COURSE 3

VORTICES AND MONOPOLES*

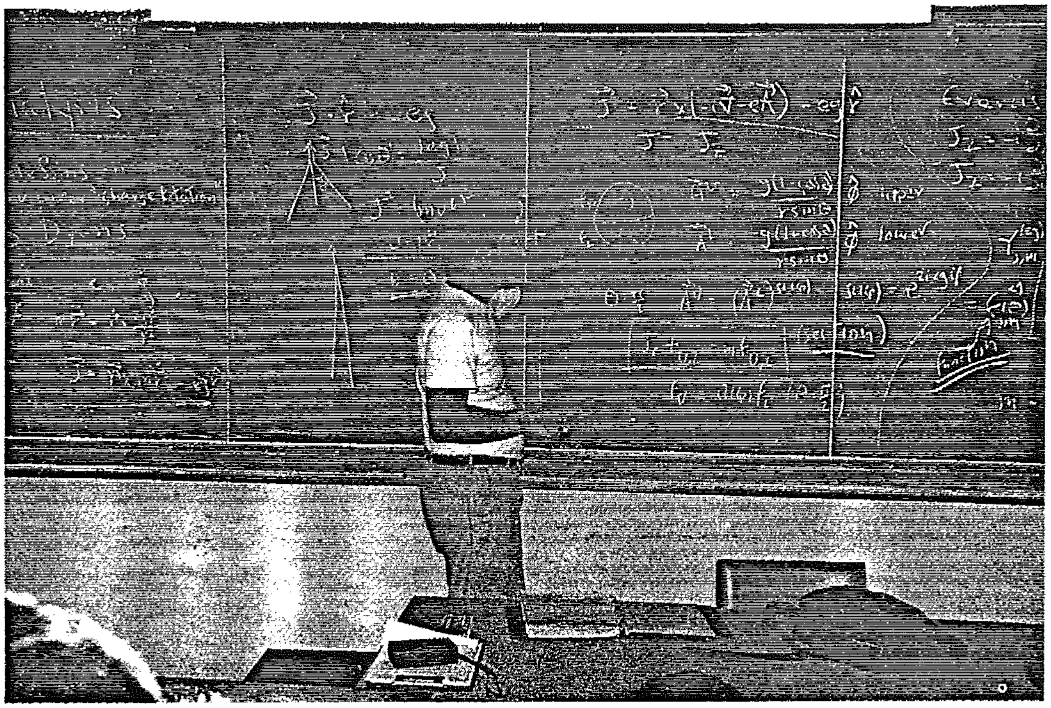
John PRESKILL**

California Institute of Technology, Pasadena, CA 91125, USA

* Work supported in part by the US Department of Energy under Contract No. DEAC-03-81-ER40050.

** Alfred P. Sloan Fellow and National Science Foundation Presidential Young Investigator.

P. Ramond and R. Stora, eds
Les Houches, Session XLIV, 1985—Architecture des interactions fondamentales à courte distance/Architecture of fundamental interactions at short distances


upper
lower

Contents

1. Vortices

1.1. *The Nielsen–Olesen vortex*

A vortex is a stable time-independent solution to a set of classical field equations that has finite energy in two spatial dimensions; it is a two-dimensional soliton. In three spatial dimensions, a vortex becomes a string, a classical solution with finite energy per unit length. A semi-classical expansion about the classical vortex or string solution can be carried out order by order in $\hbar$, but we will at first confine our attention to the classical approximation.

The prototypical example of a vortex [1] occurs in the Abelian Higgs model, a particle physicist's version of a superconductor. This model has a spontaneously broken U(1) gauge symmetry. Its Lagrange density may be written

$$L=-\tfrac{1}{4}F_{\mu\nu}F^{\mu\nu}+|D_\mu\phi|^2-V(|\phi|), \tag{1.1.1}$$

where ϕ is a charged complex scalar field and $D_\mu=\partial_\mu-\mathrm{i}eA_\mu$ is the gauge-covariant derivative. Let us suppose that $V(|\phi|)$ has its minimum at a nonzero value of $|\phi|$; if it is a quartic polynomial in ϕ and $\phi^\dagger$ (as will be required by renormalizability when we quantize this field theory in $(3+1)$ dimensions), it must have the form (up to an irrelevant additive constant)

$$V(|\phi|)=\frac{\lambda}{4}(|\phi|^2-\tfrac{1}{2}v^2)^2, \tag{1.1.2}$$

where v is real and positive.

The classical ground state of this theory is a field configuration that is constant (at least in a particular gauge) and has $|\phi|=v/\sqrt{2}$. Thus, the U(1) gauge symmetry is "spontaneously broken". There are many vacua, each labeled by the phase of the expectation value of ϕ. This apparent vacuum degeneracy is really an artifact, however, because the different vacua are related by gauge transformations. The true spectrum of the theory is most conveniently determined by choosing the unitary gauge, in which ϕ is real. Writing

$$\phi=(v+\phi')/\sqrt{2},$$

where ϕ' is a real scalar field, one can expand L to quadratic order to find that the perturbative spectrum consists of a vector boson with mass $m_V = ev$ and a scalar with mass $m_S = \sqrt{\lambda}\, v$. The Higgs mechanism has occurred; there is no Goldstone boson, but the photon has acquired a mass.

One might wish to investigate the spectrum of this theory beyond perturbation theory. More specifically, one might ask whether there exist stable time-independent solutions to the classical field equations with finite energy other than the vacuum solution. If it exists, such a solution is a localized lump of energy density known as a soliton; it behaves like a particle in the classical theory, and can be expected to survive in the spectrum of the quantum theory.

I can attempt to construct a soliton by the following strategy: Suppose I find a particular field configuration of finite energy that I know cannot be continuously deformed to the trivial vacuum configuration while the energy remains finite. Starting with that configuration, I deform it until a local minimum of the energy functional is obtained. The final configuration is a stable time-independent classical solution, guaranteed to be different from the vacuum solution.

Furthermore, a starting configuration with the required properties exists, in the Abelian Higgs model in two spatial dimensions. To see this, consider the properties of finite-energy field configurations. The energy of a time-independent field configuration is a sum of three nonnegative terms,

$$E = \int d^2 r\, [\tfrac{1}{2}(E_i E^i + B_i B^i) + D_i \phi D^i \phi^\dagger + V(|\phi|)], \tag{1.1.3}$$

each of which must be finite if the total energy is finite. In particular, for the third term to be finite, $V(|\phi|)$ must approach zero at spatial infinity, and $|\phi|$ must therefore approach $v/\sqrt{2}$. We may think of two-dimensional space as being bounded by a big circle at $r = \infty$. Finiteness of the energy requires $|\phi| = v/\sqrt{2}$ on this circle, but finiteness of the third term places no restriction on the *phase* of ϕ. We may have

$$\phi(r, \theta) \underset{r\to\infty}{\longrightarrow} (v/\sqrt{2})\, e^{i\sigma(\theta)}, \tag{1.1.4}$$

where $e^{i\sigma(\theta)}$ is an arbitrary phase factor, a periodic function of the polar angle θ with period 2π.

Thus, associated with every finite-energy field configuration is a mapping from the circle at spatial infinity to the circle defined by the phase of ϕ. A mapping from a circle to a circle has a winding number n, which we may define as

$$n=\frac{1}{2\pi}[\sigma(\theta=2\pi)-\sigma(\theta=0)]. \tag{1.1.5}$$

An important property of the winding number is that it is an integer. Because an integer cannot change continuously, the winding number must be preserved by smooth deformations of the fields that preserve the finiteness of the energy; it is a "topological invariant". Therefore a configuration with nonzero winding number cannot be continuously deformed to the vacuum, which has zero winding number. Moreover, since time evolution is continuous, the winding number must be a constant of the motion. We have a discovered a "topological conservation law" that, unlike more familiar conservation laws, is not directly associated with any symmetry of the action.

We can apparently construct a soliton by finding the configuration of lowest energy with, say, unit winding number. (This configuration is called a "vortex". The behavior of ϕ on the circle at $r=\infty$ is sketched in fig. 1.) But we must verify that it is really possible for a configuration with nonzero winding number to have finite energy. In particular, we should worry about the second term in eq. (1.1.3), which involves the covariant gradient of ϕ. ϕ surely has a nonvanishing gradient in the circumferential direction for $n\neq 0$, because σ, by eq. (1.1.5), is a nontrivial function of θ. The gradient term

$$\int \mathrm{d}^2r\left|\left(\frac{1}{r}\frac{\partial}{\partial\theta}-\mathrm{i}eA_\theta\right)\phi\right|^2 \tag{1.1.6}$$

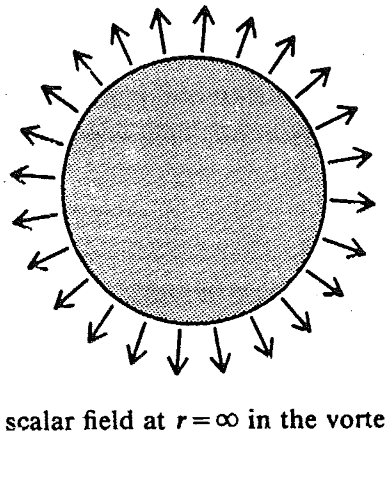


Fig. 1. The scalar field at $r=\infty$ in the vortex solution.

can be finite only if the gauge field behaves for large r like

$$A_\theta \xrightarrow[r\to\infty]{} \frac{1}{er}\frac{d\sigma}{d\theta}+\cdots, \tag{1.1.7}$$

the corrections falling off faster than $1/r$; otherwise, the energy diverges logarithmically at large r.

The gauge field eq. (1.1.7) is a "pure gauge"; thus, the prescribed asymptotic large r behavior of A_θ permits the field $F_{\mu\nu}$ to decay sufficiently rapidly at large r for the first term in eq. (1.1.3) to be finite. We have succeeded, therefore, in demonstrating the existence of a finite-energy field configuration with winding number $n=1$, and hence, of a soliton.

The gauge field cannot be pure gauge everywhere, if $n\neq 0$. The total magnetic flux through the plane is readily evaluated using Stokes' Theorem. The flux

$$\Phi=\oint r\,d\theta\,A_\theta=\frac{1}{e}[\sigma(2\pi)-\sigma(0)]=\frac{2\pi}{e}n \tag{1.1.8}$$

is quantitized, and the number of flux quanta is the winding number.

A nonsingular field configuration with $n\neq 0$ has another important property: the field ϕ must vanish somewhere. For if ϕ has no singularities and no zeros, its phase σ is well defined everywhere. By smoothly shrinking the circle at infinity to an infinitesimal circle around the origin, we can smoothly deform the mapping $\sigma(\theta)$, which has winding number $n\neq 0$, to the trivial mapping $\sigma=$ constant. This is impossible. We are forced to conclude that there is at least one point at which σ is ill-defined, because ϕ vanishes.

What does the classical vortex solution with $n=1$ look like? It is the lowest energy configuration with $n=1$, so ϕ has one zero - more zeros would cost more energy - which we may choose to lie at the origin. Since $\phi=0$ is not the minimum of the potential $V(|\phi|)$, there is a lump of energy density surrounding the origin. What can we say about the size and mass of this lump? We can easily determine the size and mass in order of magnitude without doing any detailed calculations.

Our vortex actually has two characteristic length scales. The first is the radius of the region in which $\phi(r,\theta)$ departs significantly from its vacuum value $|\phi|=v/\sqrt{2}$; call it r_S. The other length scale is the radius of the region in which the gauge field is far from its asymptotic value, eq. (1.1.7); call it r_V. To find r_S and r_V for the vortex solution, we do a variational calculation; we work out how the energy of the configuration depends on r_S and r_V, and then minimize it with respect to r_S and r_V.

In the order of magnitude, the three terms of eq. (1.1.3) become

$$E \simeq \pi v^2 \left[\frac{1}{e^2 v^2 r_V^2} + \theta(r_V - r_S)\ln(r_V/r_S) + \lambda v^2 r_S^2 \right]. \tag{1.1.9}$$

The first term is the magnetic self-energy. It favors a large value of r_V, because the magnetic flux does not like to be confined to a small region. The third term is the scalar potential energy. It favors a small value of r_S, because it costs potential energy when ϕ departs from its vacuum value. The second term, the gradient energy, ties together the two distance scales r_S and r_V.

The energy is minimized by

$$r_S \simeq (\sqrt{\lambda}\, v)^{-1} = m_S^{-1}, \qquad r_V \simeq (ev)^{-1} = m_V^{-1}, \tag{1.1.10}$$

for $m_S > m_V$; the scalar field and vector field "core sizes" correspond semiclassically to the Compton wavelength of the scalar and vector particles respectively. The minimum energy, the mass of the vortex, is

$$M_{\text{vortex}} \simeq \pi v^2 \ln(m_S/m_V), \tag{1.1.11}$$

for $m_S > m_V$.

The classical description we have given of the structure of the vortex should be appropriate for small $\hbar$. Of course, small $\hbar$ means weak coupling; the semiclassical expansion is an expansion in $e^2\hbar$ and $\lambda\hbar$ with the masses m_V and m_S fixed. Comparing eq. (1.1.10) and eq. (1.1.11), we see that in the classical (weak-coupling) limit, the vortex size becomes arbitrarily large compared to its Compton wavelength, a property we expect of an object amenable to a classical analysis.

In the Abelian Higgs model in three spatial dimensions, our time-independent vortex solution may be thought of as a cross section of an infinite "string", and eq. (1.1.11) may be interpreted as the energy per unit length of the string.

1.2. A Z_2 vortex

Vortex solutions can occur not only in Abelian gauge theories, but also in theories with simple or semi-simple gauge groups. Let us familiarize ourselves with this phenomenon by considering a simple example.

We consider a model with gauge group SO(3), spontaneously broken by the expectation value of an order parameter in the symmetric tensor (five-dimensional) representation of SO(3). The order parameter Φ can be written as a traceless 3×3 matrix, which under a gauge transformation $\Omega(x)\in \mathrm{SO}(3)$, transforms as

$$\Phi(x)\to\Omega(x)\Phi(x)\Omega^{-1}(x). \tag{1.2.1}$$

Suppose that Φ acquires the expectation value

$$\langle\Phi\rangle=\Phi_0=(v)\ \mathrm{diag}(1,1,-2), \tag{1.2.2}$$

where v is the mass scale of the symmetry breakdown, and diag denotes a diagonal matrix with the indicated eigenvalues.

The order parameter Φ_0 leaves unbroken a subgroup of SO(3) which is locally isomorphic to (has the same Lie algebra as) SO(2), the group of rotations about the z axis, generated by

$$Q=\begin{pmatrix}0 & -\mathrm{i} & 0\\ \mathrm{i} & 0 & 0\\ 0 & 0 & 0\end{pmatrix}. \tag{1.2.3}$$

But to investigate whether this model has a vortex solution, we need to know more than just the local structure of the unbroken group; we need to know its global structure. We must not fail to notice that the unbroken group contains a disconnected component generated by

$$\Omega_0=\mathrm{diag}(1,-1,-1), \tag{1.2.4}$$

a 180° rotation about the x-axis. The actual pattern of symmetry breakdown is

$$\mathrm{SO}(3)\to\mathrm{O}(2). \tag{1.2.5}$$

In the Abelian Higgs model there are many vacua, distinguished by the phase of the scalar field. In this model also, there are many vacua, which can be represented by $\Omega\Phi\Omega^{-1}$, where $\Omega\in\mathrm{SO}(3)$. The space of possible vacua, the quotient space SO(3)/O(2), is equivalent to the space of unit vectors in three-dimensional space, except that a vector pointing up cannot be distinguished from a vector pointing down. It is a two-sphere with antipodal points identified.

In the Abelian Higgs model, we demonstrated the existence of a vortex by finding a finite-energy field configuration that cannot be smoothly deformed to the vacuum solution. And we found such a configuration by exploiting the existence of loops in the space of vacua that cannot be contracted to a point. By the same reasoning, this SO(3) model will have a vortex solution if there is a loop in the manifold of vacua that

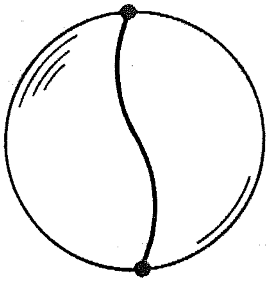


Fig. 2. A noncontractible closed path in SO(3)/O(2).

cannot be contracted to a point. Such a loop obviously exists. It can be represented as a path on the two-sphere from the south pole to the north pole (fig. 2). This is not a closed loop on the two-sphere, but is closed on the two-sphere with antipodal points identified. The behavior of the scalar field Φ in the vortex solution on the circle $r=\infty$ is indicated in fig. 3. The orientation of Φ is represented by an arrow, with the understanding that arrows pointing in opposite directions represent identical orientations.

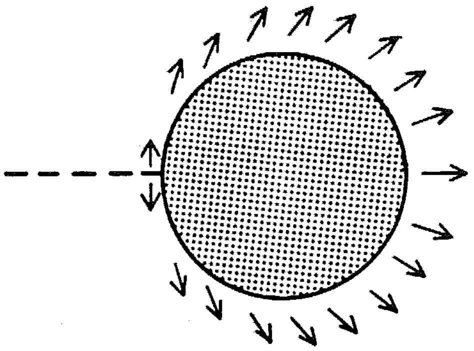


Fig. 3. The Z_2 vortex.

The noncontractible loops in the vacuum manifold of our SO(3) model differ in an important way from the noncontractible loops of the Abelian Higgs model. If we compose two such loops by tracing the two loops in succession, the result is represented by a closed loop on the two-sphere, which obviously can be contracted to a point. Thus, a configuration with two vortices can be smoothly deformed to a vacuum configuration. Rather then being an arbitrary integer, the conserved vortex number takes values in Z_2.

