0 \\ i & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \quad \lambda_3 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \\
\lambda_4 &= \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}, \quad \lambda_5 = \begin{pmatrix} 0 & 0 & -i \\ 0 & 0 & 0 \\ i & 0 & 0 \end{pmatrix}, \\
\lambda_6 &= \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}, \quad \lambda_7 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & -i \\ 0 & i & 0 \end{pmatrix}, \quad \lambda_8 = \frac{1}{\sqrt{3}} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -2 \end{pmatrix}.
\end{aligned} \tag{B.23}$$

Appendix C

Bundles

We give here just the minimal definitions that are used in the main text. A *fiber bundle* is a manifold E that looks locally like a product manifold. More precisely, there is a manifold M, called the *base space*, and a surjective map $p : E \to M$, called the *projection*, such that for every point $x \in M$ there is a neighborhood U such that the inverse image $p^{-1}(U)$ is diffeomorphic to $U \times F$, where F is called the *typical fiber*. The inverse image $p^{-1}(x)$ is called the *fiber over* x and is diffeomorphic to F. Thus, E looks locally like the product $M \times F$.

There are various types of bundles, depending on the structures that may be present in the fibers. One very useful class of bundles are the vectorbundles, whose fibers are vectorspaces. Typical examples are the tangent and cotangent bundle of M.

Another important class of bundles are those whose typical fiber is a group. Of these, the most important ones are the principal bundles, that are defined as follows. A *principal bundle* is a space P on which a group H acts freely (i.e. without fixed points) from the right. The base space is the quotient (also called the space of orbits) $M = P/H$ and the projection $p : P \to M$ maps each $y \in P$ to its equivalence class $[y] = y \,\mathrm{mod} H \in M$. Since the action is free, each orbit is diffeomorphic to H. In this case the orbits of H are the fibers of the bundle. The standard example of a principal bundle, and the one that we shall be mostly interested in, is a Lie group G, with $\iota : H \to G$ a Lie subgroup and $M = G/H$ the space of right cosets. A special case is the Hopf bundle with $P = SU(2)$, $F = U(1)$, $M = S^2$. Its projection $h : S^3 \to S^2$ is called the Hopf map and is dicussed in more details in Appendix D.4.

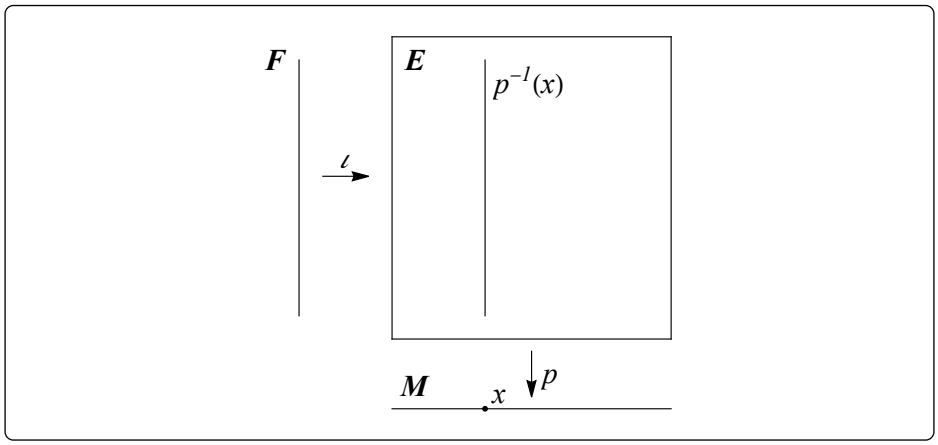


Figure 29. A fiber bundle.

A *section* of a bundle is a map $s : M \to E$ such that $p \circ s$ is the identity of M, i.e. s maps each point of M into the fiber over that point. Such maps do not always exist globally. It can be shown that a principal bundle is trivial, i.e. P is diffeomorphic to $M \times F$, if and only if it has a global section.

Appendix D

Geometry of $SU(2)$

D.1 Euler angles and covering of $SO(3)$

An element $U \in SU(2)$ is a complex 2×2 matrix

$$U = \sum_{k=1}^{3} iu_k \sigma_k + u_4 \mathbb{I} = \begin{bmatrix} u_4 + iu_3 & u_2 + iu_1 \\ -u_2 + iu_1 & u_4 - iu_3 \end{bmatrix}, \tag{D.1}$$

where σ_i are the Pauli matrices and

$$1 = \det U = u_1^2 + u_2^2 + u_3^2 + u_4^2. \tag{D.2}$$

So $SU(2)$ is the unit sphere in $\mathbb{R}^4$. The element with coordinates $(0, 0, 0, 1)$ corresponds to the unit matrix $\mathbb{I}$ and can be thought of a the north pole of the sphere. The south pole is the element with coordinates $(0, 0, 0, -1)$, corresponding to the matrix $-\mathbb{I}$.

The Lie algebra of $SU(2)$ consists of the anti-hermitian matrices. It can be identified geometrically with the plane $x_4 = 1$ in $\mathbb{R}^4$. We take as basis in the Lie algebra the matrices (B.20). The isomorphism of the Lie algebras $\mathfrak{su}(2) \to \mathfrak{so}(3)$ maps τ_i to t_i, as given in (B.17). This also defines an isomorphism of a neighbourhood of the identity in $SU(2)$ to a neighbourhood of the identity in $SO(3)$, and we can use this fact to introduce the Euler angles as coordinates on $SU(2)$. Define:

$$\begin{aligned} U(\Theta, \Phi, \Psi) &= \exp(-\Phi\tau_3) \exp(-\Theta\tau_2) \exp(-\Psi\tau_3) \\ &= \begin{bmatrix} \cos\frac{\Theta}{2} \exp\left[-\frac{i}{2}(\Phi + \Psi)\right] & -\sin\frac{\Theta}{2} \exp\left[-\frac{i}{2}(\Phi - \Psi)\right] \\ \sin\frac{\Theta}{2} \exp\left[\frac{i}{2}(\Phi - \Psi)\right] & \cos\frac{\Theta}{2} \exp\left[\frac{i}{2}(\Phi + \Psi)\right] \end{bmatrix} \end{aligned} \tag{D.3}$$

Every matrix $U \in SU(2)$ can be written in this way, provided that $0 < \Phi \leq 2\pi$, $0 \leq \Theta \leq \pi$, $0 < \Psi \leq 4\pi$. There is a homomorphism from $SU(2)$ to $SO(3)$ that maps $U(\Theta, \Phi, \Psi)$ to

$$R(\Theta,\Phi,\Psi) = \exp(-\Phi t_3)\exp(-\Theta t_2)\exp(-\Psi t_3)$$
$$= \begin{bmatrix} \cos\Theta\cos\Phi\cos\Psi - \sin\Phi\sin\Psi & -\cos\Theta\cos\Phi\sin\Psi - \sin\Phi\cos\Psi & \sin\Theta\cos\Phi \\ \cos\Theta\sin\Phi\cos\Psi + \cos\Phi\sin\Psi & -\cos\Theta\sin\Phi\sin\Psi + \cos\Phi\cos\Psi & \sin\Theta\sin\Phi \\ -\sin\Theta\cos\Psi & \sin\Theta\sin\Psi & \cos\Theta \end{bmatrix}.$$

It is a double covering because $R(\Theta, \Phi, \Psi + 2\pi) = R(\Theta, \Phi, \Psi)$ but

$$U(\Theta, \Phi, \Psi + 2\pi) = -U(\Theta, \Phi, \Psi).$$

As a consequence the range of Ψ as a coordinate in $SU(2)$ is twice the range of Ψ as a coordinate in $SO(3)$. Note also that $R(\Theta, \Phi, \Psi)$ is the adjoint representation of $U(\Theta, \Phi, \Psi)$.

D.2 Invariant forms and vectorfields

As on any Lie group, one can define the Lie-algebra-valued Maurer–Cartan forms

$$L = U^{-1}dU, \qquad R = dUU^{-1}. \tag{D.4}$$

The form L is invariant under the action of left multiplication $U \longmapsto gU$ and R is invariant under the right multiplication $U \longmapsto Ug$. They can be decomposed on the basis of the Lie algebra

$$L = L^a T_a, \qquad R = R^a T_a, \tag{D.5}$$

where L^a, R^a, with $a = 1, 2, 3$ are ordinary differental forms on $SU(2)$. Given any coordinate system $\{y^\alpha\}$, they can be decomposed on a natural basis

$$L^a = L^a_\alpha dy^\alpha, \qquad R^a = R^a_\alpha dy^\alpha. \tag{D.6}$$

The components of the Maurer–Cartan forms in Euler coordinates can be calculated directly by inserting (D.3) in (D.4) and decomposing

$$U^{-1}dU = L^a_\alpha \tau_a dy^\alpha, \qquad dUU^{-1} = R^a_\alpha \tau_a dy^\alpha$$

In this way one finds

$$L^1 = -\sin\Psi d\Theta + \sin\Theta\cos\Psi d\Phi, \tag{D.7a}$$
$$L^2 = -\cos\Psi d\Theta - \sin\Theta\sin\Psi d\Phi, \tag{D.7b}$$
$$L^3 = -d\Psi - \cos\Theta d\Phi, \tag{D.7c}$$
$$R^1 = \sin\Phi d\Theta - \sin\Theta\cos\Phi d\Psi, \tag{D.7d}$$
$$R^2 = -\cos\Phi d\Theta - \sin\Theta\sin\Phi d\Psi, \tag{D.7e}$$
$$R^3 = -d\Phi - \cos\Theta d\Psi. \tag{D.7f}$$

One can then explicitly verify the Maurer–Cartan equations:

$$dL^a + \frac{1}{2} f_{bc}{}^a L^b \wedge L^c = 0\,; \qquad dR^a - \frac{1}{2} f_{bc}{}^a R^b \wedge R^c = 0, \tag{D.8}$$

with $f_{abc} = -\epsilon_{abc}$.

The left-invariant forms L^a are linearly independent and form a global field of bases for one-forms. Then, there is a dual field of bases for vectors L_a:

$$L_a = L_a^\alpha \partial_\alpha, \tag{D.9}$$

where the matrix L_a^α is the inverse of the matrix L_α^a.

$$L_a^\alpha L_\alpha^b = \delta_a^b, \quad L_a^\alpha L_\beta^a = \delta_\beta^\alpha.$$

The vectors L_a are left-invariant. Similarly one defines a basis of right-invariant vectorfields

$$R_a = R_a^\alpha \partial_\alpha. \tag{D.10}$$

In Euler coordinates

$$L_1 = -\sin\Psi\frac{\partial}{\partial\Theta} + \frac{1}{\sin\Theta}\cos\Psi\frac{\partial}{\partial\Phi} - \cot\Theta\cos\Psi\frac{\partial}{\partial\Psi}, \tag{D.11a}$$
$$L_2 = -\cos\Psi\frac{\partial}{\partial\Theta} - \frac{1}{\sin\Theta}\sin\Psi\frac{\partial}{\partial\Phi} + \cot\Theta\sin\Psi\frac{\partial}{\partial\Psi}, \tag{D.11b}$$
$$L_3 = -\frac{\partial}{\partial\Psi}, \tag{D.11c}$$
$$R_1 = \sin\Phi\frac{\partial}{\partial\Theta} + \cot\Theta\cos\Phi\frac{\partial}{\partial\Phi} - \frac{1}{\sin\Theta}\cos\Phi\frac{\partial}{\partial\Psi}, \tag{D.11d}$$
$$R_2 = -\cos\Phi\frac{\partial}{\partial\Theta} + \cot\Theta\sin\Phi\frac{\partial}{\partial\Phi} - \frac{1}{\sin\Theta}\sin\Phi\frac{\partial}{\partial\Psi}, \tag{D.11e}$$
$$R_3 = -\frac{\partial}{\partial\Phi}. \tag{D.11f}$$

These vectorfields are the infinitesimal generators of the action of the group on itself. More precisely, the vectorfields L_a generate the right multiplication and R_a generate the left mutiplication.

A direct calculation gives the Lie brackets

$$[L_a, L_b] = -\epsilon_{abc} L_c \tag{D.12a}$$
$$[R_a, R_b] = \epsilon_{abc} R_c \tag{D.12b}$$
$$[R_a, L_b] = 0. \tag{D.12c}$$

Since $[X, Y] = \mathcal{L}_X Y$ (the Lie derivative) the last bracket expresses the fact that the vectorfields L_a are left-invariant and the R_a are right-invariant.

The adjoint representation matrices are given by:

$$Ad(g)^a{}_b = R(g)^a_\alpha L(g)^\alpha_b.$$

One can indeed check that

$$Ad(U(\Theta, \Phi, \Psi)) = R(\Theta, \Phi, \Psi).$$

D.3 Invariant metric and volume form

Given an inner product γ_{ab} in the Lie algebra, we can construct a left- and a right- invariant metric on $SU(2)$

$$h^{(L)}_{\alpha\beta} = L^a_\alpha L^b_\beta \gamma_{ab}$$

and

$$h^{(R)}_{\alpha\beta} = R^a_\alpha R^b_\beta \gamma_{ab}.$$

However, if the inner product is Ad-invariant, both metrics agree and are bi-invariant. Thus for example the inner product

$$\gamma(v, w) = C \operatorname{tr}(v\, w)$$

is Ad-invariant, because

$$\operatorname{tr}(g^{-1} v g\, g^{-1} w g) = \operatorname{tr}(v\, w).$$

The components of this inner product are

$$\gamma_{ab} = C \operatorname{tr}(\tau_a \tau_b) = -\frac{C}{2} \delta_{ab}.$$

If we choose $C = -1/2$, $\gamma_{ab} = \frac{1}{4}\delta_{ab}$, in which case $\frac{1}{2}L_a$ and $\frac{1}{2}R_a$ are orthornormal bases ("triads") for the bi-invariant metric $h_{\alpha\beta}$. Then, in Euler coordinates, the bi-invariant metric is

$$\begin{aligned} ds^2 &= \frac{1}{4}[(L^1)^2 + (L^2)^2 + (L^3)^2] \\ &= \frac{1}{4}[d\Theta^2 + d\Phi^2 + d\Psi^2 + 2\cos\Theta\, d\Phi d\Psi] \\ &= h_{\alpha\beta}\, dy^\alpha dy^\beta. \end{aligned} \tag{D.13}$$

The corresponding volume element is

$$\begin{aligned} \omega &= \frac{1}{8} L^1 \wedge L^2 \wedge L^3 \\ &= \frac{1}{8}\sin\Theta\, d\Theta \wedge d\Phi \wedge d\Psi \\ &= \sqrt{\det h}\, dy^1 \wedge dy^2 \wedge dy^3. \end{aligned} \tag{D.14}$$

With this volume form, the volume of $SU(2)$ is $\int\omega = 2\pi^2$, which is the volume of the unit three-sphere. Indeed the invariant metric h is the metric induced from the Euclidean metric in $\mathbb{R}^4$ by the embedding (D.2). The curvature scalar of this metric is $R = 6$.

D.4 The Hopf map

The Hopf map is the projection of the Hopf bundle $S^3 \to S^2$ that factors the right action of $U(1)$ generated by the vectorfield L_3. The Euler angles are well-adapted to this projection, in the sense that the orbits of $U(1)$ have constant Θ and Φ, and Ψ is a coordinate in the orbits. Thus the Hopf map maps the point with coordinates (Θ, Φ, Ψ) to the point on the sphere with spherical coordinates (Θ, Φ).

The Hopf map can be presented in terms of the Cartesian coordinates of the spaces where the spheres are embedded. It is convenient to rename the coordinates of $\mathbb{R}^4$

$$u_1 \to x_3, \quad u_2 \to x_4, \quad u_3 \to x_2, \quad u_4 \to x_1,$$

in such a way that

$$U = \begin{bmatrix} z_1 & iz_2^* \\ iz_2 & z_1^* \end{bmatrix} \qquad \text{with} \qquad z_1 = x_1 + ix_2,\ \ z_2 = x_3 + ix_4. \tag{D.15}$$

These are the complex coordinates on the sphere, when viewed as $\mathbb{CP}^1$. Comparing with (D.3) we find

$$x_1 = \cos\frac{\Theta}{2}\cos\frac{\Phi+\Psi}{2}, \tag{D.16a}$$
$$x_2 = -\cos\frac{\Theta}{2}\sin\frac{\Phi+\Psi}{2}, \tag{D.16b}$$
$$x_3 = \sin\frac{\Theta}{2}\sin\frac{\Phi-\Psi}{2}, \tag{D.16c}$$
$$x_4 = -\sin\frac{\Theta}{2}\cos\frac{\Phi-\Psi}{2}, \tag{D.16d}$$

On the other hand the Cartesian coordinates of $\mathbb{R}^3$ in which S^2 is embedded are

$$y^1 = \sin\Theta\cos\Phi, \tag{D.17a}$$
$$y^2 = \sin\Theta\sin\Phi, \tag{D.17b}$$
$$y^3 = \cos\Theta. \tag{D.17c}$$

The Hopf map is then given by

$$y^1 = 2(x_2x_3 - x_1x_4), \tag{D.18a}$$
$$y^2 = 2(x_1x_3 + x_2x_4), \tag{D.18b}$$
$$y^3 = x_1^2 + x_2^2 - x_3^2 - x_4^2. \tag{D.18c}$$

There are two very useful sections of the Hopf bundle. The map s_- maps $S^2\backslash\{N\} \to S^3$

$$z_1(\Theta,\Phi) = \cos\frac{\Theta}{2}e^{-i\varphi}, \qquad z_2(\Theta,\Phi) = -i\sin\frac{\Theta}{2}. \tag{D.19}$$

This map is regular at the South pole, since $(z_1(\pi,\Phi), z_2(\pi,\Phi)) = (0,-i)$ independently of the direction one approaches it, but singular at the North pole, since $(z_1(0,\Phi), z_2(0,\Phi)) = (e^{-i\Phi}, 0)$. In Euler coordinates, it is given by $\Psi = \Phi$.

The map s_+ maps $S^2\backslash\{S\} \to S^3$

$$z_1(\Theta,\Phi) = \cos\frac{\Theta}{2}, \qquad z_2(\Theta,\Phi) = -i\sin\frac{\Theta}{2}e^{i\Phi}. \tag{D.20}$$

This map is regular at the North pole, since $(z_1(0,\varphi), z_2(0,\varphi)) = (1,0)$ independently of the direction one approaches it, but singular at the South pole, since $(z_1(\pi,\Phi), z_2(\pi,\Phi)) = (0,-ie^{i\Phi})$. In Euler coordinates, it is given by $\Psi = -\Phi$.

The right-invariant vectorfields project on vectorfields on the sphere. In fact, the projection of $-R_a$ is precisely the vectorfield K_a in (B.15).

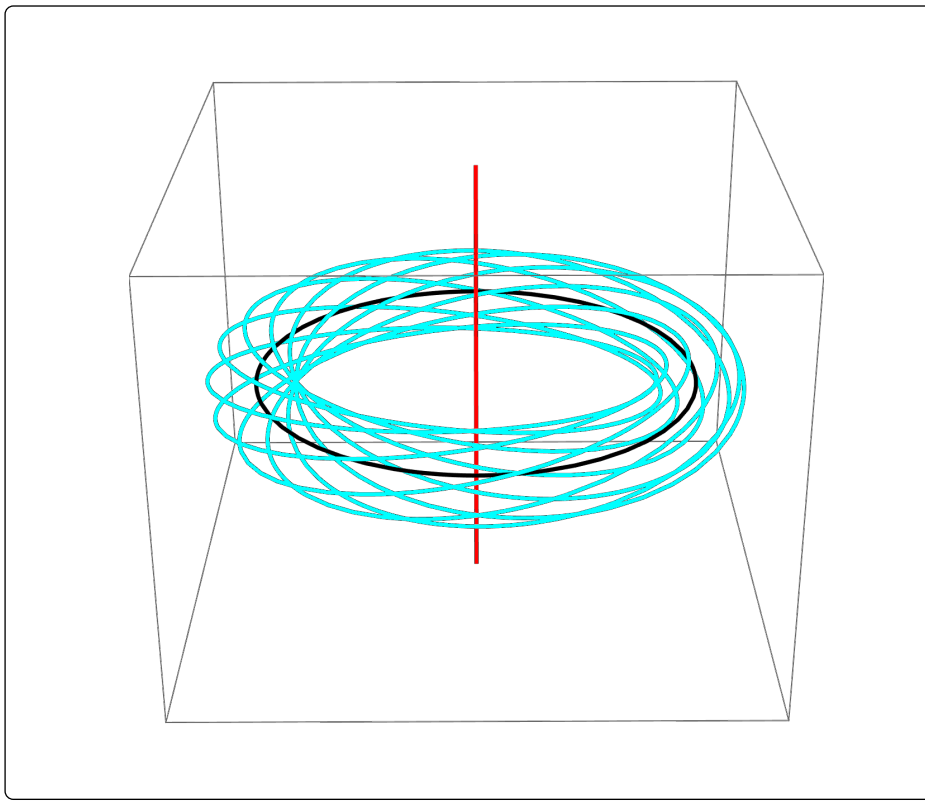


Figure 30. Twelve fibers of the Hopf bundle. The ambient space is S^3, projected stereographically to $\mathbb{R}^3$. The vertical red line is the fiber over the south pole $\theta = \pi$ (it extends to infinity and all the points at infinity are identified). The black circle in the horizontal plane is the orbit over the north pole $\theta = 0$. The ten cyan-colored curves are orbits over ten points on the parallel $\theta = \pi/8$, with different longitudes ϕ. All the fibers over a parallel span the surface of a two-dimensional torus, and the tori become thicker with increasing θ. Every fiber is linked once with every other fiber.

