10
Topological objects in field theory

Un fourmi de dix-huit mètres
Avec un chapeau sur la tête,
Ça n'existe pas;
Pourquoi pas?
Robert Desnos

No-one can deny the success which quantum field theory, in the perturbative approximation, has enjoyed over the last half century. One need only mention the interpretation of quantised fields as particles, the description of scattering processes, the precise numerical agreements in quantum electrodynamics, the successful prediction of the W particle, and the beginnings of an understanding of the strong interaction through quantum chromodynamics. Yet despite these successes, the question of how to describe *the basic matter fields of nature* has remained unanswered – except, of course, through the introduction of quantum numbers and symmetry groups. As far as field theory goes, the matter fields are treated as point objects. Even in classical field theory these present us with unpleasant problems, in the shape of the infinite self-energy of a point charge. In the quantum theory, these divergences do not disappear; on the contrary, they appear to get worse, and despite the comparative success of renormalisation theory the feeling remains that there ought to be a more satisfactory way of doing things.

Now it turns out that non-linear classical field theories possess extended solutions, commonly known as solitons, which represent stable configurations with a well-defined energy which is nowhere singular. May this be of relevance to particle physics? Since non-Abelian gauge theories are non-linear, it may well be, and the last ten years have seen the discovery of vortices, magnetic monopoles and 'instantons', which are soliton solutions to the gauge-field equations in two space dimensions (i.e. a 'string' in 3-dimensional space), three space dimensions (localised in space but not in time) and 4-dimensional space–time (localised in space and time). If gauge theories are taken seriously then so must these solutions be. It will be seen that they do give rise to new physics and there is even the hope that they may solve the problem of quark confinement.

Not the least interesting feature of this subject is the branch of mathematics which it involves; for the stability of these solitons arises from the fact that the

boundary conditions fall into distinct classes, of which the vacuum belongs only to one. These boundary conditions are characterised by a particular correspondence (mapping) between the group space and co-ordinate space, and because these mappings are not continuously deformable into one another they are *topologically* distinct. The relevant notions in topology will be developed as we go along. We begin our survey with the 'sine–Gordon' equation which has no relevance to particle physics but whose soliton solutions are quite well understood, and therefore form a good introduction to the subject.

10.1 The sine–Gordon kink

The sine–Gordon equation

$$\frac{\partial^2\phi}{\partial t^2} - \frac{\partial^2\phi}{\partial x^2} + \frac{1}{b^2}\sin(b\phi) = 0 \tag{10.1}$$

describes a scalar field in one space and one time dimension. It possesses moving, as well as stationary, solutions. To find moving solutions, we want a field of the form

$$\phi(x, t) = f(x - vt) = f(\xi).$$

It is easy to check that

$$f(\xi) = \frac{4}{b}\arctan\exp[\pm(\gamma/\sqrt{b})\xi] \tag{10.2}$$

is a solution, where $\gamma = (1 - v^2)^{-1/2}$. The appearance of this wave is shown in Fig. 10.1. It is a *solitary wave*, which moves without changing shape or size, and therefore without dissipation, in strong contrast to the waves set up when, for instance, a stone is thrown into a pond. These waves spread out and the energy is dissipated. Solitary waves (solitons) have been observed, for

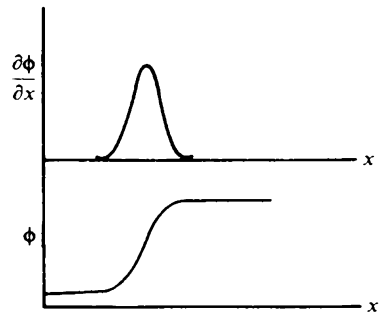

Fig. 10.1. A solitary wave (soliton).

example, moving along canals. In this case they are solutions of the Korteweg de Vries equation.

Since solitons are solutions of non-linear wave equations the superposition principle is not obeyed. This means that when two solitons meet the resultant wave form is a complicated one, but the surprising thing is that, asymptotically, the solitons separate out again – they 'pass through' one another. This property is, of course, of interest to particle physicists, though we shall not develop it any further here. Another consequence of the fact that the superposition principle does not hold is that the quantisation of solitons becomes non-trivial. We shall not follow this matter any further either. Instead, we turn to the stationary solutions of the sine–Gordon equation, which possess an interest of a different type.

It is clear that (10.1) possesses an infinite number of constant solutions (which, as we shall see in a moment, have zero energy):

$$\phi = \frac{2\pi n}{b}, \quad n = 0, \pm 1, \pm 2, \ldots ; \tag{10.3}$$

that is, the sine–Gordon equation possesses a degenerate vacuum.[‡] ('Vacuum' here does not, of course, mean the state in Hilbert space, but simply a classical field configuration of zero energy.) The Lagrangian for the sine–Gordon equation is

$$\mathscr{L} = \frac{1}{2}\left(\frac{\partial \phi}{\partial t}\right)^2 - \frac{1}{2}\left(\frac{\partial \phi}{\partial x}\right)^2 - V(\phi) \tag{10.4}$$

with

$$V(\phi) = \frac{1}{b^2}[1 - \cos(b\phi)],$$

where the constant has been chosen so that the solutions (10.3) have $V = 0$. They therefore have zero energy since the energy density of the field configuration is

$$\mathscr{H} = \frac{1}{2}\left(\frac{\partial \phi}{\partial t}\right)^2 + \frac{1}{2}\left(\frac{\partial \phi}{\partial x}\right)^2 + V(\phi). \tag{10.5}$$

Note that we may write

$$V(\phi) = \frac{1}{2}\phi^2 - \frac{b^2}{4!}\phi^4 + \cdots , \tag{10.6}$$

[‡] It is this property that is crucial in what follows, and so a corresponding analysis could be made for other field theories with degenerate vacua, for example the ϕ^4 model with $m^2 < 0$ considered in Chapter 8.

or, with $\lambda = b^2$ and unit mass m

$$V(\phi) = \frac{m^2}{2}\phi^2 - \frac{\lambda}{4!}\phi^4 + \cdots, \tag{10.7}$$

and m stands for the 'particle' mass and λ for the self-interaction coupling.

The potential V in (10.4) is shown in Fig. 10.2 with the (zero energy) ground state given by (10.3). Now construct the following configuration. Let ϕ approach one of the zeros of V (say $n = 0$) as $x \to -\infty$, but a *different* zero (say $n = 1$) as $x \to \infty$. Between these two there is clearly a region where

$$\phi \neq \frac{2\pi n}{b}, \quad \frac{\partial \phi}{\partial x} \neq 0,$$

and therefore, from (10.5), where there is a positive energy density. We assume the configuration is static, so $\partial\phi/\partial t = 0$. Because of the boundary conditions on ϕ, we expect the total energy to be finite. Let us find what it is. For a stationary solution to the sine–Gordon equation we have

$$\frac{\partial^2 \phi}{\partial x^2} = \frac{\partial V}{\partial \phi}$$

which gives on integration

$$\frac{1}{2}\left(\frac{\partial \phi}{\partial x}\right)^2 = V(\phi), \tag{10.8}$$

the integration constant being zero. From (10.5) and (10.8), the energy of the stationary soliton is

$$\begin{aligned} E &= \int \mathcal{H}\, \mathrm{d}x \\ &= \int \left[\frac{1}{2}\left(\frac{\partial \phi}{\partial x}\right)^2 + V(\phi)\right] \mathrm{d}x \\ &= \int 2V(\phi)\, \mathrm{d}x \\ &= \int_0^{2\pi/b} [2V(\phi)]^{1/2}\, \mathrm{d}\phi \end{aligned}$$

where we have put in the integration limits given by (10.3) between $n = 0$ and

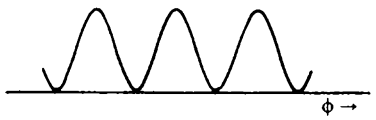

Fig. 10.2. The sine–Gordon potential $V(\phi)$.

$n = 1$. This integral is now easily performed on substituting (10.4). We have

$$\begin{aligned} E &= \frac{\sqrt{2}}{b}\int_0^{2\pi/b} [1 - \cos(b\phi)]^{1/2}\, \mathrm{d}\phi \\ &= \frac{\sqrt{2}}{b^2}\int_0^{2\pi} (1 - \cos\alpha)^{1/2}\, \mathrm{d}\alpha \\ &= \frac{8}{b^2} \\ E &= \frac{8m^3}{\lambda} \end{aligned} \tag{10.9}$$

where in the last step we have used the substitution in (10.7). So this soliton has a finite energy, with the interesting property that the energy is *inversely* proportional to the coupling constant. This may indeed be a useful property for particle physics.

There is a simple model which makes this soliton easy to visualise. Consider an infinite horizontal string with pegs attached to it at equally spaced intervals, and connect each peg to its neighbour with a small spring (the 'coupling'). Each peg is also acted on by gravity. The ground state corresponds to every peg hanging vertically. The soliton we have found, with $n = 0 \rightarrow 1$, corresponds to the situation in Fig. 10.3. This soliton – and others of this type (see below) – is called a *kink*. It should be clear from the peg model that the kink is stable, and cannot decay into the ground state with $E = 0$. This would involve a (semi-) infinite number of pegs turning over, which would need a (semi-) infinite amount of energy. But what is the *mathematical* reason for the stability of the kink? It is to be found in the boundary conditions. 'Space' in this model is an infinite line, whose boundary is two points (the end-points). At these two points the 1-kink solution has $n = 0$ and $n = 1$, and this is *not* continuously deformable into $n = 0$ and $n = 0$ (the ground state). The kink, then, is a 'topological' object. Its existence depends on the topological properties of the space (in particular, its boundary, which in this case is a discrete set). This conclusion is a general one; that is to say, the stability of soliton solutions in non-linear field theories is a consequence of topology.

Finally, the stability of the soliton (kink) obviously signals a *conservation law*: there must be a conserved charge Q, equal to an integer N (the difference between the two integers in (10.3)), and a corresponding divergenceless current

Fig. 10.3. Pegs on a line representing the kink (soliton) solution to the sine–Gordon equation.

J^μ ($\mu = 0, 1$). They are easy to construct. With

$$J^\mu = \frac{b}{2\pi}\varepsilon^{\mu\nu}\partial_\nu\phi \tag{10.10}$$

($\varepsilon^{\mu\nu}$ is antisymmetric, with $\varepsilon^{01} = 1$), we have the *identity* $\partial_\mu J^\mu = 0$, and the charge is

$$\begin{aligned} Q &= \int_{-\infty}^{\infty} J^0 \,\mathrm{d}x \\ &= \frac{b}{2\pi}\int_{-\infty}^{\infty} \frac{\partial\phi}{\partial x}\,\mathrm{d}x \\ &= \frac{b}{2\pi}[\phi(\infty) - \phi(-\infty)] = N. \end{aligned} \tag{10.11}$$

The interesting thing is that the current J^μ does *not* follow from the invariance of $\mathscr{L}$ under any symmetry transformation. It is therefore *not* a Noether current. Its divergencelessness follows independently of the equations of motion.

We consider, in the following sections, examples of solitons in gauge theories, beginning with one in two space dimensions – the vortex.