Does this vortex carry any magnetic flux? In order for the gradient energy to be finite, the gauge field at spatial infinity must be the pure gauge

$$A_\mu = \frac{1}{ie}(\partial_\mu \Omega)\Omega^{-1}, \tag{1.2.6}$$

where $\Omega(\theta)$ is a gauge transformation which describes how the order parameter is transported as we traverse the circle at infinity; that is

$$\Phi(r,\theta) \xrightarrow[r\to\infty]{} \Omega(\theta)\Phi_0\Omega(\theta)^{-1}. \tag{1.2.7}$$

By integrating eq. (1.2.6), we find that

$$\Omega(\theta = 2\pi) = \left[\mathrm{P}\exp\left(ie \oint A_\mu \, \mathrm{d}x^\mu\right)\right]\Omega(\theta = 0), \tag{1.2.8}$$

where P denotes path ordering. If Φ is a Z_2-vortex configuration, then $\Omega(\theta)$ is a path in SO(3) from the connected component to the disconnected component of the unbroken group O(2), and $\mathrm{P}\exp(ie \oint A_\mu \, \mathrm{d}x^\mu)$ must be an element of the disconnected component of O(2). This observation allows us to assign a Z_2 magnetic charge to the vortex.

This Z_2 vortex has a remarkable property which we will return to several times later in these lectures. If Q is the generator of SO(2) and Ω_0 is in the disconnected component of O(2), then

$$\Omega_0 Q \Omega_0^{-1} = -Q; \tag{1.2.9}$$

in other words, a 180° rotation about the x-axis followed by a counterclockwise rotation about the z-axis and another 180° rotation about the x-axis is equivalent to a clockwise rotation about the z-axis. Equation (1.2.9) tells us that the sign of the "electric charge" Q has no gauge-invariant meaning; charge conjugation is a gauge transformation. Furthermore, an object which is transported all the way around the string experiences a gauge transformation by Ω_0. Our Z_2 vortex might be called an "Alice vortex" (or Alice string, in three dimensions); a voyage around the string is a voyage through the charge conjugation looking-glass, interchanging matter and antimatter [2].

There can be no Alice strings in Nature. Charge conjugation is not an exact symmetry, so it cannot be a gauge symmetry. But it is at least conceivable that there is an exact discrete symmetry in Nature that interchanges ordinary matter with "shadow matter", which transforms under a mirror image of the standard SU(3) × SU(2) × U(1) gauge group.

Then one might be able, by circumnavigating an Alice string, to become the Invisible Man.

Incidentally, it is not correct, in general, to assert that, when vortices are classified by Z_2, a vortex is indistinguishable from an antivortex. While it is true that a pair of vortices, for example, can be smoothly deformed to a vortex–antivortex pair, there may be an energy barrier separating the two configurations that allows them to be unambiguously distinguished. It is therefore possible that, say, the long-range interaction between vortices is different than the vortex–antivortex interaction [3].

1.3. Topological classification of vortices

Which gauge theories, in general, contain vortices as classical solutions? We have seen that a vortex can be constructed whenever there are loops in the manifold of vacua of the theory that cannot be contracted to a point. So we wish to establish the general conditions under which such noncontractible loops exist [4].

For a theory in which the gauge symmetry G is spontaneously broken to a subgroup H, the vacuum manifold, the space of possible orientations of the order parameter is

$$G/H=\{\Phi,\ \Phi=\Omega\Phi_0,\ \Omega\in G\}. \tag{1.3.1}$$

Here Φ_0 is a standard reference position of the order parameter, which is preserved by the unbroken subgroup H. I have made the assumption that there is no "accidental" degeneracy: all vacua can be obtained from any given one by performing gauge transformations in G.

A closed loop in the space G/H, which we may choose to begin and end at the point Φ_0, can be parametrized by

$$\Phi(\theta)=\Omega(\theta)\Phi_0,\quad 0\leqslant\theta\leqslant 2\pi, \tag{1.3.2}$$

where

$$\Omega(\theta=0)=1,\qquad \Omega(\theta=2\pi)=h\in H. \tag{1.3.3}$$

Thus, the loop in G/H may be associated with a path in G that begins at the identity and ends at some element of H. This path, of course, is in the identity component of G, which by definition consists of those elements of G that can be connected to the identity by a continuous path.

In general, the identity component of G may contain several disconnected components of the subgroup H. Therefore, we may distinguish two possibilities. The endpoint h of the path $\Omega(\theta)$ is either in the identity

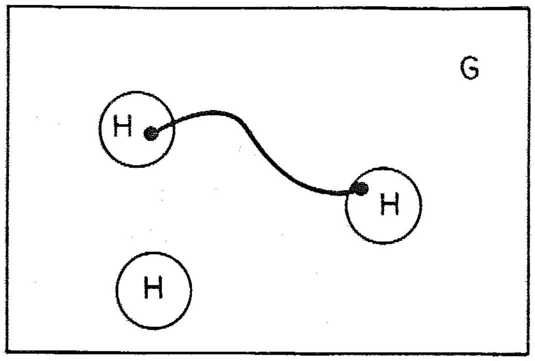


Fig. 4. A path in G associated with a noncontractible loop in G/H.

component of H, or it is not. Suppose h is not in the identity component of H. (See fig. 4.) Then the path $\Phi(\theta)$ surely cannot be contracted to a point in G/H. For $\Phi(\theta)$ can be contracted to a point in G/H only if $\Omega(\theta)$ can be deformed to a path contained entirely in H such that all intermediate paths both begin and end in H. But that is impossible, if $\Omega(\theta)$ is a path from the identity component of H to another connected component of H.

On the other hand, if the endpoint h of the path $\Omega(\theta)$ *is* in the identity component of H, then one readily sees that the loop $\Phi(\theta)$ in G/H *can* be contracted to a point, assuming that G is simply connected. (See fig. 5.) We say that G is simply connected if all loops in G are contractible. If $\Omega(\theta)$ both begins and ends in the identity component of H, then we can add a segment contained entirely in H to construct a closed loop in G. The extra segment does not modify the path $\Phi(\theta)$ in G/H. But this loop is contractible to a point, if G is simply connected, and therefore the path $\Phi(\theta)$ is also contractible.

In general, the closed paths in a space that begin and end at an arbitrarily chosen reference point fall into topological equivalence

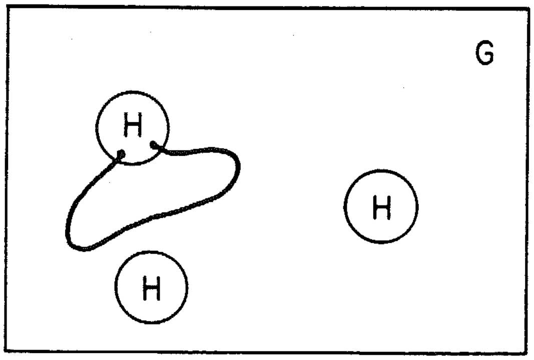


Fig. 5. A path in G associated with a contractible loop in G/H.

classes, called "homotopy" classes. Two paths are in the same class if they can be continuously deformed into one another. The classes are endowed with a natural group structure, since the composition of two paths may be defined to be a path that traces the two paths in succession. This group is called π_1, the first homotopy group of the space. It is evident from the above discussion that, if G is simply connected, the topologically distinct classes of loops in G/H are in one-to-one correspondence with the distinct connected components of H contained in the identity component of G. In an equation, this result is

$$\pi_1(G/H)=\pi_0(H)/\pi_0(G), \tag{1.3.4}$$

which holds when G is simply connected. (Here $\pi_0(H)/\pi_0(G)$ is a group whose elements are the connected components of H in the identity component of G, and the equality signifies a group isomorphism.)

There is really no loss of generality in assuming that G is simply connected; we may always regard G as a covering group which is not necessarily represented faithfully by the order parameter Φ. But it is frequently more convenient, as in the Abelian Higgs model, to choose G not to be simply connected. If G is not simply connected, there may be additional elements of $\pi_1(G/H)$, additional noncontractible loops, besides those corresponding to the elements of $\pi_0(H)/\pi_0(G)$. These additional noncontractible loops in G/H are associated with closed paths $\Omega(\theta)$ in G which are noncontractible in G. But it is also required that $\Omega(\theta)$ not be deformable in G to a path contained entirely in H; otherwise the path $\Phi(\theta)=\Omega(\theta)\Phi_0$ could evidently be contracted to a point in G/H.

For example, in the section 1.2, we considered the symmetry breaking pattern $G=\mathrm{SO}(3)\rightarrow H=\mathrm{O}(2)$, and we identified a class of noncontractible loops in G/H associated with the nontrivial connected component of O(2). There are also noncontractible loops in SO(3), which is not simply connected. However, there are no associated noncontractible loops in G/H, because a noncontractible loop in SO(3) can be deformed to a loop contained in SO(2).

You can check your understanding of this formalism by doing the following exercises.

Exercise. Consider the quotient space:

$$M^{pqr}=\frac{\mathrm{SU}(2)_1\times \mathrm{SU}(2)_2\times \mathrm{U}(1)_Y}{\mathrm{U}(1)_Q},$$

where the unbroken $U(1)_Q$ subgroup is generated by

$$Q = pT_3^{(1)} + qT_3^{(2)} + rY.$$

Here p, q, and r are integers with no common factor, and $T_3^{(1,2)}$, Y are normalized so that their smallest nonzero eigenvalue is unity. Show that

$$\pi_1(M^{pqr}) = Z_r.$$

Exercise. Show that the standard Weinberg–Salam–Glashow SU(2) × U(1) model has no topological stable string solution.

1.4. Walls bounded by strings

Discrete symmetries are frequently invoked in models of particle physics. For example, in extensions of the standard model, discrete symmetries are sometimes used to constrain the Yukawa couplings; one can thus obtain relations among the masses and mixing angles of quarks and leptons, even though the discrete symmetries are spontaneously broken at the weak-interaction mass scale by the expectation values of the scalars. For such purposes, discrete symmetries are preferable to continuous symmetries, because continuous global symmetries that get spontaneously broken at the weak-interaction scale are not phenomenologically acceptable. There would be Goldstone bosons associated with such symmetries that could be detected experimentally.

However, there is a problem with spontaneously broken discrete symmetries too, concerning not particle phenomenology but cosmology. If a discrete symmetry is spontaneously broken, it is expected that the symmetry is restored at sufficiently high temperature, and that a phase transition therefore occurred in the early universe when the symmetry breaking first turned on. In such a phase transition, domain walls would have been produced. Eventually the energy density of the universe would have become dominated by these walls; at that point, a reasonable cosmology could no longer be recovered [5].

One way to deal with this cosmological domain wall problem is to invoke inflation [6]. We can construct a model with spontaneously broken discrete symmetries in such a way that the universe enters an epoch of superluminal expansion after domain walls are produced, and the walls are "inflated away". This option is not very attractive if the symmetry breaking scale is as low as the weak-interaction scale. It is hard to concoct a reasonable scenario for inflation at such low temperature consistent with various constraints, like the baryon abundance of the universe. In this and the next section, we will see that there is another way to live with discrete symmetries. Spontaneously broken discrete symmetries can be cosmologically acceptable if they are embedded in (gauge or global) continuous symmetries that are spontaneously broken at a higher mass scale. (Spontaneously broken global symmetries are okay if the symmetry breaking scale is high enough, because the associated Goldstone bosons are then very weakly coupled.) It is then possible for a domain wall to terminate on a string [7]. The properties of these walls bounded by strings will be considered in this section, and their cosmological implications discussed in section 1.6.

First, consider a simple example of a model with a domain wall as a classical solution. It is a theory of a real scalar field ϕ with Lagrange density

$$L=\tfrac{1}{2}(\partial^{\mu}\phi)^2-V(\phi), \tag{1.4.1}$$

where the potential $V(\phi)$ has the Z_2 symmetry $\phi\to-\phi$, is minimized at $\phi=\pm v$, and satisfies $V(0)=0$, $V(v)=-\Lambda^4$ (fig. 6). By a trivial rescaling

$$L=\left(\frac{\Lambda^4}{m^2}\right)[\tfrac{1}{2}(\partial_{\mu}\tilde{\phi})^2-m^2\tilde{V}(\tilde{\phi})],\quad m^2=\Lambda^4/v^2, \tag{1.4.2}$$

where $\tilde{\phi}$ is dimensionless, and $\tilde{V}$ is a dimensionless function that is of order one when $\tilde{\phi}$ is of order one, minimized at $\tilde{\phi}=\pm1$. The discrete Z_2 symmetry is evidently spontaneously broken, and the mass of the scalar particle is of order m.

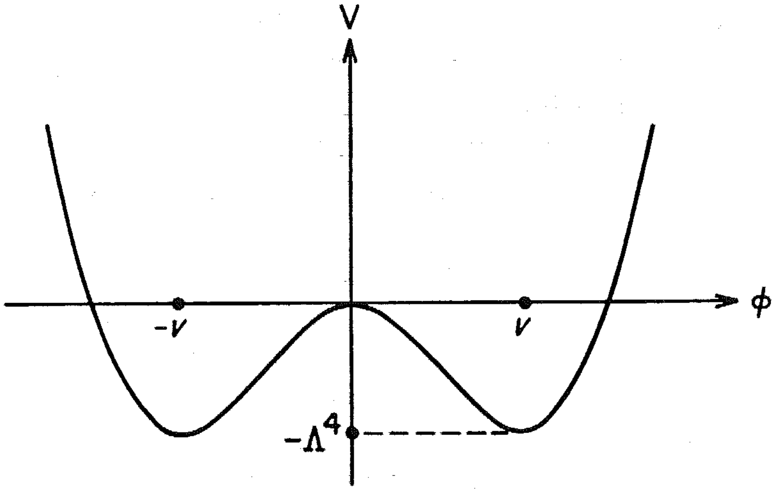


Fig. 6. A potential with a spontaneously broken Z_2 symmetry.

Field configurations with finite energy per unit area (in the y-z plane) must satisfy

$$|\tilde{\phi}| \xrightarrow[x\to\pm\infty]{} 1. \tag{1.4.3}$$

Among such configurations, those for which

$$\tilde{\phi}_+ \equiv \lim_{x\to\infty} \tilde{\phi} = -\tilde{\phi}_- \equiv -\lim_{x\to-\infty} \tilde{\phi}, \tag{1.4.4}$$

are topologically nontrivial; they cannot be smoothly deformed to the vacuum configuration $\tilde{\phi}=1$, while the energy per area remains finite. The domain wall is the configuration (independent of t, y, z) in the topologically nontrivial sector that minimizes the energy per unit area. A simple variational estimate, similar to that performed earlier for the string, shows that the domain wall has, in order of magnitude, thickness

$$b \sim m^{-1}, \tag{1.4.5}$$

and energy per area

$$\sigma \sim \Lambda^4 b. \tag{1.4.6}$$

Exercise. Show this.

Now, as an example of a model with a domain wall bounded by a string, suppose that the model of section 1.2 is modified so as to undergo a second symmetry breakdown:

$$\mathrm{SO}(3) \underset{v_1}{\rightarrow} \mathrm{O}(2) \underset{v_2}{\rightarrow} \mathrm{SO}(2), \tag{1.4.7}$$

with $v_2 \ll v_1$. The second stage of symmetry breakdown can be driven by the expectation value of an SO(3) triplet scalar field ϕ; the triplet has a component that is an SO(2) singlet, but changes sign under the discrete O(2) reflection. In the gauge in which the quintuplet Φ has vacuum expectation value eq. (1.2.2), ϕ has the expectation value

$$\langle\phi\rangle = \phi_0 = v_2(0,0,1). \tag{1.4.8}$$

From the point of view of an "effective field theory" that describes physics well below the symmetry breaking scale v_1, the expectation value of ϕ breaks an exact discrete symmetry, so there must be a domain wall configuration. (That the discrete symmetry actually anticommutes with SO(2) is irrelevant in this discussion.) But we know that this discrete symmetry is really contained in the underlying gauge symtetry SO(3) that became spontaneously broken at mass scale v_1.

To appreciate the implications of this property, consider how the field ϕ behaves in the vicinity of the string discussed in section 1.2. A long distance away from the string, the fields Φ and ϕ must approach a vacuum configuration. In the vacuum, Φ and ϕ are aligned in order to preserve the same SO(2) subgroup of SO(3). We have seen that, as a function of the polar angle θ about the string, Φ winds through the topologically nontrivial loop

$$\Phi(\theta)=\Omega(\theta)\Phi_0\Omega(\theta)^{-1}, \quad \Omega(\theta+2\pi)=\Omega_0\Omega(\theta=0), \tag{1.4.9}$$

where Ω_0 is an O(2) reflection. To remain properly aligned with Φ, ϕ must follow the path

$$\phi(\theta)=\Omega(\theta)\phi_0. \tag{1.4.10}$$

But the reflection Ω_0 changes the sign of ϕ_0; therefore $\phi(\theta)$ given by eq. (1.4.10) changes sign as it winds around the string. In order to be single-valued, ϕ must, on some surface bounded by the string, pass through a domain wall and change sign. We conclude that the string is the boundary of a domain wall [7].

In other words, if the field ϕ wants to smoothly interpolate between the vacuum values ϕ_0 and $\Omega_0\phi_0$, it has two options. It can pass through a domain wall that carries energy per unit area σ, or it can wind around a string that carries energy per unit length $\mu \sim v_2^2$. By winding around the string, ϕ can avoid the domain wall. It is pretty obvious that this feature is generic for models in which a spontaneously broken discrete symmetry is embedded in a continuous symmetry that is spontaneously broken at a larger mass scale.

Exercise. Prove this, using the topological classification of vortices described in section 1.3.