Appendix E

Homotopy

E.1 Basic definitions

Let M, N be finite dimensional manifolds. We choose a point $x_0 \in M$ and a point $y_0 \in N$; they are called the *basepoints* of M and N. We denote $\Gamma(M,N)$ the space of all smooth functions $f : M \to N$. (By smooth we mean continuous and r-times differentiable, with $0 \leq r \leq \infty$). We denote $\Gamma_*(M,N)$ the subspace of *basepoint preserving* maps, i.e. maps f such that $f(x_0) = y_0$.

We say that two maps $f, g \in \Gamma(M,N)$ are *homotopic*, and write $f \simeq g$, if there exists a continuous map $F : M \times I \to N$ such that $F(x,0) = f(x)$, $F(x,1) = g(x)$. Intuitively, F gives a one parameter family of maps, depending continuously on t, that interpolates between f and g. Sometimes it is convenient to put into evidence the dependence on the parameter, and write $f_t = F(.,t)$; then $f_0 = f$, $f_1 = g$. In the case when M, N have basepoints and $f, g \in \Gamma_*(M,N)_*$ one requires $F(x_0,t) = y_0$ for all t (this is called a *based* homotopy).

It is easy to prove that homotopies compose, in the following sense: if $f_1 \simeq f_2$ are maps from N to P and $g_1 \simeq g_2$ are map from M to N, then $f_1 \circ g_1 \simeq f_2 \circ g_2$.

The relation of being homotopic is an equivalence relation. The quotient of $\Gamma(M,N)$ by this relation, i.e. the set of homotopy classes of maps from M to N, is denoted $[M,N]$. Similarly one defines $[M,N]_*$, the set of homotopy classes of basepoint-preserving maps.

The set of homotopy classes thus defined do not depend on r, the degree of differentiability of the maps. In fact, from the mathematical point of view, it is most natural to assume that M and N are only topological spaces and that the maps are only continuous ($r = 0$).

Two spaces M and N are said to have the same *homotopy type* if there are maps $f : M \to N$ and $g : N \to M$ such that $f \circ g \simeq Id_M$ and $g \circ f \simeq Id_N$. It is easy to see that if M and N have the same homotopy type, then $[P, M] = [P, N]$ and $[M, Q] = [N, Q]$ for all spaces P, Q. A space N is said to be *contractible* if it is homotopy equivalent to a point or in other words if the identity map is homotopic to the constant map. Stated more explicitly, this means that there is a continuous map $F : I \times N \to N$ such that $F(0, y) = y$ and $F(1, y) = y_0$. For example, all vectorspaces are contractible. It is enough to take the origin as basepoint and consider $F(t, y) = ty$. If N is contractible, then $[M, N]_*$ is the trivial set consisting of a single element. To see this it is sufficient to note that for any map $f : M \to N$, $Id_N \circ f = f$ is homotopic to $y_0 \circ f = y_0$. So from the point of view of homotopy a contractible space is equivalent to a single point.

A map $p : P \to M$ is said to be a fibration if it has the *homotopy lifting property*, which means that given a homotopy $f_t : Q \to M$ and a lift of f_0, namely a map $\tilde{f}_0 : Q \to P$ such that $p \circ \tilde{f}_0 = f_0$, then there exists a homotopy $\tilde{f}_t$ such that $p \circ \tilde{f}_t = f_t$. In particular, there exist a lift of f_1 and it is homotopic to the lift of f_0. Important special cases of fibrations are fiber bundles. We shall return to this later.

In the case when M is a sphere $S^m = \{x \in \bar{R}^{m+1} \mid x_1^2 + \cdots + x_{m+1}^2 = 1\}$ with $m \geq 1$, the sets of homotopy classes can be given a group structure. This case is so important that it deserves a special name: the space $\pi_m(N) = [S^m, N]_*$ is called m-th *homotopy group* of N.

We first show how the group structure is defined in the case $m = 1$ ($\pi_1(N)$ is also called the fundamental group of N). We think of S^1 as an open interval $I = [0, 1]$ with the endpoints identified; the basepoint of S^1 corresponds to 0 (or 1). A basepoint preserving map $f : S^1 \to N$ is just a loop starting and ending at y_0. Given two loops f_1, f_2 we can define a third loop $f_1 \cdot f_2$ by "going first around f_1, then f_2, at double speed":

$$f_1 \cdot f_2(t) = \begin{cases} f_1(2t) & \text{for } 0 \leq t \leq \frac{1}{2} \\ f_2(2t-1) & \text{for } \frac{1}{2} \leq t \leq 1. \end{cases}$$

If we denote $[f] \in \pi_1(N)$ the homotopy class of the loop f, then $[f_1][f_2] = [f_1 \cdot f_2]$ defines a group multiplication in $\pi_1(N)$.

In the case $m \geq 2$, we think of S^m as the m-cube I^m with all points of the boundary identified. Note that if we call $t_1, \ldots, t_m$ the coordinates in I^m, the boundary ∂I^m of the cube consists of all points for which at least one of the coordinates is equal to 0 or 1. A map $f : I^m \to N$ such that for all $x \in \partial I^m$,

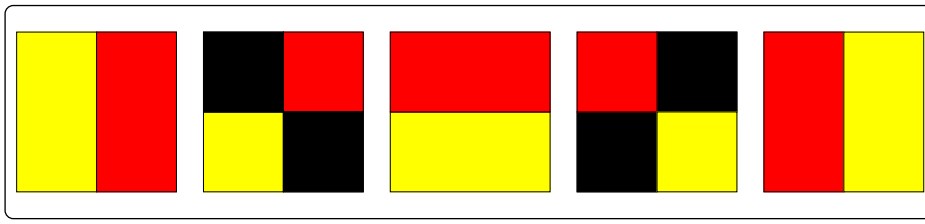


Figure 31. Proof that $\pi_2(M)$ is abelian. The first square represents the homotopy between f_1 (yellow) and f_2 (red), as given in (A.2). Black areas (including the contours of the rectangles) are points where the value of the function is y_0. By a continuous sequence of deformations one arrives at interchanging the order of f_1 and f_2 in the homotopy.

$f(x)=y_0$ can be regarded as a map $f : S^m \to N$, and thus defines a homotopy class in $\pi_m(N)$. We define $f_1 \cdot f_2$ by

$$f_1 \cdot f_2(t_1, \dots, t_m) = \begin{cases} f_1(2t_1, t_2, \dots, t_m) & \text{for } 0 \le t_1 \le \frac{1}{2} \\ f_2(2t_1 - 1, t_2, \dots, t_m) & \text{for } \frac{1}{2} \le t_1 \le 1. \end{cases}$$

The group structure in $\pi_m(N)$ is then defined as in the case $m=1$.

The groups $\pi_m(N)$ for $m \ge 2$ are always abelian, whereas $\pi_1(N)$ need not be abelian. This is illustrated in Figure 31 for the case $m=2$. It is also immediately clear why this cannot be done for $m=1$.

The definition of $\pi_m(N)$ given above works also in the case $m=0$. The sphere S^0 consists of the two points $+1$ and -1. One of them, for example $+1$, can be taken as basepoint. A basepoint-preserving map $f : S^0 \to N$ maps $+1$ to y_0 and -1 to some point y of N. Thus there is a bijective correspondence between $\Gamma_*(M,N)$ and N. Two maps f and f' are homotopic if $y = f(-1)$ and $y' = f'(-1)$ belong to the same arcwise connected component of N. Thus $\pi_0(N) = [S^0, N]_* = \{\text{arcwise connected components of } N\}$. This set does not have a group structure in general.

Summarizing, the homotopy groups give some information about the topology of a manifold: $\pi_0(N) \neq 0$ if N has more than one connected component, $\pi_1(N) \neq 0$ if N is multiply connected, $\pi_m(N) \neq 0$ if N contains non-contractible m-spheres. One can prove that if M is a smooth manifold then the homotopy groups characterize its homotopy type.

If $f : N \to Q$ is a smooth map, there are natural homomorphisms $\pi_k(f) : \pi_k(N) \to \pi_k(Q)$ for all k, defined as follows: $\pi_k(f)$ maps the homotopy class of a map $g : S^k \to N$ to the homotopy class of the map $f \circ g : S^k \to Q$. One can easily check that these are homomorphisms.

There are some results that allow us to easily calculate the homotopy groups of some spaces in terms of those of other spaces. The homotopy groups of a Cartesian product are the direct sum of the homotopy groups of the factors:

$$\pi_n(M \times N) = \pi_n(M) \oplus \pi_n(N). \tag{E.1}$$

If M is a connected but not simply connected and $p : \bar{M} \to M$ is a covering, $\pi_n(M) = \pi_n(\bar{M})$ for $n \geq 2$. If $\bar{M}$ is the universal covering of M (i.e. $\bar{M}$ is simply connected), then $\pi_1(M)$ is isomorphic to the group of deck transformations of $\bar{M}$, i.e. homeomorphisms f of $\bar{M}$ such that $p \circ f = p$.

More general relations can be obtained from the homotopy exact sequence, to be discussed in Section E.5

E.2 The winding number

Let M and N be compact, connected manifolds without boundary, both of dimension n. We denote $\omega = \frac{1}{n!}\omega_{i_1\cdots i_n} dy^{i_1} \wedge \cdots \wedge dy^{i_n}$ a volume-form on N. For example, if N is endowed with a riemannian metric $h = h_{\alpha\beta} dy^\alpha \otimes dy^\beta$ it is natural to consider the riemannian volume form $\omega = \sqrt{\det h}\, dy^1 \wedge \cdots \wedge dy^n$. Given a map $\varphi : M \to N$ we define the *winding number* of φ

$$W(\varphi) = \frac{\int_M \varphi^* \omega}{\int_N \omega} = \frac{1}{\mathrm{Vol}(N)} \int_M d^n x\, \varepsilon^{\mu_1 \cdots \mu_n} \partial_{\mu_1} \varphi^1 \cdots \partial_{\mu_n} \varphi^n\, \omega_{1\cdots n}. \tag{E.2}$$

The geometrical meaning of this quantity can be understood as follows. Recall that a point $x \in M$ is a regular point for the map φ if the tangent map $T\varphi\big|_x$ is surjective (i.e., in coordinates, if $\det(\partial_\mu \varphi^\alpha)(x) \neq 0$). A point $y \in \mathrm{Im}\varphi \subset N$ is a regular value for φ if all the points in its pre-image $\varphi^{-1}(y)$ are regular points. It can be proven that if y is any regular value of φ, then

$$\begin{aligned} W(\varphi) = &\big(\#\text{ of points in } \varphi^{-1}(y) \text{ with } \det(\partial_\mu \varphi^\alpha) > 0\big) \\ &- \big(\#\text{ of points in } \varphi^{-1}(y) \text{ with } \det(\partial_\mu \varphi^\alpha) < 0\big). \end{aligned} \tag{E.3}$$

A theorem of Hopf states that in the case when $N = S^n$, $[M, S^n]_* = \mathbb{Z}$ are classified by the winding number. In particular the winding number is the integer topological invariant that characterizes the homotopy classes of maps from S^n to S^n.

This theorem can be easily understood in the case of maps $\varphi : S^1 \to S^1$. If θ is a coordinate in the first circle and φ in the second, we must have

$$\varphi(2\pi) = \varphi(0) \mathrm{mod} 2\pi.$$

	S^1	S^2	S^3	S^4	S^5	S^6	S^7	S^8
π_1	$\mathbb{Z}$	0	0	0	0	0	0	0
π_2	0	$\mathbb{Z}$	0	0	0	0	0	0
π_3	0	$\mathbb{Z}$	$\mathbb{Z}$	0	0	0	0	0
π_4	0	$\mathbb{Z}_2$	$\mathbb{Z}_2$	$\mathbb{Z}$	0	0	0	0
π_5	0	$\mathbb{Z}_2$	$\mathbb{Z}_2$	$\mathbb{Z}_2$	$\mathbb{Z}$	0	0	0
π_6	0	$\mathbb{Z}_{12}$	$\mathbb{Z}_{12}$	$\mathbb{Z}_2$	$\mathbb{Z}_2$	$\mathbb{Z}$	0	0
π_7	0	$\mathbb{Z}_2$	$\mathbb{Z}_2$	$\mathbb{Z} \times \mathbb{Z}_{12}$	$\mathbb{Z}_2$	$\mathbb{Z}_2$	$\mathbb{Z}$	0
π_8	0	$\mathbb{Z}_2$	$\mathbb{Z}_2$	$\mathbb{Z}_2^2$	$\mathbb{Z}_{24}$	$\mathbb{Z}_2$	$\mathbb{Z}_2$	$\mathbb{Z}$

Table E.1. Homotopy groups of the low-dimensional spheres.

The winding number is given by

$$w(\varphi) = \frac{1}{2\pi} \int_0^{2\pi} d\theta \frac{d\varphi}{d\theta} = \frac{1}{2\pi}(\varphi(2\pi) - \varphi(0)).$$

By plotting the graph of $\varphi(\theta)$ one can visually see that for any regular value (i.e. point where $\frac{d\varphi}{d\theta}$ is nonzero) the winding number is equal to the number of points in the pre-image where $\frac{d\varphi}{d\theta}$ is positive minus the number of points in the pre-image where $\frac{d\varphi}{d\theta}$ is negative.

E.3 Homotopy groups of spheres

We will often need the homotopy groups of the spheres, $\pi_m(S^n)$. Some of these are given in Table E.1. All the elements above the diagonal are zero. All the elements on the diagonal are given by the theorem of Hopf mentioned in the preceding section, and the integer classifying the maps is the winding number. The part below the diagonal is quite complicated, but there are regularities. For example, the second and third column are the same from π_3 onwards. This is due to the properties of the Hopf map, as we shall see later. From the third column onwards, all the elements on the second and third lower diagonal are $\mathbb{Z}_2$; from the fifth column onwards, all the elements on the fourth lower diagonal are $\mathbb{Z}_{24}$.

E.4 Homotopy groups of Lie groups

The groups $SO(2) \approx U(1)$ and $SU(2)$ are homeomorphic to S^1 and S^3, so their homotopy groups can be read off Table E.1. The group $SO(4)$ is $SU(2) \times$

$\pi_n(.)$	$SO(3)$	$SO(4)$	$SO(5)$	$SO(6)$	$SU(3)$	$G_2, F_4, E_{6,7,8}$
π_1	$\mathbb{Z}_2$	$\mathbb{Z}_2$	$\mathbb{Z}_2$	$\mathbb{Z}_2$	0	0
π_2	0	0	0	0	0	0
π_3	$\mathbb{Z}$	$\mathbb{Z} \oplus \mathbb{Z}$	$\mathbb{Z}$	$\mathbb{Z}$	$\mathbb{Z}$	$\mathbb{Z}$
π_4	$\mathbb{Z}_2$	$\mathbb{Z}_2 \oplus \mathbb{Z}_2$	$\mathbb{Z}_2$	0	0	0
π_5	$\mathbb{Z}_2$	$\mathbb{Z}_2 \oplus \mathbb{Z}_2$	$\mathbb{Z}_2$	$\mathbb{Z}$	$\mathbb{Z}$	0
π_6	$\mathbb{Z}_{12}$	$\mathbb{Z}_{12} \oplus \mathbb{Z}_{12}$	0	0	$\mathbb{Z}_6$	0

Table E.2. Homotopy groups of some low dimensional Lie groups.

$\pi_n(.)$	$U(N)$	$O(N)$	$Sp(N)$
	$N > n/2$	$N > n+1$	$N > (n-2)/4$
π_0	0	$\mathbb{Z}_2$	0
π_1	$\mathbb{Z}$	$\mathbb{Z}_2$	0
π_2	0	0	0
π_3	$\mathbb{Z}$	$\mathbb{Z}$	$\mathbb{Z}$
π_4	0	0	$\mathbb{Z}_2$
π_5	$\mathbb{Z}$	0	$\mathbb{Z}_2$
π_6	0	0	0
π_7	$\mathbb{Z}$	$\mathbb{Z}$	$\mathbb{Z}$
π_8	0	$\mathbb{Z}_2$	0

Table E.3. Homotopy groups of the classical Lie groups for sufficiently large N.

$SU(2)/Z_2$, so its homotopy groups can be derived from the general properties mentioned in the end of Section E.1. The homotopy groups of some low dimensional Lie groups are listed in Table E.2.

Table E.3 gives the homotopy groups of the classical (unitary, orthogonal, symplectic) groups, for sufficiently large N (the condition is indicated in the second row). These homotopy groups are periodic modulo 2 (for the unitary groups) and modulo 8 (for the orthogonal and symplectic groups). This is known as Bott periodicity. The homotopy groups of the symplectic groups are the same as those of the orthogonal groups, but shifted by four.

We recall the main applications of these homotopy groups:

- $\pi_0(G)$ is the set of connected components. $\pi_0(O(3))$ is related to parity.
- $\pi_1(O(3)) = \pi_1(SO(3))$ is related to spin in dimensions $d \geq 3$.
- $\pi_3(Sp(1)) = \pi_3(SU(2))$ is related to instantons.

- $\pi_3(U(N)) = \pi_3(SU(N))$ is related to the existence of Skyrmions.
- $\pi_5(U(3)) = \pi_5(SU(3))$ is related to the spin of Skyrmions.
- $\pi_4(SU(2)) = \pi_4(Sp(1))$ is related to a global anomaly.

E.5 The homotopy exact sequence

An *exact sequence* of groups is a sequence of group homomorphisms $h_i : G_i \to G_{i+1}$ such that $\operatorname{im} h_i = \ker h_{i+1}$. When an exact sequence is known to exist, then knowledge of the properties of some of the groups or homomorphisms forming the sequence can be used to infer properties of other groups or homomorphisms.

We shall be interested in the homotopy theory of a principal bundle. As usual in homotopy theory, the total space P and the base space M are equipped with basepoints p_0 and $[p_0]$. The orbit through p_0 can be identified with the group H by identifying p_0 with the identity of H. This gives an injective map $\iota : H \to P$. The maps ι and μ are such that $\mu \circ \iota$ is the constant map $H \to M$ with image $[p_0]$. It can be shown that a fiber bundle is a fibration, so the map μ has the homotopy lifting property. Using this property, one can show that there is a long exact sequence involving the homotopy groups of H, P and M.

Recall from Section E.1 that given a map $f : M \to N$ there is an induced homomorphism of homotopy groups $f_* : \pi_n(M) \to \pi_n(N)$. Now consider a principal bundle and the homotopy groups of H, P and M. We have homomorphisms

$$\pi_n(H) \xrightarrow{\iota_*} \pi_n(P) \xrightarrow{\mu_*} \pi_n(M) \tag{E.4}$$

Since $\mu \circ \iota = [p_0]$, $\operatorname{im} \iota_* \subset \ker \mu_*$. Conversely, if $f : S^n \to P$ is such that $\mu \circ f$ is homotopic to a constant (i.e. $[f] \in \ker \mu_*$), by the homotopy lifting property there exists a map f', homotopic to f, such that $\mu \circ f' = [p_0]$. Thus $\operatorname{im} \iota_* \supset \ker \mu_*$. Altogether we have found that $\operatorname{im} \iota_* = \ker \mu_*$, therefore the sequence (E.4) is exact at $\pi_n(P)$.

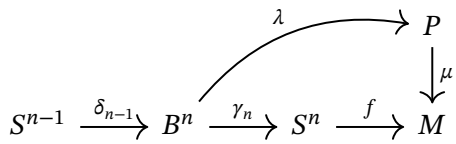
Now we can tie together the short sequences (E.4) for different n into a long exact sequence, by defining homomorphisms $\partial : \pi_n(M) \to \pi_{n-1}(H)$ and showing that $\operatorname{im} \mu_* = ker\partial$ and $\operatorname{im} \partial = \ker \iota_*$.

$$\begin{aligned} &\ldots \to \pi_{n+1}(M) \xrightarrow{\partial} \pi_n(H) \xrightarrow{\iota_*} \pi_n(P) \xrightarrow{\mu_*} \pi_n(M) \xrightarrow{\partial} \pi_{n-1}(H) \to \ldots . \\ &\ldots \xrightarrow{\partial} \pi_0(H) \xrightarrow{\iota_*} \pi_0(P) \xrightarrow{\mu_*} \pi_0(M) \end{aligned} \tag{E.5}$$

The last three sets in the sequence do not have a group structure, but the sequence is still exact if we define the kernel of a based map to consist of those elements of the domain that are mapped to the basepoint of the target.

We will now define the map ∂. Let $B^n = \{x \in \mathbb{R}^n | x_1^2 ... x_n^2 \leq 1\}$ be the closed unit ball in $\mathbb{R}^n$ and δ_n the inclusion of the unit sphere S^n as the boundary of B^{n+1}. Furthermore, let $\gamma_n : B^n \to S^n$ be the map that identifies all points of the boundary as the basepoint of S^n.

Pick a map $f : S^n \to M$. Since B^n is contractible, $f \circ \gamma_n : B^n \to M$ is homotopic to a constant. The constant map $B^n \to M$ has a lift, which is the constant map $B^n \to P$. By the homotopy lifting property, also $f \circ \gamma_n$ has a lift $\lambda : B^n \to P$. This is shown in the following commutative diagram:



Now consider the map $\lambda \circ \delta_{n-1} : S^{n-1} \to P$. Since $\mu \circ \lambda \circ \delta_{n-1}$ is the constant map, the image of $\lambda \circ \delta_{n-1}$ lies in the orbit through p_0. Thus, there must exist a map $\psi : S^{n-1} \to H$ such that $\lambda \circ \delta_{n-1} = \iota \circ \psi$. We define $\partial([f]) = [\psi]$. By repeatedly using the homotopy lifting property, it can be shown that the sequence (E.5) is exact at all groups.