10.2 Vortex lines

Now consider a scalar field in 2-dimensional space. The 'boundary' of this space is the circle at infinity, denoted S^1. We construct a field whose value on the boundary is

$$\phi = a\,\mathrm{e}^{\mathrm{i}n\theta} \quad (r \to \infty) \tag{10.12}$$

where r and θ are polar co-ordinates in the plane, a is a constant, and, to make ϕ single-valued, n is an integer. We propose this form, rather than simply $\phi = a$, because it is a generalisation to two dimensions of (10.3). ((10.3) is a solution of the sine–Gordon equation, whereas it is yet to be seen what equation describes the 2-dimensional solitons we are in the process of developing.) From (10.12), we have

$$\boldsymbol{\nabla}\phi = \frac{1}{r}(\mathrm{i}na\,\mathrm{e}^{\mathrm{i}n\theta})\,\hat{\boldsymbol{\theta}}. \tag{10.13}$$

The Lagrangian and Hamiltonian functions are

$$\mathscr{L} = \frac{1}{2}\left(\frac{\partial\phi}{\partial t}\right)^2 - \tfrac{1}{2}|\boldsymbol{\nabla}\phi|^2 - V(\phi), \tag{10.14}$$

$$\mathscr{H} = \frac{1}{2}\left(\frac{\partial\phi}{\partial t}\right)^2 + \tfrac{1}{2}|\boldsymbol{\nabla}\phi|^2 + V(\phi). \tag{10.15}$$

Now let us consider a static configuration with, for example,

$$V(\phi) = [a^2 - \phi^*\phi]^2 \tag{10.16}$$

so that $V = 0$ on the boundary. Then as $r \to \infty$

$$\mathcal{H} = \tfrac{1}{2}|\nabla\phi|^2 = \frac{n^2a^2}{2r^2}$$

and the energy (mass) of the static configuration is

$$E \approx \int^{\infty} \mathcal{H} r \, \mathrm{d}r \, \mathrm{d}\theta = \pi n^2 a^2 \int^{\infty} \frac{1}{r} \mathrm{d}r.$$

This is *logarithmically divergent*; the kink, as it stands, cannot be generalised to two dimensions – nor to more than two, for it turns out that in all these cases the energy is divergent.

To proceed, we add a gauge field, so that what counts is the covariant derivative

$$D_\mu\phi = \partial_\mu\phi + \mathrm{i}eA_\mu\phi. \tag{10.17}$$

By choosing A_μ of the form

$$\mathbf{A} = \frac{1}{e}\nabla(n\theta) \quad (r \to \infty),$$

i.e.

$$A_r \to 0, \quad A_\theta \to -\frac{n}{er} \quad (r \to \infty), \tag{10.18}$$

we find that at $r = \infty$

$$D_\theta\phi = \frac{1}{r}\left(\frac{\partial\phi}{\partial\phi}\right) + \mathrm{i}eA_\theta\phi = 0, \quad D_r\phi = 0 \tag{10.19}$$

so $D_\mu\phi \to 0$ on the boundary at infinity. The Lagrangian is now

$$\mathcal{L} = -\tfrac{1}{4}F_{\mu\nu}^2 + |D_\mu\phi|^2 - V(\phi). \tag{10.20}$$

Since (10.18) is a *pure gauge*,

$$A_\mu \to \partial_\mu\chi \quad (r \to \infty), \tag{10.21}$$

then $F_{\mu\nu} \to 0$. For a static configuration $\mathcal{H} = -\mathcal{L}$, and with $V(\phi)$ given by (10.16) we have $\mathcal{H} \to 0$ as $r \to \infty$, making possible a field configuration of *finite* energy. We shall now see that the effect of adding the gauge field is to give the soliton *magnetic flux*. Consider the integral $\oint \mathbf{A} \cdot \mathrm{d}\mathbf{l}$ round the circle S^1 at infinity. By Stokes' theorem, this is $\int \mathbf{B} \cdot \mathrm{d}\mathbf{S} = \Phi$, the flux enclosed, hence

$$\Phi = \oint \mathbf{A} \cdot \mathrm{d}\mathbf{l} = \oint A_\theta r \, \mathrm{d}\theta = -\frac{2\pi n}{e}, \tag{10.22}$$

and the flux is *quantised*. So we have, after all, constructed a 2-dimensional field configuration, consisting of a charged scalar field and a gauge field (the electromagnetic field!). It carries magnetic flux, and since $D_\mu\phi \to 0$ and $F_{\mu\nu} \to 0$ on the boundary at infinity, it appears to have finite energy. It is clear that by adding a third dimension (the z axis) on which the fields have no dependence, this configuration becomes a vortex line. Apart from the presence of the scalar field, it is the same as the solenoid discussed in §3.4 under the Bohm–Aharonov effect; and just as that effect was attributable to the topology of the gauge group $U(1)$, so here also we shall see that it is this same topology which ensures stability of the vortex.

It will not have escaped the reader's notice that the Lagrangian (10.20) with $V(\phi)$ given by (10.16) is that of the Higgs model – see (8.36) and (8.4) – that is, scalar electrodynamics with spontaneous symmetry breaking. Actually, we saw in §8.4 that this Lagrangian is the relativistic version of the Landau–Ginzburg free energy, which describes superconductivity. It is known that on the occasions when magnetic flux *does* penetrate superconductors (that is, in type II superconductors), it does so in quantised flux lines, called Abrikosov flux lines. It is these that the present solutions are describing, the field ϕ in superconductivity being the BCS condensate.

To be a little more systematic, let us start from the Higgs Lagrangian (8.36):

$$\mathscr{L} = -\tfrac{1}{4}F_{\mu\nu}F^{\mu\nu} + |(\partial_\mu + \mathrm{i}eA_\mu)\phi|^2 - m^2\phi^*\phi - \lambda(\phi^*\phi)^2. \tag{8.36}$$

Spontaneous symmetry breaking is signalled by $m^2 < 0$, and the vacuum is then given by (8.4)

$$|\phi|_{\mathrm{vac}} = a = \left(\frac{-m^2}{2\lambda}\right)^{1/2}. \tag{8.4}$$

The equations of motion obtained from (8.36) are

$$D^\mu(D_\mu\phi) = -m^2\phi - 2\lambda\phi|\phi|^2, \tag{10.23}$$

$$\mathrm{i}e(\phi\partial_\mu\phi^* - \phi^*\partial_\mu\phi) + 2e^2A_\mu|\phi|^2 = \partial^\nu F_{\mu\nu}. \tag{10.24}$$

We must first check that these equations allow the solutions (10.12) and (10.18) at infinity. Since by construction (see (10.19)) $D_\mu\phi = 0$ as $r \to \infty$, the left-hand side of (10.23) vanishes; and so does the right-hand side if ϕ takes on its vacuum value (8.4). Since A_μ is a pure gauge (see (10.21)) $F_{\mu\nu} = 0$ as $r \to \infty$, so the right-hand side of (10.24) vanishes. In view of (10.12) and (10.18) the left-hand side vanishes identically when $\mu = r$, and when $\mu = \theta$ it vanishes when ϕ assumes the vacuum value (8.4). Hence our particular choices for A_μ and ϕ are allowed by the equations of motion.

As r becomes finite, and particularly as $r \to 0$, of course, the values of A_μ and ϕ change. Let us now treat the problem as one in three dimensions, with cylindrical symmetry about the z axis. Then, since there is magnetic flux, the magnetic field component B_z must be non-zero, which means that A cannot be

a pure gauge everywhere. Also, continuity requires that $\phi \to 0$ as $r \to 0$; since this is not the vacuum value, the 2-dimensional soliton will have an energy, and the vortex will have a corresponding mass per unit length. The forms of A and ϕ are found from the equations of motion. Taking B with a z component only, and A with a θ component only, we have

$$B = B_z = \frac{1}{r}\frac{\mathrm{d}}{\mathrm{d}r}[rA(r)], \quad A(r) = A_\theta = A. \tag{10.25}$$

In addition, ϕ is of the form

$$\phi = \chi(r)\,\mathrm{e}^{in\theta} \tag{10.26}$$

with

$$\chi(r) \xrightarrow[r\to 0]{} 0, \quad \chi(r) \xrightarrow[r\to\infty]{} a. \tag{10.27}$$

In the static case, the equation of motion (10.23) then becomes

$$(\partial_i + \mathrm{i}eA_i)^2\phi - (m^2 + 2\lambda|\phi|^2)\phi = 0$$

which, on summing over the r and θ components, gives

$$\frac{1}{r}\frac{\mathrm{d}}{\mathrm{d}r}\left(r\frac{\mathrm{d}\chi}{\mathrm{d}r}\right) - \left[\left(\frac{n}{r} - eA\right)^2 + m^2 + 2\lambda\chi^2\right]\chi = 0. \tag{10.28}$$

On the other hand, taking the θ component of (10.24) gives (recall equations (2.217)–(2.221))

$$-\frac{\mathrm{i}e}{r}(2\mathrm{i}n)\chi^2 + 2e^2A\chi^2 = -\partial_i \mathrm{F}_{\theta i}$$

and hence

$$\frac{\mathrm{d}}{\mathrm{d}r}\left(\frac{1}{r}\frac{\mathrm{d}}{\mathrm{d}r}(rA)\right) - 2e\left(\frac{n}{r} + eA\right)\chi^2 = 0. \tag{10.29}$$

One should now solve the coupled non-linear equations of motion (10.28) and (10.29). No exact analytic solution, however, has yet been found. In the approximation where $\chi \simeq a$ a constant (i.e. for $r \to \infty$), Nielsen and Olesen (1973) found (with c a constant of integration and K_1 and K_0 modified Bessel functions)

$$A = -\frac{n}{er} - \frac{c}{e}K_1(|e|ar) \xrightarrow[r\to\infty]{} -\frac{n}{er} - \frac{c}{e}\left(\frac{\pi}{2|e|ar}\right)^{1/2}\mathrm{e}^{-|e|ar} + \cdots$$

with magnetic field

$$B_z = c\chi K_0(|e|ar) \to \frac{c}{e}\left(\frac{\pi a}{2|e|r}\right)^{1/2}\mathrm{e}^{-|e|ar} + \ldots. \tag{10.30}$$

To obtain the variation of the scalar field, we put

$$\chi(r) = a + \rho(r);$$

then

$$\rho(r) \simeq e^{-\sqrt{-m^2}r} \tag{10.31}$$

(recall that $-m^2 > 0$). These solutions are sketched in Fig. 10.4.

Why are these solutions stable? As with the kink, the reason is topological. The Lagrangian is invariant under a symmetry group – in this case $U(1)$, the electromagnetic gauge group. The field ϕ (with boundary value given by (10.12)) is a representation of $U(1)$. The group space of $U(1)$ is a circle S^1, since an element of $U(1)$ may be written $\exp(i\theta) = \exp[i(\theta + 2\pi)]$, so the space of all values of θ is a line with $\theta = 0$ identified with $\theta = 2\pi$, and the line becomes a circle S^1. The field ϕ in (10.12) is a representation basis of $U(1)$, but it is the boundary value of the field in a 2-dimensional space. This boundary is clearly a circle S^1 (the circle $r \to \infty$, $\theta = (0 \to 2\pi)$). Hence ϕ defines a mapping of the boundary S^1 in physical space onto the group space S^1:

$$\phi\colon S^1 \to S^1, \tag{10.32}$$

the mapping being specified by the integer n. Now a solution characterised by one value of n is stable since it cannot be continuously deformed into a solution with a different value of n (a rubber band which fits twice round a circle cannot be continuously deformed into one which goes once round the circle). This is to say (see §3.4) that the *first homotopy group* of S^1, the group space of $U(1)$, is not trivial:

$$\pi_1(Z^1) = \mathbb{Z}. \tag{10.33}$$

$\mathbb{Z}$ is the additive group of integers.