If it is possible for a domain wall to terminate on a string, then a sheet of domain wall is not absolutely stable. A hole, bounded by string, can spontaneously nucleate in the sheet. If the hole is larger than $R_c \sim \mu/\sigma$, where μ is the string tension and σ is the wall tension, then the wall tension overcomes the string tension and the hole expands catastrophically. The energy of the hole of critical size is $E_c \sim \mu^2/\sigma$; thus, a WKB estimate of the nucleation rate per unit time and area gives, in order of magnitude

$$\Gamma \propto e^{-\mu^3/\sigma^2}. \tag{1.4.11}$$

If $v_2 \ll v_1$, this rate is completely negligible, and the domain wall may be regarded as stable for all practical purposes.

We have now seen how domain walls bounded by strings arise if a spontaneously broken discrete symmetry is embedded in an exact continuous symmetry. But walls bounded by strings can also arise in another, somewhat different, way; the spontaneously broken discrete symmetry may be embedded in a continuous symmetry that is approximate rather than exact [8]. Walls bounded by strings of this second type arise in models with *axions.*

Of course, gauge symmetries are necessarily exact, so the continuous symmetry associated with the string must in this case be a global symmetry. The string arising from a spontaneously broken global U(1) symmetry is the $m_V \to 0$ limit of the Nielsen-Olesen string studied in section 1.1. As we observed there, this "global string" has a logarithmically divergent energy per unit length. It is not necessarily foolish to think about such strings, though. A finite closed loop of global string has finite energy, and a network of global strings has finite energy per unit volume, with logarithmic interactions among the strings. A network of global strings could have been produced in the early universe.

In an axion model [9] there is a scalar field ϕ that transforms under a global U(1) symmetry, the Peccei-Quinn (PQ) symmetry, under which quarks also transform. Acting on quarks, the PQ symmetry is chiral - left-handed and right-handed quarks have different charges - so the PQ current is afflicted by a chiral anomaly. Physics is left invariant by a $U(1)_{PQ}$ rotation only if the rotation is accompanied by a simultaneous rotation of the QCD angle $\bar{\theta}$,

$$\phi \to e^{i\alpha}\phi, \qquad \bar{\theta} \to \bar{\theta} + 2\pi N\alpha, \tag{1.4.12}$$

where N is an integer that depends on the PQ charges of the quarks (and of other colored fermions, if any).

Nonperturbative strong-interaction effects depend on $\bar{\theta}$, and therefore explicitly break the PQ symmetry. But $\bar{\theta}$ is a periodic variable defined modulo 2π, and therefore a PQ rotation (1.4.12) with α an integer multiple of $2\pi/N$ is a good symmetry. A Z_N subgroup of the $U(1)_{PQ}$ symmetry remains unbroken in spite of the nonperturbative effects [10].

The scalar ϕ acquires a large vacuum expectation value

$$\langle\phi\rangle = v\, e^{i\alpha}, \tag{1.4.13}$$

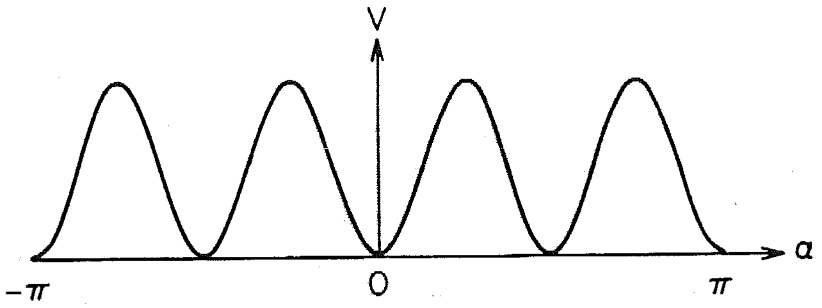


Fig. 7. Vacuum energy density in an $N=4$ axion model.

that breaks spontaneously the $\mathrm{U}(1)_{\mathrm{PQ}}$ symmetry. Because of the nonperturbative QCD effects, the vacuum energy depends on the phase α. Indeed, this dependence is the whole motivation for constructing an axion model. The true vacuum chooses α so that $\bar{\theta}=0$, because this choice minimizes the nonperturbative contribution to the vacuum energy. We are thus able to understand why the parameter $\bar{\theta}$ is observed to be very small in nature.

The vacuum energy density as a function of α is sketched in fig. 7. It has a height of order Λ^4, where $\Lambda \sim 100$ MeV is a characteristic strong-interaction scale. (Actually, the height of this potential depends on the light quark masses, but we may ignore this effect in the present discussion.) The particle arising from the oscillations in this potential is the axion, with mass $m_a \sim N\Lambda^2/v$. It is the pseudo-Goldstone boson of the spontaneously broken $\mathrm{U}(1)_{\mathrm{PQ}}$ symmetry, which has acquired its mass from the explicit symmetry breaking [11].

Associated with the breakdown of the $\mathrm{U}(1)_{\mathrm{PQ}}$ symmetry at mass scale v is a global string. As a function of polar angle around the string, the phase α of the scalar field ϕ varies from 0 to 2π. But sufficiently far from the string, it is energetically favorable for α to assume one of its vacuum values, a multiple of $2\pi/N$. Therefore, the change in α by 2π collapses to N domain walls, each with a thickness of order m_a^{-1} and an energy per area $\sigma \sim \Lambda^4/m_a$. The N domain walls meet at the string (fig. 8).

From the point of view of early cosmology, domain walls bounded by strings still cause trouble if more than one wall ends on each string. Barring the possibility of inflating the walls away, an axion model is cosmologically acceptable only if exactly one wall ends on each string. Models with this property can be constructed in two ways. The first way is to add new colored fermions to the model with appropriate PQ charges so that $N=1$ [12]. The second way is to embed the discrete Z_N symmetry

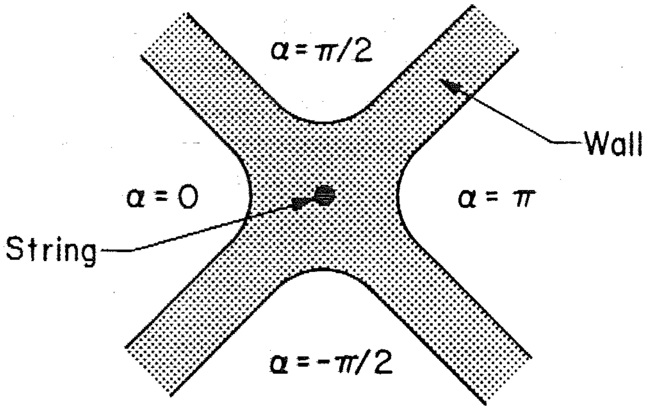


Fig. 8. Cross section of an axion string, at which $N = 4$ domain walls meet.

in an *exact* (local or global) continuous symmetry group [13]. I will describe how the second strategy works.

The idea is that the field ϕ whose expectation value breaks the $\mathrm{U}(1)_{\mathrm{PQ}}$ symmetry also transforms under an exact continuous symmetry group G and that $\mathrm{U}(1)_{\mathrm{PQ}}$ and the identity component of G intersect at the discrete Z_N subgroup. Therefore there is a minimal "hybrid" string associated with a closed path in the vacuum manifold that winds only $(1/N)$th of the way around the PQ U(1), and returns to its starting point through G. This path can be expressed as

$$\phi(\theta) = \exp(-\mathrm{i}\theta/N)\Omega_G(\theta)\phi_0, \quad 0 \leqslant \theta \leqslant 2\pi, \tag{1.4.14}$$

where $\Omega_G(\theta)$ is a path in G that begins at the identity and ends at $\mathrm{e}^{2\pi\mathrm{i}/N}$; it is represented schematically in fig. 9. The minimal string is the boundary of only one axion domain wall, because the phase of ϕ advances by only $2\pi/N$ through $\mathrm{U}(1)_{\mathrm{PQ}}$ as a function of θ.

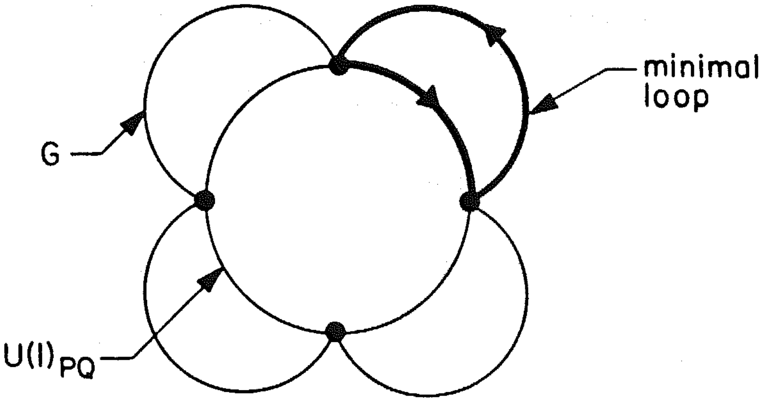


Fig. 9. The minimal noncontractible loop associated with the hybrid axion string.

It is enlightening to consider a specific example of a model of this type [13]. For this purpose, consider a grand unified model with gauge group SO(10) and one generation of fermions in the representation 16_F. Let us choose the minimal Higgs structure that permits the SO(10) symmetry to break to $SU(3)_{color} \times U(1)_{em}$; the Higgs fields are in the representations 16_H, 10_H, and 45_H of SO(10). We specify the $U(1)_{PQ}$ charges of the fields to be

$$Q_{16_F}=1, \qquad Q_{16_H}=1, \qquad Q_{10_H}=-2, \qquad Q_{45_H}=-4. \qquad (1.4.15)$$

These choices are sensible because, first, the Yukawa coupling $16_F\ 16_F\ 10_H$ is allowed, and, second, a quartic Higgs potential can be constructed that has the $U(1)_{PQ}$ symmetry but no other U(1) symmetries. (Additional U(1) symmetries might cause trouble, because there would be an exact Goldstone boson, and the symmetry breaking scale associated with the axion might turn out to be lower than desired.)

The fermion representation 16_F contains the quark fields u_L, d_L, u^c_R, d^c_R, and the $U(1)_{PQ}$ rotation $(16_F) \rightarrow e^{i\alpha}(16_F)$ is an axial rotation; it rotates left-handed and right-handed quarks by opposite phases. But the $U(1)_{PQ}$ rotation by $\alpha = \pi/2$ preserves the argument of the determinant of the quark mass matrix, and is an exact symmetry of QCD. Therefore there is an exact Z_4 subgroup of the approximate $U(1)_{PQ}$ symmetry group. The action of the generator of this Z_4 group on fields with the $U(1)_{PQ}$ charges given in eq. (1.4.15) is

$$16 \rightarrow i16, \qquad 10 \rightarrow -10, \qquad 45 \rightarrow 45. \qquad (1.4.16)$$

Furthermore, the covering group Spin(10) of SO(10) has center Z_4, and the action of the generator of the center on Spin(10) representations is precisely that prescribed in eq. (1.4.16). Therefore, the exact Z_4 subgroup of $U(1)_{PQ}$ is actually contained in the gauge group $G = \text{Spin}(10)$. This model satisfies the criterion of our earlier discussion; the minimal string is associated with a closed path in the vacuum manifold that winds only one quarter of the way around $U(1)_{PQ}$, and returns to its starting point through Spin(10). This string is the boundary of a single axion domain wall.

This model contains only a single generation of fermions, but now that we understand the idea it is not too hard to concoct analogous models with more generations. It is also possible to construct "familon" models in which the exact symmetry G is a global family symmetry, rather than a gauge symmetry [14].

1.5. Superconducting strings

We admired the peculiar properties of the "Alice" string. Now we will consider other examples of strings with exotic properties. These strings support massless excitations that propagate along the string [15].

Inside a string, a scalar field is excited; it assumes a value different than its value in the vacuum. It may happen that a fermion is Yukawa coupled to this scalar. Then the fermion will interact with the string. Let us study further the nature of this interaction.

We will treat the string as a classical background field, and consider propagation of fermions in the string background. For definiteness, suppose that Φ is a complex scalar field whose expectation value breaks a global U(1) symmetry. (It does not complicate things very much to introduce a gauge field, but we will not, to keep things as simple as possible. Thus our string is a global string, like the axion string.) A string along the z-axis is a scalar field configuration

$$\Phi(r, \theta, z) = f(r)\, \mathrm{e}^{\mathrm{i}\theta}, \tag{1.5.1}$$

where $f(0)=0$. Here r, θ, z are cylindrical coordinates.

The coupling of a four-component fermion to the string is described by a Lagrange density

$$L = \bar{\psi}_{\mathrm{L}} \mathrm{i}\partial\!\!\!/ \psi_{\mathrm{L}} + \bar{\psi}_{\mathrm{R}} \mathrm{i}\partial\!\!\!/ \psi_{\mathrm{R}} - \bar{\psi}_{\mathrm{L}} \psi_{\mathrm{R}} \Phi - \bar{\psi}_{\mathrm{R}} \psi_{\mathrm{L}} \Phi^*, \tag{1.5.2}$$

where $\psi_{\mathrm{R,L}}$ denote eigenstates of $\gamma_5 = \mathrm{i}\gamma_0\gamma_1\gamma_2\gamma_3$ with eigenvalues $+1$, -1. The Yukawa coupling has been absorbed by properly normalizing the scalar field Φ. Thus $f(r=\infty)$ is m, the fermion mass in the vacuum. The Dirac equation derived from this Lagrange density is

$$\mathrm{i}\partial\!\!\!/ \psi_{\mathrm{L}} = \Phi \psi_{\mathrm{R}}, \qquad \mathrm{i}\partial\!\!\!/ \psi_{\mathrm{R}} = \Phi^* \psi_{\mathrm{L}}. \tag{1.5.3}$$

We will try to find a zero-energy, or time-independent, solution.

If we assume that ψ is a function of r only, the Dirac equation becomes

$$\begin{aligned} &\mathrm{i}\gamma_1(\cos\theta - \gamma_1\gamma_2 \sin\theta) \frac{\partial}{\partial r} \psi_{\mathrm{L}} = f(r)\, \mathrm{e}^{\mathrm{i}\theta} \psi_{\mathrm{R}}, \\ &\mathrm{i}\gamma_1(\cos\theta - \gamma_1\gamma_2 \sin\theta) \frac{\partial}{\partial r} \psi_{\mathrm{R}} = f(r)\, \mathrm{e}^{-\mathrm{i}\theta} \psi_{\mathrm{L}}. \end{aligned} \tag{1.5.4}$$

It can be solved if

$$\mathrm{i}\gamma_1\gamma_2\psi_{\mathrm{L}} = \psi_{\mathrm{L}}, \qquad \mathrm{i}\gamma_1\gamma_2\psi_{\mathrm{R}} = -\psi_{\mathrm{R}}, \tag{1.5.5}$$

in which case we have

$$\mathrm{i}\gamma_1 \frac{\mathrm{d}}{\mathrm{d}r}\psi_{\mathrm{L}} = f(r)\psi_{\mathrm{R}}, \qquad \mathrm{i}\gamma_1 \frac{\mathrm{d}}{\mathrm{d}r}\psi_{\mathrm{R}} = f(r)\psi_{\mathrm{L}}. \tag{1.5.6}$$

Now eqs. (1.5.6) and (1.5.5) are solved by

$$\psi^0_{\mathrm{L}}(r) = \eta \exp\left[-\int_0^r f(r')\,\mathrm{d}r'\right], \qquad \psi^0_{\mathrm{R}}(r) = -\mathrm{i}\gamma_1\psi^0_{\mathrm{L}}, \tag{1.5.7}$$

where η is a constant spinor satisfying

$$-\gamma_5\eta = -\mathrm{i}\gamma_1\gamma_2\eta = \eta. \tag{1.5.8}$$

We have obtained a zero-energy solution that, in two dimensions, is normalizable, since $f(r=\infty) = m > 0$. Furthermore, the solution has two-dimensional chirality, in the sense that $\psi^0_{\mathrm{L,R}}$ are eigenstates of $\mathrm{i}\gamma_1\gamma_2$, the two-dimensional (Euclidean) analog of γ_5.

One can show explicitly that this chiral, normalizable zero-energy fermion mode survives if we now introduce a U(1) gauge field. In fact, the existence of a chiral zero mode follows from an index theorem, which states that the number of such zero modes is generically the winding number of the vortex [16].

In two spatial dimensions, there is a bound state of the fermion and vortex, because the fermion has zero energy when localized on the vortex, and mass m when far away from the vortex. In three spatial dimensions, there is again a fermion mode bound to the string, but the fermion is free to propagate along the string. We can construct a fermion wavepacket localized on the string and consider how it propagates along the string. For a fermion mode of the form

$$\psi_{\mathrm{L}} = \alpha(z, t)\psi^0_{\mathrm{L}}(r), \qquad \psi_{\mathrm{R}} = -\mathrm{i}\gamma_1\psi_{\mathrm{L}}, \tag{1.5.9}$$

the Dirac equation becomes

$$(\gamma_0\partial^0 + \gamma_3\partial^3)\alpha(z, t)\eta = 0. \tag{1.5.10}$$

It follows from the properties eq. (1.5.8) of η that $\gamma_0\gamma_3\eta = \eta$, so, eq. (1.5.10) has the general solution

$$\alpha(z, t) = \alpha(z-t). \tag{1.5.11}$$

As a consequence of the chirality of the zero-energy solution ψ^0, propagation of the fermion along the string is also chiral. The massless fermion bound to the string is a "right-mover"; it and its antiparticle propagate at the speed of light in the positive z-direction only.

We see that the low-energy excitations propagating along the string, those with wavelengths large compared to the thickness of the string, may be described by an effective (1 + 1)-dimensional field theory of chiral fermions. (By a "chiral fermion" in (1 + 1) dimensions we mean a state which propagates only to the right or only to the left. Chirality has nothing to do with helicity; there is no spin in one spatial dimension.) If we consider the antivortex $\Phi = f(r)\,\mathrm{e}^{-\mathrm{i}\theta}$ instead of a vortex, then the massless fermions propagating along the string are left-movers instead of right-movers. That conclusion is obvious: a string along the z-axis becomes an antistring if I rotate it by 180° about the x-axis, and the rotation changes the direction of propagation of the zero modes. The chirality of the fermions would also be reversed if $\bar{\psi}_{\mathrm{L}}\psi_{\mathrm{R}}$ were Yukawa coupled to Φ^* instead of Φ. All fermions that acquire mass through a Yukawa coupling to Φ are right-movers along the string, and all fermions that acquire mass through a Yukawa coupling to Φ^* are left-movers along the string.