To see what such sequences are useful for, consider first the rather trivial case of the universal covering of the circle: $\mathbb{Z} \to \mathbb{R} \to S^1$. Since all the homotopy groups of $\mathbb{R}$ are trivial, the long sequence breaks up into infinitely many short exact sequences

$$0 \to \pi_n(S^1) \to \pi_{n-1}(\mathbb{Z}) \to 0$$

for $n = 1, 2 ....$ The map in the middle is injective (because of the 0 on the left) and surjective (because of the 0 on the right), so $\pi_n(S^1)$ is isomorphic to $\pi_{n-1}(\mathbb{Z})$. The only nontrivial homotopy group of $\mathbb{Z}$ is $\pi_0(\mathbb{Z}) = \mathbb{Z}$, so the only nontrivial homotopy group of S^1 is $\pi_1(S^1) = \mathbb{Z}$, as is seen in the first column of Table E.1.

As a less trivial case consider the Hopf bundle, defined in Section (D.4), that has fiber S^1, total space S^3 and base space S^2. Since $\pi_n(S^1) = 0$ for $n \geq 2$, the long sequence breaks up into infinitely many short exact sequences

$$0 \to \pi_n(S^3) \to \pi_n(S^2) \to 0$$

for $n = 3, 4 ....$ Again, the map in the middle is injective (because of the 0 on the left) and surjective (because of the 0 on the right), so $\pi_n(S^2)$ is isomorphic to $\pi_n(S^3)$, as can indeed be seen, for the first few n, in Table E.1.

Appendix F

Basic homology and cohomology

To define the real p-th homology group of an oriented manifold M we consider the set of all p-dimensional oriented submanifolds of M. Here p is an integer smaller or equal to the dimension of M. In the case $p = 0$ an oriented submanifold is just a point of M together with a sign. If m_1, m_2, ...are p-dimensional submanifolds of M, and c_1, c_2, ...are real numbers, a formal linear combination

$$\sum_i c_i m_i,$$

is called a *real p-chain*.[1] The set of all real p-chains is denoted $C_p(M)$. It is an (infinite dimensional) real vectorspace having all p-dimensional oriented submanifolds of M as generators.

Given an oriented submanifold m one can consider its boundary ∂m. It is a $(p-1)$-dimensional oriented submanifold of M. Thus we can define a linear operator $\partial_p : C_p(M) \to C_{p-1}(M)$ by

$$\partial_p \left(\sum_i c_i m_i \right) = \sum_i c_i \partial m_i. \tag{F.1}$$

In the case $p = 0$ we make the convention that the boundary of a point is the empty set. When no confusion can arise we shall often omit the subscript p and denote the boundary operator by ∂. By construction, ∂ is a homomorphism.

[1] Later on we shall consider a more general construction where c_i are elements of an arbitrary abelian group but for the moment we stick to the reals.

The boundary of a manifold of dimension p is a manifold of dimension $(p-1)$ without boundary. Therefore

$$\partial \circ \partial = 0. \tag{F.2}$$

Let $Z_p(M) = ker\partial_p \subset C_p(M)$ and $B_p(M) = im\partial_{p-1} \subset C_p(M)$. The elements of $Z_p(M)$ are called *p-cycles* and the elements of $B_p(M)$ are called *p-boundaries*. Because of (F.2), every boundary is a cycle, i.e. $B_p(M) \subset Z_p(M)$, but not every cycle is necessarily a boundary.

Two cycles are said to be *homologous* if their difference is a boundary. For example consider two cycles that consist just of two p-dimensional submanifolds m_1 and m_2 without boundary, each with coefficient 1. These two cycles are homologous provided there exists a $(p+1)$-dimensional submanifold n whose boundary is given by the union of m_1 and m_2, with the appropriate orientation.

The relation of being homologous is an equivalence relations and we are interested in the equivalence classes of p-cycles. The *p-th homology group of* M is the quotient $H_p(M) = Z_p(M)/B_p(M)$.[2] If we consider the sequence of vectorspaces

$$... \to C_{p+1} \xrightarrow{\partial_{p-1}} C_p \xrightarrow{\partial_p} C_{p-1} \xrightarrow{\partial_{p+1}} ...$$

$H_p(M)$ is a measure of the failure of the sequence being exact at C_p.

It is remarkable that although $C_p(M)$, $Z_p(M)$ and $B_p(M)$ are infinite dimensional groups, the quotient $H_p(M)$ is finite dimensional. The dimension of $H_p(M)$ is called the p-th *Betti number* of M, denoted $b_p(M)$. We will not prove that b_p are finite, but we can make this plausible by considering some examples.

Example F.1

Since the boundary of a point is empty, every 0-chain is a 0-cycle: $Z_0(M) = C_0(M)$. On the other hand B_0 consists of linear combinations of points which arise by taking the boundary of a one-chain. For example, consider two zero-cycles each consisting of a single point, with coefficients 1 and -1 respectively. These cycles are homologous if and only if there exist a curve segment joining them. If M is connected, any two points can be joined by a curve, so all homology classes are just multiples of the homology class of a point: $H_0(M) = \mathbb{R}$ and $b_0 = 1$. If M consists of several connected components, choose a point p_i in each

[2] This terminology emphasizes the additive abelian group structure.

one. Every 0-cycle is homologous to a linear combination of p_i's. In this case b_0 is equal to the number of connected components of M.

Example F.2

Every loop embedded in a two-sphere divides the sphere in two disks. Therefore every one-cycle on the sphere is homologous to zero, and $H_1(S^2) = 0$.

Example F.3

The torus can be seen as the product of two unit intervals $I \times I$ (a square) with opposing sides identified (i.e. $(x, 0) \equiv (x, 1)$ and $(0, y) \equiv (1, y)$). The loops $t \longmapsto (x_0, t)$ and $t \longmapsto (t, y_0)$, for fixed x_0 and y_0, are one-dimensional submanifolds that have no boundary and are not the boundary of any two-dimensional submanifold. Thus they define nontrivial homology classes. Every 1-cycle is homologous to a linear combination of these two loops, so they generate $H_1(S^1 \times S^1) = \mathbb{R}^2$. The first Betti number of the 2-dimensional torus is $b_1(S^1 \times S^1) = 2$.

Example F.4

By a similar argument, the first Betti number of the n-dimensional torus is n and the p-th Betti number of the n-dimensional torus is $\binom{n}{p}$, the number of ways of choosing p directions out of n, each choice giving a different p-dimensional cycle that is not a boundary.

Example F.5

Let M be a compact, connected n-dimensional manifold without boundary. There are no $(n + 1)$-chains, so $B_n(M) = 0$ and $H_n(M) = Z_n(M)$. Every cycle is a multiple of M itself, so $H_n(M) = \mathbb{R}$, the generator being M itself, and $b_n = 1$.

Example F.6

There are no $(m + 1)$-chains, so $H_p(M) = 0$ for $p > m$.

Let us now define the real cohomology groups. We use de Rham's

definition, which is based on the use of differential forms. We denote $C^p(M)$ the space of smooth p-forms (totally antisymmetric p-tensors) on M. In this context the p-forms are also called *p-cochains*. the exterior differential d is a map from $C^p(M)$ to $C^{p+1}(M)$, with the property that

$$d \circ d = 0. \tag{F.3}$$

A p form α is said to be *closed*, or a *p-cocycle*, if $d\alpha = 0$, and *exact*, or a *p-coboundary*, if there exists a $(p+1)$-form β such that $\alpha = d\beta$. The space of all closed forms is denoted $Z^p(M)$ and the space of all exact forms is denoted $B^p(M)$. Because of (F.3), we have $B^p(M) \subset Z^p(M)$. Two closed forms which differ by an exact form are said to be cohomologous. This is an equivalence relation and the equivalence classes of closed p-forms modulo exact p-forms are called cohomology classes. The *p-th (de Rham) cohomology group of M* is $H^p(M) = Z^p(M)/B^p(M)$.

Also in this case, C^p, Z^p and B^p are infinite dimensional vectorspaces, but the cohomology groups are finite dimensional. Their dimensions are denoted b^p.

Example F.7

The space $C^0(M)$ is the space of real functions on M and $Z^0(M)$ is the subspace of locally constant functions. The space $B^0(M)$ is empty, since there are no forms of order -1. If M is connected, every locally constant function is constant, so $H^0(M) = Z^0(M) = \mathbb{R}$. If M has several connected components, a locally constant function is constant on each connected component, so $H^0(M) = \mathbb{R}^{b^0}$, where b^0 is the number of connected components of M.

Example F.8

Every one-form on a sphere of dimension $m > 1$ is exact, so $H^1(S^m) = 0$.

Example F.9

Parametrize a circle S^1 with an angle $0 \leq \varphi < 2\pi$. The form $d\varphi$ is closed and locally exact, but it is not globally exact, because φ is not a (single-valued) function on the circle. Given a one-form α on the

circle consider the integral $f(\varphi) = \int_0^\varphi \alpha$. We have locally $\alpha = df$. If $f(2\pi) = 0$, then f is a well-defined function on the circle and α is exact. Thus two one-forms α_1 and α_2 are cohomologous if and only if $\int_{S^1} \alpha_1 = \int_{S^1} \alpha_2$. Suppose that $\int_{S^1} \alpha = c$. Then α is cohomologous to $(c/2\pi)d\varphi$ and we find that $H^1(S^1) = \mathbb{R}$, generated by (the cohomology class of) $\omega = \frac{1}{2\pi} d\varphi$.

Example F.10

Generalizing the previous example, on an n-dimensional compact, connected manifold without boundary, all forms of degree n are closed, because there are no forms of degree $n+1$. Thus, $Z^n = C^n$. From Stokes' theorem, the integral of any exact n-form on M must be zero. Thus a volume form defines a nontrivial cohomology class, that generates $H^n(M) = \mathbb{R}$.

The most important property of the homology and cohomology groups is that they are topological invariants, i.e. if M and N are homeomorphic, they have the same homology and cohomology groups. However, in general, two manifolds with the same homology and cohomology groups need not be homeomorphic.

Comparing the results in the given examples one notices that $H^p(M)$ and $H_p(M)$ are the same. This is not casual. Consider the real number defined by

$$\langle \alpha | m \rangle = \int_m \alpha, \tag{F.4}$$

where α is a p form and m is a p dimensional submanifold of M. This defines a bilinear pairing $C^p(M) \times C_p(M) \to \mathbb{R}$. If m is a cycle and α is a cocycle, the pairing depends only on the homology class of m and the cohomology class of α. In fact, using Stokes' theorem,

$$\langle \alpha + d\beta | m \rangle = \langle \alpha | m \rangle$$

and

$$\langle \alpha | m + \partial n \rangle = \langle \alpha | m \rangle.$$

So we actually have a bilinear pairing $H^p(M) \times H_p(M) \to \mathbb{R}$.

One can prove that this pairing is nondegenerate, in the sense that $\langle \alpha | m \rangle = 0$ for all $\alpha \in Z^p(M)$ implies $m = \partial n$ and $\langle \alpha | m \rangle = 0$ for all $m \in Z_p(M)$ implies

$\alpha = d\beta$. Thus, $H^p(M)$ is isomorphic to the dual space $H_p(M)^*$. In particular, they have the same dimensions $b^p = b_p$.

This shows that there is no more information in the cohomology groups than there is in the homology groups. However, the direct sum $\oplus_p H^p(M)$ can be given an algebra structure, with the product coming from the exterior product of forms. This whole algebra is a topological invariant, and it does not have a counterpart in homology.

One can define homology groups with coefficients in any abelian group G. In the definition of a p-chain given above one just reinterprets the coefficients c_i as elements of G instead of real numbers. The resulting homology groups are denoted $H_p(M, G)$. The most important case is $G = \mathbb{Z}$, the group of the integers. The integer homology group $H_p(M, \mathbb{Z})$ can be shown to be a finitely generated abelian group, and has the general structure

$$\mathbb{Z} \oplus ... \oplus \mathbb{Z} \oplus \mathbb{Z}_{n_1} \oplus ... \mathbb{Z}_{n_k},$$

where there are b_p direct addends $\mathbb{Z}$ and k addends which are cyclic groups (of order $n_1, ..., n_k$). The direct sum of the $\mathbb{Z}$ groups forms the so-called free part, while the direct sum of the cyclic groups is called the torsion part.

The integer homology groups are the ones that contain most information. The homology groups with other coefficients can be obtained from the ones with integer coefficients by using the so-called universal coefficient theorem. For example, the real homology groups are obtained by replacing every addend $\mathbb{Z}$ by an addend $\mathbb{R}$ and dropping the torsion part. Therefore, they contain less information than the integer homology groups.

One can also define cohomology groups with arbitrary coefficients. In general it is not possible to use differential forms. In the real case, one can regard a differential form as a linear map from $C_p(M)$ to $\mathbb{R}$. In general one can define $C^p(M, G)$ to be the space of all homomorphisms from $C_p(M, \mathbb{Z})$ to G. The differential d is defined in this case by the requirement that for every cochain α and chain m, $d\alpha(m) = \alpha(\partial m)$. The resulting cohomology groups $H^p(M, G)$ are again related to the corresponding homology groups. For example

$$H^p(M, \mathbb{Z}) = \text{free}(H_p(M, \mathbb{Z})) \oplus \text{tor}(H_{p-1}(M, \mathbb{Z})).$$

It is possible to represent the integer cohomology classes by means of singular differential forms [AlE57]. For our purposes it will be enough to note that the homomorphism $\mathbb{Z} \to \mathbb{R}$ gives rise to a homomorphism $H^p(M, \mathbb{Z}) \to H^p(M, \mathbb{R})$, and that the latter group can be represented in the de Rham way by differential forms. A de Rham cohomology class is in the image of this

homomorphism if and only if

$$\langle \alpha | m \rangle = \int_m \alpha \in \mathbb{Z} \quad \forall m \in Z_p(M, \mathbb{Z}).$$

Finally we mention a connection between homology and homotopy groups, known as the Hurewicz theorem: if $\pi_1(M) = ... = \pi_r(M) = 0$, then $H_1(M, \mathbb{Z}) = ... = H_r(M, \mathbb{Z}) = 0$ and $H_{r+1}(M, \mathbb{Z}) = \pi_{r+1}(M)$.

Appendix G

Manifolds of maps

G.1 Geometry of spaces of maps

The spaces of maps $\Gamma(M,N)$ introduced in Section E.1 can be given the structure of topological spaces and differentiable manifolds. The purpose of a differentiable structure on some space X is to be able to define smooth functions on X. This is done by postulating that every point in X has a neighborhood that is homeomorphic to an open set in some vector space V, and then patching together sufficiently many such neighborhoods to construct a global atlas. For finite dimensional real manifolds, $V = \mathbb{R}^n$. We know how to do calculus on $\mathbb{R}^n$ and we use the atlas to define calculus on X. One then says that $\mathbb{R}^n$ is the model for M. If M is m-dimensional and N is n-dimensional, a suitable manifold model for $\Gamma(M,N)$ will be something akin to $\Gamma(\mathbb{R}^m,\mathbb{R}^n)$. Functions on an infinite dimensional manifold are often called functionals and the notation $f[\varphi]$ is used instead of $f(\varphi)$ to make this clear. In the main text we have not used this convention, that is mostly superfluous, but we will find it convenient to use it in this appendix. We will not deal with any of the subtleties involved in defining smooth function(al)s on $\Gamma(M,N)$, but show instead how, following the standard constructions for finite dimensional manifolds, one can define tensors on $\Gamma(M,N)$ and do some actual calculations with them. Physicists will find that this is mostly an application of familiar rules from the calculus of variations.

We assume coordinate systems $\{x^\mu\}$ on M and $\{y^\alpha\}$ on N, One can take as coordinates on $\Gamma(M,N)$ the values $\varphi^\alpha(x)$, for all α and x. Therefore the coordinates are indexed by the infinite set $\{\alpha, x\}$.

A real functional is a real-valued function on $\Gamma(M,N)$. Most functionals of interest in physics are integrals of scalars constructed with the field and its derivatives. Functionals that are integrals of scalars formed with φ and

a finite number of derivatives of φ are called *local functionals*. Note that in general, even though their argument is a function on M, functionals have no dependence on a point in M. However, particular functionals may have such dependence, for example one may define a functional $ev_x = \int d^n y \delta(x-y)$, whose value on a function is the value of that function at x:

$$ev_x[\varphi] = \varphi(x). \tag{G.1}$$

Let φ be a point on $\Gamma(M,N)$ and let $c(t)$ be a curve on $\Gamma(M,N)$, parametrized by a real parameter t, with $c(0) = \varphi$. The vector tangent to this curve at the point φ is a linear first order differential operator v that acting upon a real-valued functional f gives the real number:

$$v(f) = \frac{df(c(t))}{dt}\Big|_{t=0}. \tag{G.2}$$

We can then write

$$v = \int_x v^\alpha(x) \frac{\delta}{\delta\varphi^\alpha(x)}, \tag{G.3}$$

where $\frac{\delta}{\delta\varphi^\alpha(x)}$ are the basis vectors and we use the shorthand notation

$$\int_x = \int_M d^n x.$$

The sum over the indexing set now consists of a sum over α and an integral over x.[1]

The space of all vectors at φ is a linear space called the tangent space at φ, denoted $T_\varphi\Gamma(M,N)$. The union of all tangent spaces is another manifold $T\Gamma(M,N)$ called the tangent bundle. It is a bundle in the sense that there is a projection from $T\Gamma(M,N)$ to $\Gamma(M,N)$, namely the map that associates to any vector the point at which it is attached.

One can visualize a vector tangent to $\Gamma(M,N)$ at a point φ as follows. A curve $c(t)$ in $\Gamma(M,N)$ is a one-parameter deformation of the map $\varphi = c(0)$. For every point $x \in M$ we have a curve c_x in N defined by $c_x(t) = (c(t))(x)$. Every such curve has a tangent vector at $\varphi(x)$. Therefore a vector tangent to φ is an assignment to each point $x \in M$ of a vector tangent to N at $\varphi(x)$. In other words a vector tangent to $\Gamma(M,N)$ at a point φ is a map v from M to TN which projects onto φ. This is called a *vectorfield along* φ.

[1] De Witt proposed that the coordinate x be subsumed in the index α and that the summation convention should be extended to include integration over x [DeW64]. Then this expression would be written $v^\alpha \cdot \frac{\delta}{\delta\varphi^\alpha}$. We shall not use this condensed notation here.

A one-form on $\Gamma(M,N)$ at φ is a linear map from the tangent space at φ to the reals. The space of all one-forms at φ is the cotangent space $T^*_\varphi\Gamma(M,N)$.

If $f[\varphi]$ is a functional on $\Gamma(M,N)$, its differential is the one-form δf, defined by

$$\delta f(v) = v(f). \tag{G.4}$$

In the coordinates defined above, the differential is

$$\delta f = \int_x \frac{\delta f}{\delta\varphi^\alpha(x)}\delta\varphi^\alpha(x). \tag{G.5}$$

The natural basis for forms consists of the functional differentials $\delta\varphi^\alpha(x)$. It is dual to the natural basis for vectors

$$\delta\varphi^\alpha(x)\left(\frac{\delta}{\delta\varphi^\beta(y)}\right) = \frac{\delta\varphi^\alpha(x)}{\delta\varphi^\beta(y)} = \delta^\alpha_\beta\delta(x-y). \tag{G.6}$$

Thus, a one-form ω can be expanded in local coordinates,

$$\omega = \int_x \omega_\alpha(x)\delta\varphi^\alpha(x). \tag{G.7}$$

In this way the value of the form ω on the vector v is given by

$$\omega(v) = \int_x \omega_\alpha(x)v^\alpha(x). \tag{G.8}$$

As usual, tensors can be defined as multilinear maps from the tensor products of tangent and cotangents spaces to the reals. We will only need differential forms, that are totally antisymmetric covariant tensors. In general, a p-form can be written

$$\omega = \frac{1}{p!}\int_{x_1}\cdots\int_{x_p}\omega_{\alpha_1\ldots\alpha_p}(x_1,\ldots,x_p)\delta\varphi^{\alpha_1}(x_1)\wedge\cdots\wedge\delta\varphi^{\alpha_p}(x_p). \tag{G.9}$$

The value of ω on vectors $v_1,\ldots v_p$ is

$$\omega(v_1,\ldots,v_p) = \frac{1}{p!}\int_{x_1}\cdots\int_{x_p}\omega_{\alpha_1\ldots\alpha_p}(x_1,\ldots,x_p)v_1^{\alpha_1}(x_1)\ldots v_p^{\alpha_p}(x_p). \tag{G.10}$$

Note that the differentials in (G.9) in general are evaluated at the different points: the components $\omega_{\alpha_1\ldots\alpha_p}$ are multilocal. In special cases, some or all the arguments $(x_1,\ldots,x_p)$ may coincide, in which case the number of integrations is correspondingly reduced.

Also note that in general the components of tensors (vectors, forms…) are functionals with a dependence on the point in M. In this respect they are somewhat similar to (G.1). The dependence on x has to be understood in the same sense as an index. Thus the components are not genuine scalar functionals, just like the components of a vector on a finite dimensional manifold are not scalar functions on the manifold. It is only when they are contracted with other objects and integrated that they become true scalar functionals (without a dependence on points in M).