The status of a topological argument like this is that it provides a very general condition which *must* be fulfilled in order that solitons exist in a particular model. If, as in the model above, the topological argument indicates that soliton solutions are possible in principle then one goes to the equations of motion to find them. Topology therefore provides *existence arguments*. As an example, let us enquire whether stringlike solutions to (spontaneously broken)

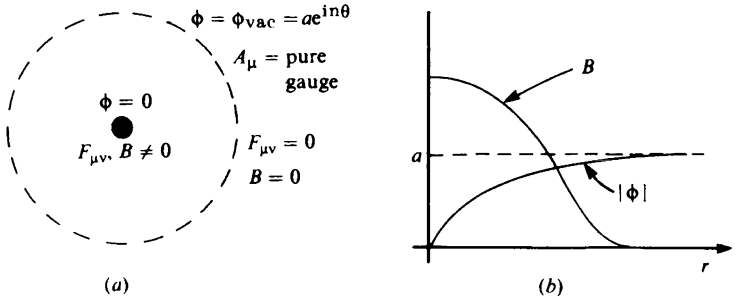

Fig. 10.4. The variations of the scalar and magnetic fields in the Nielsen–Olesen solution.

gauge theories exist when the gauge group is $SU(2)$. This is the group of 2×2 matrices.

$$U = u_0 + \mathrm{i}\sum_{j=1}^{3} u_j \sigma_j$$

where σ_j are the Pauli matrices, and the condition that U is unitary and has unit determinant is

$$u_0^2 + u_1^2 + u_2^2 + u_3^2 = 1. \tag{10.34}$$

Now this is the equation for the unit sphere S^3 in 4-dimensional Euclidean space E^4; that is, the group space of $SU(2)$ is S^3. There will exist stable vortices in an $SU(2)$ gauge theory if the mappings of the group onto the S^1 boundary of the two-dimensional parameter space fall into distinct classes; that is, if $\pi_1(S^3)$ is non-trivial. But $\pi_1(S^3)$ is, in fact, trivial, for S^3 is a simply connected space; every closed curve S^1 on S^3 may be shrunk to a point, so the boundary conditions may all be shrunk to the trivial constant condition $\phi = \text{const.}$, and no vortices exist.

The group $O(3)$, on the other hand, is not simply connected but doubly connected. For example, corresponding to the $O(3)$ matrix

$$\begin{pmatrix} \cos\alpha & -\sin\alpha & 0 \\ \sin\alpha & \cos\alpha & 0 \\ 0 & 0 & 1 \end{pmatrix}, \tag{10.35}$$

which corresponds to rotation about the z axis through an angle $-\alpha$, is the $SU(2)$ matrix

$$\begin{pmatrix} \mathrm{e}^{\mathrm{i}\alpha/2} & 0 \\ 0 & \mathrm{e}^{-\mathrm{i}\alpha/2} \end{pmatrix}. \tag{10.36}$$

Now $\alpha = 0$ clearly gives the identity matrix in both cases, but $\alpha = 2\pi$ gives the identity again in $O(3)$ and minus the identity in $SU(2)$. This is the origin of the well-known statement that vectors do not change sign on rotation through 2π but spinors do. In other words, corresponding to two elements of $SU(2)$ (the identity, and minus the identity) there is only one element of $O(3)$:

$$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \quad \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix} \quad \Big\rangle \quad \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

$SU(2)$ $\qquad\qquad$ $O(3)$

There is a two-to-one mapping of $SU(2)$ onto $O(3)$. The group space of $O(3)$, accordingly, is obtained from that of $SU(2)$ by *identifying opposite points* on the 3-space S^3, since they correspond to the same $O(3)$ transformation. This space is *doubly connected*, as we shall now show. We consider *closed curves* S^1 in the group space of $O(3)$. Each curve corresponds to a continuous set of rotations, starting (say) from the identity 0, and returning there. One possible type of closed curve is the path c_1 in Fig. 10.5. This corresponds to a series of rotations, the angle of which nowhere exceeds π. If the angle does exceed π, then the path in group space becomes like c_2. On reaching the angle π at the point A, the path reappears at the opposite point A', and eventually returns to the origin O. It is clear that c_1 is homotopic (may be shrunk) to a point, where c_2 is homotopic to a line. Readers may convince themselves that a closed path in which the angle of rotation exceeds 2π reappears at opposite points on the surface of S^3 *twice*, and is therefore homotopic to a point. Similarly, one in which the angle exceeds 3π is homotopic to a straight line. Consequently, there are only two types of closed path S^1 in the group space of $O(3)$: those homotopic to a point and those homotopic to a line. This means there is one non-trivial vortex in an $O(3)$ gauge theory. The vortices may have 'charges' (flux) 1 or 0, with the algebra $0 + 0 = 0$, $1 + 0 = 1$, $1 + 1 = 0$, so two non-trivial vortices will annihilate each other. (It may be parenthetically remarked that whether the gauge group is $SU(2)$ or $O(3)$ depends on *what particles exist*: if there are particles with 'isospin' of $\frac{1}{2}$, $\frac{3}{2}$, $\frac{5}{2}$, etc. then the gauge group is $SU(2)$, but if all the particles have integral isospin the group is $O(3)$.)

One way of making lines of magnetic flux is to place two opposite magnetic charges close together. So an obvious question is: if gauge theories allow flux lines, do they also allow magnetic charges? In fact they do, and these are called 't Hooft–Polyakov magnetic monopoles, after their discoverers. Like vortices, these monopoles owe their stability (and therefore existence) to the non-trivial topological properties of the gauge group. In this respect they are completely different from the 'ordinary' magnetic monopoles, which are *point* magnetic charges, and which may be introduced into Maxwell's equations to make them symmetric between electricity and magnetism. Dirac showed that the prescriptions of quantum theory imply a remarkable quantisation condition for point

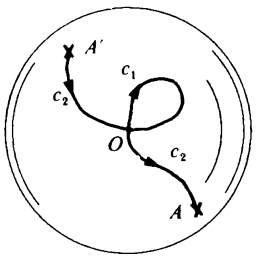

Fig. 10.5. Two types of closed path in the group space of $O(3)$.

magnetic charges, and for this reason they are sometimes referred to as Dirac monopoles. We shall study Dirac monopoles in the next section, and 't Hooft–Polyakov monopoles after that. This will serve to acquaint the reader with the idea of magnetic charge, as well as to demonstrate the difference between the two types of monopole.

10.3 The Dirac monopole

Consider a magnetic monopole of strength g at the origin. The magnetic field is radial and is given by a Coulomb-type law

$$\mathbf{B} = \frac{g}{r^3}\mathbf{r} = -g\nabla\left(\frac{1}{r}\right) \tag{10.37}$$

(we are using Gaussian units). Since $\nabla^2(1/r) = -4\pi\delta^3(r)$, we have

$$\nabla \cdot \mathbf{B} = 4\pi g\delta^3(r) \tag{10.38}$$

corresponding to a point magnetic charge, as desired. Since $\mathbf{B}$ is radial, the total flux through a sphere surrounding the origin is

$$\Phi = 4\pi r^2 B = 4\pi g. \tag{10.39}$$

Consider a particle with electric charge e in the field of this monopole. The wave function for a free particle is

$$\psi = |\psi| \exp\left[\frac{\mathrm{i}}{\hbar}(\mathbf{p}\cdot\mathbf{r} - Et)\right].$$

In the presence of an electromagnetic field, $\mathbf{p} \to \mathbf{p} - (e/c)\mathbf{A}$, so

$$\psi \to \psi \exp\left(-\frac{\mathrm{i}e}{\hbar c}\mathbf{A}\cdot\mathbf{r}\right);$$

or the phase α changes by

$$\alpha \to \alpha - \frac{e}{\hbar c}\mathbf{A}\cdot\mathbf{r}.$$

Consider a closed path at fixed r, θ, with ϕ ranging from 0 to 2π. The total change in phase is

$$\begin{aligned}
\Delta\alpha &= \frac{e}{\hbar c}\oint \mathbf{A}\cdot\mathrm{d}\mathbf{l} \\
&= \frac{e}{\hbar c}\int \operatorname{curl}\mathbf{A}\cdot\mathrm{d}\mathbf{S} \\
&= \frac{e}{\hbar c}\int \mathbf{B}\cdot\mathrm{d}\mathbf{S} \\
&= \frac{e}{\hbar c}(\text{Flux through cap}) = \frac{e}{\hbar c}\Phi(r,\theta);
\end{aligned} \tag{10.40}$$

$\Phi(r, \theta)$ is the flux through the cap defined by a particular r and θ, as shown by the shaded area in Fig. 10.6. As θ is varied the flux through the cap varies. As $\theta \to 0$ the loop shrinks to a point and the flux passing through the cap approaches zero:

$$\Phi(r, 0) = 0.$$

As the loop is lowered over the sphere the cap encloses more and more flux until, eventually, at $\theta \to \pi$ we should have, from (10.39),

$$\Phi(r, \pi) = 4\pi g. \tag{10.41}$$

However, as $\theta \to \pi$ *the loop has again shrunk to a point* so the requirement that $\Phi(r, \pi)$ is finite entails, from (10.40), that A *is singular at* $\theta = \pi$. This argument holds for all spheres of all possible radii, so it follows that **A** is singular along *the entire negative z axis*. This is known as the *Dirac string*. It is clear that by a suitable choice of coordinates the string may be chosen to be along any direction, and, in fact, need not be straight, but must be continuous.

The singularity in **A** gives rise to the so-called Dirac veto – that the wave function vanish along the negative z axis. Its phase is therefore indeterminate there and referring to (10.40) there is *no necessity* that as $\theta \to \pi$, $\Delta\alpha \to 0$. We must have $\Delta\alpha = 2\pi n$, however, in order for ψ to be single-valued. From (10.40) and (10.41) we then have

$$2\pi n = \frac{e}{\hbar c} 4\pi g,$$

■

$$eg = \tfrac{1}{2} n \hbar c. \tag{10.42}$$

This is the Dirac quantisation condition. It implies that the product of *any* electric with *any* magnetic charge is given by the above. Then, in principle, if there exists a magnetic charge anywhere in the universe all electric charges will be quantised:

$$e = n \frac{\hbar c}{2g}.$$

This is a possible explanation for the observed 'quantisation' of electric charge (see the footnote on p. 85), though nowadays this is more commonly

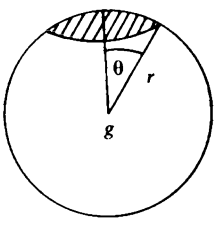

Fig. 10.6.

ascribed to the existence of quarks and non-Abelian symmetry groups. Note, however, that the quantisation condition has an explicit dependence on Planck's constant, and therefore on the quantum theory. In units $\hbar = c = 1$ (10.42) becomes

$$eg = \tfrac{1}{2}n. \tag{10.43}$$

Let us now derive an expression for the vector potential A_μ. As seen above, it is singular. This much is clear from (10.38), for if $\mathbf{B} = \operatorname{curl}\mathbf{A}$ and $\mathbf{A}$ is regular $\operatorname{div}\mathbf{B} = 0$, and no magnetic charges may exist. From the argument above, $\mathbf{A}$ is constructed by considering the pole as the end-point of a string of magnetic dipoles whose other end is at infinity. This gives

$$A_x = g\frac{-y}{r(r+z)}, \quad A_y = g\frac{x}{r(r+z)}, \quad A_z = 0 \tag{10.44}$$

or

$$A_r = A_\theta = 0, \quad A_\phi = \frac{g}{r}\frac{1-\cos\theta}{\sin\theta}. \tag{10.45}$$