Now, to understand at last why the word "superconducting" appears in the title of this section, let us imagine that our fermions carry non-vanishing charges under both the $\mathrm{U}(1)_Y$ gauge group, which is spontaneously broken by the expectation value of Φ, and also another, unbroken, electromagnetic gauge group $\mathrm{U}(1)_Q$. Let us ask what happens if an electric field in the positive z-direction is applied along a string that is initially in its fermionic ground state. In the ground state, the Dirac sea is filled for each of the fermion modes. But constant electric field of strength E applied for a time t will increase the Fermi level for each right-mover, and decrease the Fermi level for each left-mover by an amount

$$P_{\mathrm{F}} = qEt, \tag{1.5.12}$$

where q is the charge of the fermion in units of e. Thus the Dirac vacuum is not preserved by the applied electric field; fermions or antifermions are created (fig. 10). From the one-dimensional density of states $\mathrm{d}p/2\pi$, we infer that the number per unit length of right-moving fermions or left-moving antifermions produced is $qeEt/2\pi$, and that the total electric charge density ρ on the string changes at the rate

$$\frac{\mathrm{d}\rho}{\mathrm{d}t} = \left(\sum_{\mathrm{R}} q_{\mathrm{R}}^2 - \sum_{\mathrm{L}} q_{\mathrm{L}}^2\right)\frac{e^2E}{2\pi}, \tag{1.5.13}$$

if the contributions of all right-moving and left-moving fermion modes are summed. We find that the electric charge on the string is not conserved

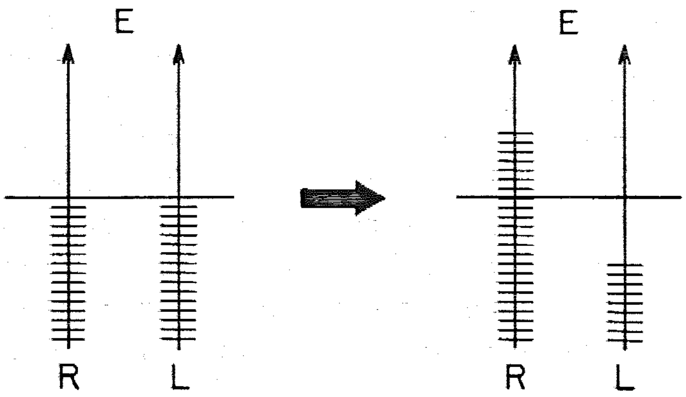


Fig. 10. Right-moving fermions and left-moving holes are created when an electric field is applied along a string.

unless

$$\sum_{\mathrm{R}} q_{\mathrm{R}}^2 = \sum_{\mathrm{L}} q_{\mathrm{L}}^2. \tag{1.5.14}$$

We have encountered the well-known chiral anomaly, which renders inconsistent a two-dimensional gauge theory in which eq. (1.5.14) is not satisfied.

In fact, though, this anomaly does not occur in our effective two-dimensional field theory if the four-dimensional theory we started with is itself anomaly free. The scalar field Φ has a $\mathrm{U}(1)_Y$ charge, which we may normalize to one, and vanishing $\mathrm{U}(1)_Q$ charge. There will be a right-moving fermion mode on the string if $\bar{\psi}_{\mathrm{L}}\psi_{\mathrm{R}}$ is Yukawa coupled to Φ. Such a coupling is invariant under $\mathrm{U}(1)_Q \times \mathrm{U}(1)_Y$ if ψ_{L} and ψ_{R} have charge assignments

$$\psi_{\mathrm{L}}\colon q, y \qquad \psi_{\mathrm{R}}\colon q, y-1. \tag{1.5.15}$$

The pair of fermions ψ_{L}, ψ_{R} therefore make a contribution proportional to q^2 to the QQY triangle anomaly in four dimensions. There will be a left-moving fermion mode on the string if $\bar{\psi}_{\mathrm{L}}\psi_{\mathrm{R}}$ couples to Φ^*, in which case the allowed charge assignments are

$$\psi_{\mathrm{L}}\colon q, y, \qquad \psi_{\mathrm{R}}\colon q, y+1, \tag{1.5.16}$$

and the pair ψ_{L}, ψ_{R} make a contribution proportional to $-q^2$ to the QQY anomaly. Thus, the condition for cancellation of QQY anomalies in the four-dimensional theory becomes precisely eq. (1.5.14) in the effective two-dimensional theory.

You might wonder about the case of the axion string. If $\mathrm{U}(1)_Y$ is the Peccei–Quinn symmetry, then there *is* a QQY anomaly in the four-dimensional theory, and electric charge is *not* conserved in the effective two-dimensional theory. Apparently, electric charge can flow onto and off the string. This observation seems puzzling at first, because low-energy fermions are firmly bound to the string; they cannot leave it without acquiring large masses. But the paradox is resolved when we recall that axion strings are necessarily the boundaries of domain walls. The domain walls also have normalizable zero-energy nodes, and electric charge can flow from wall to string and back again [17].

Anyway, let us suppose that both $\mathrm{U}(1)_Q$ and $\mathrm{U}(1)_Y$ are gauged in the four-dimensional theory, and that both are free of anomalies, so that the effective two-dimensional theory is also anomaly free. Then the right-movers and left-movers make equal and opposite contributions to the rate of change of the charge density ρ on the string. But they make contributions of the same sign to the rate of change of the electric current J flowing on the string. After a constant field E has been applied for a time t, the current flows in the same direction as the applied field and has magnitude

$$J=\left(\sum_{R} q_{\mathrm{R}}^{2}+\sum_{\mathrm{L}} q_{\mathrm{L}}^{2}\right)\frac{e^{2}E}{2\pi}\,t. \tag{1.5.17}$$

The signal that the string behaves like a superconducting wire is that it is $\mathrm{d}J/\mathrm{d}t$, rather than J, which is proportional to the applied field E. If the electric field is turned off, the current persists indefinitely.

The supercurrent eventually saturates. When the fermions have an energy comparable to their mass m, they are no longer bound to the string. Thus, the maximum current due to a single fermion mode is roughly $J_{\max}=qem/2\pi$. If m is the electron mass and $q=1$, this current is about 20 A.

We can easily imagine a grand unified theory in which there are superheavy charged fermions which acquire their mass from the expectation value of a certain scalar field. If the theory has a string solution for which that scalar field has a nontrivial winding number, then the superheavy fermions have zero-energy modes confined to the string. The current does not saturate until it is truly enormous, if the fermion masses are very large. Furthermore, Nature provides a convenient mechanism for driving the current; such a string crossing the magnetic field lines of our galaxy would be subject to an effective electric field along the string. If such strings exist, and are produced in the early universe, then we might not have to build the SSC. There would not be fermion–antifermion annihilations, because fermions and their antiparticles move in the same direction along the string. But there would be hard-scattering events between right-movers and left-movers that could produce lots of stuff. We would still need to build detectors, but the accelerator would be provided for free.

It is also fun to contemplate a string such that the electroweak Higgs doublet H has a nontrivial winding number. The light fermions, ordinary quarks and leptons, acquire their masses from the expectation value of H. Therefore, a light fermion supercurrent would flow along this string.

For each generation of quarks and leptons, the up quark u and its antiquark $\bar{\mathrm{u}}$ get mass from a Yukawa coupling to H, while the down quark d, its antiquark $\bar{\mathrm{d}}$, and the charged leptons e^+, e^- get mass from the charge conjugate scalar H^{c}. Thus, the fermions in the effective two-dimensional theory describing propagation along the string are

$$\text{right-movers: u}, \bar{\mathrm{u}}, \qquad \text{left-movers: d}, \bar{\mathrm{d}}, \mathrm{e}^+, \mathrm{e}^-. \tag{1.5.18}$$

When an electric field is applied along the string, each fermion species is produced at a rate proportional to its electric charge. Recalling the three-fold color degeneracy of the quarks, we see that the fermions created on the string have the quantum numbers

$$uude^-, \tag{1.5.19}$$

the same quantum numbers as a hydrogen atom. Predictably, no electric charge or $B-L$ are produced; these are good anomaly-free symmetries of the standard electroweak model. However, baryon number is not conserved, reflecting the BQY anomaly of the standard model, where Y is weak hypercharge. An electroweak string carries a weak hypercharge magnetic flux. When an electric field is applied along the string, there is a nonzero $\boldsymbol{E}_Q \cdot \boldsymbol{B}_Y$, which can act as a source of baryon number, as 't Hooft pointed out long ago. When the Fermi level in the string has reached a few hundred MeV, hadrons that have been produced are able to leave the string. This string is thus able to convert the energy stored in the galactic magnetic field into matter.

Are there realistic grand unified theories that exhibit the phenomena we have been talking about? I will describe just one example, which nicely illustrates the possibilities.

The example is an E_6 model [15]. Let us suppose that E_6 is first broken down to $\mathrm{SO}(10) \times \mathrm{U}(1)$ by, say, a Higgs field in the 78 representation, and that the U(1) symmetry is subsequently broken at a lower mass scale by the expectation value of the SO(10) singlet contained in a Higgs transforming as a 27 representation:

$$E_6 \xrightarrow[\langle 78_H \rangle]{} SO(10) \times U(1) \xrightarrow[\langle 27_H \rangle]{} SO(10). \qquad (1.5.20)$$

There is a string associated with the breakdown of the U(1) symmetry at the second stage of symmetry breakdown. (Actually, as we will discuss later, this string is not really topologically stable. There are also magnetic monopoles in this theory, and it is possible for the string to break by the nucleation of a monopole–antimonopole pair. But we may choose the two symmetry breaking scales to differ by orders of magnitude, and in that case the probability that the string will break is so small that it can be safely neglected.)

The only thing we will need to know about E_6 is that the 27 representation of E_6 decomposes under the $SO(10) \times U(1)$ subgroup as

$$27 \rightarrow 1^1 + 10^{-1/2} + 16^{1/4}.$$

It is the 1^1_H component of the Higgs field 27_H that acquires the vev that breaks the U(1) symmetry, and it is this field which has a nontrivial winding number in the string solution. Now, fermions in this model are also in the 27 representation, and fermion masses are generated by an E_6-invariant Yukawa coupling

$$27_F\, 27_F\, 27_H. \qquad (1.5.21)$$

Decomposed with respect to $SO(10) \times U(1)$, this coupling contains a piece of the form

$$10_F^{-1/2}\, 10_F^{-1/2}\, 1_H^1. \qquad (1.5.22)$$

Thus, there is a superheavy fermion transforming as $10^{-1/2}$ under $SU(10) \times U(1)$ that acquires its mass from the Higgs field 1^1_H. This fermion has a zero mode confined to the string.

Does the string also have light fermion zero modes? The light fermions are contained in the 16 representations of SO(10), and acquire mass from a term in (1.5.21) of the form

$$16_F^{1/4}\, 16_F^{1/4}\, 10_H^{-1/2}, \qquad (1.5.23)$$

which, when decomposed with respect to representations of the SU(5) subgroup of SO(10) becomes

$$10_F\, 10_F\, 5_H + 10_F\, \bar{5}_F\, \bar{5}_H + \cdots. \qquad (1.5.24)$$

The electroweak doublet is a linear combination of the doublet contained in 5_H and $\bar{5}_H$. The u acquires its mass from the expectation value of 5_H, and d and e acquire mass from the expectation value of $\bar{5}_H$.

Let us consider how the 5_H and $\bar{5}_H$ fields behave in the vicinity of the string, once their vacuum expectation values turn on. As the 5_H or $\bar{5}_H$ is transported around the string, it undergoes a U(1) gauge transformation. This gauge transformation rotates the phase of the 1_H^1 Higgs field by 2π, but because the 5_H and $\bar{5}_H$ have U(1) charge $-\frac{1}{2}$, their phases are rotated by only $-\pi$. In order for these fields to be single-valued, the string must carry a $U(1)_Y$ weak hypercharge magnetic flux, so that 5_H and $\bar{5}_H$ also undergo a $U(1)_Y$ gauge transformation that rotates their phases by $\pm\pi$. And since 5_H and $\bar{5}_H$ have opposite weak hypercharge, the $U(1)_Y$ transformation rotates their phases in opposite directions. Thus, in the actual string solution, either the 5_H or the $\bar{5}_H$ Higgs field, but not both, will have a nontrivial winding number; which case is realized depends on details of the Higgs potential. The up quark u or the down quark d and charged lepton e are trapped on the string. The low-energy effective theory describing the fermion modes propagating on the string contains light fermions and superheavy fermions moving in opposite directions.

The string, in this case, has a very heavy, compact core associated with the superheavy scale of symmetry breakdown, surrounded by a much lighter envelope with a thickness determined by the electroweak scale. The superheavy fermion current is confined to the core, and the light fermion current flows in the envelope.

Exercise. I mentioned that the strings of the E_6 model can end on monopoles. What do you think would happen if a fermion bound to the string were to encounter a monopole?

So far we have considered string superconductivity due to fermionic charge carriers, but it is also possible for a bosonic charge carrier to be confined to a string. This possibility is illustrated by the Alice string of section 1.2.

When the symmetry breakdown $SO(3) \to O(2)$ occurs, the charged vector boson fields have vanishing expectation values in the vacuum, of course. But inside the string the heavy vector bosons are excited, and the fields have non-vanishing expectation values. As a result, the electromagnetic $U(1)_Q$ symmetry is in effect spontaneously broken inside the string, and the string behaves like a superconducting wire [18].

Since the charged fields are excited inside the string, the string is not invariant under a global $U(1)_Q$ rotation; the strings form a degenerate set labeled by a charge rotation angle σ. We can introduce a local field $\sigma(z, t)$ on the string by performing charge rotations that vary as a function of position along the string. But a global color rotation costs no energy; therefore σ is a massless scalar field on the string. The massless bosonic excitations that carry the supercurrent are the σ excitations.

We saw that when a charged particle circles an Alice string, its charge changes sign. Electric charge is conserved, so it is natural to wonder what happened to the charge. We can now understand that the charge is transferred to the string in the form of a σ excitation, and is then carried away along the string at the speed of light.

To establish that the charged fields are really excited inside the string, we may argue as follows: We noted earlier that the asymptotic gauge field at a long distance from the string is a pure gauge. And from the gauge field on the circle at infinity, we can construct the path-ordered exponential

$$\mathrm{P}\exp\left(\mathrm{i}\oint A_\mu\,\mathrm{d}x^\mu\right) = h, \tag{1.5.25}$$

where h is an element of the component of O(2) not connected to the identity. Now imagine shrinking the circle at infinity down to a point at the origin. As the circle shrinks to a point, the path-ordered exponential must become the identity; otherwise there would be a finite magnetic flux at a point singularity, which would surely cost infinite energy. Since it begins in the component of O(2) not connected to the identity, and ends up at the identity, the path-ordered exponential must take values in SO(3) which are not in O(2) as the circle shrinks. This means that the fields coupled to the broken SO(3) generators must take nonvanishing values inside the core, as we wanted to show.

1.6. Cosmic strings

The strings that appear in spontaneously broken gauge theories have drawn increasing attention from cosmologists in recent years. The reason for this interest is that strings might have been produced in a phase transition in the very early universe, and these "cosmic strings" are the basis of a quite attractive, though still very speculative, theory of the formation of galaxies and other large-scale structures in the universe.

The main idea [19, 20] of the string picture of galaxy formation is that closed loops of cosmic string served as seeds onto which matter accreted, which led to the formation of galaxies, clusters of galaxies, and other structures. Several features of the observed large-scale structure in the universe lend support to this view.

For one thing, the spatial positions of loops of cosmic strings are not expected to be randomly distributed; they are correlated, and these correlations are inherited by the objects seeded by the loops. Furthermore, as will be explained in more detail below, the distribution of loops of string should be scale invariant. The correlations among loops with a characteristic size of order R_1 have the same form as the correlation among loops with size R_2, aside from a trivial redefinition of the length scale. Loops of different size seed features of different mass. We are thus led to predict, for example, that the two-point correlation function for galaxies should have roughly the same form as the two-point function for rich clusters of galaxies, if the unit of length in both cases is chosen to be the mean separation between the objects being considered [21]. This prediction is confirmed reasonably well by observation, especially if one allows for an enhancement of the galaxy correlations due to nonlinear gravitational effects [22]. The traditional view of the origin of large-scale structure, in which structures evolve from a Gaussian distribution of small fluctuations in the energy density, has been less successful in explaining the relation between the correlations of galaxies and the correlations of clusters.

A related observation is that the virial and peculiar velocities of objects in the universe, ranging from stars in galaxies to galaxies in superclusters, are always roughly $v \sim 10^{-3}c$, independent of length scale. This scale independence might be explained by the cosmic string picture, since the escape velocity from a loop of string is independent of the linear size of the loop; it is

$$v \sim (\sqrt{G\mu}),$$

where μ is the energy per unit length of the string and G is Newton's constant. To obtain $v \sim 10^{-3}c$ we require $\mu \sim (10^{16}\,\mathrm{GeV})^2$. The mass 10^{16} GeV is not implausible as a scale of symmetry breakdown in a grand unified theory.

Attempts to derive detailed predictions concerning the evolution of large-scale structure from the cosmic string scenario are just beginnning, and will not be described here. I will briefly describe, though, the basic picture of how strings might have been produced in the early universe, and how they subsequently evolve.

We should note first of all that the proposal that galaxies are seeded by loops of cosmic string implicitly assumes that no density fluctuations other than those produced by the strings have an important influence on large-scale structure. Therefore, the universe must have been homogeneous to very high accuracy before the strings were produced. The string scenario therefore requires inflation prior to the production of strings, or some other means of establishing a very homogeneous initial state.