The contraction of the p-form ω with a vector v is the $p-1$-form

$$i_v\omega = \frac{1}{(p-1)!}\int_{x_1}\cdots\int_{x_p} v^{\alpha_1}(x_1)\omega(x_1,\dots,x_p)_{\alpha_1\dots\alpha_p}\delta\varphi^{\alpha_2}(x_2)\wedge\cdots\wedge\delta\varphi^{\alpha_p}(x_p). \tag{G.11}$$

Next we define tensor fields. A vector field v is a section of $T\Gamma(M,N)$, i.e. the assignment of a tangent vector to each point of $\Gamma(M,N)$. A tangent vector field can be written as in (G.3), but now with the components v^α being also functionals of φ. Then, (G.3) becomes

$$v[\varphi] = \int dx\, v[\varphi]^\alpha(x)\frac{\delta}{\delta\varphi^\alpha(x)}.$$

For example, the Lie bracket of two vector fields v and w is given by

$$[v,w] = \int_x\int_y\Big(v[\varphi]^\alpha(x)\frac{\delta w[\varphi]^\beta(y)}{\delta\varphi^\alpha(x)} - w[\varphi]^\alpha(x)\frac{\delta v[\varphi]^\beta(y)}{\delta\varphi^\alpha(x)}\Big)\frac{\delta}{\delta\varphi^\beta(y)}. \tag{G.12}$$

Tensor fields are defined in a similar way. In particular one can define fields of differential forms. Then, the exterior derivative is defined as usual by

$$\begin{aligned} d\omega(v_1,\dots v_{k+1}) = &\sum_i(-)^{i+1}v_i\big(\omega(v_1,\dots,\widehat{v_i},\dots v_{k+1})\big)\\ &-\sum_{i<j}(-)^{i+j}\omega([v_i,v_j],v_1,\dots\widehat{v_i}\dots,\widehat{v_j},\dots v_{k+1}), \end{aligned} \tag{G.13}$$

where v_i on the r.h.s. has to be regarded as a directional derivative acting on the function in brackets, and the hat denotes that one argument is missing. In particular, if ω is a one-form on $\Gamma(M,N)$

$$\begin{aligned} d\omega(v,w) &= v(\omega(w)) - w(\omega(v)) - \omega([v,w])\\ &= \int_x\int_y v[\varphi]^\alpha(x)w[\varphi]^\beta(y)\Big(\frac{\delta\omega[\varphi]_\beta(y)}{\delta\varphi^\alpha(x)} - \frac{\delta\omega[\varphi]_\alpha(x)}{\delta\varphi^\beta(y)}\Big). \end{aligned} \tag{G.14}$$

G.2 Homotopy of spaces of maps

Next, we describe some relations between the homotopy, and cohomology groups of N and those of the space of maps $\Gamma_*(M,N)$. We assume for simplicity that M and N are compact, connected manifolds without boundary, of dimension m and n respectively.

We begin by the following result, of which special cases are used several times in this book:

$$\pi_k(\Gamma_*(S^m,N)) = \pi_{k+m}(N). \tag{G.15}$$

To understand this result it is convenient to represent the sphere S^k as a k-dimensional cube I^k with the points on the boundary identified. We shall denote $0 \le t_i \le 1$, $i = 1,\dots,k$ the coordinates on this cube. The same can be done with S^m. The space N has a basepoint denoted y_0; we also choose as a basepoint in $\Gamma_*(S^m,N)$ the constant map y_0 (the distinction between these two meanings of the symbol y_0 will be clear from the context).

An element of $\Gamma_*(S^m,N)$ is a map φ from I^m to N such that $\varphi(x^1,\dots,x^m) = y_0$ whenever one of its arguments is either 0 or 1. An element of $\pi_k(\Gamma_*(S^m,N))$ is (the homotopy class of) a map f from I^k to $\Gamma_*(S^m,N)$, such that $f(t_1,\dots,t_k) = y_0$ whenever one of its arguments is either 0 or 1. Let us define a map $\widehat{f} : I^{k+m} \to N$ by

$$\widehat{f}(t_1,\dots,t_k,x^1,\dots,x^m) = (f(t_1,\dots,t_k))(x^1,\dots,x^m).$$

We have $\widehat{f} = y_0$ whenever one of its arguments is either 0 or 1. So $\widehat{f}$ defines a map from S^{k+m} to N, and hence an element of $\pi_{k+m}(N)$.

Continuous deformations of f correspond to continuous deformations of $\widehat{f}$. Therefore the correspondence of the homotopy class of f to the homotopy class of $\widehat{f}$ is bijective.

G.3 Cohomology of spaces of maps

Insofar as $\Gamma_*(M,N)$ is a topological space, one can define its cohomology groups. In the main text, and in particular in Sections 3.3 and 4.2, we have used some relations between the cohomology groups of these spaces and those of N. In the rest of this section we outline these constructions in some detail and in greater generality.

We begin by defining linear maps $h^p : C^{p+m}(N) \to C^p(\Gamma(M,N))$, where $p = 0,1,2,\dots$. Given a $p+m$-form ω on N, a map $\varphi \in \Gamma(M,N)$ and tangent vectors $v_1,\dots,v_p \in T_\varphi\Gamma(M,N)$, we define the p-form $h^p\omega$ on $\Gamma(M,N))$ by

giving its value at the point φ on the vectors $v_1 \dots v_p$:

$$(h^p\omega)_\varphi(v_1,\dots,v_p) = \int_M \varphi^* i_{v_p}\dots i_{v_1}\omega. \tag{G.16}$$

The forms on $\Gamma(M,N)$ obtained in this way are local, in the sense that they depend only on a finite number of derivatives of φ. The homomorphisms h^p intertwine the action of the exterior differentials on N and $\Gamma(M,N)$:

$$d\circ h^p = h^{p+1}\circ d. \tag{G.17}$$

Thus h^p induce homomorphisms on the cohomology groups $h^p : H^{p+m}(N) \to H^p(\Gamma(M,N))$. In the rest of the section we give the proof of (G.17), by explicit calculation.

Let ω be a field of $p+m$-forms and $v_i,\ i = 1,\dots,p+1$ be arbitrary vectorfields, all defined in the neighbourhood of a map φ in $\Gamma(M,N)$. Thus v_i can be decomposed as in (G.3), the components $v^\alpha(x)$ now being arbitrary functionals of φ. From (G.13) and (G.16) we have

$$\begin{aligned} d(h^p\omega)_\varphi(v_1,\dots,v_{p+1}) = &\sum_i (-)^{i+1} v_i \int_M \varphi^* i_{v_{p+1}}\dots\widehat{i}_{v_i}\dots i_{v_1}\omega \\ &+ \sum_{i<j}(-)^{i+j}\int_M \varphi^* i_{v_{p+1}}\dots\widehat{i}_{v_j}\dots\widehat{i}_{v_i}\dots i_{v_1} i_{[v_i,v_j]}\omega. \end{aligned} \tag{G.18}$$

In the first term, v_i acts as a directional derivative, and the Lie bracket of vectorfields is defined in (G.12). The i-th term in the first sum is

$$\begin{aligned} &\int_x v_i^\alpha(x)\frac{\delta}{\delta\varphi^\alpha(x)}\int_y \varepsilon^{i_1\dots i_m}\frac{\partial\varphi^{\alpha_1}}{\partial y^{i_1}}\dots\frac{\partial\varphi^{\alpha_m}}{\partial y^{i_m}} v_1^{\beta_1}(y)\dots\widehat{v_i^{\beta_i}}(y)\dots v_{p+1}^{\beta_{p+1}}(y) \\ &\quad\times\omega_{\beta_1\dots\widehat{\beta}_i\dots\beta_{p+1}\alpha_1\dots\alpha_m}(\varphi(y)) \end{aligned} \tag{G.19}$$

that evaluates to

$$\begin{aligned} &\int_x\int_y \varepsilon^{i_1\dots i_m} v_i^\alpha(x)\Big[m\frac{\partial}{\partial y^{i_1}}\delta(x-y)\frac{\partial\varphi^{\alpha_2}}{\partial y^{i_2}}\dots\frac{\partial\varphi^{\alpha_m}}{\partial y^{i_m}} v_1^{\beta_1}(y)\dots\widehat{v_i^{\beta_i}}(y)\dots v_{p+1}^{\beta_{p+1}}(y) \\ &\quad\times\omega_{\beta_1\dots\widehat{\beta}_i\dots\beta_{p+1}\alpha\alpha_2\dots\alpha_m}(\varphi(y)) \\ &+\sum_{j\neq i}\frac{\partial\varphi^{\alpha_1}}{\partial y^{i_1}}\dots\frac{\partial\varphi^{\alpha_m}}{\partial y^{i_m}} v_1^{\beta_1}(y)\dots\widehat{v_i^{\beta_i}}(y)\dots\frac{\delta v_j^{\beta_j}(y)}{\delta\varphi^\alpha(x)}\dots v_{p+1}^{\beta_{p+1}}(y)\,\omega_{\beta_1\dots\widehat{\beta}_i\dots\beta_{p+1}\alpha_1\dots\alpha_m}(\varphi(y)) \\ &+\frac{\partial\varphi^{\alpha_1}}{\partial y^{i_1}}\dots\frac{\partial\varphi^{\alpha_m}}{\partial y^{i_m}} v_1^{\beta_1}\dots\widehat{v_i^{\beta_i}}\dots v_{p+1}^{\beta_{p+1}}\delta(x-y)(\partial_\alpha\omega_{\beta_1\dots\widehat{\beta}_i\dots\beta_{p+1}\alpha_1\dots\alpha_m})(\varphi(y))\Big]. \end{aligned} \tag{G.20}$$

In the first term we replace $\frac{\partial}{\partial y^{i_1}}\delta(x-y)$ by $-\frac{\partial}{\partial x^{i_1}}\delta(x-y)$, integrate by parts and change the dummy index α to β_i. Then one can rewrite this as

$$\begin{aligned}
&\int_x \varepsilon^{i_1\dots i_m}\partial_{i_2}\varphi^{\alpha_2}\dots\partial_{i_m}\varphi^{\alpha_m} m v_1^{\beta_1}\dots\frac{\partial v_i^{\beta_i}}{\partial x^{i_1}}\dots v_{p+1}^{\beta_{p+1}}\omega_{\beta_1\dots\widehat{\beta_i}\dots\beta_{p+1}\beta_i\alpha_2\dots\alpha_m}\\
&+\int_x\int_y \varepsilon^{i_1\dots i_m}\frac{\partial\varphi^{\alpha_1}}{\partial y^{i_1}}\dots\frac{\partial\varphi^{\alpha_m}}{\partial y^{i_m}}\sum_{j\neq i} v_i^{\alpha}(x)\frac{\delta v_j^{\beta_j}(y)}{\delta\varphi^{\alpha}(x)}\\
&\quad\times v_1^{\beta_1}(y)\dots\widehat{v_i^{\beta_i}}(y)\dots\widehat{v_j^{\beta_j}}(y)\dots v_{p+1}^{\beta_{p+1}}(y)\omega_{\beta_1\dots\widehat{\beta_i}\dots\beta_{p+1}\alpha_1\dots a_m}(\varphi(y))\\
&+\int_x \varepsilon^{i_1\dots i_m}\partial_{i_1}\varphi^{\alpha_1}\dots\partial_{i_m}\varphi^{\alpha_m}v_1^{\beta_1}\dots v_i^{\beta_i}\dots v_{p+1}^{\beta_{p+1}}\partial_{\beta_i}\omega_{\beta_1\dots\widehat{\beta_i}\dots\beta_{p+1}\alpha_1\dots\alpha_m}.
\end{aligned}\tag{G.21}$$

When this expression is inserted into the first sum on the r.h.s. of (G.18), the alternating sum over i of the second term gives Lie-brackets that cancel the second sum on the r.h.s. of (G.18). Thus, (G.18) is equal to the alternating sum over i of the first and third term in this last expression. Using

$$\omega_{\beta_1\dots\widehat{\beta_i}\dots\beta_{p+1}\beta_i\alpha_2\dots\alpha_m}=(-)^{i+p-1}\omega_{\beta_1\dots\beta_{p+1}\alpha_2\dots\alpha_m},$$

the alternating sum of the first term in (G.21) can be rewritten:

$$\begin{aligned}
&(-)^p\sum_i\int_x \varepsilon^{i_1\dots i_m}\partial_{i_2}\varphi^{\alpha_2}\dots\partial_{i_m}\varphi^{\alpha_m}m v_1^{\beta_1}\dots\frac{\partial v_i^{\beta_i}}{\partial x^{i_1}}\dots v_{p+1}^{\beta_{p+1}}\omega_{\beta_1\dots\beta_{p+1}\alpha_2\dots\alpha_m}\\
&=(-)^p\int_x \varepsilon^{i_1\dots i_m}\partial_{i_2}\varphi^{\alpha_2}\dots\partial_{i_m}\varphi^{\alpha_m}m\frac{\partial}{\partial x^{i_1}}\left(v_1^{\beta_1}\dots v_{p+1}^{\beta_{p+1}}\right)\omega_{\beta_1\dots\beta_{p+1}\alpha_2\dots\alpha_m}\\
&=(-)^{p+1}\int_x \varepsilon^{i_1\dots i_m}\partial_{i_1}\varphi^{\alpha_1}\dots\partial_{i_m}\varphi^{\alpha_m}v_1^{\beta_1}\dots v_{p+1}^{\beta_{p+1}}m\,\partial_{\alpha_1}\omega_{\beta_1\dots\beta_{p+1}\alpha_2\dots\alpha_m}.
\end{aligned}\tag{G.22}$$

Altogether, (G.18) is equal to

$$\begin{aligned}
&\int_x \varepsilon^{i_1\dots i_m}\partial_{i_1}\varphi^{\alpha_1}\dots\partial_{i_m}\varphi^{\alpha_m}v_1^{\beta_1}\dots v_{p+1}^{\beta_{p+1}}\Big[\sum_i(-)^{i+1}\partial_{\beta_i}\omega_{\beta_1\dots\widehat{\beta_i}\dots\beta_{p+1}\alpha_1\dots\alpha_m}\\
&\quad+(-)^{p+1}m\partial_{\alpha_1}\omega_{\beta_1\dots\beta_{p+1}\alpha_2\dots\alpha_m}\Big].
\end{aligned}\tag{G.23}$$

Using total antisymmetry in $\alpha_1\dots\alpha_m$, the quantity in square bracket can be replaced by $(m+p+1)\partial_{[\beta_1}\omega_{\beta_2\dots\beta_{p+1}\alpha_1\dots\alpha_m]}=(d\omega)_{\beta_1\dots\beta_{p+1}\alpha_1\dots\alpha_m}$. Thus (G.18) is equal to

$$\int_M\varphi^* i_{v_{p+1}}\dots i_{v_1}d\omega=h^{p+1}(d\omega)_\varphi(v_1\dots v_{p+1}).\tag{G.24}$$

Since $v_1,\dots,v_{p+1}$ are arbitrary, this concludes the proof of (G.17).

Appendix H

Solutions to selected exercises

H.1 Exercise 1.2: Noether currents of the $O(N)$ model

The generators of $O(N)$ are $N \times N$ matrices $(T_a)_{mn}$ with $a = 1, \dots, \frac{N(N-1)}{2}$, $m, n = 1, \dots, N$, antisymmetric in (m, n). Hence the infinitesimal transformation with small parameters ϵ_a acts as

$$\delta\phi_m = \epsilon_a T_{amn}\phi^n.$$

According to the general formula (1.7), the Noether current is

$$j_a^\mu = -\partial_\mu \phi^m T_{amn}\phi^n.$$

The divergence of the current is

$$\partial_\mu j_a^\mu = -\Box\phi^m T_{amn}\phi^n - \partial_\mu\phi^m T_{amn}\partial^\mu\phi^n,$$

The second term is identically zero, because T_a is antisymmetric. The equations of motion for ϕ^m is

$$\Box\phi_n = (m^2 + \lambda\phi^2)\phi_n.$$

When we use this equation, the first term becomes proportional to $\phi^m\phi^n$ and also vanishes because of antisymmetry. Hence the current j_a^μ is conserved on shell.

Let's consider now the broken phase: after the field redefinition, the Lagrangian becomes

$$-\frac{1}{2}\partial_\mu\pi^m\partial^\mu\pi^m - \frac{1}{2}\partial_\mu\chi\partial^\mu\chi - m^2\chi^2 - \frac{\lambda}{4}(\pi^m\pi^m + \chi^2)^2 - \lambda f\chi(\pi^m\pi^m + \chi^2).$$

We consider separately the action of the unbroken generators of $O(N-1)$, with $a = 1, \dots, \frac{(N-1)(N-2)}{2}$ and acting as

$$\delta\pi_m = \epsilon_a T_{amn}\pi^n, \quad \delta\chi = 0,$$

and the remaining $N-1$ generators acting as

$$\delta\pi_m = \epsilon_b T_{bmN}(f+\chi), \qquad \delta\chi = \epsilon_b T_{bNm}\pi^m.$$

The conserved currents are

$$j_a^\mu = -\partial_\mu \pi^m T_{amn}\pi^n$$

and

$$j_b^\mu = -T_{bmN}[(f+\chi)\partial_\mu\pi^m - \pi^m\partial_\mu\chi],$$

respectively. The equations of motion for π and χ are:

$$\begin{aligned}\Box\pi^m &= \lambda\pi^m(\pi^n\pi^n + \chi^2 + 2f\chi)\\ \Box\chi &= \lambda(\pi^m\pi^m + \chi^2 + 2f\chi)(f+\chi).\end{aligned}$$

The divergences of the two currents

$$\begin{aligned}\partial_\mu j_a^\mu &= -\Box\pi^m T_{amn}\pi^n\\ \partial_\mu j_b^\mu &= -T_{bmN}[(f+\chi)\Box\pi^m - \pi^m\Box\chi]\end{aligned}$$

are found again to vanish on-shell.

H.2 Exercise 1.4: alternative chiral Lagrangian

Consider first the bosonic part (the first line) of the Lagrangian (1.218):

$$-\frac{1}{4}\mathrm{tr}\partial_\mu\Sigma^\dagger\partial^\mu\Sigma - \frac{\lambda}{4}\left(\frac{1}{2}\mathrm{tr}\Sigma^\dagger\Sigma - f^2\right)^2$$

Using $\mathrm{tr}(\sigma_a\sigma_b) = 2\delta_{ab}$ we find

$$\mathrm{tr}(\partial_\mu\Sigma^\dagger\partial_\mu\Sigma) = 2(\partial_\mu\sigma\partial^\mu\sigma + \partial_\mu\pi^a\partial^\mu\pi^a)$$

and

$$\mathrm{tr}\Sigma^+\Sigma = 2(\pi^a\pi^a + \sigma^2),$$

that immediately reproduce the bosonic part of (1.59).

From the definitions

$$P_{\pm} = \frac{1 \pm \gamma_5}{2}$$

and

$$N_L = P_- N, \qquad N_R = P_+ N,$$

one has for the conjugate spinors

$$\bar{N}_L = iN_L^+\gamma^0 = iN^+P_-\gamma^0 = iN^+\gamma^0 P_+ = \bar{N}P_+$$

and

$$\bar{N}_R = iN_R^+\gamma^0 = iN^+P_+\gamma^0 = iN^+\gamma^0 P_- = \bar{N}P_-,$$

where we have used the fact that γ_5 anticommutes with γ^μ. Thus

$$\bar{N}_L\gamma^\mu\partial_\mu N_L = \bar{N}P_+\gamma^\mu\partial_\mu P_- N = \bar{N}\gamma^\mu\partial_\mu P_- N$$

and

$$\bar{N}_R\gamma^\mu\partial_\mu N_R = \bar{N}P_-\gamma^\mu\partial_\mu P_+ N = \bar{N}\gamma^\mu\partial_\mu P_+ N,$$

so the free fermionic part of the Lagrangian (1.218) reduces to

$$-\bar{N}\gamma^\mu\partial_\mu N.$$

For the fermion-scalar interaction term, using the same machinery, we have

$$\begin{aligned}\bar{N}_L\Sigma N_R + \bar{N}_R\Sigma^\dagger N_L &= \bar{N}[(\sigma + i\pi^a\sigma_a)P_+ + (\sigma - i\pi^a\sigma_a)P_-]N \\ &= \bar{N}(\sigma + i\pi^a\sigma_a\gamma_5)N \\ &= \bar{N}(\sigma + 2\pi^a\tau_a\gamma_5)N.\end{aligned}$$

H.3 Exercise 1.5: coordinates on the sphere

In each case, the simplest way to calculate the metric is to start from the expressions $z_i(y^\alpha)$ of the embedding coordinates as functions of the coordinates on the sphere. Differentiate

$$dz_i = \frac{\partial z_i}{\partial y^\alpha} dy^\alpha,$$

and insert in the line element of the embedding space $\mathbb{R}^N$

$$ds^2 = dz_1^2 + ... + dz_N^2.$$

Comparing with the line element on the sphere

$$ds^2 = h_{\alpha\beta} dy^\alpha dy^\beta$$

one reads off the components $h_{\alpha\beta}$.