$\mathbf{A}$ is clearly singular along $r = -z$. If, on the other hand, the Dirac string were chosen to be along $r = z$, we should have

$$A_r = A_\theta = 0, \quad A_\phi = -\frac{g}{r}\frac{1+\cos\theta}{\sin\theta}. \tag{10.46}$$

The rationale for writing the alternative expressions (10.45) and (10.46) is that the Dirac string singularity is clearly unphysical, and in these expressions it is in different places. The only *physical* singularity in $\mathbf{A}$ is at the origin, where, from (10.38), $\operatorname{div}\mathbf{B} = \operatorname{div}(\operatorname{curl}\mathbf{A})$ is singular. Since it is obviously desirable to get rid of unphysical singularities, this suggests the following construction. Divide the space surrounding the monopole – the sphere, essentially – into two overlapping regions R_a and R_b, as shown in Fig. 10.7. R_a excludes the negative

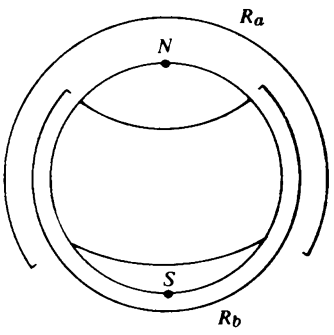

Fig. 10.7. R_a and R_b are overlapping domains on the sphere. R_a excludes the S pole, R_b the N pole.

z axis (S pole) and R_b excludes the positive z axis (N pole). In each region $\mathbf{A}$ is defined differently:

$$A_r^a = A_\theta^a = 0, \quad A_\phi^a = \frac{g}{r}\frac{1-\cos\theta}{\sin\theta}. \tag{10.47}$$

$$A_r^b = A_\theta^b = 0, \quad A_\phi^b = -\frac{g}{r}\frac{1+\cos\theta}{\sin\theta}. \tag{10.48}$$

Referring to (10.45) and (10.46), it is clear that $\mathbf{A}^a$ and $\mathbf{A}^b$ are *both finite in their own domain*. In the region of overlap, however, they are not the same, but are related by a *gauge transformation* ($\hbar = c = 1$):

$$A_\phi^b = A_\phi^a - \frac{2g}{r\sin\theta} = A_\phi^a - \frac{i}{e}S\nabla_\phi S^{-1} \tag{10.49}$$

with

$$S = \exp(2ige\phi) \tag{10.50}$$

The covariant form of (10.49) is

$$A_\mu^b = A_\mu^a - \frac{\mathrm{i}}{e}S\partial_\mu S^{-1}. \tag{10.51}$$

The requirement that the gauge transform function S be single-valued as $\phi \to \phi + 2\pi$ is clearly the Dirac quantisation condition (10.43). To check that (10.47) and (10.48) really do represent a monopole, we calculate the total magnetic flux through a sphere surrounding the origin.

$$\begin{aligned}\Phi &= \int F_{\mu\nu}\,\mathrm{d}x^{\mu\nu} \\ &= \oint \operatorname{curl}\mathbf{A}\cdot\mathrm{d}\mathbf{S} \\ &= \int_{R_a}\operatorname{curl}\mathbf{A}\cdot\mathrm{d}\mathbf{S} + \int_{R_b}\operatorname{curl}\mathbf{A}\cdot\mathrm{d}\mathbf{S}.\end{aligned}$$

Here we take R_a and R_b as not actually overlapping, but having a common boundary, which for convenience is taken to be the equator $\theta = \pi/2$. Since R_a and R_b have boundaries Stokes' theorem is applicable, and since the equator bounds R_a in a positive orientation and R_b in a negative one we have

$$\begin{aligned}\Phi &= \oint_{\theta=\pi/2}\mathbf{A}^a\cdot\mathrm{d}\mathbf{l}^a - \oint_{\theta=\pi/2}\mathbf{A}^b\cdot\mathrm{d}\mathbf{l}^b \\ &= \frac{\mathrm{i}}{e}\oint\frac{\mathrm{d}}{\mathrm{d}\phi}(\ln S^{-1})\,\mathrm{d}\phi \\ &= 4\pi g\end{aligned}$$

from (10.50), and using (10.47) and (10.48). This agrees with (10.41).

This construction is due to Wu and Yang, and is, in essence, a fibre bundle formulation of the magnetic monopole. The base space (3-dimensional space R^3 minus the origin $\approx R^3 - (\text{point}) \approx S^2 \times R^1$) is parameterised in two independent ways, corresponding to two overlapping but not identical regions. In each region the vector potential is given by a different expression. Readers familiar with the Möbius strip will recognize a similarity here. There is no unique parameterisation of the Möbius strip; *locally* it is the direct product of an interval $(0, 1)$ and a circle, but globally the circle has to be divided into two distinct overlapping regions, with a different parameterisation of the strip in each region.

There is thus a fibre-bundle formulation of the Dirac monopole. The base space is essentially S^2 (the sphere surrounding the monopole) and the group space is S^1 (since the gauge group is $U(1)$). The fibre bundle is not $S^2 \times S^1$ but S^3, which is *locally* the same as $S^2 \times S^1$ but is globally distinct. For further details on the fibre-bundle formulation, the reader is referred to the literature.

10.4 The 't Hooft–Polyakov monopole

In the context of Maxwell's electrodynamics, with Abelian gauge group $U(1)$, it is clear that although magnetic charges may be 'added' to the theory, there is no necessity for doing this. A theory with monopoles is more symmetric between electricity and magnetism than one without, but this does not amount to a *requirement* that monopoles exist. They may or may not; the above considerations do not allow us to decide. When the gauge symmetry is enlarged to a non-Abelian group, however, and spontaneous symmetry breaking is introduced, the field equations yield a solution which corresponds to a magnetic charge. If such theories are correct, then, magnetic monopoles *must* exist, and should therefore be looked for. It is a matter of natural curiosity to enquire where the magnetic charge in this model comes from, since the matter and gauge fields in the theory carry electric charge *only*. It will not surprise the reader to hear that the origin of the magnetic charge is topological. The theoretical possibility of monopoles of this type was discovered in 1974 by t' Hooft and Polyakov.

We consider a theory with an $O(3)$ symmetry group, containing the gauge held $F^a_{\mu\nu}$ (a is the group index) and an isovector Higgs field ϕ^a. The Lagrangian is (cf. (8.42))

$$\mathcal{L} = -\tfrac{1}{4}F^a_{\mu\nu}F^{\mu\nu a} + \tfrac{1}{2}(D_\mu\phi^a)(D^\mu\phi^a) - \frac{m^2}{2}\phi^a\phi^a - \lambda(\phi^a\phi^a)^2 \qquad (10.52)$$

where

$$\left.\begin{aligned} F^a_{\mu\nu} &= \partial_\mu A^a_\nu - \partial_\nu A^a_\mu + e\varepsilon^{abc}A^b_\mu A^c_\nu, \\ D_\mu\phi^a &= \partial_\mu\phi^a + e\varepsilon^{abc}A^b_\mu\phi^c. \end{aligned}\right\} \qquad (10.53)$$

We are interested in static solutions in which the gauge potentials have the non-trivial form

$$\left.\begin{aligned} A_i^a &= -\varepsilon_{iab}\frac{r^b}{er^2} \quad (r \to \infty), \\ A_0^a &= 0. \end{aligned}\right\} \tag{10.54}$$

and the scalar field is

$$\phi^a = F\frac{r^a}{r} \quad (r \to \infty) \tag{10.55}$$

with $F^2 = -m^2/4\lambda$. These expressions have a remarkable form because of the mixing they employ between space and isospace indices. For example, (10.55) describes a field which, in the x direction in space, has only an isospin '1' component, in the y direction, only a '2' component, and in the z direction, only a '3' component. In a manner of speaking, it is 'radial' – Polyakov calls it a 'hedgehog' solution. It can be shown ('t Hooft 1974) that there exist regular solutions to the field equations derived from (10.52), which have the asymptotic form (10.54), (10.55). For example, the equation of motion of ϕ is

$$-(m^2 + 4\lambda\phi^b\phi^b)\phi^a = D_\mu(D^\mu\phi^a).$$

Equation (10.55) implies $|\phi| = F$, so the left-hand side of the above equation vanishes at infinity. It is easy to see that $D_\mu\phi^a$ also vanishes; for with $i = x, y, z$ we have

$$\begin{aligned} D_i\phi^a &= F\partial_i\left(\frac{r^a}{r}\right) + e\varepsilon^{abc}A_i^b F\frac{r^c}{r} \\ &= F\left(\frac{\delta^{ia}}{r} - \frac{r^i r^a}{r^3}\right) - \varepsilon^{abc}\varepsilon_{ibm}\frac{Fr^m r^c}{r^3} \\ &= 0. \end{aligned}$$

Hence, at infinity, ϕ takes on its vacuum value and is covariantly constant, but has the non-trivial boundary condition (10.55), rather than the more usual ('Abelian') condition $\phi^{1,2} = 0$, $\phi^3 \neq 0$. On the other hand, $F_{\mu\nu}^a$ is not zero at infinity. We shall see below that there is a radial magnetic field. This solution is sketched in Fig. 10.8.

Now let us generalise the definition of the electromagnetic field $F_{\mu\nu}$ so that it reduces to the usual one when the scalar field ϕ has only a third component. We put

$$F_{\mu\nu} = \frac{1}{|\phi|}\phi^a F_{\mu\nu}^a - \frac{1}{e|\phi|^3}\varepsilon_{abc}\phi^a(D_\mu\phi^b)(D_\nu\phi^c). \tag{10.56}$$

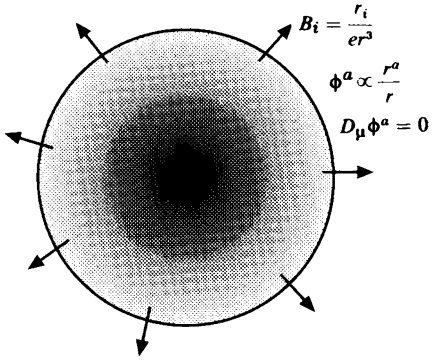

Fig. 10.8. The asymptotic forms of the gauge and scalar fields constituting a 't Hooft–Polyakov monopole. Polyakov calls it a 'hedgehog' solution.

It is quite clear that when

$$\left.\begin{aligned} A_\mu^{1,2} &= 0, \quad A_\mu^3 \equiv A_\mu \neq 0, \\ \phi^{1,2} &= 0, \quad \phi^3 = F \neq 0 \end{aligned}\right\} \tag{10.57}$$

this gives the usual $F_{\mu\nu}$, so long as $A_\mu^3 = A_\mu$, the Maxwell vector potential. Now, defining

$$A_\mu = \frac{1}{|\phi|}\phi^a A_\mu^a \tag{10.58}$$

a straightforward calculation gives

$$F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu - \frac{1}{e|\phi|^3}\varepsilon_{abc}\phi^a(\partial_\mu\phi^b)(\partial_\nu\phi^c). \tag{10.59}$$

This is similar to, but more complicated than, the usual definition of the electromagnetic field, but it reduces to it when ϕ becomes fixed in isospace. Inserting the asymptotic conditions (10.54) and (10.55), it is easily seen that $A_\mu = 0$, so all the electromagnetic field is contributed by the Higgs field; and we find

$$F_{0i} = 0, \quad F_{ij} = -\frac{1}{er^3}\varepsilon_{ijk}r^k. \tag{10.60}$$

This corresponds to a radial magnetic field (see (2.221))

$$B_k = \frac{r^k}{er^3}. \tag{10.61}$$

The magnetic flux is, from (10.39),

$$\Phi = \frac{4\pi}{e},$$

so by comparison with (10.41) the magnetic charge g is such that

$$eg = 1. \tag{10.62}$$

From (10.43), this is twice the Dirac unit. We conclude that the configuration of gauge and scalar fields with asymptotic form (10.54–10.55) carries a magnetic charge – i.e. when viewed from infinity, there is a radial magnetic field. It has been shown by 't Hooft that this configuration is everywhere non-singular, and therefore has a finite energy. He estimates the monopole mass to be of the order $137M_W$, where M_W is a typical vector boson mass, so the monopoles are extremely heavy. The mass is inversely proportional to e^2 (cf. (10.9)).