Strings arise as a consequence of spontaneous symmetry breakdown. But we typically expect that spontaneously broken symmetry is restored at sufficiently high temperature. There is a critical temperature T_c comparable to the symmetry-breaking mass scale v, and for temperature T above T_c, the scalar field Φ that acts as an order parameter for the symmetry breakdown has a vanishing expectation value. In the early universe, T was initially above T_c, but as the universe expanded and cooled, T eventually fell below T_c, and a phase transition occurred; the expectation value of Φ turned on, and strings were produced. (Incidentally, it is essential that the symmetry breakdown induced by the expectation value of Φ not admit magnetic monopoles as well as strings. Otherwise, an unacceptably large abundance of monopoles would also be produced in the phase transition, which would radically alter the evolution of the universe. The cosmic string scenario requires that monopole production occurs before inflation, while string production occurs after.)

Since the temperature of the universe was very uniform, the phase transition occurred everywhere at roughly the same time. But when the expectation value of Φ turned on, it chose its orientation in the vacuum manifold at random. Furthermore, for two regions separated by a distance greater than t, the time since the universe reheated after inflation, the choices made by Φ in the two regions were essentially uncorrelated; these regions had not communicated since prior to inflation. We may thus regard Φ as having a domain structure soon after the phase transition, with the characteristic correlation length ξ, the size of a domain, satisfying $\xi < t$. When the domains coalesce, topological defects are sometimes frozen in; these are the strings [23].

This process is simulated in fig. 11. Domains are represented by the sites of a triangular lattice in the plane. Suppose that a U(1) symmetry is spontaneously broken in the phase transition, giving rise to Nielsen–Olesen vortices. The order parameter may take any value on the unit

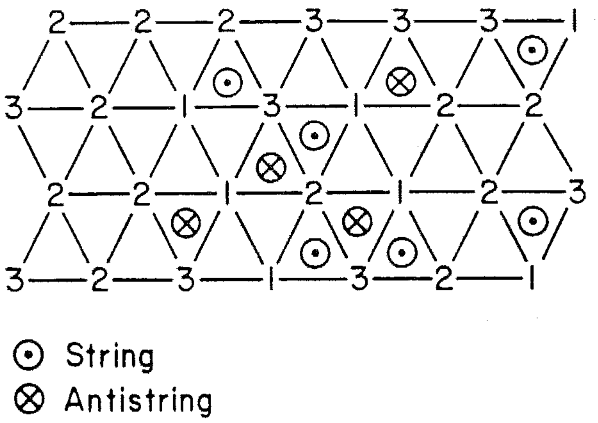


Fig. 11. A simulation of string production in a cosmological phase transition.

circle, but for purposes of illustration, we divide the circle into three equal segments, and assign to each domain the value 1, 2, or 3, chosen randomly. A vortex appears on each triangular plaquette of the lattice for which the sites of the plaquette take the values 1–2–3 in the clockwise sense; an antivortex appears if the sites take the values 1–2–3 in the counterclockwise sense. In this model, $\frac{1}{9}$ of the plaquettes contain vortices and $\frac{1}{9}$ contain antivortices, on the average.

Such a simulation can be extended to three-dimensional space, by filling space with triangular simplices. The plaquettes with vortices can be joined together, defining a self-avoiding network of strings. Each string is either infinite in extent, or forms a finite closed loop. The trajectory of a string is essentially Brownian: the distance the trajectory travels from its starting points increases like the square root of the length of the trajectory, on the average. One might expect the trajectories to deviate from Brownian, because the string network is self-avoiding. Indeed, a single self-avoiding random walk is not Brownian; the self-avoiding condition acts like an effective repulsive force that causes the trajectory to tend to straighten and expand. But in a self-avoiding *network* of strings, this repulsion becomes isotropic for trajectories that are sufficiently long. The self-avoiding condition is as likely to compress a trajectory as stretch it, and the net result is that sufficiently long trajectories are Brownian [24].

In two dimensions, a random walk always returns to its starting point, but in three dimensions there is a finite probability that the walk never returns to its starting point. Thus, in three dimensions, a finite fraction of order one of the length of string in a self-avoiding network consists of infinite strings, rather than closed loops [25, 26]. The precise fraction depends on the lattice chosen for the simulation, because most of the length of string in closed loops consists of small loops with a size comparable to the lattice spacing. But the scaling behavior of the probability distribution for large loops is easily found, because a large loop is well approximated by a random walk. The probability that a three-dimensional random walk returns to the origin on its kth step is $(\pi k)^{-3/2}$ for k large; this is the familiar "spreading of the wave packet" in the solution to the diffusion equation. The number density $\mathrm{d}n$ of closed loops with length between l and $l+\mathrm{d}l$ is proportional to the probability that a random walk returns to its origin after traveling a distance between l and $l+\mathrm{d}l$, or

$$\mathrm{d}n \propto \mathrm{d}l/l^{5/2}. \tag{1.6.1}$$

Since the loops are Brownian, a loop of length l typically fits inside a sphere of radius $R \approx l^{1/2}$. Expressing the distribution eq. (1.6.1) in terms of the radius R of the loop, we have

$$\mathrm{d}n \propto \mathrm{d}R/R^4. \tag{1.6.2}$$

From the dimensional consistency of eq. (1.6.2), we see that the step length of the walk has dropped out of the relation between n and R. The distribution of loop sizes does not depend on any intrinsic length scale; it is scale invariant [25, 27].

Having established some of the statistical properties of the initial configuration of strings produced in the phase transition, let us now consider how the string network subsequently evolves. When first formed, the strings have many kinks and wiggles, and the string tension causes the wiggles to vibrate. At first, friction due to the surrounding radiation gas may impede the vibrations, but as the universe expands and the radiation density decreases, friction quickly becomes negligible, and the strings are soon moving with velocity of order c [23]. Causality requires that wiggles with wavelength larger than the "horizon" size t_{H} (where t_{H}^{-1} is the Hubble parameter) remain frozen in; they are merely conformally stretched as the universe expands. But as the horizon size rapidly increases, the wiggles eventually come within the horizon, and begin to vibrate. As they vibrate, they are reduced in amplitude by the cosmological red shift. Thus, the strings tend to straighten out on distance scales smaller than the horizon, and the step length of the random walk remains comparable to the Hubble length [27].

As the network vibrates, strings inevitably collide. How the string network evolves depends crucially on how colliding strings behave. When two strings collide they may either pass through each other intact, or break and rejoin with new partners, a process called intercommuting. The likelihood that two colliding strings intercommute can be parametrized by an intercommuting probability p. It seems reasonable to expect that p is of order one. In particular, one expects that intercommuting is possible in the classical, or weak-coupling limit. In this limit, intercommuting is a deterministic process; whether it occurs or not is determined by initial conditions, the relative velocity of the strings and the angle at which they cross. There is no reason to expect the phase space of the initial conditions for which intercommuting occurs to be small, or for the intercommuting process to be suppressed by quantum corrections. Nevertheless, for the purpose of discussing the evolution of the string network, it is useful to imagine that the intercommuting probability p is small, so that it is sensible to expand in powers of p.

To begin with, consider the case of noninteracting strings, $p=0$. We wish to determine how the contribution to the energy density due to the infinite strings evolves in this case. The key observation is that the number of open strings crossing a horizon volume increases with time. To see this, suppose that the string network initially has a step length s_0. Counting only the infinite strings, and ignoring the loops, we may identify the mean number of pieces of infinite string that cross a cubic cell with edge length s_0; call this number m_0. Now consider a larger cell, with side Ns_0. How many open strings m cross this cell? The total length of string inside the cell is $N^3 m_0 s_0$, while each string crossing the cell has a length of order $N^2 s_0$; thus

$$m = Nm_0. \tag{1.6.3}$$

In order to appreciate the implications of this observation, it is convenient to introduce a "conformal time" variable τ such that the space–time metric can be expressed as

$$\mathrm{d}s^2 = a^2(\tau)[\mathrm{d}\tau^2 - \mathrm{d}x^2]. \tag{1.6.4}$$

Conformal time is convenient because the *coordinate* horizon size increases linearly with τ. Since features of the string network larger than the horizon size are conformally stretched, it follows from eq. (1.6.3) that the number of open strings crossing the horizon volume increases linearly with τ.

To find how the energy density due to open strings evolves, recall that the persistence length of the string remains comparable to the horizon size. If the universe is radiation dominated, then $a(\tau) \propto \tau$, and each open string crossing a horizon volume contributes $\mu[\tau a(\tau)]^{-2}$ to the energy density, where μ is the mass per unit length of the string ($\tau a(\tau)=t$ is the horizon size). If the number of such open strings increases like τ, we have

$$\rho_{\substack{\text{open}\\ \text{string}}} \propto [a(\tau)]^{-3}, \tag{1.6.5}$$

the same behavior as for nonrelativistic matter; the length of string per comoving volume is preserved. (The expansion of the universe merely straightens the strings; it does not create new string by stretching the network.) Meanwhile, because of the red shift, the energy density due to radiation decreases like $\rho_{\text{rad}} \propto a^{-4}$. If the strings were really noninteracting, they would eventually dominate the energy density of the universe [28].

But if the strings intercommute with probability p, then closed loops of string can be produced by the intercommutation of open strings, and the mean number m of open strings crossing a horizon volume eventually stabilizes. Loops with a size comparable to the Hubble size can be produced by various means. A single open string might self-intersect and break off a loop. Since the open strings have a persistence length comparable to the Hubble length, and move with velocities of order c, roughly m such self-intersections typically occur in each Hubble volume per Hubble time. For each self-intersection, the probability is p that a loop forms, so loops are produced by this mechanism at a rate of order pm per Hubble volume and Hubble time. A loop can also be produced by a *pair* of open strings that intercommute twice; this mechanism produces Hubble-size loops at a rate of order $\frac{1}{2}(pm)^2$ per Hubble volume and Hubble time. Adding the production rates due to interactions of three, four, and more open strings, one finds that the result exponentiates; a crude estimate of the rate of loop production for small p is $e^{pm}-1$ per Hubble volume and Hubble time. Meanwhile, due to the effect described above, the number m of open strings per Hubble volume tends to increase at a rate of order m per Hubble time. But the length of string converted into closed loops is removed from the network of open strings, and we see that this tendency of m to increase is in equilibrium with the tendency of closed-loop formation to decrease m for $m \sim (1/p)\ln(1/p)$. Of course, in view of the crudeness of this discussion, it is evident that the logarithm should not be taken seriously, but it seems safe to conclude that m will stop increasing when it reaches an equilibrium value of order $1/p$. When m attains this equilibrium value, Hubble-size loops are forming at a rate of order $1/p$ per Hubble volume and Hubble time. (Loops are also being destroyed at a comparable rate, breaking open in collisions of loops with open strings.)

We conclude that, regardless of the detailed configuration of the initial string network (which might depend on the nature of the phase transition), the network approaches a steady state in which of order $1/p$ open strings cross each horizon volume, and of order $1/p$ loops with a radius comparable to the Hubble length are produced per Hubble volume and Hubble time [29]. A few Hubble times after they form, these loops are no longer likely to encounter an open string; they have become isolated from the open-string network.

Since the number of open strings per horizon volume stabilizes, the energy density due to open strings approaches a fixed fraction of the radiation density; in order of magnitude this is

$$\rho_{\mathrm{string}}/\rho_{\mathrm{rad}} \approx (\mu/pt_{\mathrm{H}}^2)/(3/32\pi G t_{\mathrm{H}}^2) = \tfrac{32}{3}\pi p^{-1}(G\mu). \tag{1.6.6}$$

A horizon volume typically contains $m \sim 1/p$ open strings, but there are of course $\sqrt{m}$ fluctuations when we compare different horizon volumes. Since the string formation process is causal, and involves no transfer of energy over distances greater than t_{H}, there must be compensating fluctuations in the radiation energy density. But as a comoving volume comes within the horizon, the motion of the strings at velocities of order c destroys the delicate balance between the string and radiation fluctuations, and, within a Hubble time genuine energy density fluctuations are established [30], with

$$(\delta\rho/\rho)_{\mathrm{Horizon}} \sim \mathrm{O}(30)p^{-1/2}G\mu. \tag{1.6.7}$$

Using our earlier estimate $G\mu \sim 10^{-6}$, and taking $p \sim 1$, these fluctuations are just barely small enough to be consistent with current observational bounds on the anisotropy of the microwave background [31].

More novel from the point of view of the formation of large-scale structure are the fluctuations produced by the closed loops of string. After a loop forms, its string tension causes it to oscillate with a period comparable to its size. The oscillating loop is likely to intersect itself, and, by intercommuting, to cut itself into smaller loop fragments. But it has been suggested [19, 32], and this is the crucial assumption of the cosmic string model of galaxy formation, that this self-intersection process may typically terminate after a parent loop has broken into a small number (of order ten) of daughter loops; the daughter loops will assume configurations that do not self-intersect as they oscillate. The daughter loops may then survive for a long time, long enough to seed large-scale structures. And daughters of the same parent loop will have correlated positions; these correlations will be inherited by the seeded structures [21, 22, 29]. Since the production and fragmentation of loops of all sizes occur by the same process - the only relevant length scale in the process is the size of the loop itself - the distribution of loops has the scale-invariance property stressed at the beginning of this section.

Even if loops of string cease to self-intersect, they cannot survive forever. (If they could, the *loops* would eventually dominate the energy density of the universe.) The most efficient means by which a string loop can lose energy is the emission of gravitational radiation [27]. Other types of radiation that might conceivably be emitted by an oscillating loop of radius R are quite inefficient when the frequency R^{-1} of the radiation is very small compared to the energy μR of the loop. But the rate of emission of gravitational radiation does not depend only on R; it is proportional to $G\mu^2$. Thus, an oscillating loop loses an appreciable fraction of its initial energy to gravitational radiation in of order $(G\mu)^{-1}$ oscillations. The gravitational radiation emitted by the loops is a very important feature of the cosmic string scenario because it should be experimentally detectable [33]. The contribution from decaying loops to the stochastic gravitation wave background ought to have an observable influence on pulsar timing measurements within a decade [34].

While on the subject of the evolution of a system of cosmic strings, let us consider how a system of walls bounded by strings should be expected to evolve. In a model (like an axion model) in which the symmetry-breaking scale associated with the domain walls is much lower than the symmetry-breaking scale associated with the strings, the walls can have no appreciable influence on the dynamics of the string network until the characteristic distance between the strings is larger than the wall thickness, and the energy density inside the walls is greater than the energy density of the surrounding radiation. The distribution of the walls when they finally do appear can be modeled by an extension of the simulation of the string network outlined earlier. For the case of axion strings, we may imagine that the spontaneously broken U(1) symmetry is not really exact, and that the energetically preferred value of the order parameter is in region 3 of the unit circle; then we should place the domain wall so that each link of the lattice that connects 1 and 2 pierces the wall. Sliced through a plane, the domain walls are curves

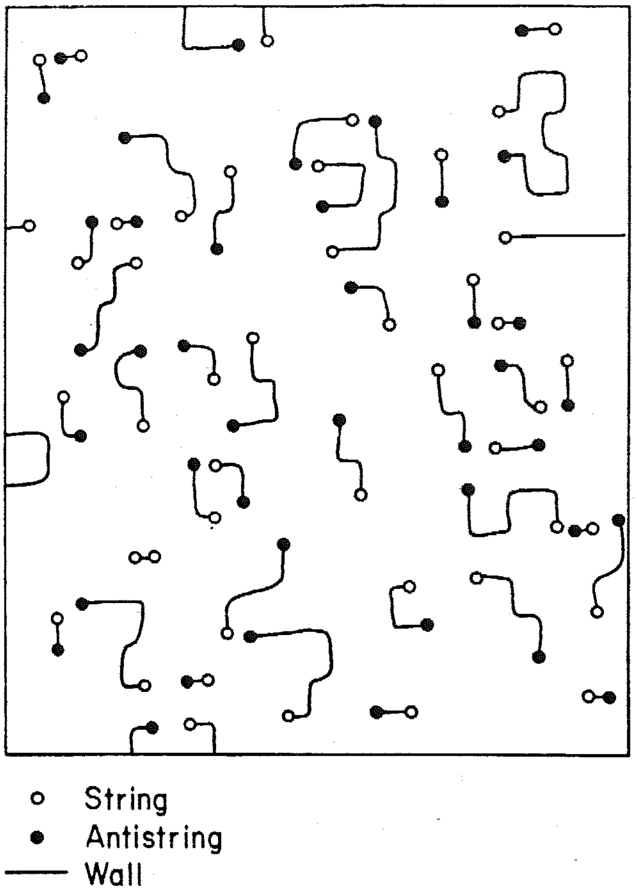


Fig. 12. A simulated two-dimensional slice of a system of domain walls bounded by strings.

that connect each vortex to an antivortex. A typical distribution of domain wall slices is shown in fig. 12. One sees that, in any two-dimensional slice through the system of walls bounded by strings, slices of wall much larger than the characteristic distance between strings are not common. If I start at a vortex and walk along the wall, each time I advance by one lattice spacing the probability that I encounter an antivortex, and the end of the wall, is $\frac{1}{3}$. Thus, the abundance of long wall slices is exponentially small [35].