In the case of the stereographic coordinates, the relations $z_i(\omega_\alpha)$ are derived as follows. From elementary geometry,

$$\frac{z_i}{r - z_N} = \frac{\omega_i}{2r} \qquad \text{for} \quad i = 1, \dots, N-1.$$

Define

$$\rho^2 = \omega_1^2 + \dots + \omega_{N-1}^2 = 4r^2 \frac{r + z_N}{r - z_N}.$$

Inverting we obtain

$$z_N = r \frac{\rho^2 - 4r^2}{\rho^2 + 4r^2}$$

and therefore

$$z_i = \frac{4r^2}{\rho^2 + 4r^2} \omega_i \qquad \text{for} \quad i = 1, \dots, N-1.$$

H.4 Exercise 1.6: Noether's theorems for Yang–Mills fields

Let's consider YM theory minimally coupled with scalars and fermions

$$\mathcal{L} = -\frac{1}{4} F^a_{\mu\nu} F_a^{\mu\nu} - \bar\psi_n (\gamma^\mu D_\mu + m) \psi^n - \frac{1}{2} D_\mu \phi_m D^\mu \phi^m, \tag{H.1}$$

where m and n are indices in the (generally different) representation of G associated with the matter fields ψ and ϕ. The conjugated momenta of the various fields are

$$\begin{aligned}
\pi^{\mu\nu}_{A\,a} &= \frac{\delta S}{\delta \partial_\mu A^a_\nu} = -F^{\mu\nu}_a \\
\pi^{\mu m}_{\phi} &= \frac{\delta S}{\delta \partial_\mu \phi_m} = -D^\mu \phi^m \\
\pi^{\mu n}_{\psi} &= \frac{\delta S}{\delta \partial_\mu \psi_n} = -\bar\psi^n \gamma^\mu
\end{aligned}$$

So, using field transformations (1.117), the current K^μ defined in (1.126) is

$$K^\mu_\epsilon = -\frac{1}{e} F^{\mu\nu}_a D_\nu \epsilon^a + D^\mu \phi_m \epsilon^a \rho(T_a)^m{}_l \phi^l + \bar\psi_n \gamma^\mu \epsilon^a \rho(T_a)^n{}_l \psi^l.$$

The conserved current from the second Noether theorem is

$$k^\mu_\epsilon = K^\mu_\epsilon + \epsilon^a R^\mu_a \cdot E,$$

where R^μ_a is defined in eq. (1.125), while E is the left hand side of the Euler–Lagrange equation $E = 0$. The only field with $R^\mu_a \neq 0$ is the gauge field and its equation of motion is

$$D_\nu F_a^{\nu\mu} - e\phi^l \rho(T_a)_l{}^m D^\mu \phi_m - e\bar{\psi}_n \gamma^\mu \rho(T_a)^n{}_l \psi^l = 0,$$

Hence

$$k^\mu_\epsilon = \frac{1}{e}\epsilon^a D_\nu F_a^{\nu\mu} - \frac{1}{e} F_a^{\mu\nu} D_\nu \epsilon^a = \frac{1}{e}\partial_\nu(\epsilon^a F_a^{\nu\mu}).$$

The conservation of this current is a trivial consequence of F being antisymmetric.

If ϵ is constant, K^μ_ϵ reduces to

$$j^\mu_\epsilon = -\frac{1}{e} F_a^{\mu\nu} f_{bc}{}^a A^b_\mu \epsilon^c + D^\mu \phi_m \epsilon^a \rho(T_a)^m{}_l \phi^l + \bar{\psi}_n \gamma^\mu \epsilon^a \rho(T_a)^n{}_l \psi^l, \qquad \text{(H.2)}$$

which is just the Noether current associated to invariance of (H.1) under global gauge transformations. For $A_\mu = 0$ this is identical to the sum of the Noether currents (1.12) and (1.18). In the presence of a nonzero gauge field, we can write

$$j^\mu = [A_\nu, F^{\mu\nu}] + J^\mu, \qquad \text{(H.3)}$$

where J^μ is the covariantly conserved matter current defined by (1.135).

Then, removing the parameter ϵ and suppressing the algebra index, we have

$$\begin{aligned}
\partial_\mu j^\mu &= [\partial_\mu A_\nu, F^{\mu\nu}] + [A_\nu, \partial_\mu F^{\mu\nu}] + \partial_\mu J^\mu \\
&= \frac{1}{2}[F_{\mu\nu} - [A_\mu, A_\nu], F^{\mu\nu}] + [A_\nu, -[A_\mu, F^{\mu\nu}] + J^\nu] + D_\mu J^\mu - [A_\mu, J^\mu] \\
&= -\frac{1}{2}[[A_\mu, A_\nu], F^{\mu\nu}] - [A_\nu, [A_\mu, F^{\mu\nu}]] = 0, \qquad \text{(H.4)}
\end{aligned}$$

where we used the YM equation and covariant conservation of J^μ.

H.5 Exercise 1.7: covariant derivatives of nonlinear fields

Apply the covariant derivatives to the inverse of Equations (1.69):

$$\rho = \sqrt{\phi_1^2 + \phi_2^2 + \phi_3^2},$$
$$\Theta = \arctan\left(\frac{\sqrt{\phi_1^2 + \phi_2^2}}{\phi_3}\right),$$
$$\Phi = \arctan\left(\frac{\phi_2}{\phi_1}\right)$$

and use the Leibnitz rule $D_\mu f(\phi) = f'(\phi) D_\mu \phi$ to find

$$D_\mu \rho = \partial_\mu \rho,$$
$$D_\mu \Theta = \frac{1}{\phi_1^2 + \phi_2^2 + \phi_3^2}\left[\frac{\phi_3}{\sqrt{\phi_1^2 + \phi_2^2}}(\phi_1 D_\mu \phi_1 + \phi_2 D_\mu \phi_2) - \sqrt{\phi_1^2 + \phi_2^2} D_\mu \phi_3\right],$$
$$D_\mu \Phi = \frac{1}{\phi_1^2 + \phi_2^2}(\phi_1 D_\mu \phi_2 - \phi_2 D_\mu \phi_1).$$

Then, expanding the covariant derivatives and using (1.69) one gets

$$D_\mu \rho = \partial_\mu \rho,$$
$$D_\mu \Theta = \partial_\mu \Theta + \sin\Phi A_\mu^1 - \cos\Phi A_\mu^2,$$
$$D_\mu \Phi = \partial_\mu \Phi + \cot\Theta(\cos\Phi A_\mu^1 + \sin\Phi A_\mu^2) - A_\mu^3$$

This agrees with (1.143), when we use for K_a^α the explicit formula (B.15).

H.6 Exercise 1.8: London penetration depth

The equations of motion of the Proca Lagrangian (1.146) are

$$\Box A_\nu - \partial_\nu \partial_\mu A_\mu - m_A^2 A_\nu = 0,$$

thus the magnetic field $H_i = \epsilon_{ijk}\partial_j A_k$ must satisfy

$$\Box H_i = m_A^2 H_i,$$

whose general solution is

$$H^i = c_1^i e^{m_A \bar{x}\cdot\hat{r}_1} + c_2^i e^{-m_A \bar{x}\cdot\hat{r}_2}$$

with c_1 and c_2 constant vectors and r_1 and r_2 constant unit vectors ($|r_{1,2}| = 1$). Moreover, the divergence of H must be zero:

$$\partial_i H^i = m_A \hat{r}_1 \cdot \bar{c}_1 e^{m_A \bar{x}\cdot\hat{r}_1} - m_A \hat{r}_2 \cdot \bar{c}_2 e^{-m_A \bar{x}\cdot\hat{r}_2} = 0$$

From the boundary conditions, H must be constant on the plane $x = 0$, hence $\hat{r}_1 = \hat{r}_2 = (1, 0, 0)$. A field configuration with $c_1 \neq 0$ would lead to infinite energy, so we are forced to set $c_1 = 0$. From the null divergence condition there immediately follows $c_{2x} = 0$, therefore the component of the magnetic field orthogonal to the surface $x = 0$ does not penetrate in the superconductor. Without loss of generality we can therefore assume that the magnetic field for $x < 0$ is $\bar{H} = (0, 0, H)$. We have $c_2^i = H^i$, so the unique solution of the equations of motion is

$$H_x = 0, \qquad H_y = 0, \qquad H_z = He^{-m_A x},$$

where m_A is the inverse of the London penetration depth λ_L.

H.7 Exercise 1.9: weakly vanishing functions

The statement that F is weakly zero means that $F(q^i, p_a, \mathcal{A}_m(q)) = 0$. Thus

$$\delta F\big|_\Sigma = \left(\frac{\partial F}{\partial q^i} + \frac{\partial F}{\partial p_m}\frac{\partial \mathcal{A}_m}{\partial q^i}\right)\delta q^i + \frac{\partial F}{\partial p_a}\delta p_a.$$

The coefficients of δq^i and δp_a have to vanish separately. In the former we can replace $\mathcal{A}_m$ by $-\Phi_m$ and we get

$$\frac{\partial}{\partial q^i}\left(F - \Phi_m \frac{\partial F}{\partial p_m}\right) \approx 0.$$

Since the primary constraints do not depend on p_a at all, the second also gives

$$\frac{\partial}{\partial p_a}\left(F - \Phi_m \frac{\partial F}{\partial p_m}\right) \approx 0.$$

Finally, since $\frac{\partial \Phi_m}{\partial p_n} = \delta_m^n$ we also have

$$\frac{\partial}{\partial p_n}\left(F - \Phi_m \frac{\partial F}{\partial p_m}\right) \approx 0.$$

Thus the function $F - \Phi_m \frac{\partial F}{\partial p_m}$ is a constant, and the constant is zero, because it is zero on Σ.

H.8 Exercise 2.1: Bogomol'nyi bound for the kink

In terms of the prepotential $W(\phi)$, the static energy is

$$\int dx \left(\frac{1}{2}\frac{d\phi}{dx} + \frac{1}{2}W'^2\right) = \int dx \frac{1}{2}\left(\frac{d\phi}{dx} + W'\right)^2 - \int dx \frac{d\phi}{dx} W'.$$

The second integrand is equal to $\frac{dW}{dx}$, so the static energy is

$$E_S = \int dx \frac{1}{2}\left(\frac{d\phi}{dx} + W'\right)^2 - (W(\infty) - W(-\infty)).$$

In order to reproduce the explicit potential (2.2) one needs

$$W' = \sqrt{\frac{\lambda}{2}}\left(\phi^2 - f^2\right).$$

We can simply integrate this expression in ϕ to find

$$W = \sqrt{\frac{\lambda}{2}}\left(\frac{1}{3}\phi^3 - f^2\phi\right)$$

In this form

$$W(\infty) - W(-\infty) = -\sqrt{\frac{\lambda}{2}}\frac{2}{3}f^2(\phi(\infty) - \phi(-\infty)),$$

that is proportional to the topological charge (2.10).

It is then straightforward to prove that the kink solves the equation

$$\frac{d\phi}{dx} + W' = 0.$$

H.9 Exercise 2.2: interactions between kinks

The momentum of a generic field configuration in the half line to the left of b is

$$P = -\int_{-\infty}^{b} dx\, \dot{\phi}\phi',$$

so the force that it exerts is

$$\begin{aligned} F \equiv \frac{dP}{dt} &= -\int_{-\infty}^{b} dx\,(\ddot{\phi}\phi' + \dot{\phi}\dot{\phi}') \\ &= -\int_{-\infty}^{b} dx\,(\phi''\phi' - V_{,\phi}\phi' + \dot{\phi}\dot{\phi}') \\ &= -\int_{-\infty}^{b} dx\left(\frac{1}{2}\phi'^2 - V + \frac{1}{2}\dot{\phi}^2\right)' \\ &= -\frac{1}{2}\phi'^2 + V - \frac{1}{2}\dot{\phi}^2\Big|_{-\infty}^{b}. \end{aligned} \tag{H.5}$$

In passing to the second line, we have used the equation of motion

$$\ddot{\phi} = \phi'' - V_{,\phi}.$$

If we now let ϕ_1 and ϕ_2 be respectively an antikink at $x = -a$ and a kink at $x = a$ with a much larger than their linear dimension, we can write

$$\phi(x) = \phi_1(x) + \phi_2(x) + 1,$$

where 1 is a minimum of the potential. Since a kink rapidly approaches -1 for $x \ll a$, we can treat $\phi_2 + 1$ as a small perturbation. For this field, the force at $-a \ll b \ll a$ is

$$\begin{aligned} F &= -\frac{1}{2}(\phi_1' + \phi_2')^2 + V(\phi_1 + \phi_2 + 1)\Big|_{-\infty}^{b} \\ &= -\frac{1}{2}\phi_1'^2 - \frac{1}{2}\phi_2'^2 - \phi_1'\phi_2' + V(\phi_1) + V_{,\phi}(\phi_1)(\phi_2 + 1)\Big|_{-\infty}^{b} \\ &= -V(\phi_2) - \phi_1'\phi_2' + V_{,\phi}(\phi_1)(\phi_2 + 1)\Big|_{-\infty}^{b} \\ &= -\phi_1'\phi_2' + \phi_1''(\phi_2 + 1)\Big|_{-\infty}^{b}, \end{aligned}$$

where we used the individual energy equipartition for the kink and the antikink, the equation of motion (H.9) and considered that $V(\phi_2) \approx V(-1) = 0$ for $x \leq b$.

Looking at the specific forms of ϕ_1 and ϕ_2,

$$\phi_1(x) = -\frac{m}{\sqrt{\lambda}}\tanh\left[\frac{m}{\sqrt{2}}(x + a)\right], \qquad \phi_2(x) = \frac{m}{\sqrt{\lambda}}\tanh\left[\frac{m}{\sqrt{2}}(x - a)\right],$$

we can use the following approximations for $x \approx b$:

$$\phi_1(x) \approx \frac{m}{\sqrt{\lambda}}\left(-1 + 2e^{-\sqrt{2}m(x+a)}\right), \qquad \phi_2(x) \approx \frac{m}{\sqrt{\lambda}}\left(-1 + 2e^{\sqrt{2}m(x-a)}\right),$$

and we find

$$F \approx \frac{m^4}{\lambda} 16 e^{-2\sqrt{2}ma},$$

which is exponentially small for $a \gg 1/m$.

H.10 Exercise 2.4: critical vortices

At the critical point the static energy is

$$E_s = \int d^2x \left[\frac{1}{2}B^2 + \frac{1}{2}|D_i\phi|^2 + \frac{e^2}{8}(|\phi|^2 - f^2)^2\right] \tag{H.6}$$

In the second term we can write

$$|D_i\phi|^2 = |(D_1 + iD_2)\phi|^2 - i(D_1\phi)^* D_2\phi + iD_1\phi(D_2\phi)^*$$

which, integrating by parts (without discarding the surface term) and using the boundary conditions for ϕ, becomes

$$|(D_1 + iD_2)\phi|^2 - eB(f^2 - |\phi|^2) - i\varepsilon^{ij}(\partial_i\phi)^*\partial_j\phi.$$

Note that the last term is a total derivative and its boundary contribution is πf^2 times the winding number. On the other hand, the first and third term in (H.6) give

$$\frac{1}{2}\left[B + \frac{e}{2}(|\phi|^2 - f^2)\right]^2 - \frac{e}{2}B(|\phi|^2 - f^2),$$

thus the static energy becomes

$$E_s = \int d^2x \left\{\frac{1}{2}\left[B + \frac{e}{2}(|\phi|^2 - f^2)\right]^2 + \frac{1}{2}|(D_1 + iD_2)\phi|^2\right\} + \pi n f^2.$$

The two terms in the integral are manifestly positive, so

$$E_s \geq \pi n f^2 \tag{H.7}$$

and the inequality is saturated only if (2.147) are satisfied.

Using the Ansatz (2.99) with $\phi = F(r)e^{in\varphi}$ in order to describe vortices with arbitrary winding number, the Bogomol'nyi equations reduce to

$$A'(r) + \frac{A(r)}{r} + \frac{e}{2}(F(r)^2 - f^2) = 0 \tag{H.8}$$

$$F'(r) - n\frac{F(r)}{r} + eA(r)F(r) = 0, \tag{H.9}$$

while the equations of motion are

$$-e^2A^2F - \frac{1}{2}e^2F(F^2 - f^2) - n^2\frac{F}{r^2} + \frac{F'}{r} + F'' + 2en\frac{AF}{r} = 0$$

for the scalar field and

$$A'' + \frac{A'}{r} - \frac{A}{r^2} - e^2F^2A + en\frac{F^2}{r} = 0$$

for the vector potential A^μ.

If we call B_1 the l.h.s. eq. (H.8) and B_2 the l.h.s. of eq. (H.9), it is straightforward to see that the equation of motion for the scalar field is equal to $B_2' - eAB_2 + (n+1)\frac{B_2}{r} - eFB_1$, while the equation for the electromagnetic field is $B_1' - eFB_2$, thus the Bogomol'nyi equations directly imply the equations of motion.

From boundary conditions, the asymptotic behaviour of F and A for $r \to \infty$ is

$$F \sim f, \qquad A \sim \frac{n}{er},$$

hence we can linearize the equations of motion at large r for small perturbations δA and δF

$$r^2\partial_r^2\delta A + r\partial_r\delta A - (e^2f^2r^2 + 1)\delta A = 0$$
$$r^2\partial_r^2\delta B + r\partial_r\delta B - (e^2f^2r^2)\delta B = 0.$$

After a rescaling $efr \to x$, the two expressions reduce to particular cases of the defining equation of modified Bessel functions K_α

$$x^2\partial_x^2 y(x) + x\partial_x y(x) - (x^2 + \alpha^2)y(x) = 0.$$

So, finally,

$$F \sim f - k_S K_0(efr)$$
$$A \sim \frac{n}{er} - k_A K_0(efr),$$

with k_A and k_S constants.

H.11 Exercise 2.5: interaction of vortices

We write

$$\phi = Fe^{i\chi}.$$

where F is gauge invariant and χ transforms by a shift under gauge transformations. Whenever $F \neq 0$, we can define the gauge invariant vector field

$$a_\mu = A_\mu - \frac{1}{e}\partial_\mu \chi,$$

which is related to the covariant derivative of χ:

$$D_\mu \chi = \partial_\mu \chi - eA_\mu = -\frac{1}{e}a_\mu.$$

The relation between A_μ and a_μ has the form of a gauge transformation, and since the abelian magnetic field is gauge invariant,

$$\partial_i a_j - \partial_j a_i = \epsilon_{ij} B.$$

Since the action and the energy are gauge invariant, they can be rewritten entirely in terms if the gauge invariant fields F and a_μ:

$$E_S = \int d^2v \left[\frac{1}{2}B^2 + \frac{1}{2}e^2F^2a_i a_i + \frac{1}{2}\partial_i F \partial_i F + \frac{\lambda}{4}(F^2 - f^2)^2\right].$$

In these variables, the field equations are

$$\begin{aligned} \partial_i F_{ij} &= e^2F^2a_j \\ \partial^2 F - e^2F\,|a|^2 &= \lambda F(F^2 - f^2). \end{aligned}$$

The vacuum solution is defined by $F = f$, $a_i = 0$. A small deviation from the vacuum can be parameterized as

$$F = f(1 - \sigma).$$

In these variables the EOMs are

$$\partial_i F_{ij} - e^2f^2a_j = e^2f^2(\sigma^2 - 2\sigma)a_j \tag{H.10}$$

$$\partial^2\sigma - 2\lambda f^2\sigma = -e^2|a|^2 + e^2\sigma|a|^2 - 3\lambda f^2\sigma^2 + \lambda f^2\sigma^3. \tag{H.11}$$

The linear terms have been written on the l.h.s., the interactions in the r.h.s. . The energy functional is

$$\begin{aligned} E_S = \int d^2x \Big[&\frac{1}{2}B^2 + \frac{1}{2}e^2f^2|a|^2 - e^2f^2|a|^2\sigma \\ &+ \frac{1}{2}f^2(\partial\sigma)^2 + \lambda f^4\sigma^2 + \frac{1}{2}e^2f^2|a|^2\sigma^2 - \lambda f^4\sigma^3 + \frac{\lambda}{4}f^4\sigma^4\Big]. \end{aligned} \tag{H.12}$$

Now the vacuum is $\sigma = 0$, $a_i = 0$, and its energy is zero. We will be interested in calculating the energy of certain solutions of the equations of motion. If σ and a_i are small, one would naively be tempted to approximate the energy by keeping only terms quadratic in these variables, namely the first two terms in each line above. However, this would be incorrect. Notice that in (H.11) $|a|^2$ acts as a source term for σ. Thus a term that looks already quadratic in the small variables, is seen to be only linear on shell.