What is the origin of this magnetic charge? How does it come about that a configuration of fields carrying electric charge only can arrange itself in such a way as to simulate a magnetic charge? To answer this, we write the magnetic current K_μ as

$$\begin{aligned} K^\mu &= \partial_\nu \tilde{F}^{\mu\nu} \\ &= \tfrac{1}{2}\varepsilon^{\mu\nu\rho\sigma}\partial_\nu F_{\rho\sigma} \end{aligned} \tag{10.63}$$

where $\tilde{F}_{\mu\nu}$ is the dual of $F_{\mu\nu}$ – cf. equation (2.236) which holds when no magnetic sources are present. From (10.59) we then have

$$K^\mu = -\frac{1}{2e}\varepsilon^{\mu\nu\rho\sigma}\varepsilon_{abc}\partial_\nu\hat{\phi}^a\partial_\rho\hat{\phi}^b\partial_\sigma\hat{\phi}^c \tag{10.64}$$

where

$$\hat{\phi}^a = \frac{1}{|\phi|}\phi^a.$$

We see that the magnetic current *depends on the Higgs field only*, as noticed above in (10.60). Moreover, this current is *identically conserved*:

$$\partial_\mu K^\mu = 0. \tag{10.65}$$

This property is reminiscent of the current (10.10) for the sine–Gordon kink. The conservation of both these currents does not follow from a symmetry of the Lagrangian, so they are not Noether currents. It will be recalled that the sine–Gordon charge (10.11) – the 'kink number' – depends simply on the non-trivial boundary conditions. We anticipate the same phenomenon here. The conserved magnetic charge is

$$\begin{aligned} M &= \frac{1}{4\pi}\int K^0 \,\mathrm{d}^3x \\ &= -\frac{1}{8\pi e}\oint_{S^2}\varepsilon_{ijk}\varepsilon_{abc}\hat{\phi}^a\partial_j\hat{\phi}^b\partial_k\hat{\phi}^c(\mathrm{d}^2S)_i. \end{aligned} \tag{10.66}$$

Here the integral is taken over the sphere S^2 at infinity, which, of course, is the

boundary of the static field configuration ϕ. Since ϕ must be *single-valued*, as $(\mathrm{d}S)_i$ covers the sphere once, the vector ϕ will be covered an *integral number* of times, say d. It then follows (Arafune, Freund & Goebel 1975) that the integral in (10.66) is $8\pi d$, hence

$$M = \frac{d}{e}, \quad d \text{ integer.} \tag{10.67}$$

Since ϕ^a is an isovector, the unit vector $\hat{\phi}$ describes a sphere S^2 in field space (isospace), so the boundary describes a mapping of the sphere S^2 in coordinate space onto the $\hat{\phi}$ manifold, which is S^2.

$$\hat{\phi}\text{: } S^2 \text{ in field space} \rightarrow S^2 \text{ in coordinate space;} \tag{10.68}$$

d is called the *Brouwer degree* of this mapping. It is necessarily integral. So equation (10.67) displays explicitly the topological nature of the 't Hooft–Polyakov monopole.

In the model considered by 't Hooft, the non-Abelian group is $SO(3)$, electromagnetism being represented by the Abelian subgroup $U(1)$. An interesting question is: how does the existence of magnetic charge in non-Abelian gauge theories depend on the gauge group? To answer this we begin by reflecting on equation (10.68). It is obvious that in general terms what is important is the $\hat{\phi}$ manifold (so the gauge theory *must* be spontaneously broken). What is the space of the $\hat{\phi}$ manifold in general? In Chapter 8 we learned that it is the vacuum manifold. If the symmetry group of the theory is G (in this case $SO(3)$), and the unbroken subgroup is H (in this case $U(1)$), then transformations belonging to H leave the vacuum manifold invariant. So the space of $\hat{\phi}$ is the set of transformations in G which are *not* related by a transformation belonging to H. This is the definition of a coset space. In schematic terms the elements of the gauge G may be written

$$G = H + HM_1 + HM_2 + \cdots \tag{10.69}$$

where H denotes the elements of the subgroup H, and M_1, M_2, ... belong to G but not to H, and are all different. The vacuum manifold is essentially the space of the elements M_i, and that is the coset space of G/H. Consulting (10.68) again, the existence of magnetic monopoles requires a non-trivial mapping of G/H onto S^2, the boundary in co-ordinate space. As we saw in Chapter 3 (equation (3.114)) these mappings form a group, in this case the *second homotopy group of* G/H, $\pi_2(G/H)$. Magnetic monopoles will exist if this group is non-trivial. We now invoke a mathematical theorem involving homotopy groups (Coleman in Zichichi 1977; Tyupkin, Fateev & Shvarts 1975; Monastyrskii & Perelomov 1975).

Theorem. $\pi_2(G/H)$ is isomorphic to the kernel of the natural homomorphism of $\pi_1(H)$ into $\pi_1(G)$. (10.70)

To explain the terms in this theorem: $\pi_1(H)$ and $\pi_1(G)$ are the first homotopy groups of H and G. They are trivial if the groups are simply connected, isomorphic to $\mathbb{Z}_2(C_2)$ if the groups are doubly connected, etc. Since every closed path in H is also a closed path in G, there is a natural mapping of $\pi_1(H)$ into $\pi_2(G)$; this is called a homomorphism. The kernel of the homomorphism is the set of elements of $\pi_1(H)$ which are mapped onto the *identity* of $\pi_1(G)$.

Let us watch this theorem in action by applying it to the 't Hooft case, where $G = SO(3)$, $H = U(1)$. Since $SO(3)$ is doubly connected (see above), $\pi_1(G) = \mathbb{Z}_2$. On the other hand, $U(1)$ is infinitely connected (the group space is a circle, and a closed curve going n times round a circle cannot be continuously deformed into one going $m(\neq n)$ times round), so $\pi_1(H) = \mathbb{Z}$, the additive group of integers. So the kernel of the mapping of $\pi_1(H)$ into $\pi_1(G)$ is the additive group of *even* integers, hence

$$\pi_2(SO(3)/U(1)) = \text{additive group of even integers.} \tag{10.71}$$

This is consistent with what we found; the monopole charge was *twice* the Dirac quantum.

The trouble is that the true non-Abelian electroweak group is not $SO(3)$, but $SU(2) \times U(1)$, given by the Weinberg-Salam model. (The $SO(3)$ model is the Georgi–Glashow model (Georgi & Glashow 1972). Its salient characteristic is that the only neutral current in it is the electromagnetic current. It was therefore rendered obsolete by the discovery of weak neutral current events, such as $\nu_e + p \to \nu_e + p + \pi^0$.) Moreover, the electromagnetic subgroup, although given by $U(1)$, is irregularly embedded in $SU(2) \times U(1)$, and so is *non-compact*, with the consequence that magnetic monopoles do not exist in the Weinberg–Salam model. To see this argument, note that in this model there are two $U(1)$ subgroups, so that a particle with a third component of weak isospin I_3^{w} and weak hypercharge Y^{w}, will transform under these $U(1)$ groups by

$$\exp(\mathrm{i}\alpha I_3^{\mathrm{w}}) \exp(\mathrm{i}\beta Y^{\mathrm{w}}). \tag{10.72}$$

The group space of $U(1)$ is a circle, or, equivalently, a line with the points 0 and 2π identified. Hence the group space of $U(1) \times U(1)$ may be represented, as in Fig. 10.9, by a square $ABCD$, with the edges AC and AB identified with BD and CD respectively. This is a *torus* T^2. (In general the group space of the direct product of n groups $U(1)$ is a toroid T^n.) The group element (10.72) will correspond to a point (α, β) in the group space T^2 of Fig. 10.9. Electric charge Q in the Weinberg–Salam model is given by

$$Q = \sin\theta_{\mathrm{W}} I_3^{\mathrm{w}} + \cos\theta_{\mathrm{W}} Y^{\mathrm{w}}$$

where θ_{W} is the Weinberg angle (cf. equation (8.82)). Under an electromagnetic gauge transformation through an angle γ, the state vector for a particle

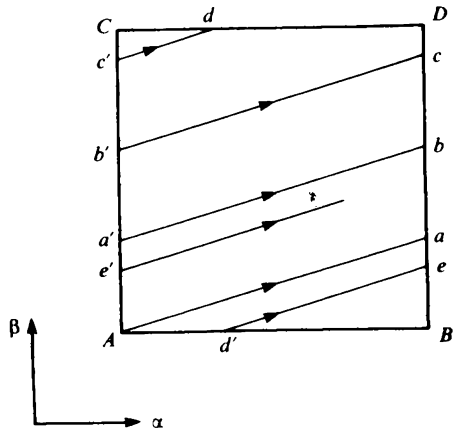

Fig. 10.9. The group space of $U(1) \otimes U(1)$ is a square with opposite edges identified, hence a torus. An electromagnetic gauge transformation traces out the line $Aaa'bb'cc'dd'ee' \ldots$.

with charge Q is multiplied by

$$\exp(i\gamma Q) = \exp[i(\gamma \sin\theta_W I_3^W + \gamma \cos\theta_W Y^W)],$$

and to this transformation corresponds a point in group space given by

$$\alpha = \gamma \sin\theta_W, \quad \beta = \gamma \cos\theta_W$$

hence

$$\alpha/\beta = \tan\theta_W = \text{irrational}. \tag{10.73}$$

The above condition corresponds to a line in group space $Aaa'bb'cc'dd'ee' \ldots$ (see Fig. 10.9), which, since α/β is an irrational number, is a line of *infinite length*. It winds round the torus without ever meeting itself again. Hence the electromagnetic gauge group in the Weinberg–Salam model has infinite volume, and is *non-compact*. It follows that $\pi_1(H)$ does not exist (or is trivial) so $\pi_2(G/H)$ is also trival and no monopoles exist. If nature is 'grand-unified', however, and the electroweak group $SU(2) \times U(1)$ is a subgroup of a grand-unified semisimple group, say $SU(5)$, then this argument no longer holds, and monopoles may exist. These questions have recently come to life following a claim that a monopole has been discovered (Cabrera 1982; see also Cabrera *et al.* 1983).