In any two-dimensional slice of the wall–string system, each string is connected by wall to a nearby "antistring". In three dimensions, the system looks like a network of branching ribbons, depicted in fig. 13. Two strings closely approach each other at one point, and are connected by a wall. Eventually, these strings wander apart, and another string assumes the role of partner to these strings. The ribbon of wall connecting two strings appears to bifurcate; it branches into two ribbons. A ribbon can also form a "dangling end" if a string folds back on itself. The

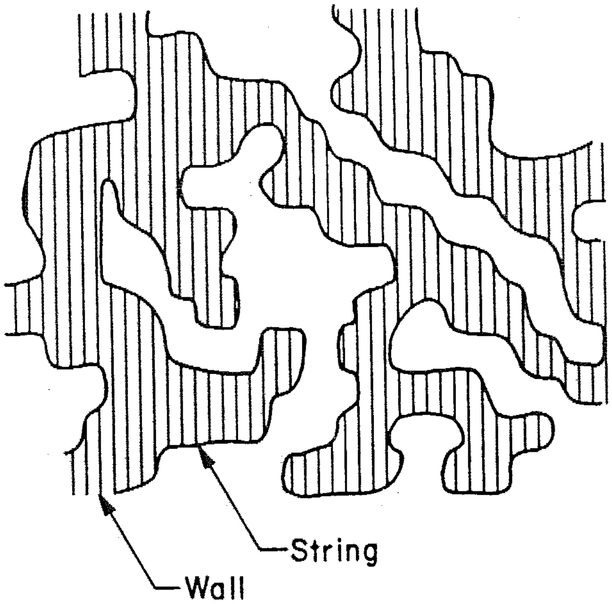


Fig. 13. Three-dimensional view of walls bounded by strings.

branching ribbons form a percolating network that fills space. In the language of polymer physics, a system of cosmic strings resembles a polymer *melt*, and a system of walls bounded by strings resembles a polymer *gel*.

The tension in the walls and strings causes the network to jiggle. In particular, the wall tension induces frequent crossings between the strings that bound a ribbon. When these strings cross, they sometimes intercommute and sever the ribbon. Rather quickly, unless the intercommuting probability is very small, the infinite network of ribbons breaks up into finite pieces. (The system no longer percolates when the mean distance between breaks in the ribbon becomes comparable to the distance between branchings.) These finite pieces fragment further, and eventually decay by emission of gravitational radiation [7, 8].

From a cosmological viewpoint, the wall-bounded-by-string system is not very interesting. It disappears with hardly a trace, and has little influence on the evolution of the universe. It is amusing, though, to reflect once more upon the cosmological status of discrete symmetries in particle physics. A spontaneously broken exact discrete symmetry causes trouble. The energy density of the universe would become dominated by a system of cosmic domain walls, unless the walls are "inflated away". But if there is an effective discrete symmetry of low-energy physics that is spontaneously broken, it need not cause trouble if the discrete symmetry is embedded in a continuous symmetry that became spontaneously broken at a larger mass scale. There may be strings generated by the higher symmetry breaking scale that act as boundaries of the domain walls, and render them harmless. On the other hand, if the strings are inflated away before the domain walls form, the walls are again troublesome.

2. Monopoles

2.1. *The quantization condition*

Magnetic monopoles, like vortices, arise as time-independent solutions with finite energy to the classical field equations of a spontaneously broken gauge theory. But a monopole has finite energy in three spatial dimensions, instead of two dimensions. And, unlike a vortex, a monopole has a long-range (magnetic) gauge field, from which it gets its name.

Most of the mass of the monopole is concentrated in a core with a size characterized by the scale of the spontaneous symmetry breakdown. We will return to a more detailed consideration of the structure of the core later. For now, let us restrict our attention to how the long-range gauge field might behave.

Suppose that the unbroken gauge group is $H=\mathrm{U}(1)$, and that the long-range ($r\to\infty$) U(1) gauge field is that of a magnetic monopole with magnetic charge g,

$$\boldsymbol{B}=\frac{g\hat{\boldsymbol{r}}}{r^2}, \qquad \boldsymbol{E}=0. \tag{2.1.1}$$

A charged particle with electric charge e interacting with the magnetic monopole satisfies the classical equation of motion

$$m\ddot{\boldsymbol{r}}=e\dot{\boldsymbol{r}}\times\boldsymbol{B}. \tag{2.1.2}$$

This equation of motion is gauge invariant and the classical dynamics it defines is perfectly sensible whatever the values of e and g. But to define the quantum mechanics of a charged particle interacting with a magnetic monopole, we need to introduce the vector potential $\boldsymbol{A}$ such that $\boldsymbol{B}=\boldsymbol{\nabla}\times\boldsymbol{A}$. The vector potential of a magnetic monopole is necessarily singular; this singularity leads to trouble, and the result is a restriction on the magnetic charge g.

To define quantum mechanics, we introduce an action functional S, and sum over classical histories weighted by the phase e^{iS}. For a charged particle in a magnetic field, the action is

$$S = S_{\text{kin}} + S_{\text{int}},$$

where S_{kin} is the action of a free particle, and

$$S_{\text{int}} = e\int_1^2 \mathrm{d}t\frac{\mathrm{d}\boldsymbol{r}}{\mathrm{d}t}\cdot\boldsymbol{A} = e\int_1^2 \mathrm{d}\boldsymbol{r}\cdot\boldsymbol{A}. \tag{2.1.3}$$

The interaction term in the action depends only on the path traveled by the particle, not on its velocity along the path.

The vector potential A cannot be smoothly defined on a sphere surrounding a magnetic monopole, but does it matter? In quantum mechanics, we care only about the relative phase associated with two paths, not about the overall phase. For two paths Γ and Γ' with the same endpoints, this relative phase is

$$(S_{\text{int}})_\Gamma - (S_{\text{int}})_{\Gamma'} = e\oint_{\Gamma-\Gamma'} \mathrm{d}\boldsymbol{r}\cdot\boldsymbol{A} = e\int_{S_{\Gamma-\Gamma'}} \mathrm{d}^2\boldsymbol{S}\cdot\boldsymbol{B} = e\Phi_{\Gamma-\Gamma'}. \tag{2.1.4}$$

By applying Stokes' Theorem, the relative phase has been expressed as the magnetic flux through a surface bounded by the closed loop $\Gamma-\Gamma'$ (fig. 14). All reference to the vector potential has disappeared, and the relative phase therefore appears to be well defined.

But in fact there is still a problem, because the phase is multi-valued. If the path Γ' is permitted to sweep once around a closed surface surrounding the monopole and return to its initial position, the action changes by

$$\Delta S_{\text{int}} = e\Phi_{\text{sphere}} = 4\pi eg. \tag{2.1.5}$$

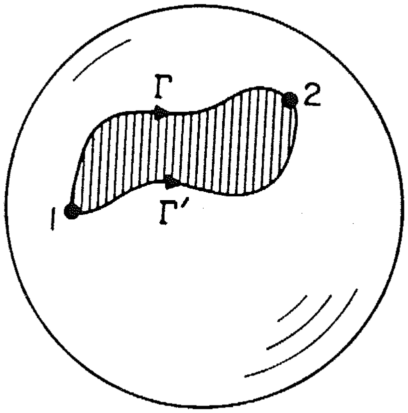


Fig. 14. Two possible trajectories with given endpoints for a charged particle on a closed surface surrounding a monopole.

The relative phase associated with two paths is unambiguously defined only if $\exp(\mathrm{i}\Delta S_{\rm int}) = 1$, or

$$eg = n/2, \tag{2.1.6}$$

where n is an integer. Equation (2.1.6) is Dirac's quantization condition [36]. The minimum allowed nonvanishing magnetic charge $g_{\rm D} = 1/2e$ is called the Dirac magnetic charge.

The Dirac quantization condition can be viewed as a consequence of gauge invariance. As we have seen, it is required for consistency in the quantum mechanics of a charged particle interacting with a magnetic monopole that the phase

$$\exp\left[\mathrm{i}e \oint_{\Gamma} \boldsymbol{A} \cdot \mathrm{d}\boldsymbol{r}\right]$$

associated with a given closed path Γ is well defined. Although it is not possible to smoothly define a vector potential everywhere on a closed surface surrounding a monopole, it is always possible to find a smooth vector potential on a *disk*, a surface with boundary. (This follows from the Poincaré lemma.) Let us therefore imagine that the closed path Γ divides a surface S surrounding the monopole into two disks U ("upper") and L ("lower") and that each disk is equipped with its own vector potential, denoted $\boldsymbol{A}_{\rm U}$ and $\boldsymbol{A}_{\rm L}$ respectively [37]. Consistency requires that the phases determined by $\boldsymbol{A}_{\rm U}$ and $\boldsymbol{A}_{\rm L}$ agree for the path Γ, or

$$\exp\left(\mathrm{i}e \oint_{\Gamma} \boldsymbol{A}_{\rm U} \cdot \mathrm{d}\boldsymbol{r}\right) = \exp\left(\mathrm{i}e \oint_{\Gamma} \boldsymbol{A}_{\rm L} \cdot \mathrm{d}\boldsymbol{r}\right). \tag{2.1.7}$$

This can be rewritten as

$$1 = \exp\left(\mathrm{i}e \oint_{\Gamma} (\boldsymbol{A}_{\rm U} - \mathrm{A}_{\rm L}) \cdot \mathrm{d}\boldsymbol{r}\right) = \mathrm{e}^{\mathrm{i}e(\Phi_{\rm U}+\Phi_{\rm L})} = \mathrm{e}^{\mathrm{i}4\pi eg}, \tag{2.1.8}$$

where $\Phi_{\rm U,L}$ is the magnetic flux through the disk U, L; we have obtained again the Dirac quantization condition.

But now eq. (2.1.8) can be reinterpreted as the statement that $\boldsymbol{A}_{\rm U}$ and $\boldsymbol{A}_{\rm L}$ are related on the boundary Γ common to both disks by a single-valued gauge transformation on Γ. Defining the gauge transformation

$$\Omega(\boldsymbol{r}_{\rm f}) = \exp\left[\mathrm{i}e \int_{\boldsymbol{r}_{\rm i}}^{\boldsymbol{r}_{\rm f}} (\boldsymbol{A}_{\rm U} - \boldsymbol{A}_{\rm L}) \cdot \mathrm{d}\boldsymbol{r}\right] \Omega(\boldsymbol{r}_{\rm i}), \tag{2.1.9}$$

with the line integral performed along Γ, we have

$$A_{\mathrm{U}}=A_{\mathrm{L}}+\frac{1}{\mathrm{i}e}(\nabla_{\Gamma}\Omega)\Omega^{-1}, \tag{2.1.10}$$

with the gradient taken along Γ. That is, the vector potentials A_{U} and A_{L} defined on the upper and lower disks are gauge equivalent on the boundary Γ where the two disks intersect, and therefore describe the same physics there. Equation (2.1.8) is just the statement that the gauge transformation Ω relating A_{U} and A_{L} on Γ is single-valued on Γ.

The gauge transformation Ω maps the closed path Γ to U(1), and it has a winding number

$$n=2eg; \tag{2.1.11}$$

this winding number is the integer that appears in the Dirac quantization condition. We have thus discovered that the Dirac quantization condition has a topological origin. Magnetic charge is quantized because the winding number must be an integer. Furthermore, since the winding number is a topological invariant, it is unaffected by deformations of the closed surface S or the loop Γ; the winding number is intrinsic to the monopole, and independent of the choice of the surface S enclosing the monopole, or the loop Γ contained in the surface.

To be more explicit, let us choose the surface S to be a sphere centered on the monopole, and the loop Γ to be the equator of the sphere (fig. 15). Then a monopole with $\boldsymbol{B}=g\hat{\boldsymbol{r}}/r^2$ can be represented by [37]

$$\begin{aligned} A_{\mathrm{U}}\cdot \mathrm{d}\boldsymbol{r}&=g(1-\cos\theta)\,\mathrm{d}\phi, \quad \text{upper } (0\leqslant\theta\leqslant\pi/2),\\ A_{\mathrm{L}}\cdot \mathrm{d}\boldsymbol{r}&=-g(1+\cos\theta)\,\mathrm{d}\phi, \quad \text{lower } (\pi/2\leqslant\theta\leqslant\pi). \end{aligned} \tag{2.1.12}$$

At the equator ($\theta=\pi/2$), where the two hemispheres intersect, A_{U} and A_{L} are related by

$$(A_{\mathrm{U}}-A_{\mathrm{L}})\cdot\mathrm{d}\boldsymbol{r}=2g=\frac{1}{\mathrm{i}e}(\mathrm{d}_{\phi}\Omega)\Omega^{-1}, \tag{2.1.13}$$

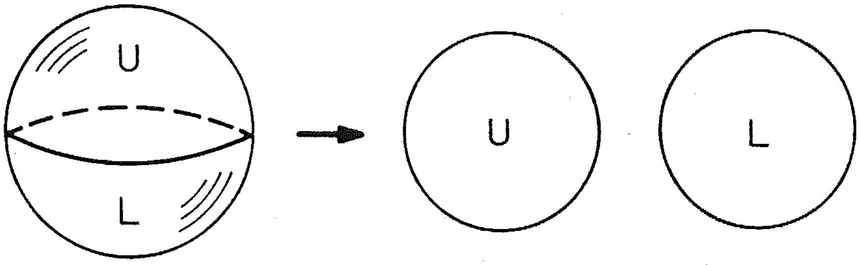


Fig. 15. Cutting a sphere at the equator reduces it to two disks.

where

$$\Omega(\phi)=\exp(\mathrm{i}2eg\phi). \qquad (2.1.14)$$

The winding number of $\Omega(\phi)$ is evidently $n=2eg$.

Now imagine that our sphere, initially very large, smoothly shrinks to infinitesimal size. As the radius of the sphere shrinks, other multipoles of the $\boldsymbol{B}$ field other than the monopole may become important, and the field may become very complicated. But as long as the sphere encounters no singularities of the $\boldsymbol{B}$ field, the winding number n must remain constant, independent of the radius of the sphere. If n is nonzero, we are forced to conclude that the magnetic charge g is contained in an arbitrarily small sphere; the monopole is a point singularity. In order to avoid this singularity the gauge transformation Ω must be allowed to wander through a larger gauge group containing U(1), in which it can "unwind". This is precisely the option exercised by the nonsingular monopole solution to be described in the next section.

Before proceeding to the discussion of the nonsingular monopole, let us quickly note that our observations concerning the U(1) monopole can be easily generalized to apply to configurations with non-Abelian long-range gauge fields. We can thus obtain a topological definition of magnetic charge appropriate for the non-Abelian case [38–40].

We may consider gauge fields, defined on a sphere, in the Lie algebra of an arbitrary Lie group H. As before, we describe the gauge field configuration by specifying nonsingular gauge potentials A_{U} and A_{L} on the upper and lower hemispheres, and a single-valued gauge transformation $\Omega(\phi)\in H$ that relates A_{U} and A_{L} on the equator. The gauge transformation $\Omega(\phi)$ is a loop in the gauge group H, classified by the first homotopy group $\pi_1(H)$. We define the magnetic charge enclosed by the sphere to be the "winding number" of $\Omega(\phi)$, the associated element of $\pi_1(H)$. This is the natural generalization of the Abelian magnetic charge.

For example, suppose that the gauge group is $H=\mathrm{SO}(3)$. SO(3) is topologically equivalent to a three-sphere with antipodal points identified; therefore there are closed paths in SO(3), those beginning at one point of the three-sphere and ending at the antipodal point, that cannot be smoothly contracted to a point. But a path that begins and ends at the same point of the three-sphere *can* be contracted to a point; it has winding number zero. Thus, the winding number of a loop in SO(3) can have only two possible values, 0 and 1, and the magnetic charge in an SO(3) gauge theory can have only two possible values. In particular, a magnetic monopole is indistinguishable from an antimonopole.

More generally, the gauge fields always transform as the adjoint representation of the gauge group, which is a faithful representation of $H=\bar{H}/K$, where $\bar{H}$ is the simply connected covering group of H, and K is a subgroup of the center of $\bar{H}$. Magnetic monopoles are classified by $\pi_1(\bar{H}/K)=K$. (We may think of $\bar{H}/K$ as the group $\bar{H}$, but with elements differing by multiplication by an element of K identified as the same element.) For SU(N) the gauge fields transform as a representation of SU(N)/Z$_N$, and the allowed magnetic charges take values in Z$_N$.

Our topological definition of non-Abelian magnetic charge is sensible. As long as the gauge fields are nonsingular and Ω is an element of H, the winding number must be a constant, independent of the radius of the sphere. So the magnetic charge is not carried by the long-range field of the monopole; it either resides on a point singularity (Dirac monopole) or a core in which gauge fields other than H gauge fields are excited (nonsingular monopole). And this magnetic charge is obviously conserved. It is a discrete quantity. But time evolution is continuous, so the total magnetic charge must be time independent.

While other gauge-invariant definitions of magnetic charge are possible, only the topological definition, which requires the monopole to have a point singularity or a core, can guarantee the stability of the monopole. If we assign "magnetic charge" to an H gauge field that is nonsingular everywhere in space, nothing can prevent this "magnetic charge" from propagating to spatial infinity as non-Abelian radiation [39, 41].

So far we have considered magnetic monopole configurations in a classical gauge field theory, but eventually we must worry about quantum mechanical effects on the magnetic field. There is really something to worry about, because non-Abelian gauge theories are believed to be confining, and to have no massless excitations. Therefore, the magnetic field must be screened by gluon fluctuations at distances large compared to the confinement distance scale [39, 42]. Fortunately, since our definition of magnetic charge is topological, it can be applied to the quantum theory. The gluon fluctuations that cause the magnetic screening cannot change the winding number of the field configuration.

2.2. *Monopoles as solitons*

The finite-energy field configurations of a spontaneously broken gauge theory in *three* dimensions are subject to a topological classification closely analogous to the classification of vortices in section 1.3. In carrying out this classification, we will discover a close connection between the topological charge and magnetic charge of a soliton [4, 39, 40].

For a theory in which the gauge group G is spontaneously broken to the subgroup H, the vacuum manifold is

$$G/H = \{\Phi, \Phi = \Omega\Phi_0, \Omega \in G\}, \tag{2.2.1}$$

where Φ_0 is a standard reference vacuum preserved by the subgroup H. For any field configuration of finite energy, the order parameter must assume a value in the vacuum manifold at each point on the two-sphere at spatial infinity. Thus to each finite-energy field configuration we may assign a mapping from S^2 into the vacuum manifold G/H. If this mapping cannot be smoothly deformed to the trivial constant mapping, there is an associated topological soliton.