In practice, to calculate the energy of a solution we can use the EOMs. Thus we add to the energy integrand the EOM of σ, multiplied by $f^2\sigma$, in such a way as to get rid of the term linear in σ in the energy functional (the "source" term). Integrating by parts the term $f^2\sigma\partial^2\sigma$ (this is allowed if σ goes to zero sufficiently fast at infinity), we get

$$E_S = \int d^2x \Big[\frac{1}{2}B^2 + \frac{1}{2}e^2 f^2 |a|^2 - \frac{1}{2}f^2(\partial\sigma)^2 - \lambda f^4\sigma^2 - \frac{1}{2}e^2 f^2 |a|^2\sigma^2 + 2\lambda f^4\sigma^3 - \frac{3\lambda}{4}f^4\sigma^4 \Big]. \quad \text{(H.13)}$$

If we now neglect terms of order cubic and quartic in the variables σ and a_i, we see that the part of the energy due to the scalar field has exactly the opposite sign as in the original expression. This result holds for the single vortex solutions $\sigma^{(i)}$, $\chi^{(i)}$ and $a^{(i)}$, but also for the Abrikosov ansatz with $\sigma = \sigma^{(1)} + \sigma^{(2)}$, $\chi = \chi^{(1)} + \chi^{(2)}$ and $a = a^{(1)} + a^{(2)}$, so we can compute the interaction energy

$$E_{\text{int}} = \Delta E = E_S - E_S^{(1)} - E_S^{(2)} \quad \text{(H.14)}$$

starting from (H.13). The fields $\sigma^{(i)}$ and $a^{(i)}$ are exponentially decreasing functions of the distance from the respective vortex cores, and we retain terms linear in each of these fields. Let us consider separately the three terms

$$\frac{1}{2}B^2 + \frac{1}{2}|D\phi|^2 + V.$$

For the magnetic energy $E_{\text{mag}} = \frac{1}{2}\int d^2x\, B^2$, we have

$$\Delta E_{\text{mag}} = \int d^2x\, B^{(1)}B^{(2)}$$

The magnetic field for a vortex located in the origin is

$$B(r) = -a'(r) - \frac{a(r)}{r}.$$

Using the explicit form (2.148a)

$$B(r) = -k_A m_A \left(K_1'(m_A r) + \frac{1}{m_A r} K_1(m_A r) \right) = k_A m_A K_0(m_A r).$$

Thus finally shifting the origins of the coordinates

$$\Delta E_{\text{mag}} = k_A^2 m_A^2 \int d^2x \, K_0(m_A |x - x_{(1)}|) K_0(m_A |x - x_{(2)}|).$$

Now we come to the kinetic term of the scalars. The integrand in ΔE_1 is

$$-f^2 \partial_i \sigma^{(1)} \partial_i \sigma^{(2)} + f^2 e^2 a_i^{(1)} a_i^{(2)}.$$

Now inserting the asymptotic forms,

$$\begin{aligned} \Delta E_1 = \int d^2x \Big[& - k_S^2 \partial_i K_0(m_S |x - x_{(1)}|) \partial_i K_0(m_S |x - x_{(2)}|) \\ & + \cos(\varphi_{(1)} - \varphi_{(2)}) f^2 e^2 k_A^2 K_1(m_A |x - x_{(1)}|) K_1(m_A |x - x_{(2)}|) \Big]. \end{aligned}$$

Where $\varphi_{(1)}$and $\varphi_{(2)}$ are the angular coordinates of x in the reference frames centered respectively in $x_{(1)}$ and $x_{(2)}$. Since

$$\partial_x K_0(mx) = -m K_1(mx),$$

we can easily recast it in the form

$$\Delta E_1 = \int d^2x \Big[(f^2 e^2 k_A^2 - k_S^2) \partial_i K_0(m_S |x - x_{(1)}|) \partial_i K_0(m_S |x - x_{(2)}|) \Big].$$

For the potential energy $E_2 = \int d^2x V$ we get

$$\Delta E_2 = -2\lambda f^2 \int d^2x \, \sigma^{(1)} \sigma^{(2)}.$$

and using the explicit form (2.148b)

$$\Delta V = -2\lambda f^2 k_S^2 \int d^2x \, K_0(m_S |x - x_{(1)}|) K_0(m_S |x - x_{(2)}|).$$

Summing up the different contributions,

$$\begin{aligned} E_{\text{int}} = \int d^2x \Big\{ & k_A^2 \big[m_A^2 K_0(m_A |x - x_{(1)}|) K_0(m_A |x - x_{(2)}|) \\ & + \partial_i K_0(m_S |x - x_{(1)}|) \partial_i K_0(m_S |x - x_{(2)}|) \big] \\ & - k_S^2 \big[\partial_i K_0(m_S |x - x_{(1)}|) \partial_i K_0(m_S |x - x_{(2)}|) \\ & + m_S^2 K_0(m_S |x - x_{(1)}|) K_0(m_S |x - x_{(2)}|) \big] \Big\} \end{aligned}$$

In $d = 2$, a relevant property of K_0 is

$$m^2 K_0(m|x|) = 2\pi\delta^2(x) + \partial^2 K_0(m|x|),$$

so, after an integration by parts, the internal energy reduces to

$$E_{\text{int}} = 2\pi k_A^2 K_0(m_A|x_{(1)} - x_{(2)}|) - 2\pi k_S^2 K_0(m_A|x_{(1)} - x_{(2)}|),$$

where $C_A = 2\pi k_A^2$ and $C_\phi = 2\pi k_S^2$. Notice that at criticality, due to the Bogomol'nyi equations, $k_S = k_A$, so the interaction energy is zero.

H.12 Exercise 2.7: formulae for the monopole

Consider Equation (2.127)

$$\mathcal{F}_{\mu\nu} = \hat{\phi}^a F_{\mu\nu} - \frac{1}{e}\varepsilon_{abc}\hat{\phi}^a D_\mu \hat{\phi}^b D_\nu \hat{\phi}^c.$$

It is equal to

$$\begin{aligned}\mathcal{F}_{\mu\nu} = {} & \hat{\phi}^a \partial_\mu A^a_\nu - \hat{\phi}^a \partial_\nu A^a_\mu + e\varepsilon_{abc}\hat{\phi}^a A^b_\mu A^b_\nu \\ & - \frac{1}{e}\varepsilon_{abc}\hat{\phi}^a(\partial_\mu \hat{\phi}^b + e\varepsilon_{bde}A^d_\mu \hat{\phi}^e)(\partial_\nu \hat{\phi}^c + e\varepsilon_{ckl}A^k_\nu \hat{\phi}^l)\end{aligned}$$

and since $\varepsilon_{ijk}\varepsilon_{imn} = \delta_{jm}\delta_{kn} - \delta_{jn}\delta_{km}$, it can be expanded into

$$\begin{aligned}& \hat{\phi}^a \partial_\mu A^a_\nu - \hat{\phi}^a \partial_\nu A^a_\mu - \frac{1}{e}\varepsilon_{abc}\hat{\phi}^a \partial_\mu \hat{\phi}^b \partial_\nu \hat{\phi}^c \\ & + \hat{\phi}^a \hat{\phi}^a A^b_\nu \partial_\mu \hat{\phi}^b - \hat{\phi}^a \hat{\phi}^a A^b_\mu \partial_\nu \hat{\phi}^b \\ & - \hat{\phi}^a A^a_\nu \hat{\phi}^b \partial_\mu \hat{\phi}^b + \hat{\phi}^a A^a_\mu \hat{\phi}^b \partial_\nu \hat{\phi}^b.\end{aligned}$$

The last two terms are identically zero because $\hat{\phi}^a \hat{\phi}^a = 1$, so $0 = \partial_\mu(\hat{\phi}^a \hat{\phi}^a) = 2\hat{\phi}^a \partial_\mu \hat{\phi}^a$, while the rest becomes

$$\partial_\mu(\hat{\phi}^a A^a_\nu) - \partial_\nu(\hat{\phi}^a A^a_\mu) - \frac{1}{e}\varepsilon_{abc}\hat{\phi}^a \partial_\mu \hat{\phi}^b \partial_\nu \hat{\phi}^c,$$

that is exactly Equation (2.126).

H.13 Exercise 2.8: monopole in unitary gauge

The Higgs field transforms as $\phi'^a = T_{ab}\phi^b$ and in the ansatz (2.130) it is proportional to $\frac{x^a}{r} = \hat{x}^a$, so we just need to compute

$$T\hat{x} = \begin{pmatrix} \hat{x}_1 \frac{\hat{x}_2^2+\hat{x}_1^2\hat{x}_3}{1-\hat{x}_3^2} - \frac{\hat{x}_1\hat{x}_2^2}{1+\hat{x}_3} - \hat{x}_1\hat{x}_3 \\ -\frac{\hat{x}_1\hat{x}_2^2}{1+\hat{x}_3} + \hat{x}_2 \frac{\hat{x}_1^2+\hat{x}_2^2\hat{x}_3}{1-\hat{x}_3^2} - \hat{x}_2\hat{x}_3 \\ \hat{x}_1^2 + \hat{x}_2^2 + \hat{x}_3^2 \end{pmatrix}$$

Taking in account that $\hat{x}_1^2+\hat{x}_2^2+\hat{x}_3^2 = 1$, the last expression can be simplified to

$$T\hat{x} = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}.$$

So the scalar field after this transformation is $\phi'^a = \delta_{a3}F(x)$, that corresponds to the unitary gauge.

H.14 Exercise 3.7: symmetric gauge fields

Under an infinitesimal rotation $\delta x_a = \epsilon_{ab}x_b$, the gauge field transforms as

$$\begin{aligned} \delta_\epsilon A_a &= \frac{\partial A_a}{\partial x_b}\delta x_b - \epsilon_{ab}A_b \\ &= (\Xi_{ab}\epsilon_{bc} - \epsilon_{ab}\Xi_{bc})x_c f(r). \end{aligned}$$

On the other hand, under an infinitesimal gauge transformation with constant gauge parameter $\omega = \frac{1}{2}\omega_{cd}\Xi_{cd}$, the gauge field transforms as

$$\begin{aligned} \delta_\omega A_a &= [A_a, \omega] \\ &= -(\Xi_{ab}\omega_{bc} - \omega_{ab}\Xi_{bc})x_c f(r). \end{aligned}$$

Thus, if we choose $\omega_{ab} = \epsilon_{ab}$, we have a symmetry $\delta A_a = 0$.

The monopole ansatz (2.130) is of the form (3.137), with the generators in the spinor representation given by $\Xi_{ab} = \epsilon_{abc}\tau_c = \frac{i}{2}\epsilon_{abc}\sigma_c$.

The instanton/anti-instanton ansatz (3.53), (3.55) is also of this form, in the case $N = 4$. Recall that $\mathfrak{so}(4) = \mathfrak{su}(2) \oplus \mathfrak{su}(2)$ and that in the adjoint

representation, the two $\mathfrak{su}(2)$ subalgebras correspond to the self-dual and anti-self-dual 4×4 real matrices respectively. The full algebra is generated by

$$\Xi_{\mu\nu} = \begin{pmatrix} \Sigma_{\mu\nu} & 0 \\ 0 & \bar{\Sigma}_{\mu\nu} \end{pmatrix}$$

that satisfy the commutation relations (3.136).

H.15 Exercise 3.8: the BPST instanton on the sphere

We use x^μ, with $\mu = 1,2,3,4$, as coordinates in Euclidean space $\mathbb{R}^4$ and z^a with $a = 1,2,3,4,5$ as coordinates in $\mathbb{R}^5$. The sphere is embedded in $\mathbb{R}^5$ by the condition

$$(z^1)^2 + (z^2)^2 + (z^3)^2 + (z^4)^2 + (z^5)^2 = \lambda^2.$$

The stereographic map is

$$z^\mu = \frac{2\lambda^2 x^\mu}{r^2+\lambda^2} \qquad z^5 = \frac{\lambda^2 - r^2}{\lambda^2 + r^2}\lambda \tag{H.15}$$

where $r^2 = (x^1)^2 + (x^2)^2 + (x^3)^2 + (x^4)^2$. (This agrees with the results of Exercise 1.5 up to a trivial rescaling.)

Under this map a gauge field A_μ in $\mathbb{R}^4$ becomes a gauge field $\hat{A}_a$ on S^4. They are related by

$$\begin{aligned} A_\mu &= \frac{\partial z^a}{\partial x^\mu}\hat{A}_a \\ &= \frac{1}{\lambda^2+x^2}\left(2\lambda^2\hat{A}_\mu - \frac{4\lambda^2 x_\mu x^\nu}{\lambda^2+x^2}\hat{A}_\nu - 2\lambda x_\mu \hat{A}_5 - 2\lambda x_\mu \frac{\lambda^2 - x^2}{\lambda^2+x^2}\hat{A}_5\right) \\ &= \frac{1}{\lambda^2+x^2}\left(2\lambda^2\hat{A}_\mu - 2\lambda x_\mu \hat{A}_5 - 2\lambda^2 x_\mu z^a \hat{A}_a\right) \\ &= \frac{2\lambda^2}{\lambda^2+x^2}\left(2\hat{A}_\mu - \frac{x_\mu}{\lambda}\hat{A}_5\right). \end{aligned} \tag{H.16}$$

In the last step we used that $\hat{A}$ is tangential to the sphere. To obtain the inverse relation one can contract relation (H.16) with x^μ, giving

$$x^\mu A_\mu = \frac{2\lambda}{\lambda^2+x^2}\left(\lambda x^\mu \hat{A}_\mu - x^2 \hat{A}_5\right) = z^\mu \hat{A}_\mu - \frac{2\lambda x^2}{\lambda^2+x^2}\hat{A}_5$$

On the sphere $z^\mu \hat{A}_\mu = -z^5 \hat{A}_5$, hence

$$\hat{A}_5 = -\frac{x^\mu}{\lambda} A_\mu \qquad \text{and} \quad \hat{A}_\mu = \frac{\lambda^2 + x^2}{2\lambda^2} A_\mu - \frac{x^\nu x_\mu}{\lambda^2} A_\nu$$

We take the generators of $\mathfrak{so}(5)$ to be

$$\Xi_{\mu\nu} = \begin{pmatrix} \Sigma_{\mu\nu} & 0 \\ 0 & \bar{\Sigma}_{\mu\nu} \end{pmatrix} \qquad \text{for} \quad \mu, \nu = 1, 2, 3, 4$$

and

$$\Xi_{i5} = \frac{1}{2} \begin{pmatrix} 0 & \sigma^i \\ \sigma^i & 0 \end{pmatrix}, \qquad \Xi_{45} = \frac{i}{2} \begin{pmatrix} 0 & -I \\ I & 0 \end{pmatrix}$$

with $i = 1, 2, 3$. They satisfy the commutation relations (3.136).

The ansatz (3.137) gives us a $SO(5)$ Yang–Mills field in $\mathbb{R}^5$

$$\hat{A}_a = c\, \Xi_{ab} x^b, \tag{H.17}$$

that is tangential to S^4 and is invariant under simultaneous rotations and gauge transformations, with the same parameter. (Since the radius of the sphere is fixed, the function $f(r)$ becomes just a constant c).

A gauge transformation of $\hat{A}$ has the form

$$\hat{A}_a \to \hat{A}'_a = U^{-1} \hat{A}_a U + U^{-1} i \frac{z^b}{\lambda^2} M_{ba} U$$

where $z^b M_{ba} = z^2 \partial_a - z^a z^b \partial_b$ is the derivative tangential to the sphere. We can eliminate the components of $\hat{A}$ in the directions $\Sigma_{\mu 5}$ by a transformation of the form

$$U = e^{i f(z_5) \Xi_{\mu 5} z^\mu}.$$

The fifth component is

$$\begin{aligned} \hat{A}'_5 &= U^{-1} \hat{A}_5 U + U^{-1} i \frac{z^b}{\lambda^2} (-i z_b \partial_5 + i z_5 \partial_b) U \\ &= U^{-1} \frac{1}{\lambda^2} \left(-i z^\mu + i z^\nu z_\nu f'(z_5) z^\mu - i z_5 f(z_5) z^\mu \right) \Xi_{\mu 5} U \end{aligned}$$

Demanding that this vanishes results in a differential equation for f

$$f'(z_5)(\lambda^2 - z_5^2) - z_5 f(z_5) - 1 = 0,$$

where we have used $z^\nu z_\nu = \lambda^2 - z_5^2$ on the hypersphere. This is solved by

$$f(z_5) = \frac{\arccos \frac{z_5}{\lambda^2}}{\sqrt{\lambda^2 - z_5^2}}.$$

After the gauge fixing $\hat{A}_5 = 0$, the field configuration has a residual $SO(4)$ symmetry, and takes the form

$$\hat{A}_\mu = -\frac{ic}{\lambda(\lambda + z_5)} \Xi_{\mu\nu} z^\nu.$$

Using the transformation (H.15), we immediately see that this corresponds to

$$A_\mu = -\frac{2ic}{\lambda^2 + x^2} \Xi_{\mu\nu} x^\nu$$

in $\mathbb{R}^4$. When we put $c = 1$, this is recognized as the $SO(4)$ field of an instanton and anti-instanton put together. The choice $c = 1$ is the only one for which the ansatz (H.17) solves the YM equations [JaR76]. Also note that we have chosen the radius of the sphere to be exactly the size of the instanton. In this way the action density is constant.

H.16 Exercise 3.9: quantum fluctuations around the instanton

We will compute $W_1(\bar{A}) = -\log Z_1(\bar{A})$ at zero external source. In this case the expectation value of the quantum field a_μ will be zero and we can identify $D_\mu = \bar{D}_\mu$. Thus the ghost operator is simply the Laplacian on scalars $\Delta^{(0)} = -\bar{D}^2$. Then, the effective action is formally

$$\begin{aligned} W_1(\bar{A}) = S_{YM}(\bar{A}) + \frac{1}{2}\mathrm{Tr}\log\left(\lambda^2\Delta^{(1)}(\bar{A})\right) - \frac{1}{2}\mathrm{Tr}\log\left(\lambda^2\Delta^{(1)}(0)\right) \\ - \mathrm{Tr}\log\left(\lambda^2\Delta^{(0)}(\bar{A})\right) + \mathrm{Tr}\log\left(\lambda^2\Delta^{(0)}(0)\right). \end{aligned} \tag{H.18}$$

We work on a sphere of radius λ, so we have naturally inserted factors λ^2 for dimensional reasons. For an operator $\mathcal{O}$, with eigenvalues λ_n and multiplicities m_n, we have

$$\mathrm{Tr}\log\left(\lambda^2\mathcal{O}\right) = \sum m_n \log\left(\lambda^2\lambda_n\right). \tag{H.19}$$

We thus need the spectra of these operators.

Since our problem has $O(5)$ symmetry, the calculation of the eigenvalues and their multiplicities is essentially a problem of group theory: the eigenvalues are the Casimirs of the representations and the multiplicities their dimensions. We can split the field a_μ^a into a pure gauge (longitudinal) part and a "transverse" part.

From the definitions (3.109), (3.111) one easily sees that given a Lie-algebra valued function ϕ,

$$\Delta^{(1)} D_\mu \phi = D_\mu \Delta^{(0)} \phi. \tag{H.20}$$

Therefore, $\Delta^{(1)}$ acting on longitudinal vectors has the same spectrum as $\Delta^{(0)}$ on scalars. The derivation of this spectrum, as well as the spectrum of $\Delta^{(1)}$ acting on transverse vectors, can be found in [BeP77]. For the operators of interest, the eigenvalues and multiplicities are given in the following table:

operator	$\lambda_n \lambda^2$	m_n	$n_{\min}$	N.Z.
	$n(n+3)-4$	$\frac{1}{6}(n+1)(n+2)(2n+3)$	1	5
	$n(n+3)-2$	$\frac{1}{2}n(n+3)(2n+3)$	2	0
$\Delta^{(1T)}(\bar{A})$	$n(n+3)+2$	$\frac{5}{6}(n-1)(n+4)(2n+3)$	2	0
$\Delta^{(1T)}(0)$	$n(n+3)+2$	$\frac{3}{2}n(n+3)(2n+3)$	1	0
$\Delta^{(1L)}(\bar{A})$	$n(n+3)-2$	$\frac{1}{2}n(n+3)(2n+3)$	1	0
$\Delta^{(1L)}(0)$	$n(n+3)$	$\frac{1}{2}(n+1)(n+2)(2n+3)$	1	0
$\Delta^{(0)}(\bar{A})$	$n(n+3)-2$	$\frac{1}{2}n(n+3)(2n+3)$	1	0
$\Delta^{(0)}(0)$	$n(n+3)$	$\frac{1}{2}(n+1)(n+2)(2n+3)$	0	3

All the series of eigenvalues start from $n_{\min}$; N.Z. is the number of zero modes. We write $\Delta^{(1T)}$ and $\Delta^{(1L)}$ for the operator $\Delta^{(1)}$ acting on transverse and longitudinal fields. As mentioned before, the spectra of $\Delta^{(1L)}$ and $\Delta^{(0)}$ are the same, except that the mode $n=0$ does not exist for the operator $\Delta^{(1L)}(0)$, because the three zero modes of $\Delta^{(0)}$ are constants, and then $D_\mu \phi_0 = \partial_\mu \phi_0 = 0$. Because of this, when we expand $\det\Delta^{(1)} = \det\Delta^{(1)T} \det\Delta^{(1)L}$ we have some partial

cancellations with the ghost operators and we can write formally

$$Z_1 = e^{-S_{YM}(\bar{A})} \frac{\sqrt{\det\Delta^{(1T)}(0)}}{\sqrt{\det\Delta^{(1T)}(\bar{A})}} \frac{\sqrt{\det'\Delta^{(0)}(0)}}{\sqrt{\det\Delta^{(0)}(\bar{A})}} \frac{\det\Delta^{(0)}(\bar{A})}{\det\Delta^{(0)}(0)}$$
$$= \frac{\sqrt{\det\Delta^{(1T)}(0)}}{\sqrt{\det\Delta^{(1T)}(\bar{A})}} \frac{\sqrt{\det\Delta^{(0)}(\bar{A})}}{\sqrt{\det'\Delta^{(0)}(0)}} \frac{1}{0^3}. \tag{H.21}$$

In the first line, the three fractions are the contributions of transverse vectors, longitudinal vectors and ghosts, respectively and primes indicate that zero modes are omitted. The term $1/0^3$ is due to the zero modes of $\Delta^{(0)}(0)$ and will be dealt with below, together with the zero modes of $\Delta^{(1T)}(\bar{A})$.