Comparing the 't Hooft–Polyakov monopole with the Dirac monopole of §10.3, it will seem as if they have almost nothing in common; to be more precise, nothing at all except that they both possess magnetic charge. This is not quite true, however, and it may be helpful to conclude this section by showing how the two may be related. We start with a Dirac monopole with a string singularity along the negative z axis. The vector potential is therefore given by (10.45). Now we submerge this in an $SU(2)$ theory, with the vector

potential aligned in the third direction in isospin space. Using the matrix potential $A_\mu = A_\mu^a \tau^a$ then gives

$$A_0 = A_r = A_\theta = 0, \quad A_\phi = \tau_3\left(-\frac{g}{r}\right)\left(\frac{1-\cos\theta}{\sin\theta}\right). \tag{10.74}$$

In addition, we introduce a scalar field ϕ, with vacuum expectation value F, also aligned along the third direction in isospin space:

$$\phi = \tau_3 F. \tag{10.75}$$

Now we transform A_μ and ϕ by a space-dependent isospin gauge transformation. A general $SU(2)$ gauge transformation may be characterised by the Euler angles (α, β, γ) and written

$$\begin{aligned} S &= e^{(i/2)\alpha\tau_3} e^{(i/2)\beta\tau_2} e^{(i/2)\gamma\tau_3} \\ &= \begin{pmatrix} \cos\beta/2\, e^{i(\alpha+\gamma)/2} & \sin\beta/2\, e^{i(-\gamma+\alpha)/2} \\ -\sin\beta/2\, e^{i(\gamma-\alpha)/2} & \cos\beta/2\, e^{-i(\gamma+\alpha)/2} \end{pmatrix} \end{aligned}$$

Now we put $\gamma = -\alpha = \phi$, $\beta = -\theta$, giving

$$S = \begin{pmatrix} \cos\theta/2 & -e^{-i\phi}\sin\theta/2 \\ e^{i\phi}\sin\theta/2 & \cos\theta/2 \end{pmatrix} \tag{10.76a}$$

hence

$$S^{-1} = \begin{pmatrix} \cos\theta/2 & e^{-i\phi}\sin\theta/2 \\ -e^{i\phi}\sin\theta/2 & \cos\theta/2 \end{pmatrix}. \tag{10.76b}$$

The transformation law for A_μ is (as in (3.162), but with $A_\mu = A_\mu^a \tau^a$ and $g \to e$)

$$A'_\mu = SA_\mu S^{-1} + \frac{2i}{e} S\partial_\mu S^{-1}. \tag{10.77}$$

From (10.76*b*) it follows that

$$\begin{aligned} \partial_r S^{-1} &= 0, \quad \partial_\theta S^{-1} = \frac{1}{2r}\begin{pmatrix} -\sin\theta/2 & e^{-i\phi}\cos\theta/2 \\ -e^{i\phi}\cos\theta/2 & -\sin\theta/2 \end{pmatrix}, \\ \partial_\phi S^{-1} &= \frac{-i}{r\sin\theta}\begin{pmatrix} 0 & e^{-i\phi}\sin\theta/2 \\ e^{i\phi}\sin\theta/2 & 0 \end{pmatrix}. \end{aligned}$$

Substituting these into (10.77), using (10.74) and putting $g = 1/e$ (from (10.62)) gives, after straighforward manipulations,

$$\begin{aligned} A'_0 &= A'_r = 0, \\ A'_\theta &= \frac{1}{er}(\tau_1 \sin\phi - \tau_2 \cos\phi), \\ A'_\phi &= \frac{1}{er}(\tau_1 \cos\theta\cos\phi + \tau_2 \cos\theta \sin\phi - \tau_3 \sin\theta). \end{aligned}$$

The Cartesian components of A may then be found; for example,

$$\begin{aligned} A'_x &= A'_r \cos\phi \sin\theta + A'_\theta \cos\theta \cos\phi - A'_\phi \sin\phi \\ &= \frac{1}{er}\left[\tau_2\left(\frac{-z}{r}\right) + \tau_3\left(\frac{y}{r}\right)\right]. \end{aligned} \tag{10.78}$$

This is the 'hedgehog' form (10.54). Under the same transformation (10.76) the Higgs field (10.75) becomes

$$\begin{aligned} \phi' &= S\phi S^{-1} \\ &= F\begin{pmatrix} \cos\theta & e^{-i\phi}\sin\theta \\ e^{i\phi}\sin\theta & -\cos\theta \end{pmatrix} \\ &= F(\sin\theta\cos\phi\tau_1 + \sin\theta\sin\phi\tau_2 + \cos\theta\tau_3), \end{aligned} \tag{10.79}$$

i.e.

$$\phi'^a = F\frac{r^a}{r} \tag{10.80}$$

as in (10.55). As a result of this transformation, the string singularity of the Dirac potential *disappears*, and the source of the monopole resides, as it were, in the Higgs field. From equation (10.59) we may say that the gauge transformation transfers responsibility for the monopole from the first (Dirac) term, to the second (topological, Higgs) one. Thus the Dirac and 't Hooft–Polyakov monopoles are not so unconnected as they at first appear.

10.5 Instantons

Our final example of soliton solutions is concerned with those which are localised in time as well as in space, and which 't Hooft has therefore christened 'instantons'. (An alternative name, suggested by Polyakov, is 'pseudo particles'.) It is not surprising that such solutions exist, since the gauge-field equations are fully relativistic, so allow a topological non-triviality in time as well as in space. Moreover, the gauge group $SU(2)$ plays a rather special role as may be seen from the following consideration. To begin with, space–time is considered to be 'Euclideanised' so that it becomes E^4. Its boundary is then S^3, the 3-sphere. On the other hand, it was seen in (10.34) above that the group space of $SU(2)$ is also S^3. Hence topologically non-trivial solutions to the $SU(2)$ gauge-field equations are possible if there exist non-trivial (non-homotopic) mappings of S^3 onto S^3, that is if $\pi_3(S^3)$ is non-trivial (see (3.114)). And indeed it is:

$$\pi_3(S^3) = \mathbb{Z}. \tag{10.81}$$

It follows that instantons are therefore possible in the *pure* gauge theory; spontaneous symmetry breaking is unnecessary. This distinguishes instantons

from monopoles. The plan in this section will be to write down the instanton solution, exhibiting its topological nature, and then to mention briefly the physical consequences that follow from the existence of instantons. There is a considerable amount of literature on this topic, so our treatment will be very introductory and readers are referred to the many excellent reviews to broaden their knowledge. In addition, instantons have aroused the interest of a number of pure mathematicians, and many papers explore their connections with topology and algebraic geometry. But these matters are beyond the scope of this book, and readers are again referred elsewhere.

We begin with some mathematical preliminaries. Euclidean space has coordinates (x_1, x_2, x_3, x_4) with (see (6.16))

$$x_0 = -\mathrm{i}x_4 \tag{10.82}$$

(and $x_0 = ct$). The Euclidean field tensor $F^a_{\mu\nu}$ is defined (Vainshtein *et al.* 1982) in the same way as the Minkowski tensor (see (3.169)):

$$F^a_{\mu\nu} = \partial_\mu A^a_\nu - \partial_\mu A^a_\mu + g\varepsilon^{abc} A^b_\mu A^c_\nu \tag{10.83}$$

with

$$A_\mu = \tfrac{1}{2}\sigma^a A^a_\mu, \quad F_{\mu\nu} = \tfrac{1}{2}\sigma^a F^a_{\mu\nu}. \tag{10.84}$$

This takes the form (see (3.166))

$$F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu - \mathrm{i}g[A_\mu, A_\nu]. \tag{10.85}$$

Defining

$$\partial_{[\mu} A_{\nu]} \equiv \partial_\mu A_\nu - \partial_\nu A_\mu \tag{10.86}$$

this becomes

$$F_{\mu\nu} = \partial_{[\mu} A_{\nu]} - \mathrm{i}g[A_\mu, A_\nu]. \tag{10.87}$$

The *dual* of $F_{\mu\nu}$, denoted $\tilde{F}_{\mu\nu}$, is defined by

$$\tilde{F}_{\mu\nu} = \tfrac{1}{2}\varepsilon_{\mu\nu\rho\sigma} F_{\rho\sigma} \tag{10.88}$$

(remembering that in Euclidean space there is no need to distinguish upper and lower indices). With $\varepsilon_{1234} = 1$ this yields

$$\tilde{\tilde{F}}_{\mu\nu} = F_{\mu\nu}, \tag{10.89}$$

whereas in Minkowski space, since when $\varepsilon^{0123} = 1$, then $\varepsilon_{0123} = -1$, so

$$\tilde{\tilde{F}}_{\mu\nu} = -F_{\mu\nu} \quad \text{(in Minkowski space)}. \tag{10.90}$$

Under gauge transformations

$$A'_\mu = SA_\mu S^{-1} - \frac{\mathrm{i}}{g}(\partial_\mu S)S^{-1}, \tag{10.91}$$

$$F'_{\mu\nu} = SF_{\mu\nu}S^{-1}. \tag{10.92}$$

Now we define

$$K_\mu = \tfrac{1}{4}\varepsilon_{\mu\nu\kappa\lambda}\left(A^a_\nu \partial_\kappa A^a_\lambda + \frac{g}{3}\varepsilon_{abc}A^a_\nu A^b_\kappa A^c_\lambda\right)$$
$$= \varepsilon_{\mu\nu\kappa\lambda}\,\mathrm{Tr}\left(\tfrac{1}{2}A_\nu \partial_\kappa A_\lambda - \frac{\mathrm{i}g}{3}A_\nu A_\kappa A_\lambda\right). \tag{10.93}$$

Then

$$\partial_\mu K_\mu = \tfrac{1}{4}\,\mathrm{Tr}\,\tilde{F}_{\mu\nu}F_{\mu\nu} = \tfrac{1}{8}\tilde{F}^a_{\mu\nu}F^a_{\mu\nu} \tag{10.94}$$

so that $\mathrm{Tr}\,\tilde{F}F$ is a total divergence.

Proof. Because of the cyclic property of the trace

$$\partial_\mu K_\mu = \varepsilon_{\mu\nu\kappa\lambda}\,\mathrm{Tr}\,[\tfrac{1}{2}(\partial_\mu A_\nu)(\partial_\kappa A_\lambda) - \mathrm{i}g(\partial_\mu A_\nu)A_\kappa A_\lambda]$$

On the other hand,

$$\mathrm{Tr}\,F_{\mu\nu}\tilde{F}_{\mu\nu} = \tfrac{1}{2}\varepsilon_{\mu\nu\kappa\lambda}\,\mathrm{Tr}\,\{\partial_{[\mu}A_{\nu]} - \mathrm{i}g[A_\mu, A_\nu]\}\{\partial_{[\kappa}A_{\lambda]} - \mathrm{i}g[A_\kappa, A_\nu]\}$$
$$= 2\varepsilon_{\mu\nu\kappa\lambda}\,\mathrm{Tr}\,(\partial_\mu A_\nu)(\partial_\kappa A_\lambda) - 2\mathrm{i}g\varepsilon_{\mu\nu\kappa\lambda}\,\mathrm{Tr}\,A_\mu A_\nu(\partial_\kappa A_\lambda)$$
$$- 2\mathrm{i}g\varepsilon_{\mu\nu\kappa\lambda}\,\mathrm{Tr}\,(\partial_\mu A_\nu)A_\kappa A_\lambda - 2g^2\varepsilon_{\mu\nu\kappa\lambda}\,\mathrm{Tr}\,A_\mu A_\nu A_\kappa A_\lambda.$$

Because of the cyclic trace property, the second two terms above are equal and the last one vanishes. Hence (10.94) is proved.