By multiplying by an appropriate constant element of G, we may turn any mapping from S^2 into G/H into a mapping that takes an arbitrarily chosen reference point, say the north pole, to Φ_0. The mappings from S^2 into G/H that take the north pole to Φ_0 fall into topological equivalence classes, such that two mappings are in the same class if they can be smoothly deformed into one another. These classes are endowed with a natural group structure, since there is a natural way to compose two mappings that both take the north pole to Φ_0. This group is $\pi_2(G/H)$, the "second homotopy group" of G/H.

The group $\pi_2(G/H)$ is discrete; its elements are the possible "topological charges" of finite-energy field configurations. The discrete topological charge is preserved by continuous time evolution, and the classical field theory has a topological conservation law.

How can we compute $\pi_2(G/H)$? Mappings from S^2 into G/H are not so easy to visualize. But fortunately, we can, by a trick, reduce the topological classification of two-spheres in G/H to the topological classification of loops in H. This reduces the calculation of $\pi_2(G/H)$ to the calculation of $\pi_1(H)$, which we already know how to do.

The trick is to cut the sphere into two hemispheres, along the quator. Given a mapping $\Phi(\theta, \phi)$ from S^2 into G/H it is possible to find smooth gauge transformations Ω_U and Ω_L on each hemisphere that rotate the order parameter to the reference position Φ_0:

$$\begin{aligned} &\Omega_U(\theta, \phi)\Phi(\theta, \phi) = \Phi_0, \quad \text{upper } (0 \leq \theta \leq \pi/2), \\ &\Omega_L(\theta, \phi)\Phi(\theta, \phi) = \Phi_0, \quad \text{lower } (\pi/2 \leq \theta \leq \pi). \end{aligned} \tag{2.2.2}$$

On the equator $\theta = \pi/2$, where the two hemispheres intersect, the gauge transformation $\Omega=\Omega_U\Omega_L^{-1}$ can be defined; it preserves Φ_0 and is thus in the subgroup H. So

$$\Omega_U(\theta=\pi/2,\phi)\Omega_L^{-1}(\theta=\pi/2,\phi)=\Omega(\phi)\in H. \tag{2.2.3}$$

Now, either this loop in H can be contracted to a point in H, or it cannot be. If the loop $\Omega(\phi)$ in H *is* contractible in H, then the mapping $\Phi(\theta,\phi)$ can be deformed to a trivial constant mapping. To see this, note that eq. (2.2.2) is still satisfied if we make the replacement

$$\Omega_U(\theta,\phi)\to\Omega'_U(\theta,\phi)=\tilde{\Omega}^{-1}(\theta,\phi)\Omega_U(\theta,\phi), \tag{2.2.4}$$

where $\tilde{\Omega}(\theta,\phi)\in H$. We may choose $\tilde{\Omega}(\theta,\phi)$ to be the homotopy that contracts $\Omega(\phi)$ in H; that is

$$\tilde{\Omega}(\theta=\pi/2,\phi)=\Omega(\phi),\quad \Omega(\theta=0)=1. \tag{2.2.5}$$

Now $\Omega'_U=\Omega_L$ at $\theta=\pi/2$, and we have found a gauge transformation smoothly defined on the whole sphere that takes $\Phi(\theta,\phi)$ to Φ_0. Furthermore, it is known that $\pi_2(G)=0$ for any compact Lie group G. Thus, this gauge transformation can be deformed to the trivial gauge transformation, and the mapping $\Phi(\theta,\phi)$ can be smoothly deformed to Φ_0.

If, on the other hand, the loop $\Omega(\phi)$ is *not* contractible in H, then it is clear that the mapping $\Phi(\theta,\phi)$ *cannot* be deformed to the trivial mapping; the deformation of $\Phi(\theta,\phi)$ to Φ_0 would necessarily define a homotopy in H that shrinks $\Omega(\phi)$ to the identity. Notice, though, that our loop $\Omega(\phi)$, which is not contractible in H, *can* be contracted to a point in G. As θ varies from $\pi/2$ to 0 (or π), $\Omega_U(\theta,\phi)$ provides a homotopy that shrinks $\Omega_U(\pi/2,\phi)$ (or $\Omega_L(\pi/2,\phi)$) to a point. Since both $\Omega_U(\theta=\pi/2,\phi)$ and $\Omega_L(\theta=\pi/2,\phi)$ are contractible loops in G, so is $\Omega=\Omega_U\Omega_L^{-1}$.

We have now seen that to every class of topologically nontrivial maps $\Phi(\theta,\phi)$ from S^2 to G/H there corresponds a class of loops $\Omega(\theta)$ in H that cannot be contracted to a point in H, but can be contracted in G. It is not hard to see that this correspondence is actually one-to-one; in an equation,

$$\pi_2(G/H)=\pi_1(H)/\pi_1(G). \tag{2.2.6}$$

It only remains to show that, given any noncontractible loop in H that is contractible in G, there is a corresponding noncontractible two-sphere in G/H. Indeed, this two-sphere is generated by the homotopy that shrinks the loop $\Omega(\phi)\in H$ (represented by a point in G/H) to a point in G (fig. 16). Given the loop $\Omega(\phi)$ in H, contractible in G, we can find smooth gauge transformations Ω_U and Ω_L in G defined on the upper

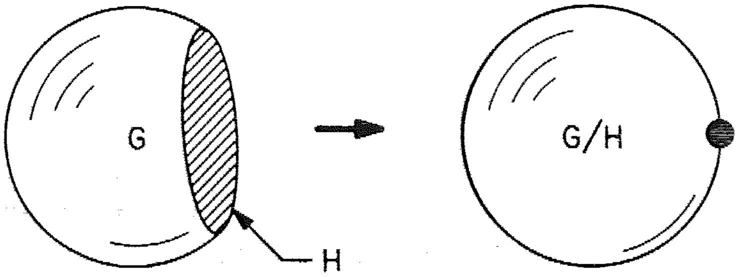


Fig. 16. The homotopy that shrinks a loop in H to a point in G defines a two-sphere in G/H.

and lower hemispheres such that

$$\Omega_{\mathrm{U}}(\theta=\pi/2,\phi)=\Omega(\phi), \qquad \Omega_{\mathrm{L}}(\theta=\pi/2,\phi)=1; \tag{2.2.7}$$

we simply choose $\Omega_{\mathrm{U}}(\theta,\phi)$ to be the smooth deformation in G of the loop $\Omega_{\mathrm{U}}(\theta=\pi/2,\phi)$ to the point $\Omega_{\mathrm{U}}(\theta=0)$. Now,

$$\begin{aligned}\Phi(\theta,\phi)&=\Omega_{\mathrm{U}}^{-1}(\theta,\phi)\Phi_0 \quad &(0\leqslant\theta\leqslant\pi/2),\\ \Phi(\theta,\phi)&=\Phi_0, &(\pi/2\leqslant\theta\leqslant\pi),\end{aligned} \tag{2.2.8}$$

is the smooth mapping from S^2 into G/H corresponding to the loop $\Omega(\phi)$.

We saw earlier that the magnetic charge of a configuration with a long-range H gauge field is specified by an element of $\pi_1(H)$. And we noted that, for a *nonsingular* monopole, it must be possible to unwind the noncontractible loop in H through a larger group G. Now we have found that the topological charge of a finite-energy field configuration in a theory with gauge group G spontaneously broken to the subgroup H is an element of $\pi_1(H)/\pi_1(G)$. We cannot help but suspect that the topological and magnetic charges of a finite-energy configuration precisely coincide. To verify this conjecture, we must consider the long-range gauge field of the soliton.

As in our analysis of vortices, we must require of a finite-energy configuration that the covariant gradient of the order parameter falls off sufficiently rapidly at large distances,

$$D_i\Phi=(\partial_i-\mathrm{i}eA_i)\Phi \xrightarrow[r\to\infty]{} 0. \tag{2.2.9}$$

In the gauge constructed in eq. (2.2.2), for which $\Phi=\Phi_0$ on the sphere at $r=\infty$, the only gauge fields that can be excited at large r are the H gauge fields, those associated with the generators of G that annihilate Φ_0. If the gauge field A is nonsingular on the sphere in the gauge for which $\Phi(\theta, \phi)$ is nonsingular, then the gauge fields A_U and A_L defined on each hemisphere are nonsingular in the gauge defined by eq. (2.2.2). Furthermore, at the equator, A_U and A_L are related by the gauge transformation $\Omega(\phi)$. The winding number of $\Omega(\phi)$, which is the topological charge of the soliton, is also the magnetic charge as defined in the previous section. Topological charge does indeed equal magnetic charge.

In any unified gauge theory, the electromagnetic $U(1)_{em}$ gauge group is embedded in a semisimple group that is spontaneously broken. The analysis of this section shows that every unified gauge theory contains magnetic monopoles as topological solitons. We will encounter some examples in the next two sections.

2.3. *The classical monopole solution*

A monopole, like a vortex, has a *core* with a characteristic finite size. The size of the core and the mass of the monopole are determined by the classical field equations. We would like to estimate the size and mass, as we did for a vortex. For this purpose we will consider the simplest unified gauge theory that contains a monopole solution; it was in the context of this model that the non-singular monopole was first discovered by 't Hooft [43] and Polyakov [44]. More complicated models will be described in the next section.

The model has the gauge group $G = SU(2)$ and a Higgs field Φ in the triplet representation of the group; its Lagrangian is

$$L = -\tfrac{1}{4}F^a_{\mu\nu}F^{\mu\nu a} + \tfrac{1}{2}D_\mu\Phi^a D^\mu\Phi^a - U(\Phi), \tag{2.3.1}$$

where

$$U(\Phi) = \tfrac{1}{8}\lambda(\Phi^a\Phi^a - v^2), \tag{2.3.2}$$

$$D_\mu\Phi^a = \partial_\mu\Phi^a - e\varepsilon^{abc}A^b_\mu\Phi^c, \tag{2.3.3}$$

$$F^a_{\mu\nu} = \partial_\mu A^a_\nu - \partial_\nu A^a_\mu - e\varepsilon^{abc}A^b_\mu A^c_\nu, \tag{2.3.4}$$

and $a = 1, 2, 3$.

The potential $U(\Phi)$ is minimized by

$$\Phi = (0, 0, v) \tag{2.3.5}$$

(in a particular gauge), and the SU(2) gauge symmetry is evidently spontaneously broken to $H = U(1)$. In perturbation theory, the spectrum of the model consists of a vector boson with mass $m_V = ev$ and a scalar with mass $m_S = \sqrt{\lambda}\ v$.

The vacuum manifold G/H, the space of values of Φ that are gauge equivalent to (2.3.5), is obviously isomorphic to the two-sphere S^2. The simplest topologically nontrivial mapping from S^2 into S^2 is the identity map. Thus, the monopole configuration, in a particular gauge, is one for which the order parameter Φ on the sphere at spatial infinity takes the form

$$\Phi^a = v\hat{r}^a. \tag{2.3.6}$$

This configuration is also called a "hedgehog", because the order parameter points radially outward.

The classical solution with the asymptotic behavior (2.3.7) has two characteristic length scales. These are the radii r_S and r_V of the regions in which the scalar field and vector field respectively depart significantly from their asymptotic values. (Compare the discussion of the vortex in section 1.1.) These lengths are chosen to minimize the energy

$$E = \int d^3x \, [\tfrac{1}{2} E_i^a E_i^a + \tfrac{1}{2} B_i^a B_i^a + \tfrac{1}{2} D_i\Phi^a D_i\Phi^a + U(\Phi)] \tag{2.3.7}$$

of the configuration. For a spherically symmetrical configuration, E is given in order of magnitude by

$$E = \sim \frac{4\pi}{e^2} m_V \left[\frac{1}{m_V r_V} + \frac{e}{\sqrt{\lambda}} m_S^3 r_S^3 \right. \\ \left. + \left(m_V r_V - \frac{e}{\sqrt{\lambda}} m_S r_S \right) \theta(r_V - r_S) + \frac{e}{2\sqrt{\lambda}} m_S r_S \right]. \tag{2.3.8}$$

The first term is the magnetic self-energy of a magnetically charged sphere with radius r_V; it favors expansion of r_V. The second term is the energy stored in the potential $U(\Phi)$; it encourages r_S to shrink. The third term is the energy due to the circumferential gradient of the scalar field Φ. This term ties together the two length scales r_V and r_S, because the gradient becomes substantial for $r > r_S$, and is eventually "screened" by the gauge field at $r \sim r_V$. (This term is not present for $r_S > r_V$.) The fourth term is the energy due to the radial gradient of Φ.

Now r_S and r_V can be chosen to minimize the energy. We find:

$$m_S > m_V: \quad r_S \sim m_S^{-1}, \qquad r_V \sim m_V^{-1}, \\ E = m_{\text{monopole}} \sim (4\pi/e^2) m_V, \tag{2.3.9}$$

$$m_S < m_V: \quad r_S \sim m_V^{-1}, \qquad r_V \sim m_V^{-1}, \\ E = m_{\text{monopole}} \sim (4\pi/e^2) m_V. \tag{2.3.10}$$

The monopole mass is not sensitively dependent on the ratio m_S/m_V;

when m_S is large, the scalar core radius is small, and the contribution of the scalar core energy to the total energy is not significant.

Comparing r_V and m_{monopole}, we see that the size of the monopole core is larger by the factor $\alpha^{-1} = (4\pi/e^2)$ than the monopole Compton wavelength. As a result, the quantum corrections to the structure of the monopole are under control, if α is small. Even though the coupling $g = 1/e$ is large, the effects of virtual monopole pairs are small, because the monopole is a complicated coherent excitation that cannot be easily produced as a quantum fluctuation.

This situation should be contrasted with the quantum mechanics of a point monopole. Virtual monopole pairs have a drastic effect on the structure of the point monopole, for which g is a genuine strong coupling. In fact, the vacuum-polarization cloud of a point monopole must extend out to distances of order $(\alpha m)^{-1}$, because the magnetic self-energy of a monopole of that size is of order m. Thus, both the nonsingular monopole and the point monopole have a complicated structure in a region with radius of order $(\alpha m)^{-1}$. But for the non-singular monopole, we have an explicit classical description of this structure, and quantum corrections are small and calculable if α is small. The point monopole, on the other hand, is a genuine strong-coupling problem. We cannot calculate anything.

The estimate $m_{\text{monopole}} \sim (4\pi/e^2) m_V$, where m_V is the mass of a heavy vector boson, also applies to more complicated unified gauge theories (see section 2.4). In a typical grand unified theory, we might have $m_V \sim 10^{14}$ GeV, and thus $m_{\text{monopole}} \sim 10^{16}$ GeV. A monopole might therefore be a spectacularly heavy elementary particle; $10^{16}\ \text{GeV} \sim 10^{-8}\ \text{g} \sim 10^{6}\ \text{J}$ is comparable to the mass of a bacterium, or the kinetic energy of a charging rhinoceros.

2.4. Examples

In order to gain a deeper understanding of the topological formalism that we have developed, we will now apply this formalism to a number of model gauge theories that contain monopoles [40]. In the process, we will learn much that is interesting about the properties of the monopoles in the various models.

2.4.1. A symmetry-breaking hierarchy

Our first example illustrates the importance in monopole theory of the global structure of the unbroken gauge group. Consider a model with gauge group $G = \mathrm{SU}(3)$ and a scalar field Φ transforming as the adjoint (octet) representation of $G: \Phi$ can be written as a hermitan traceless 3×3 matrix, which, under a gauge transformation $\Omega(x)$, transforms according to

$$\Phi(x) \to \Omega(x)\Phi(x)\Omega^{-1}(x). \tag{2.4.1}$$

Suppose that Φ acquires the expectation value

$$\langle\Phi\rangle = \Phi_0 = (v)\ \mathrm{diag}(\tfrac{1}{2}, \tfrac{1}{2}, -1), \tag{2.4.2}$$

where v is the mass scale of the symmetry breakdown, and the $\mathrm{diag}(\frac{1}{2}, \frac{1}{2}, -1)$ notation denotes a diagonal matrix with the indicated eigenvalues.

The unbroken subgroup H of G, the stability group of Φ_0, is locally isomorphic to $\mathrm{SU}(2)\times\mathrm{U}(1)$. "Locally isomorphic" means that H has the same Lie algebra of infinitesimal generators as $\mathrm{SU}(2)\times\mathrm{U}(1)$. The generators of H are the SU(3) generators that commute with Φ_0. These are the SU(2) generators that mix the two degenerate eigenstates of Φ_0, and also the U(1) generator

$$Q = \mathrm{diag}(\tfrac{1}{2}, \tfrac{1}{2}, -1), \tag{2.4.3}$$

which is proportional to Φ_0, and obviously commutes with it. (The eigenvalues of Q are the U(1) electric charges of the members of the SU(3) triplet, in units of e.)

To perform the topological classification of monopole solutions in this model, we need to determine $\pi_2(G/H) = \pi_1(H)$. So it is not sufficient to know that H has the local structure of the direct product $\mathrm{SU}(2)\times\mathrm{U}(1)$; we must know its global structure. For this purpose, we check to see whether the U(1) subgroup of G generated by Q has any elements in common with the unbroken SU(2) subgroup, other than the identity. And, indeed

$$\exp(\mathrm{i}2\pi Q) = \mathrm{diag}(-1, -1, 1) \tag{2.4.4}$$

is the nontrivial element of the center Z_2 of SU(2). We conclude that

$$H = [\mathrm{SU}(2)\times\mathrm{U}(1)]/\mathrm{Z}_2, \tag{2.4.5}$$

where "=" denotes a global isomorphism; there are two elements of $\mathrm{SU}(2)\times\mathrm{U}(1)$ corresponding to each element of H.