Since the multiplicities grow like n^3, each sum is quartically divergent. The sums can be regulated using the Schwinger proper time method

$$\mathrm{Tr}\log\left(\lambda^2\mathcal{O}\right) = -\int_0^\infty \frac{ds}{s} K_{\mathcal{O}}(s), \tag{H.22}$$

where

$$K_{\mathcal{O}}(s) = \mathrm{Tr}e^{-s\mathcal{O}} = \sum_n m_n e^{-s\lambda_n} \tag{H.23}$$

is the heat kernel of the operator $\mathcal{O}$.

The ultraviolet corresponds to the lower end of the s-integration and the infrared to the upper end, as one understands by noting that s has dimension of length squared. Infrared divergences are only associated to the eight zero modes, and are dealt with by going to collective coordinates. In particular, the 5 zero modes of $\Delta^{(1T)}$ can be related to translations and dilatations in flat space, while the three zero modes of $\Delta^{(0)}$ are related to the rotational moduli and produce an unimportant factor of the volume of $SU(2)$. Remember that for each collective coordinate we have a Jacobian, that is given by $\sqrt{S_{cl}}$, and that the action of the instanton is $\frac{8\pi^2}{g_B^2}$, where we have affixed a "B" to the YM coupling to remind us that it is the "bare" coupling, since the instanton is a solution of the classical YM equations. Thus the zero modes give rise to integrals

$$\text{constant} \times \left(\frac{8\pi^2}{g_B^2}\right)^4 \int d^4x_0 \int \frac{d\lambda}{\lambda^5}.$$

The factor λ^{-5} has been inserted on the basis of dimensional analysis.

For small s the heat kernel can be expanded as

$$\begin{aligned} TrK_{\mathcal{O}}(s) &\sim \frac{1}{(4\pi s)^2}\int d^4x\left[b_0(\mathcal{O}) + sb_2(\mathcal{O}) + s^2 b_4(\mathcal{O}) + ...\right] \\ &= \frac{a}{s^2} + \frac{b}{s} + c + O(s). \end{aligned} \tag{H.24}$$

and the traces are regulated by putting a cutoff Λ_{UV}^2 in the lower end of the s-integration. Again for dimensional reasons, and since we are now interested in isolating the ultraviolet divergences, we also put an infrared cutoff λ^2 on the upper end of the integration. Then one sees that the first three terms give quartic, quadratic and logarithmic divergences, repectively. In particular, the logarithmic divergence is

$$\frac{1}{2}\mathrm{Tr}\log\left(\lambda^2\mathcal{O}\right) = -c\log\left(\lambda^2\Lambda_{UV}\right). \tag{H.25}$$

The so-called Seeley–DeWitt coefficients $b_k(\mathcal{O})$ are gauge invariant combinations of the background field and its derivatives, for which general formulas are available. However, given that we know the spectra exactly, it is more convenient to calculate directly the coefficients a, b, c by evaluating the traces (H.19) with the Euler–Maclaurin formula

$$\sum_{n=n_{\min}}^{\infty} f(n) = \int_{n_{\min}}^{\infty} dx f(x) + \frac{1}{2}(f(n_{\min}) + f(\infty)) + \sum_{k=2}^{\infty}\frac{B_k}{k!}\left(f^{(k-1)}(x)\right)\Big|_{n_{\min}}^{\infty}, \tag{H.26}$$

where B_n are the Bernoulli numbers. Since we are only interested in the leading terms of the expansion in s, it is enough to keep the terms $k = 2, 3$ in the last sum. The zero modes have to be retained in this sum, as explained above.

The results of these sums are reported in following table for each series of eigenvalues:

operator	a	b	c_1	c_0
$\Delta^{(1)T}(\bar{A})$	$\frac{1}{6}$	1	$-\frac{271}{90}$	5
	$\frac{1}{2}$	1	$-\frac{281}{30}$	0
	$\frac{5}{6}$	-5	$\frac{239}{18}$	0
$\Delta^{(1)T}(0)$	$\frac{3}{2}$	-3	$-\frac{281}{10}+30$	0
$\Delta^{(0)}(\bar{A})$	$\frac{1}{2}$	1	$\frac{19}{30}$	0
$\Delta^{(0)}(0)$	$\frac{1}{2}$	1	$-\frac{61}{30}$	3

We have split $c = c_0 + c_1$: the column c_0 gives the contribution to c of the zero modes and the column c_1 gives the contribution of the nonzero modes. In the fourth line, 30 is the contribution of the mode $n = 1$. All the remaining sums for the vector operator start from $n = 2$ and those for the ghost operator start from $n = 1$.

We observe that in (H.18) the coefficients a and b sum up to zero, separately for the vectors and ghosts. This means that there are no quartic and quadratic divergences. This is the desired effect of having normalized the functional integral with the integral without instanton. There remain the logarithmic divergences. From the preceding table, we find

$$c_{\text{tot}} = -\frac{271}{90} + 5 - \frac{281}{90} + \frac{239}{18} + \frac{281}{10} - 30 - \left(-\frac{19}{30} + \frac{61}{30}\right) + 6 = \frac{22}{3}$$

so the effective action is

$$\begin{aligned} W_1(\lambda) &= S_{YM}(\bar{A}) + c_{\text{tot}} \log\left(\lambda^2 \Lambda_{UV}\right) + \text{constant} \\ &= \frac{8\pi^2}{g_B^2} - \frac{11}{3} \log\left(\lambda^2 \Lambda_{UV}^2\right) + \text{constant} \end{aligned} \tag{H.27}$$

and the amplitude is

$$Z_1 = \text{constant} \times \left(\frac{8\pi^2}{g_B^2}\right)^4 \int d^4 x_0 \int \frac{d\lambda}{\lambda^5} e^{-W_1(\lambda)}. \tag{H.28}$$

H.17 Exercise 6.1: the ABJ anomaly in $d = 4$

Going to momentum space and performing the trace and the integration over y we arrive at the following expression:

$$-4e^2\varepsilon^{\mu\nu\rho\lambda}F_{\alpha\mu}\epsilon^{\alpha}\int\frac{d^4k}{(2\pi)^4}k_\nu\tilde{A}_\rho(k)e^{-ik\left(x+\frac{\epsilon}{2}\right)}\int\frac{d^4p}{(2\pi)^4}e^{ip\epsilon}\frac{p_\lambda}{p^2(p+k)^2}.$$

The last integral is equal to

$$-i\frac{\epsilon^\beta}{\epsilon^2}\int\frac{d^4p}{(2\pi)^4}\frac{\partial e^{ip\epsilon}}{\partial p^\beta}\frac{p_\lambda}{p^2(p+k)^2}=i\frac{\epsilon^\beta}{\epsilon^2}\int\frac{d^4p}{(2\pi)^4}e^{ip\epsilon}\frac{\partial}{\partial p^\beta}\frac{p_\lambda}{p^2(p+k)^2}$$

so we remain with

$$-4ie^2\varepsilon^{\mu\nu\rho\lambda}F_{\alpha\mu}\frac{\epsilon^\alpha\epsilon^\beta}{\epsilon^2}\int\frac{d^4k}{(2\pi)^4}k_\nu\tilde{A}_\rho(k)e^{-ik\left(x+\frac{\epsilon}{2}\right)}\int\frac{d^4p}{(2\pi)^4}e^{ip\epsilon}\frac{\partial}{\partial p^\beta}\frac{p_\lambda}{p^2(p+k)^2}.$$

At this point we perform the average over the directions of ϵ

$$-ie^2\varepsilon^{\mu\nu\rho\lambda}F_{\alpha\mu}\int\frac{d^4k}{(2\pi)^4}k_\nu\tilde{A}_\rho(k)e^{-ikx}\int\frac{d^4p}{(2\pi)^4}\frac{\partial}{\partial p^\alpha}\frac{p_\lambda}{p^2(p+k)^2}.$$

The dependence on ϵ has disappeared, so this expression survives in the limit $\epsilon\to 0$. The second integral is a surface term. It is independent of k, giving

$$\int\frac{d^4p}{(2\pi)^4}\frac{\partial}{\partial p^\alpha}\frac{p_\nu}{p^2(p-k)^2}=\frac{1}{8\pi^2}\delta^\alpha_\nu.$$

The remaining integral is the derivative of the electromagnetic field

$$\int\frac{d^4k}{(2\pi)^4}k_\nu\tilde{A}_\rho(k)e^{-ikx}=i\partial_\nu A_\rho,$$

so, collecting all the pieces, the divergence of the current is

$$\frac{e^2}{16\pi^2}\varepsilon^{\mu\nu\rho\sigma}F_{\mu\nu}F_{\rho\sigma}.$$

H.18 Exercise 6.3: anomalies in commutators

Using the representation of the gamma matrices $\gamma^0 = i\sigma_2$, $\gamma^1 = \sigma_1$ and $\gamma^A = \sigma_3$, the fermionic action can be split into

$$-\int d^2x\,\bar{\psi}\gamma^\mu\partial_\mu\psi=i\int d^2x\left[\psi_L^\dagger\partial_-\psi_L+\psi_R^\dagger\partial_+\psi_R\right],\tag{H.29}$$

where $\partial_\pm = \partial_0 \pm \partial_1$. We see that the two chiralities correspond to left- and right-movers. We restrict our attention to the right-handed sector and we drop the subscript R from the fermions from now on.

The canonical quantization of the resulting theory, is obtained by imposing the canonical anti-commutation relation at fixed t:

$$\{\psi^\dagger_\alpha(x), \psi^\beta(y)\} = \delta^\beta_\alpha \delta(x-y). \tag{H.30}$$

The equation of motion for ψ_α is solved by:

$$\psi_\alpha = \int_0^{+\infty} \frac{dk}{2\pi} [b_\alpha(k) e^{ik(x-t)} + a^\dagger_\alpha(k) e^{-ik(x-t)}]. \tag{H.31}$$

The relation (H.30), leads to

$$\{b_\alpha(k), b^\dagger_\beta(k')\} = 2\pi \delta_{\alpha\beta} \delta(k-k'), \tag{H.32}$$

$$\{a_\alpha(k), a^\dagger_\beta(k')\} = 2\pi \delta_{\alpha\beta} \delta(k-k'), \tag{H.33}$$

a, b and $a^\dagger$, $b^\dagger$ are the usual annihilation and creation operators, defining the Fock space:

$$b_\alpha(k)|0\rangle = 0, \qquad b^\dagger_\alpha(k)|0\rangle = |k, \alpha, +\rangle, \tag{H.34}$$

$$a_\alpha(k)|0\rangle = 0, \qquad a^\dagger_\alpha(k)|0\rangle = |k, \alpha, -\rangle, \tag{H.35}$$

$|0\rangle$ is the Fock vacuum and $|k, \alpha, \pm\rangle$ one-particle states.

Let us consider the current operator $J_a = -i\psi^\dagger_\alpha T_a{}^\alpha{}_\beta \psi^\beta$, It will be useful to smear it with a test function $\epsilon(x)$ with values in the Lie algebra of $SU(N)$: $J_\epsilon = \int dx\, \epsilon^a J_a$. The current generates global $SU(N)$ transformations and obeys the classical current algebra

$$[J_{\epsilon_1}, J_{\epsilon_2}] = J_{[\epsilon_1, \epsilon_2]} \tag{H.36}$$

However at quantum level J_ϵ is not a well defined operator: we are just considering two quantum operator at the same point and ultraviolet divergences arise in this limit. In two dimensions a simple solution is normal ordering

$$J^a(x) = -i : \psi^\dagger_\alpha T^a_{\alpha\beta} \psi_\beta :.$$

This is sufficient to avoid all the ultraviolet divergences in two dimensions.

With this definition let us computethe commutator of two charge densities $\rho_{\alpha\alpha'} =: \psi^\dagger_\alpha \psi_{\alpha'} :$. Wick's theorem gives

$$\begin{aligned}[\rho_{\alpha\alpha'}(x), \rho_{\beta\beta'}(y)] = &\{\psi_{\alpha'}(x), \psi^\dagger_\beta(y)\} : \psi^\dagger_\alpha(x)\psi_{\beta'}(y) : \\ &- \{\psi_{\beta'}(y), \psi^\dagger_\alpha(x)\} : \psi^\dagger_\beta(y)\psi_{\alpha'}(x) : \\ &+ \langle 0|[\rho_{\alpha\alpha'}(x), \rho_{\beta\beta'}(y)]|0\rangle. \end{aligned} \tag{H.37}$$

The first two pieces are the ones we expect from the classical algebra; the third term derives from the ordering we have used to define the currents. The expectation value of the commutator can be evaluated using (H.31) and the distributional identity

$$-\frac{1}{(2\pi)^2}\left(\frac{1}{(x-y+i\epsilon)^2} - \frac{1}{(x-y-i\epsilon)^2}\right) = \frac{1}{2\pi i}\frac{d}{dx}\delta(x-y).$$

One obtains

$$\langle 0|[\rho_{\alpha\alpha'}(x), \rho_{\beta\beta'}(y)]|0\rangle = \delta_{\alpha\beta}\delta_{\alpha'\beta'}\frac{1}{2\pi i}\frac{d}{dx}\delta(x-y),$$

Contracting (H.37) with $-iT^a_{\alpha\alpha'}$ and $-iT^b_{\beta\beta'}$ we obtain the algebra of the quantum current operators:

$$[J^a(x), J^b(y)] = if^{abc}J^c(x)\delta(x-y) - \frac{C}{2\pi i}\delta^{ab}\frac{d}{dx}\delta(x-y), \tag{H.38}$$

where C is defined by $\mathrm{tr}\, T_a T_b = C\delta_{ab}$. The new term commutes with all operators in the theory, so it is called a *central extension*. The resulting algebra is called a Kac–Moody algebra.

The same calculation for the left-handed fermion yields the same algebra but with the opposite sign for the central term.

In the presence of the chiral coupling to a gauge field, one can show by different methods that, with suitable choice of regularization method, the central extension is the same as in the free case.

H.19 Exercise 6.4: the two-dimensional WZ functional

We compute

$$-\frac{1}{4\pi}\int d^2x\, \varepsilon^{\mu\nu}\mathrm{tr}\left[\partial_\mu(g^{-1}U)(g^{-1}U)^{-1}(g^{-1}A_\nu g + g^{-1}\partial_\nu g)\right] + S_{WZW}(g^{-1}U).$$

The first integral gives

$$\frac{1}{4\pi}\int d^2x\,\varepsilon^{\mu\nu}\mathrm{tr}\left[R^g_\mu A_\nu\ + R^g_\mu R^g_\nu - R^U_\mu A_\nu - R^U_\mu R^g_\nu\right] \tag{H.39}$$

where we denote $R^g_\mu = \partial_\mu g g^{-1}$ and $R^U_\mu = \partial_\mu U U^{-1}$. The second term gives

$$-\frac{1}{12\pi}\int d^3x\,\varepsilon^{\lambda\mu\nu}\mathrm{tr}\,\left[-\bar{R}^g_\lambda\bar{R}^g_\mu\bar{R}^g_\nu + 3\bar{R}^g_\lambda\bar{R}^g_\mu\bar{R}^U_\nu - 3\bar{R}^g_\lambda\bar{R}^U_\mu\bar{R}^U_\nu + \bar{R}^U_\lambda\bar{R}^U_\mu\bar{R}^U_\nu\right], \tag{H.40}$$

The first and last terms in this expression are equal to $-S_{WZW}(\bar{g})$ and $S_{WZW}(\bar{U})$ respectively. The two middle terms add up to a total derivative that exactly cancels the last term in (H.39). The second term in (H.39) vanishes identically. The remaining terms exactly reconstruct $-\Gamma_{WZ}(A,g)+\Gamma_{WZ}(A,U)$.

H.20 Exercise 6.6: anomalies in the Standard Model

The group $SU(2)$ is safe by itself, but problems can arise when $U(1)$ generators are present in the trace defining d_{abc}. Since the $SU(2)$ generators are traceless, we only need to worry about terms with one and three $U(1)$ insertions.

If we consider one $U(1)$ contribution, since the right-handed fermions are $SU(2)$-singlets, we get the condition

$$0 = \sum Y_L Tr[T^a T^b] = -\frac{1}{2}\delta_{ab}\sum Y_L, \tag{H.41}$$

where the sum is over doublets and Y_L is the $U(1)$ hypercharge of the (left-handed) doublet. This is equivalent to

$$\sum Y_L = 0. \tag{H.42}$$

Three $U(1)$ insertions yield the condition

$$0 = \sum_L Y_L^3 - \sum_R Y_R^3, \tag{H.43}$$

where the sums are over both left and right fields. In the first generation we have the following hypercharge assignments:

$$Y_{e_L} = -\frac{1}{2}, \qquad Y_{\nu_L} = -\frac{1}{2}, \qquad Y_{e_R} = -1, \qquad Y_{\nu_R} = 0,$$
$$Y_{u_L} = \frac{1}{6}, \qquad Y_{d_L} = \frac{1}{6}, \qquad Y_{u_R} = \frac{2}{3}, \qquad Y_{d_R} = -\frac{1}{3},$$

These values satisfy both (H.42) and (H.43). Note that (H.42) is satisfied separately by the leptons and quarks while (H.43) holds due to a more intricate cancellation.

Finally we observe that in each family there are four doublets of chiral fermions (one leptonic doublet and three quark doublets) and therefore the condition for the cancellation of the global $SU(2)$ anomaly is also satisfied.

H.21 Exercise 6.7: the Schwinger model

1. The identity (6.146), which is easily checked for our choice of gamma matrices, is the key to the peculiar properties of this model. The first such property is that the axial invariance of massless fermions extends to a local invariance. The fermionic action is invariant, by construction, under the "vector" gauge transformations

$$\psi'(x) = e^{-i\alpha(x)}\psi(x), \quad \bar{\psi}'(x) = \bar{\psi}(x)e^{i\alpha(x)}, \quad A'_\mu(x) = A_\mu(x) - \frac{1}{e}\partial_\mu\alpha.$$

but using (6.146) and

$$\gamma^\mu e^{-i\beta\gamma^A} = e^{i\beta\gamma^A}\gamma^\mu,$$

it is also invariant under an additional "axial" local gauge invariance

$$\psi'(x) = e^{i\beta(x)\gamma_A}\psi(x), \quad \bar{\psi}'(x) = \bar{\psi}(x)e^{i\beta(x)\gamma_A}, \quad A'_\mu(x) = A_\mu(x) + \frac{i}{e}\varepsilon_{\mu\nu}\partial^\nu\beta. \tag{H.44}$$

Since every gauge potential can be decomposed in its longitudinal and transverse parts

$$A_\mu = \frac{1}{e}\left(\partial_\mu\alpha + \varepsilon_{\mu\nu}\partial^\nu\beta\right). \tag{H.45}$$

we see that by the transformations written above we can eliminate A_μ from the action, which then reduces to that of free fermions. Define the Clifford algebra-valued functions

$$\phi = \alpha\mathbb{I} + \beta\gamma^A, \qquad \bar{\phi} = \alpha\mathbb{I} - \beta\gamma^A.$$

where the scalars α and β can be written as non-local functions of the gauge potential:

$$\begin{aligned}\alpha &= \frac{e}{\partial^2}\partial^\mu A_\mu \\ \beta &= \frac{e}{\partial^2}\epsilon^{\rho\sigma}\partial_\rho A_\sigma.\end{aligned} \tag{H.46}$$

The gauge potential can be rewritten as a kind of pure gauge field:

$$\gamma^\mu A_\mu = \frac{1}{e}\gamma^\mu \partial_\mu \phi.$$

Then one can check the operator identity

$$\gamma^\mu(\partial_\mu - ieA_\mu) = e^{i\bar{\phi}}\gamma^\mu \partial_\mu\, e^{-i\phi}.$$

On the r.h.s. the derivative is supposed to act on everything that is to its right. One can then write an exact formula for the propagator in the background electromagnetic field:

$$\begin{aligned} G(x,y) &= e^{i\phi(x)} S(x-y)\, e^{-i\bar{\phi}(y)} \\ &= S(x-y) e^{i(\bar{\phi}(x)-\bar{\phi}(y))}, \end{aligned} \tag{H.47}$$

where $S(x-y)$ is the propagator of the free fermion.

Using the methods of Section 6.1, the VEV of the vector current can be defined as the $\epsilon \to 0$ limit of

$$\langle J_V^\mu(x,\epsilon)\rangle = -\mathrm{Tr}\gamma^\mu G\left(x-\frac{\epsilon}{2}, x+\frac{\epsilon}{2}\right) e^{ie\int A},$$

For small ϵ we then have

$$G\left(x-\frac{\epsilon}{2}, x+\frac{\epsilon}{2}\right) = S(-\epsilon)\left[1 - i\epsilon^\mu \partial_\mu \bar{\phi}(x) + \ldots\right],$$

where the free propagator is given by (6.17)

$$S(-\epsilon) = -\frac{i}{2\pi}\frac{\gamma^\mu \epsilon_\mu}{|\epsilon|^2}.$$

Using this expansion, expanding also the parallel transport operator to first order in ϵ, the terms that contain up to two ϵ's in the numerator are

$$\begin{aligned} \langle J_V^\mu\rangle &= -\mathrm{tr}\,\gamma^\mu \frac{\gamma^\rho \epsilon_\rho}{2\pi\epsilon^2}\left(\mathbb{I} - i\epsilon^\nu \partial_\nu \bar{\phi} + ie\epsilon^\lambda A_\lambda(x)\right) \\ &= -\frac{\epsilon_\nu}{\pi\epsilon^2}\left(\eta^{\mu\nu} + 2\varepsilon^{\mu\nu}\epsilon_\rho \partial_\rho \beta\right). \end{aligned}$$

The first term vanishes upon averaging over directions. The remaining two terms, using (6.18) and then the definition (H.45), become

$$\langle J_V^\mu\rangle = \frac{1}{\pi}\epsilon^{\mu\nu}\partial_\nu \beta.$$

The vector current is conserved, as desired. Finally using (H.46) we obtain

$$\langle J_V^\mu\rangle = \frac{e}{\pi}\left(\delta_\nu^\mu - \frac{\partial^\mu \partial_\nu}{\partial^2}\right) A^\nu. \tag{H.48}$$

2. While vector gauge invariance is maintained, the classical invariance under the axial gauge transformations (H.44) is broken. In fact in two dimensions, due to the identity (6.146), the axial current is dual of the vector current
$$J^\mu_A = \varepsilon^{\mu\nu} J_{V\nu}$$
and therefore using (H.48) one finds
$$\langle \partial_\mu J^\mu_A \rangle = \frac{e}{\pi} \varepsilon^{\mu\nu} \partial_\mu A_\nu,$$
in agreement with our previous result (6.19).