Now consider a 4-dimensional volume V^4 in E^4, with boundary $\partial V^4 \sim S^3$. Suppose it is a pure vacuum, $A_\mu = 0$, $F_{\mu\nu} = 0$. Then $K_\mu = 0$. The field equations (in the absence of matter)

$$D_\mu F_{\mu\nu} = 0 \tag{10.95}$$

are clearly satisfied over the whole region V^4, as is the Bianchi identity

$$D_\mu \tilde{F}_{\mu\nu} = 0 \tag{10.96}$$

which, of course, *must* be satisfied. Applying Gauss' theorem to (10.94) gives

$$\int_{V^4} \mathrm{Tr}\,F_{\mu\nu}\tilde{F}_{\mu\nu}\,\mathrm{d}^4x = 4\int_{V^4}\partial_\mu K_\mu\,\mathrm{d}^4x$$
$$= 4\oint_{\partial V^4} K_\perp\,\mathrm{d}^3x. \tag{10.97}$$

This is trivially satisfied if V^4 is a pure vacuum.

Now we perform a (space–time-dependent) gauge transformation at the *boundary* S^3

$$A_\mu \to -\frac{\mathrm{i}}{g}(\partial_\mu S)S^{-1} \quad (\text{on } S^3), \tag{10.98}$$

i.e.

$$F_{\mu\nu} = 0$$

so the boundary becomes a 'pure gauge' vacuum; and take

$$S = \frac{x_4 + \mathrm{i}\mathbf{x}\cdot\boldsymbol{\sigma}}{\sqrt{\tau^2}} \tag{10.99}$$

where

$$\tau^2 = x_4^2 + \mathbf{x}^2. \tag{10.100}$$

Then after some straightforward (but lengthy) algebra we find

$$\left.\begin{aligned} A_i &= \frac{\mathrm{i}}{g\tau^2}[x_i - \sigma_i(\boldsymbol{\sigma}\cdot\mathbf{x} + \mathrm{i}x_4], \\ A_4 &= -\frac{1}{g\tau^2}\boldsymbol{\sigma}\cdot\mathbf{x}, \end{aligned}\right\} \tag{10.101}$$

and

$$K_\mu = \frac{2x_\mu}{g^2\tau^4}. \tag{10.102}$$

Equation (10.97) then yields

$$\begin{aligned} \int \mathrm{Tr}\, F_{\mu\nu}\tilde{F}_{\mu\nu}\,\mathrm{d}^4x &= 4\oint_{S^3} K_\perp\,\mathrm{d}^3\sigma \\ &= \frac{8\tau}{g^2\tau^4}\oint_{S^3}\mathrm{d(area)} \\ &= \frac{16\pi^2}{g^2}, \end{aligned} \tag{10.103}$$

using the fact that the area of the 3-sphere of radius τ is $2\pi^2\tau^3$. (The area of the unit sphere S^n is $\pi^{n/2}2^{n+1}(n/2)!/(n!)$.) We see immediately that *$F_{\mu\nu}$ cannot be zero over the whole volume V^4*, although it does vanish on the boundary. It will be appreciated that this is a consequence of the fact that K_μ is not gauge invariant.

The above situation is sketched in Fig. 10.10. The field strength $F_{\mu\nu}$ is non-zero inside the volume V^4, but vanishes on the boundary S^3, where A_μ becomes a pure gauge. It is clear that (10.98) is *not* a solution to the gauge-field equations over the whole space, but is simply the asymptotic form as $\tau^2 \to \infty$. How are we to understand this? We first show that the integral (10.103) above defines a *topological index*. It is called the *Pontryagin index* (or Pontryagin class), and denoted q:

$$q = \frac{g^2}{16\pi^2}\mathrm{Tr}\int F_{\mu\nu}\tilde{F}_{\mu\nu}\,\mathrm{d}^4x. \tag{10.104}$$

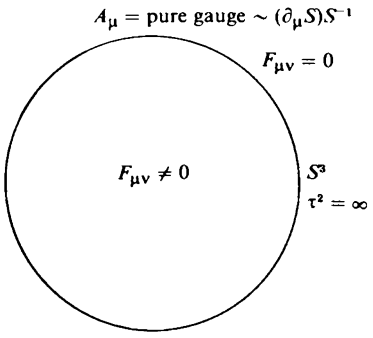

Fig. 10.10. The instanton. Inside the volume V^4 the field strength $F_{\mu\nu}$ is non-vanishing, but $F_{\mu\nu}$ vanishes on the boundary S^3.

Then in the case we are considering we have

$$q = \frac{g^2}{4\pi^2}\int \partial_\mu K_\mu \, \mathrm{d}^4x = 1. \tag{10.105}$$

We shall show that q is the degree of the mapping of the group space, S^3, of $SU(2)$ onto the co-ordinate space boundary S^3. Putting (10.98) into (10.93) gives

$$K_\mu = \frac{1}{6g^2}\varepsilon_{\mu\nu\kappa\lambda}\,\mathrm{Tr}\,(S^{-1}\partial_\nu S)(S^{-1}\partial_\lambda S)(S^{-1}\partial_\kappa S),$$

hence

$$\begin{aligned} q &= \frac{1}{24\pi^2}\oint_{S^3} \varepsilon_{\mu\nu\kappa\lambda}\hat{n}_\mu\,\mathrm{Tr}\,(S^{-1}\partial_\nu S)(S^{-1}\partial_\lambda S)(S^{-1}\partial_\kappa S)\,\mathrm{d}^3\sigma \\ &= \frac{1}{24\pi^2}\oint_{S^3}\frac{\partial(g)}{\partial(\sigma)}\,\mathrm{d}^3\sigma \\ &= \frac{1}{24\pi^2}\int_G \mathrm{d}^3 g \end{aligned} \tag{10.106}$$

where $\mathrm{d}^3 g$ is the invariant element of volume in group space. Hence q gives the (Brouwer) degree of the mapping $S^3 \to S^3$.

This solution, then, is like the soliton solution, except that E^4 has one time and three space dimensions. The similarity is that as one of these co-ordinates passes from $-\infty$ to $+\infty$ the field configuration changes, so that the boundary conditions at $-\infty$ and $+\infty$ are different, rather as in Figs. 10.1 and 10.3 for the sine–Gordon kink. In that case, of course, the relevant co-ordinate is a spatial one. An obvious way to interpret the present solution is as an evolution in *time*, rather than space. This suggests redrawing the boundary S^3 as in Fig.

10.11. I and II are the hypersurfaces $x_4 \to \infty$ and $x_4 \to -\infty$ and III is the hypercylindrical surface joining them. Then

$$q = \frac{1}{24\pi^2}\left[\int_{\text{I-II}} \mathrm{d}^3\sigma\varepsilon_{4ijk}\,\mathrm{Tr}\,(\bar{A}_i\bar{A}_j\bar{A}_k) + \int_{-\infty}^{\infty}\mathrm{d}x_4\int_{\text{III}}\mathrm{d}^2\sigma_i\varepsilon_{i\nu\kappa\lambda}\,\mathrm{Tr}\,(\bar{A}_\nu\bar{A}_\kappa\bar{A}_\lambda)\right] \tag{10.107}$$

where $\bar{A}_\mu = S^{-1}(\partial_\mu S) = \mathrm{i}gA_\mu$.

Now, as remarked above, A_μ is not a pure gauge over the whole volume. The required expression for A_μ is

$$A_\mu = \frac{\tau^2}{\tau^2 + \lambda^2}\left(\frac{-\mathrm{i}}{g}\right)(\partial_\mu S)S^{-1} \tag{10.108}$$

where λ is a constant (Belavin *et al.* 1975). As $x_4 \to \pm\infty$ this tends to the pure gauge form (10.98), but in the 'interior' of the 4-volume V^4 is such that $F_{\mu\nu} \neq 0$, as required. This expression for A_μ is a solution of the field equations, and is the one which should be used in the expression for q above. However, q is gauge invariant so it is convenient to choose a gauge in which $A' = 0$ so that the integral over the 'cylinder' III in (10.107) vanishes (since the condition for a non-vanishing integral is that one of the indices ν, κ, λ should be 4). Such a gauge transformation is‡

$$A'_\mu = UA_\mu U^{-1} - \mathrm{i}(\partial_\mu U)U^{-1} \tag{10.109}$$

when

$$\left.\begin{aligned} U &= \exp\left[\frac{\mathrm{i}\mathbf{x}\cdot\boldsymbol{\sigma}}{(\tau^2+\lambda^2)^{1/2}}\theta\right], \\ \theta &= \tan^{-1}\left[\frac{x_4}{(\tau^2+\lambda^2)^{1/2}} - \frac{\pi}{2}\right] \end{aligned}\right\} \tag{10.110}$$

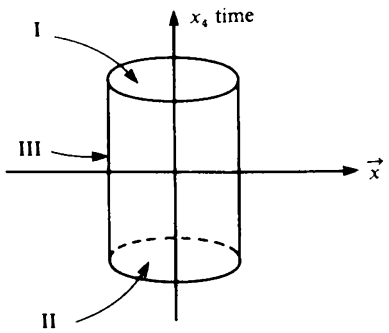

Fig. 10.11. The instanton boundary. I is $x_4 \to \infty$, II is $x_4 \to -\infty$.

‡ This is taken from A. Chakrabarti, talks given at Rencontre de Rabat, May 1978 (unpublished). See also Jackiw & Rebbi (1977).

and A_μ is equal to $\tau^2(\tau^2+\lambda^2)^{-1}$ times the expression (10.98) (with (10.99)). It is seen that

$$A_4' = 0$$

so that q reduces to the difference between two integrals, on the surfaces $x_4 \to -\infty$ and $x_4 \to \infty$. A_i' is a complicated expression which in the cases $x_4 \to \pm\infty$ reduces to

$$\left.\begin{aligned} x_4 &\to \infty, \quad A_i' \to \mathrm{i}(g_n)^{-1}(\partial_i g_n), \\ x_4 &\to -\infty, \quad A_i' \to \mathrm{i}(g_{n-1})^{-1}(\partial_i g_{n-1}), \end{aligned}\right\} \tag{10.111}$$

with

$$g_n = (g_1)^n, \quad g_1 = \exp\left[-\mathrm{i}\pi \frac{\mathbf{x}\cdot\boldsymbol{\sigma}}{(\tau^2+\lambda^2)^{1/2}}\right] \tag{10.112}$$

g_n is clearly an element of the group $SU(2)$, but g_n and g_m $(n \neq m)$ are *not homotopic*. In particular g_1 and $g_0 = 1$ are not homotopic; that is, it is not possible to find a function $g(g_1, a)$ with a a continuous variable between 0 and 1 such that $g(g_1, 1) = g_1$ and $g(g_1, 0) = 1$ (see §3.4). *The instanton therefore describes a solution of the gauge-field equations in which, as x_4 evolves from $-\infty$ to $+\infty$, a vacuum (belonging to homotopy class $n-1$) evolves into another vacuum (belonging to homotopy class n) and the Pontryagin index is*

$$q = n - (n-1) = 1.$$

In between these vacua is a region when the field tensor $F_{\mu\nu}$ is non-vanishing, and therefore there is *positive field energy*. The Yang–Mills vacuum is therefore infinitely degenerate, consisting of an infinite number of homotopically non-equivalent vacua. The instanton solution represents a transition from one vacuum class to another. Physics enters the scene when we ask what is the *amplitude* for this transition. Classically, of course, it is zero, since there is an energy hump in between two vacua. But because of quantum mechanics there is a *barrier penetration factor*. We now consider this.