3. We can now easily integrate (6.147) to obtain the effective action
$$W_F(A) = \frac{e^2}{2\pi} \int d^2x \, A_\mu \left(\eta^{\mu\nu} - \frac{\partial^\mu \partial^\nu}{\partial^2} \right) A_\nu. \tag{H.49}$$
Integrating by parts, one can also rewrite this expression as
$$W_F[A] = \frac{e^2}{\pi} \int dx \, F^{\mu\nu} \frac{1}{\partial^2} F_{\mu\nu},$$
that is manifestly gauge invariant.

4. Adding the fermionic effective action to the Maxwell action the partition function is
$$Z = \int (dA) \exp \left[i \int d^2x \left(-\frac{1}{4} F_{\mu\nu} F^{\mu\nu} + \frac{e^2}{2\pi} A^\mu \left(g_{\mu\nu} - \frac{\partial^\mu \partial_\nu}{\partial^2} \right) A^\nu \right) \right]. \tag{H.50}$$
It is gauge invariant and describes a free massive boson field. The equation of motion is:
$$\partial_\mu F^{\mu\nu} - \frac{e^2}{\pi} \left(g_{\mu\nu} - \frac{\partial^\mu \partial_\nu}{\partial^2} \right) A^\nu;$$
multiplying by ∂^2 the above equation we get:
$$\partial^2 (\partial_\mu F^{\mu\nu}) + \frac{e^2}{\pi} (\partial_\mu F^{\mu\nu}) = 0.$$
The physical degree of freedom is $V^\mu = (\partial_\mu F^{\mu\nu})$, with mass $m^2 = \frac{e^2}{\pi}$; actually, due to the condition $\partial_\mu V^\mu = 0$, only one component is independent. One can regard the massive boson as a bound state of the original fermion-antifermion degrees of freedom.

5. The effective action in the chiral Schwinger model can be easily obtained from (H.49). Using (6.146) we have

$$\gamma_\mu \left(\frac{1-\gamma_A}{2} \right) A^\mu = \gamma_\mu \left(\frac{g^{\mu\nu} - \varepsilon^{\mu\nu}}{2} \right) A_\nu$$

and therefore the partition function of the chirally coupled model can be obtained from the one of the vectorially coupled model by replacing

$$A^\mu \to \frac{g^{\mu\nu} - \varepsilon^{\mu\nu}}{2} A_\nu,$$

This leads to

$$W = \frac{e^2}{8\pi} A_\mu \left(\eta^{\mu\nu} - 2\frac{\partial^\mu \partial^\nu}{\partial^2} - \frac{\varepsilon^{\mu\lambda}\partial^\nu \partial_\lambda + \varepsilon^{\nu\lambda}\partial^\mu \partial_\lambda}{\partial^2} \right) A_\nu,$$

whose gauge variation is

$$\delta_\epsilon W = \frac{e^2}{4\pi} \int d^2x \left(\epsilon \partial^\mu A_\mu - \varepsilon^{\mu\nu} \partial_\mu \epsilon A_\nu \right).$$

The first term is the gauge variation of $A_\mu A^\mu$. It could be modified or even removed by adding a local counterterm. The second term has the same form as the consistent anomaly.

Bibliography

[ABa69] S.L. Adler and W.A. Bardeen, *Absence of higher order corrections in the anomalous axial vector divergence equation*, *Phys. Rev.* **182** (1969) 1517.

[Adl69] S.L. Adler, *Axial vector vertex in spinor electrodynamics*, *Phys. Rev.* **177** (1969) 2426.

[AFG74] J. Arafune, P.G.O. Freund and C.J. Goebel, *Topology of Higgs Fields*, *J. Math. Phys.* **16** (1975) 433.

[AhB59] Y. Aharonov and D. Bohm, *Significance of electromagnetic potentials in the quantum theory*, *Phys. Rev.* **115** (1959) 485.

[AhB61] Y. Aharonov and D. Bohm, *Further Considerations on Electromagnetic Potentials in the Quantum Theory*, *Phys. Rev.* **123** (1961) 1511.

[AlE57] C.B. Allendoerfer and J. Eells, *On the cohomology of smooth manifolds*, *Comment. Math. Helv.* **32** (1958) 165.

[AMi85] M. Asorey and P.K. Mitter, *Cohomology of the Gauge Orbit Space and (2+1)-dimensional Yang–Mills Theory With the Chern-simons Term*, *Phys. Lett. B* **153** (1985) 147.

[ApB80] T. Appelquist and C.W. Bernard, *Strongly Interacting Higgs Bosons*, *Phys. Rev. D* **22** (1980) 200.

[ApC74] T. Appelquist and J. Carazzone, *Infrared Singularities and Massive Fields*, *Phys. Rev. D* **11** (1975) 2856.

[Bar69] W.A. Bardeen, *Anomalous Ward identities in spinor field theories*, *Phys. Rev.* **184** (1969) 1848.

[BaZ84] W.A. Bardeen and B. Zumino, *Consistent and Covariant Anomalies in Gauge and Gravitational Theories*, Nucl. Phys. B **244** (1984) 421.

[BeJ69] J.S. Bell and R. Jackiw, *A PCAC puzzle: $\pi^0 \to \gamma\gamma$ in the σ model*, Nuovo Cim. A **60** (1969) 47.

[BeP77] A.A. Belavin and A.M. Polyakov, *Quantum Fluctuations of Pseudoparticles*, Nucl. Phys. B **123** (1977) 429.

[BIM72] C. Bouchiat, J. Iliopoulos and P. Meyer, *An Anomaly Free Version of Weinberg's Model*, Phys. Lett. B **38** (1972) 519.

[BKW86] M.J. Bowick, D. Karabali and L.C.R. Wijewardhana, *Fractional spin via canonical quantization of the O(3) nonlinear sigma model*, Nucl. Phys. B **271** (1986) 417.

[Bog75] E.B. Bogomolny, *Stability of Classical Solutions*, Sov. J. Nucl. Phys. **24** (1976) 449.

[BoT82] R. Bott and L.W. Tu, *Differential Forms in Algebraic Topology*, Springer (1982) [DOI: 10.1007/978-1-4757-3951-0].

[BPST75] A.A. Belavin, A.M. Polyakov, A.S. Schwartz and Y.S. Tyupkin, *Pseudoparticle Solutions of the Yang–Mills Equations*, Phys. Lett. B **59** (1975) 85.

[Car96] J.L. Cardy, *Scaling and renormalization in statistical physics*, Cambridge University Press (1996).

[CCWZ69] C.G. Callan Jr., S.R. Coleman, J. Wess and B. Zumino, *Structure of phenomenological Lagrangians. 2*, Phys. Rev. **177** (1969) 2247.

[CDDN77] S. Chadha, P. Di Vecchia, A. D'Adda and F. Nicodemi, *Zeta Function Regularization of the Quantum Fluctuations Around the Yang–Mills Pseudoparticle*, Phys. Lett. B **72** (1977) 103.

[CGM77] S.R. Coleman, V. Glaser and A. Martin, *Action Minima Among Solutions to a Class of Euclidean Scalar Field Equations*, Commun. Math. Phys. **58** (1978) 211.

[ChM96] Y.M. Cho and D. Maison, *Monopoles in Weinberg–Salam model*, Phys. Lett. B **391** (1997) 360 [arXiv: hep-th/9601028].

[Col85] S.R. Coleman, *Aspects of Symmetry: Selected Erice Lectures*, Cambridge University Press, Cambridge, U.K. (1985) [DOI: 10.1017/CBO9780511565045].

[Col73] S.R. Coleman, *There are no Goldstone bosons in two-dimensions*, *Commun. Math. Phys.* **31** (1973) 259.

[CWZ69] S.R. Coleman, J. Wess and B. Zumino, *Structure of phenomenological Lagrangians. 1*, *Phys. Rev.* **177** (1969) 2239.

[DeW64] B. DeWitt, *Dynamical theory of groups and fields*, lectures given at *Les Houches summer school*, Les Houches, 1963, in B. DeWitt and C. DeWitt Morette eds., *Relativity, groups and topology*, Gordon and Breach, New York, U.S.A. (1964).

[DGH22] J.F. Donoghue, E. Golowich and B.R. Holstein, *Dynamics of the Standard Model: Second edition*, Cambridge University Press (2022) [DOI: 10.1017/9781009291033].

[DHF84a] E. D'Hoker and E. Farhi, *Decoupling a Fermion Whose Mass Is Generated by a Yukawa Coupling: The General Case*, *Nucl. Phys. B* **248** (1984) 59.

[DHF84b] E. D'Hoker and E. Farhi, *Decoupling a Fermion in the Standard Electroweak Theory*, *Nucl. Phys. B* **248** (1984) 77.

[DHN74a] R.F. Dashen, B. Hasslacher and A. Neveu, *Nonperturbative Methods and Extended Hadron Models in Field Theory 1. Semiclassical Functional Methods*, *Phys. Rev. D* **10** (1974) 4114.

[DHN74b] R.F. Dashen, B. Hasslacher and A. Neveu, *Nonperturbative Methods and Extended Hadron Models in Field Theory 2. Two-Dimensional Models and Extended Hadrons*, *Phys. Rev. D* **10** (1974) 4130.

[DiZ84] A.M. Din and W.J. Zakrzewski, *Spin and Statistics of CP^1 Skyrmions*, *Phys. Lett. B* **146** (1984) 341.

[DJT81] S. Deser, R. Jackiw and S. Templeton, *Topologically Massive Gauge Theories*, *Annals Phys.* **140** (1982) 372 [*Erratum ibid.* **185** (1988) 406].

[DLD78] A. D'Adda, M. Luscher and P. Di Vecchia, *A 1/n Expandable Series of Nonlinear Sigma Models with Instantons*, *Nucl. Phys. B* **146** (1978) 63.

[Eic79] H. Eichenherr and M. Forger, *On the Dual Symmetry of the Nonlinear Sigma Models*, *Nucl. Phys. B* **155** (1979) 381.

[Fad84] L.D. Faddeev, *Operator Anomaly for the Gauss Law*, *Phys. Lett. B* **145** (1984) 81.

[Fer93] F. Feruglio, *The chiral approach to the electroweak interactions*, *Int. J. Mod. Phys. A* **8** (1993) 4937 [arXiv: hep-ph/9301281].

[FiR68] D. Finkelstein and J. Rubinstein, *Connection between spin, statistics, and kinks*, *J. Math. Phys.* **9** (1968) 1762.

[Fri80] D. Friedan, *Nonlinear Models in Two Epsilon Dimensions*, *Phys. Rev. Lett.* **45** (1980) 1057.

[Fuj80] K. Fujikawa, *Path Integral for Gauge Theories with Fermions*, *Phys. Rev. D* **21** (1980) 2848 [*Erratum ibid.* **22** (1980) 1499].

[GaL83] J. Gasser and H. Leutwyler, *Chiral Perturbation Theory to One Loop*, *Annals Phys.* **158** (1984) 142.

[GJa72] D.J. Gross and R. Jackiw, *Effect of anomalies on quasirenormalizable theories*, *Phys. Rev. D* **6** (1972) 477.

[GML60] M. Gell–Mann and M. Levy, *The axial vector current in beta decay*, *Nuovo Cim.* **16** (1960) 705.

[GoJ74] J. Goldstone and R. Jackiw, *Quantization of Nonlinear Waves*, *Phys. Rev. D* **11** (1975) 1486.

[Gol92] N. Goldenfeld, *Lectures on phase transitions and the renormalization group*, Westview Press (1992).

[Gol23] N. Goldenfeld, *There's plenty of room in the middle: The unsung revolution of the renormalization group*, *Int. J. Mod. Phys. B* **38** (2024) 2430002 [arXiv: 2306.06020].

[HeR94] M.J. Herrero and E. Ruiz Morales, *The electroweak chiral Lagrangian for the Standard Model with a heavy Higgs*, *Nucl. Phys. B* **418** (1994) 431 [arXiv: hep-ph/9308276].

[Jac72] R. Jackiw, *Field theoretic investigations in current algebra* in S. Treiman, R. Jackiw and D. Gross eds., *Lectures in current algebra and its applications*, Princeton University Press (1972).

[Jac83] R. Jackiw, *Topological investigations of quantized gauge theories*, contribution to *Les Houches Summer School on Theoretical Physics: Relativity, Groups and Topology II*, Les Houches, 1983, *Conf. Proc. C* **8306271** (1983) 221.

[JaR76] R. Jackiw and C. Rebbi, *Conformal Properties of a Yang–Mills Pseudoparticle*, *Phys. Rev. D* **14** (1976) 517.

[Jo851] S.G. Jo, *Commutators in an Anomalous Nonabelian Chiral Gauge Theory*, *Phys. Lett. B* **163** (1985) 353.

[Jo852] S.G. Jo, *Commutator of Gauge Group Generators in Nonabelian Chiral Theory*, *Nucl. Phys. B* **259** (1985) 616.

[Kli90] F.R. Klinkhamer, *Another look at the $SU(2)$ anomaly*, *Phys. Lett. B* **256** (1991) 41.

[KoN63] S.Kobayashi and K.Nomizu *Foundations of differential geometry*, Interscience Publishers vol. I (1963) and vol. II (1969).

[LaD70] M.G.G. Laidlaw and C.M. DeWitt, *Feynman functional integrals for systems of indistinguishable particles*, *Phys. Rev. D* **3** (1971) 1375.

[Lon80] A.C. Longhitano, *Heavy Higgs Bosons in the Weinberg–Salam Model*, *Phys. Rev. D* **22** (1980) 1166.

[Man04] N.S. Manton and P. Sutcliffe, *Topological solitons*, Cambridge University Press (2004) [DOI: 10.1017/CBO9780511617034].

[Mer66] N.D. Mermin and H. Wagner, *Absence of ferromagnetism or antiferromagnetism in one-dimensional or two-dimensional isotropic Heisenberg models*, *Phys. Rev. Lett.* **17** (1966) 1133.

[Mic79] L. Michel, *Minima of Higgs–Landau polynomials*, CERN-TH-2716, CERN, Geneva (1979).

[MiV81] P.K. Mitter and C.M. Viallet, *On the Bundle of Connections and the Gauge Orbit Manifold in Yang–Mills Theory*, *Commun. Math. Phys.* **79** (1981) 457.

[MoF53] P.M. Morse and H, Feshbach, *Methods of theoretical physics*, McGraw Hill (1953).

[Nak03] M. Nakahara, *Geometry, topology and physics*, 2nd edition, Taylor and Francis (2003).

[Col77] S.R. Coleman, *There Are No Classical Glueballs*, *Commun. Math. Phys.* **55** (1977) 113.

[NiO73] H.B. Nielsen and P. Olesen, *Vortex Line Models for Dual Strings*, *Nucl. Phys. B* **61** (1973) 45.

[Ore76] F.R. Ore Jr., *Quantum Field Theory About a Yang–Mills Pseudoparticle*, *Phys. Rev. D* **15** (1977) 470.

[PaP90] N.K. Pak and R. Percacci, *Topology and fractional spin in the (2+1)-dimensional sigma model*, *Phys. Rev. D* **43** (1991) 1375.

[PeR88] R. Percacci and R. Rajaraman, *Constrained Hamiltonian Structure of the Chirally Gauged Wess–Zumino–Witten Model*, *Int. J. Mod. Phys. A* **4** (1989) 4177.

[Pol74] A.M. Polyakov, *Particle Spectrum in Quantum Field Theory*, *JETP Lett.* **20** (1974) 194.

[Pol75] A.M. Polyakov and A.A. Belavin, *Metastable States of Two-Dimensional Isotropic Ferromagnets*, *JETP Lett.* **22** (1975) 245.

[PSo75] M.K. Prasad and C.M. Sommerfield, *An Exact Classical Solution for the 't Hooft Monopole and the Julia–Zee Dyon*, *Phys. Rev. Lett.* **35** (1975) 760.

[Raj82] R. Rajaraman, *Solitons and instantons. An introduction to solitons and instantons in quantum field theory*, North Holland (1982).

[Ram84] T.R. Ramadas, *The Wess–Zumino term and fermionic solitons*, *Commun. Math. Phys.* **93** (1984) 355.

[Sch14] M. Schwartz, *Quantum Field Theory and the Standard Model*, Cambridge University Press (2014).

[Shi12] M. Shifman, *Advanced topics in quantum field theory.: A lecture course*, Cambridge University Press, Cambridge, UK (2012) [DOI: 10.1017/9781108885911].

[SiT67] I.M. Singer and J.M. Thorpe, *Lecture Notes on Elementary Topology and Geometry*, 1st ed., Scott, Foresman (1967).

[Sky61] T.H.R. Skyrme, *A nonlinear field theory*, *Proc. Roy. Soc. Lond. A* **260** (1961) 127.

[Sky62] T.H.R. Skyrme, *A Unified Field Theory of Mesons and Baryons*, *Nucl. Phys.* **31** (1962) 556.

[tHo74] G. 't Hooft, *Magnetic Monopoles in Unified Gauge Theories*, *Nucl. Phys. B* **79** (1974) 276.

[tHo76] G. 't Hooft, *Computation of the Quantum Effects Due to a Four-Dimensional Pseudoparticle*, *Phys. Rev. D* **14** (1976) 3432 [*Erratum ibid.* **18** (1978) 2199].

[tHo79] G. 't Hooft, *Naturalness, chiral symmetry, and spontaneous chiral symmetry breaking*, *NATO Sci. Ser. B* **59** (1980) 135.

[Tre86] S.B. Treiman, E. Witten, R. Jackiw and B. Zumino, *Current Algebra and Anomalies*, Princeton University Press (2014), ISBN 978-0-691-61089-4.

[Wei86] S. Weinberg, *Superconductivity for Particular Theorists*, *Prog. Theor. Phys. Suppl.* **86** (1986) 43.

[Wei78] S. Weinberg, *Phenomenological Lagrangians*, *Physica A* **96** (1979) 327.

[Wei87] E.J. Weinberg and A.-Q. Wu, *Understanding complex perturbative effective potentials*, *Phys. Rev. D* **36** (1987) 2474.

[Wei95] S. Weinberg, *The quantum theory of fields*, Cambridge University Press (1995).

[Wei96] S. Weinberg, *What is quantum field theory, and what did we think it is?*, in the proceedings of the *Conference on Historical Examination and Philosophical Reflections on the Foundations of Quantum Field Theory*, Boston, U.S.A., March 01–03 (1996) [arXiv: hep-th/9702027].

[Wei09] S. Weinberg, *Effective Field Theory, Past and Future*, *PoS* **CD09** (2009) 001 [arXiv: 0908.1964].

[Wil73] K.G. Wilson and J.B. Kogut, *The renormalization group and the epsilon expansion*, *Phys. Rept.* **12** (1974) 75.

[WiZ83] F. Wilczek and A. Zee, *Linking Numbers, Spin, and Statistics of Solitons*, *Phys. Rev. Lett.* **51** (1983) 2250.

[Wit79] E. Witten, *Dyons of Charge e theta/2 pi*, *Phys. Lett. B* **86** (1979) 283.

[Wit82] E. Witten, *An $SU(2)$ Anomaly*, *Phys. Lett. B* **117** (1982) 324.

[Wit83a] E. Witten, *Current Algebra, Baryons, and Quark Confinement*, *Nucl. Phys. B* **223** (1983) 433.

[Wit83b] E. Witten, *Global Aspects of Current Algebra*, *Nucl. Phys. B* **223** (1983) 422.

[WuY75] T.T. Wu and C.N. Yang, *Concept of Nonintegrable Phase Factors and Global Formulation of Gauge Fields*, *Phys. Rev. D* **12** (1975) 3845.

[WuZ84] Y.-S. Wu and A. Zee, *Comments on the Hopf Lagrangian and Fractional Statistics of Solitons*, *Phys. Lett. B* **147** (1984) 325.

Index

Non-Perturbative Quantum Field Theory

An Introduction to Topological and Semiclassical Methods

Roberto Percacci

This book presents in a systematic fashion a number of quantum field theoretic phenomena that have a topological underpinning.

The systematics is provided by the homotopy groups of the configuration space: solitons and instantons are related to the zeroth and first homotopy groups respectively, and quantized parameters to the second. The close relation of some of these notions to anomalies is also discussed. These concepts have many applications, from particle physics to statistical and condensed matter physics. The focus is mainly on the former, but some particularly instructive examples of the latter are also described.

SISSA Medialab

ICTP The Abdus Salam International Centre for Theoretical Physics

SCOAP3 Sponsoring Consortium for Open Access Publishing in Particle Physics

ISBN 9788898587056