Quantum tunnelling, θ-vacua and symmetry breaking

What we shall argue is that the barrier penetration amplitude is

$$\mathrm{e}^{-S_\mathrm{E}}, \quad S_\mathrm{E} = \text{Euclidean action}. \tag{10.113}$$

To see this consider the problem of the motion of a single particle through a 1-dimensional potential well, in the quasi-classical (WKB) approximation. If $V > E$ the process is classically ($h = 0$) forbidden, but the actual tunnelling amplitude is

$$\exp\left\{-\frac{1}{\hbar}\int_a^b [2m(V-E)]^{1/2}\,\mathrm{d}x\right\} \equiv \exp\left(-\frac{1}{\hbar}S_\mathrm{E}\right) \tag{10.114}$$

where S_E is *defined* by the integral above. We shall proceed to show that S_E is, in fact, the action for imaginary times. For consider the case where $E > V$, and the transition is classically allowed. In this case the wave function oscillates, and the number of oscillations is given by

$$\frac{1}{h}\int_a^b p\,dx = \frac{1}{h}\int_a^b [2m(E - V)]^{1/2}\,dx. \tag{10.115}$$

On the other hand,

$$\int p\,dx = \int p\dot{x}\,dt = \int (H + L)\,dt = \int (E + L)\,dt.$$

If the total energy is normalised to zero, then

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

which is the total action for the motion from a to b. Now the only difference between (10.114) and (10.115) is that the sign of $E - V$ is reversed. However, the sign of V in the equation of motion

$$m\ddot{x} = -\frac{\partial V}{\partial x}$$

is reversed if we replace t by it. Hence S_E, defined in (10.114), is the action for imaginary times. In the case of field theory, this becomes the action in Euclidean space. The tunnelling amplitude, then, is given by (10.113).

What is the action for our instanton? It is easily calculated from the inequality

$$\mathrm{Tr}\,(F_{\mu\nu} - \tilde{F}_{\mu\nu})^2 \geq 0. \tag{10.116}$$

Noting that

$$\varepsilon_{\mu\nu\rho\sigma}\varepsilon_{\mu\nu\kappa\lambda} = 2(\delta_{\rho\kappa}\delta_{\sigma\lambda} - \delta_{\rho\lambda}\delta_{\sigma\kappa})$$

it follows immediately that

$$\tilde{F}_{\mu\nu}\tilde{F}_{\mu\nu} = F_{\mu\nu}F_{\mu\nu}$$

so that (10.116) yields

$$\mathrm{Tr}\,F_{\mu\nu}F_{\mu\nu} \geq \mathrm{Tr}\,\tilde{F}_{\mu\nu}F_{\mu\nu}. \tag{10.117}$$

The solution (10.108), however, possesses the property of *self-duality* (see (10.89)):

$$F_{\mu\nu} = \tilde{F}_{\mu\nu}. \tag{10.118}$$

(This is a crucial property of instantons, and many treatments use it as a starting point. In view of the Bianchi identity (10.96) self-duality guarantees

that the field equations (10.95) are satisfied.) It follows that (10.117) becomes an equality for instantons. Noting that the action (in Euclidean space) is

$$S = -\frac{1}{4}\int F^a_{\mu\nu}F^a_{\mu\nu}\,\mathrm{d}^4x$$

$$= -\frac{1}{2}\int \mathrm{Tr}\, F_{\mu\nu}F_{\mu\nu}\,\mathrm{d}^4x, \tag{10.119}$$

the equality (10.117) together with (10.104) yields

$$S = -\frac{8\pi^2}{g^2}q = -\frac{8\pi^2}{g^2} \tag{10.120}$$

since $q = 1$. Hence the tunnelling amplitude between the pure vacuum and the gauge rotated vacuum is of the order

$$\mathrm{e}^{-8\pi^2/g^2}. \tag{10.121}$$

Now that we have established that in a Yang–Mills quantum theory the vacuum is infinitely degenerate, with non-zero transition amplitudes between the gauge rotated vacua belonging to different homotopy classes, it follows that the true ground state of Hilbert space may be written

$$|\mathrm{vac}\rangle_\theta = \sum_{n=-\infty}^{\infty} \mathrm{e}^{\mathrm{i}n\theta}|\mathrm{vac}\rangle_n \tag{10.122}$$

where n is an integer labelling the homotopy class. It is characterised by a particular value of θ, and the coefficients $\mathrm{e}^{\mathrm{i}n\theta}$ ensure the invariance (up to a phase) of $|\mathrm{vac}\rangle_\theta$ under gauge transformations g_1 (see (10.112)). They have the effect

$$|\mathrm{vac}\rangle_n \xrightarrow{g_1} |\mathrm{vac}\rangle_{n+1} \tag{10.123}$$

and hence

$$|\mathrm{vac}\rangle_\theta \xrightarrow{g_1} \mathrm{e}^{-\mathrm{i}\theta}|\mathrm{vac}\rangle_\theta. \tag{10.124}$$

Gauge transformations of the type g_1 (or g_n), which change the homotopy class, are sometimes called 'large' gauge transformations. 'Small' gauge transformations are those continuously deformable to the identity (for example, infinitesimal ones), which do not change the homotopy class.

Vacua of the type (10.117) are known as *θ-vacua*, and they have several important consequences in particle physics. If $\theta \neq 0$ the vacuum state is complex, and time reversal invariance is violated. From the CPT theorem, it then follows that CP invariance is violated. Further, since under parity $g_1 \to (g_1)^{-1}$, unless $\theta = 0$, P is also violated. The observed scale of T violation in physics requires $\theta < 10^{-5}$ (Wilczek 1978). A satisfactory explanation of why θ is so small yet not zero is yet to be found.

Finally, 't Hooft has drawn attention to a remarkable consequence of the existence of instantons when fermions are also present. Consider a theory with N massless quarks, where N is a 'flavour' index. It has a chiral symmetry $SU(N)_L \otimes SU(N)_R \otimes U(1)$. The axial current J^5_μ, however, has an anomaly (cf. equation (9.265))

$$\partial_\mu J^5_\mu = \frac{Ng^2}{16\pi^2} F^a_{\mu\nu} \tilde{F}^a_{\mu\nu}.$$

Comparing with (10.104), however, gives

$$\int d^4x \partial_\mu J^5_\mu = 2Nq$$

so that in the field of an instanton with $q = 1$ there is a violation of axial charge Q^5 by

$$\Delta Q^5 = 2N.$$

This results in decays such as

$$p + n \rightarrow e^+ + \bar{\nu}_\mu \quad \text{or} \quad \mu^+ + \bar{\nu}_e$$

which violate baryon and lepton number (which are not gauge symmetries). The probability of these decays is, however,

$$\begin{aligned} e^{-16\pi^2/g^2} &= e^{-16\pi^2/e^2 \sin^{-2}\theta_W} \\ &= e^{-4\pi \times 137 \times \sin^2\theta_W} \\ &= e^{-602.6} = 10^{-262} \end{aligned}$$

if $\sin^2 \theta_W \approx 0.35$. This gives a deuteron lifetime of the order of 10^{225} s $\approx 10^{218}$ yr. Such an enormously large number is typical of the results of instanton calculations. It would be interesting if some of the large numbers in physics owed their origin to considerations of this type.

The methods of the present chapter are in essence *non-perturbative*; firstly, because a perturbation around a pure vacuum will never produce an excitation above a vacuum belonging to a *different homotopy class*; and secondly, because the semi-classical approximation is also non-perturbative. This has given a large measure of impetus and excitement to topological methods in the last few years, because of the knowledge that areas of physics are being explored which are completely inaccessible to perturbation theory. Some hope is held out, for example, that quark confinement may be explained by these methods. In any case, a new spectre has opened on the world, and non-Abelian gauge theories like electroweak theory, QCD and grand unification (and gravity?) are now seen to have a much richer structure than had hitherto been dreamed of.

Summary

[1]The kink solution to the sine–Gordon equation is exhibited. The stability of the kink is due to the topology of the boundary conditions. [2]It is shown that in two (or more) space dimensions finite energy solitons may only exist if there is also a gauge field. The corresponding solution in 2-dimensional space (or 3-dimensional space with cylindrical symmetry) is a line carrying magnetic flux, identified with the Abrikosov flux line in superconductivity. Such vortex lines exist when the gauge group is $U(1)$, but not when it is $SU(2)$. In the case of $O(3)$, there is only one value for the charge per unit length of the vortex. [3]The magnetic monopole is introduced, and Dirac's quantisation condition derived. Wu and Yang's fibre bundle formulation of the Dirac monopole is briefly outlined. [4]Certain spontaneously broken non-Abelian gauge theories possess solutions with magnetic charge, the so-called 't Hooft–Polyakov monopoles. If the gauge group is G, and the unbroken subgroup is H, the condition that monopoles exist is that $\pi_2(G/H)$ is non-trivial. Hence 't Hooft–Polyakov monopoles do not exist in the Weinberg–Salam model. It is shown how a gauge transformation relates the Dirac and 't Hooft–Polyakov monopoles. [5]The instanton is a topologically non-trivial solution to the pure (not spontaneously broken) gauge-field equations. It describes a configuration with energy localised in time as well as in space. Its topological nature is described. The vacuum is infinitely degenerate, and some physical consequences of this, depending on quantum tunnelling, are outlined.

Guide to further reading

For comprehensive reviews of solitons in non-linear theories (excluding gauge theories), see Scott, Chu & McLaughlin (1973), Whitham (1974). For short reviews, see Coleman in Zichichi (1977), Wick in Zichichi (1978). Early examples of kinks in physics are discussed in Finkelstein & Misner (1959), Finkelstein (1966). For an example of possible application to elementary particles, see Faddeev (1976). Quantisation of solitons is discussed by Coleman in Zichichi (1977) and in Neveu (1977). Vortex lines were first shown to exist in gauge theories by Nielsen & Olesen (1973). For an application to the Weinberg–Salam model see Nambu (1977). For a review, see Jaffe & Taubes (1980). Dirac's paper on magnetic monopoles is Dirac (1931); see also Wentzel (1966). The Wu–Yang formulation of Dirac's theory is found in Wu & Yang (1975). For further developments of the fibre bundle formulation of the Dirac monopole, see Trautman (1977), Minami (1979), Ryder (1980). Reviews of the Dirac monopole are to be found in Felsager (1981, ch. 9), Coleman in Zichichi (1983), Goddard & Olive (1978), Craigie, Goddard & Nahm (1982). The original papers on the 't Hooft and Polyakov monopoles are: 't Hooft (1974), Polyakov (1974, 1976). The topological origin of these monopoles was pointed

out by Arafune, Freund & Goebel (1975), Tyupkin, Fateev & Shvarts (1975), Monastyrskii & Perelomov (1975). For reviews, see refs. Nambu (1979), Rajaraman (1982), S. Coleman in Zichichi (1977), Jaffe & Taubes (1980), Coleman (1983), Goddard & Olive (1978), Craigie *et al.* (1982), Huang (1982). Instantons were discovered by Belavin *et al.* (1975). Important developments were made by 't Hooft (1976, 1978), Jackiw & Rebbi (1976), Callan, Dashen & Gross (1976). Mathematical aspects of instantons are explored in Jackiw & Rebbi (1977); Atiyah, *et al.* (1978), Drinfeld & Manin (1978), Atiyah, Hitchin & Singer (1978). General topological aspects of instantons and monopoles are outlined in Nowakowski & Trautman (1978), Trautman (1979). Reviews of instantons are to be found in Coleman in Zichichi (1979), Vainshtein *et al.* (1982), Crewther & Schroer in Urban (1978). See also Felsager (1981, ch. 5), Rajaraman (1982), and Huang (1992). A review of the mathematical aspects of instantons is to be found in Drinfeld & Manin (1980). A very readable introductory account of the differential geometry of gauge fields, including the topology of monopoles and instantons, is Eguchi, Gilkey & Hanson (1980).