The topologically nontrivial loops in H consist of loops winding around the U(1) subgroup of H, and also of loops traveling through the U(1) subgroup from the identity to the element in eq. (2.4.4) and returning

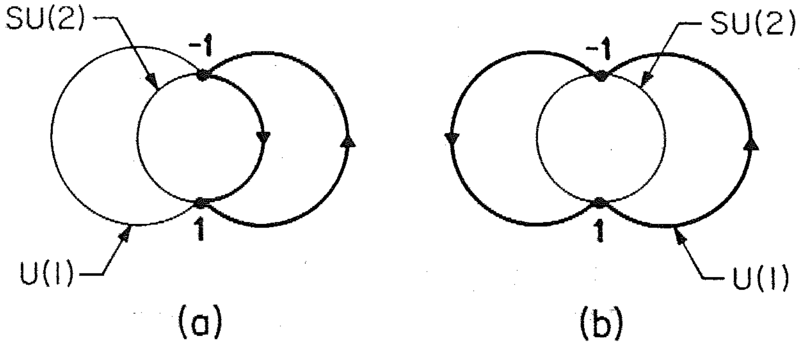


Fig. 17. An (a) minimal and (b) nonminimal loop in $H_1 = [\mathrm{SU}(2)\times \mathrm{U}(1)]/\mathrm{Z}_2$.

to the identity through the SU(2) subgroup of H (fig. 17). If we failed to recognize that H is not globally the direct product $\mathrm{SU}(2)\times\mathrm{U}(1)$, we would have missed the latter set of nontrivial loops, and thus missed half of the monopole solutions in this model.

The monopole with minimal U(1) magnetic charge is associated with a loop that winds only half-way around U(1); it necessarily also has a Z_2 non-Abelian magnetic charge. It is very instructive to examine closely the Dirac quantization condition satisfied by this monopole. By an appropriate gauge choice, the long-range gauge field of the monopole can be chosen to have the form [45, 39]

$$\begin{aligned} A_{\mathrm{U}}\cdot \mathrm{d}\boldsymbol{r} &= \frac{1}{2e} Q_{\mathrm{M}}(1-\cos\theta)\,\mathrm{d}\phi, \quad \text{upper } (0\leq\theta\leq\pi/2),\\ A_{\mathrm{L}}\cdot \mathrm{d}\boldsymbol{r} &= -\frac{1}{2e} Q_{\mathrm{M}}(1+\cos\theta)\,\mathrm{d}\phi, \quad \text{lower } (\pi/2\leq\theta\leq\pi), \end{aligned} \tag{2.4.6}$$

where Q_{M} is a constant generator of H, and e is the gauge coupling. The gauge transformation that relates A_{U} and A_{L} at the equator is

$$\Omega(\phi) = \mathrm{e}^{\mathrm{i}Q_{\mathrm{M}}\phi}. \tag{2.4.7}$$

Since Ω is required to be a single-valued gauge transformation, Q_{M} must have integer eigenvalues. This is the Dirac quantization condition.

The matrix Q_{M} can be chosen to be diagonal; hence, it can be expressed as a linear combination of Q and the SU(2) generator

$$T_3 = \mathrm{diag}(\tfrac{1}{2}, -\tfrac{1}{2}, 0). \tag{2.4.8}$$

Comparing (2.1.12) and (2.4.6), one sees that the U(1) magnetic charge g of the monopole is the coefficient of $2eQ$ in this expression for Q_{M}. Since Q has the eigenvalue $\frac{1}{2}$, one's naive expectation may be that the minimal magnetic charge allowed by the Dirac quantization condition is $g=1/e$. But this expectation is wrong for a monopole that carries both a U(1) magnetic charge and an SU(2) magnetic charge [46]. The choice of Q_M for which $e^{iQ_M\phi}$ is the *minimal* loop in H is not $Q_M=2Q$, but

$$Q_M=Q'\equiv T^3+Q=\mathrm{diag}(1,0,-1); \tag{2.4.9}$$

the associated monopole has U(1) magnetic charge $g=1/2e$.

Equation (2.4.4) implies that objects with trivial SU(2) "duality" have integer U(1) charge Q, although objects with nontrivial duality can have half-integer charge. Thus, the Dirac quantization condition can still be expressed as $n=2eg$, as in section 2.1, but e must now be interpreted as the minimal U(1) charge carried by an SU(2) *singlet* object.

In "realistic" unified gauge theories, spontaneous symmetry breakdown typically occurs at two or more scales differing by many orders of magnitude. To illustrate the effect of such a symmetry-breaking hierarchy on magnetic monopoles, let us imagine that the $G=\mathrm{SU(3)}$ gauge symmetry of our model breaks down in two stages, first to $H_1=[\mathrm{SU(2)}\times \mathrm{U(1)}]/Z_2$ at mass scale v_1, then to $H_2=\mathrm{U(1)}$ at mass scale $v_2 \ll v_1$,

$$G=\mathrm{SU(3)}\overset{v_1}{\rightarrow} H_1=[\mathrm{SU(2)}\times\mathrm{U(1)}]/Z_2\overset{v_2}{\rightarrow} H_2=\mathrm{U(1)}. \tag{2.4.10}$$

The effect of the second stage of symmetry breakdown on the monopoles generated by the first stage depends on which U(1) subgroup of H_1 remains unbroken at the second stage [47].

First, suppose that H_2 is the U(1) subgroup generated by

$$Q_2=Q'=\mathrm{diag}(1,0,-1). \tag{2.4.11}$$

Since this is the same charge as that carried by the monopole associated with the $G\rightarrow H_1$ breakdown at mass scale v_1, the breakdown at the much lower mass scale v_2 has no significant effect on the monopole.

But if H_2 is the U(1) subgroup generated by

$$Q_2=Q=\mathrm{diag}(\tfrac{1}{2},\tfrac{1}{2},-1), \tag{2.4.12}$$

the monopole is significantly affected, for the only monopole solutions now have twice the U(1) magnetic charge allowed by the $G\rightarrow H_1$ breakdown.

What would happen to the minimal G/H_1 monopole if we varied the parameters of the model so as smoothly to turn on the second symmetry-breaking scale v_2? This question is not entirely academic, because the H_1 symmetry is expected to be restored at sufficiently high temperature,

$T \gg v_2$. As the temperature is lowered, a phase transition occurs at $T \sim v_2$ in which H_1 becomes spontaneously broken. We might be interested in what happens to the minimal G/H_1 monopoles during this phase transition, especially since a phase transition like this one may have occurred in the very early universe.

A reasonable guess is that pairs of minimal G/H_1 monopoles or monopole–antimonopole pairs become connected by magnetic flux tubes, and form composite objects with either twice the minimal U(1) magnetic charge or zero magnetic charge. To verify that this guess is correct, we note that the U(1) factor of H_1 is unaffected by the second stage of symmetry breakdown, and that the flux tubes associated with the second stage are classified by

$$\pi_1(\mathrm{SU}(2)/\mathrm{Z}_2)=\mathrm{Z}_2. \tag{2.4.13}$$

Thus, the SU(2) magnetic flux emanating from the minimal G/H_1 monopole does indeed become confined to a Z_2 flux tube. It may be helpful to restate this argument slightly differently: Associated with the G/H_1 monopole is the noncontractible loop in H_1 depicted in fig. 17a. Since this loop cannot be deformed to a loop contained entirely in H_2, there is also an associated H_1/H_2 vortex. But the composition of two such loops *is* homotopic to the loop in H_2 depicted in fig. 17b. (It is equivalent to the composition of a loop in H_2 and a loop in SU(2), and the loop in SU(2) can be shrunk to the identity, because SU(2) is simply connected. See fig. 18.) Therefore, the vortex is a Z_2 vortex, and the Z_2 magnetic flux confined to the vortex is precisely the Z_2 magnetic flux carried by the G/H_1 monopole.

The flux tubes link each G/H_1 monopole with minimal H_1 magnetic charge to either another monopole or an antimonopole, since the monopole and antimonopole carry the same Z_2 charge. The bound pairs of monopoles have the minimal H_2 magnetic charge allowed by the Dirac

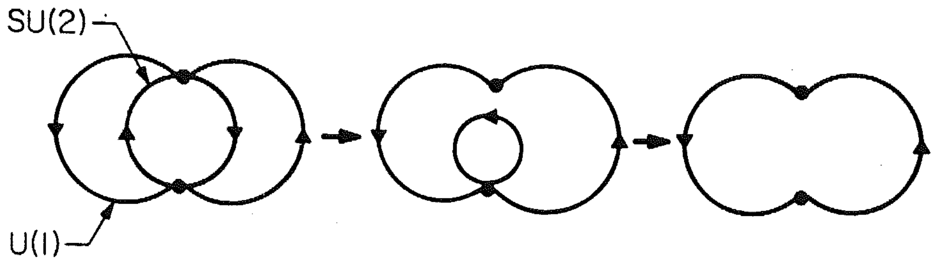


Fig. 18. The composition of two minimal loops in $H_1=[\mathrm{SU}(2)\times \mathrm{U}(1)]/\mathrm{Z}_2$ is homotopic to a loop in $H_2=\mathrm{U}(1)$.

quantization condition. The thickness and energy per unit length of the flux tubes are determined by the lower symmetry-breaking scale v_2; the thickness is of order $(ev_2)^{-1}$, and the energy per unit length is of order v_2^2.

Note that the flux tubes in this model are not absolutely stable; they can break in two via the nucleation of a monopole pair. But this process is a highly unlikely quantum tunneling event if $v_1 \gg v_2$. The barrier that must be penetrated has height of order m and width of order m/μ, where $m \sim v_1/e$ is the monopole mass and $\mu \sim v_2^2$ is the tension in the tube. Thus, the probability per unit length and time of pair nucleation is, in order of magnitude,

$$\Gamma \propto \exp(-m^2/\mu) \sim \exp(-v_1^2/e^2 v_2^2). \tag{2.4.14}$$

This probability is absolutely negligible for $v_1 \gg v_2$.

Finally, suppose that the unbroken U(1) group H_2 is generated by

$$Q_2 = T_3 = \mathrm{diag}(\tfrac{1}{2}, -\tfrac{1}{2}, 0). \tag{2.4.15}$$

In this case H_2 is contained in $\mathrm{SU}(2) \subset H_1$ and the symmetry breakdown $H_2 \to H_1$ can be represented by

$$\begin{tikzcd} H_1 = \mathrm{SU}(2) \arrow[d] & \times & \mathrm{U}(1) \arrow[d] \\ H_2 = \mathrm{U}(1) & & 1 \end{tikzcd} \tag{2.4.16}$$

The flux tubes associated with the breakdown of H_1 are classified by

$$\pi_1[U(1)] = \mathrm{Z}. \tag{2.4.17}$$

These are the Z flux tubes to which the U(1) magnetic flux becomes confined, and therefore no heavy monopoles with mass of order v_1/e can survive when v_2 turns on; all heavy monopoles become bound to antimonopoles by the flux tubes. Since $\pi_2(G/H_2) = Z$, there must still be stable, but light (mass of order v_2/e), monopoles associated with the symmetry breakdown $H_1 \to H_2$.

We see that magnetic monopoles generated at a large symmetry-breaking mass scale may be affected by a small symmetry-breaking mass scale in various ways. The monopoles may survive intact, may become bound by flux tubes into monopole-antimonopole pairs, or may become bound into both monopole-antimonopole pairs and clusters of n monopoles. And, of course, new monopoles might also be generated at the smaller mass scale.

Exercise. Show that for any symmetry breaking hierarchy of the form $G \to H_1 \to H_2$, if a monopole generated by the first stage of symmetry breakdown is unable to survive at the second stage, then a flux tube is generated at the second stage that can end on the monopole. (Use the topological classification of vortices and monopoles in sections 1.3 and 2.2.)

2.4.2. *The* SU(5) *model*

The SU(5) model is a realistic grand unified theory that has many features in common with the simpler model considered above.

The SU(5) model is the simplest gauge theory uniting the $\mathrm{SU}(3)_c$ gauge group of the strong interactions with the $[\mathrm{SU}(2)\times\mathrm{U}(1)]_{ew}$ gauge group of the electroweak interaction. This model undergoes symmetry breakdown at two different mass scales.

$$\begin{aligned} G=\mathrm{SU}(5) &\overset{v_1}{\rightarrow} H_1=\{\mathrm{SU}(3)_c\times[\mathrm{SU}(2)\times\mathrm{U}(1)]_{ew}\}/Z_6 \\ &\overset{v_2}{\rightarrow} H_2=[\mathrm{SU}(3)_c\times\mathrm{U}(1)_{em}]/Z_3. \end{aligned} \tag{2.4.18}$$

Here $v_2 \sim 250$ GeV is the mass scale of the electroweak symmetry breakdown, and $v_1 \sim 10^{15}$ GeV is the mass scale of unification.

The order parameter for the symmetry breakdown at mass scale v_1 is a scalar field Φ transforming as the adjoint representation of G, which acquires the expectation value

$$\langle\Phi\rangle=\Phi_0=v_1\,\mathrm{diag}(\tfrac{1}{3},\tfrac{1}{3},\tfrac{1}{3},-\tfrac{1}{2},-\tfrac{1}{2}). \tag{2.4.19}$$

The stability group H of G is locally isomorphic to $\mathrm{SU}(3)\times\mathrm{SU}(2)\times\mathrm{U}(1)$, where SU(3) acts on the three degenerate eigenvectors of Φ_0/v_1 with eigenvalue $\frac{1}{3}$, and SU(2) acts on the two degenerate eigenvectors with eigenvalue $-\frac{1}{2}$. The unbroken U(1) is generated by

$$Q=\mathrm{diag}(\tfrac{1}{3},\tfrac{1}{3},\tfrac{1}{3},-\tfrac{1}{2},-\tfrac{1}{2}), \tag{2.4.20}$$

and, since

$$\exp(i2\pi Q)=\mathrm{diag}[\exp(i2\pi/3),\exp(i2\pi/3),\exp(i2\pi/3),-1,-1], \tag{2.4.21}$$

we see that this U(1) contains the center of $\mathrm{SU}(3)\times\mathrm{SU}(2)$, so that the unbroken group is actually $H_1=[\mathrm{SU}(3)\times\mathrm{SU}(2)\times\mathrm{U}(1)]/Z_6$.

Equation (2.4.21) ensures that any object with trivial SU(3) triality and SU(2) duality has integer U(1) charge, in units of e. Thus, there exists a magnetic monopole in this model with the Dirac U(1) magnetic charge $g_D = 1/2e$, which also carries a Z_3 color magnetic charge and a Z_2 SU(2) magnetic charge. In an appropriate gauge, we may regard the magnetic charge carried by the monopole to be a U(1)′ charge generated by

$$Q' = Q + Q_{\text{weak}} + Q_{\text{color}} = \text{diag}(0, 0, 1, 0, -1), \tag{2.4.22}$$

where

$$Q_{\text{weak}} = \text{diag}(0, 0, 0, \tfrac{1}{2}, -\tfrac{1}{2}), \tag{2.4.23}$$

is an SU(2) generator and

$$Q_{\text{color}} = \text{diag}(-\tfrac{1}{3}, -\tfrac{1}{3}, \tfrac{2}{3}, 0, 0,), \tag{2.4.24}$$

is an SU(3) generator. Since Q' has integer eigenvalues, a monopole with U(1) magnetic charge $g = g_D = 1/2e$ is consistent with the Dirac quantization condition.

The electroweak symmetry breakdown at mass scale v_2 leaves unbroken the $U(1)_{em}$ subgroup of $[SU(2) \times U(1)]_{ew}$ generated by

$$Q_{em} = Q + Q_{\text{weak}} = \text{diag}(\tfrac{1}{3}, \tfrac{1}{3}, \tfrac{1}{3}, 0, -1). \tag{2.4.25}$$

Since $\exp(i2\pi Q_{em})$ is a nontrivial element of the center of $SU(3)_c$, the unbroken subgroup is $H_2 = [SU(3) \times U(1)]/Z_3$, and the monopole with minimal $U(1)_{em}$ magnetic charge still carries the U(1)′ charge generated by Q'. (Although a quark can carry electric charge 1/3, a monopole with magnetic charge g_D is consistent with the Dirac quantization condition because color-singlet objects carry integer charge.)

The structure of the SU(5) monopole is not much affected by the electroweak symmetry breakdown, because the magnetic charge carried by the monopole is not changed by this breakdown. There are no W and Z fields excited inside an electroweak core with a radius of order $(ev_2)^{-1} \sim M_W^{-1}$, at least in the classical approximation. The true core of the monopole has a radius of order $(ev_1)^{-1} \sim 10^{-28}$ cm and the mass of the monopole is of order $(v_1/e) \sim 10^{16}$ GeV.

That the electroweak $SU(2) \times U(1)$ gauge symmetry is restored within a distance M_W^{-1} of the center of the monopole has some important consequences, though. For one thing, two monopoles with a separation much less than M_W^{-1} may orient their magnetic charges in orthogonal directions in $SU(3) \times SU(2) \times U(1)$, and reduce their Coulomb repulsion to zero. For an appropriate choice of parameters, it is then possible for the attractive force between the monopoles generated by scalar exchange to cause a stable two-monopole bound state to form, with twice the minimal $U(1)_{em}$ magnetic charge [48].

2.4.3. A Z_2 monopole

In the previous examples, we encountered monopoles that carry both a U(1) magnetic charge and a non-Abelian magnetic charge. It is possible, of course, for a monopole to carry a pure non-Abelian charge.

For example, consider a model with gauge group $G=SU(3)$, and a scalar field Φ transforming as the symmetric tensor representation of G. Φ can be written as a symmetric 3×3 matrix, which, under a gauge transformation $\Omega(x)$, transforms according to

$$\Phi(x)\rightarrow\Omega(x)\Phi(x)\Omega^{T}(x). \tag{2.4.26}$$

If Φ acquires the expectation value

$$\langle\Phi\rangle=\Phi_0=v\mathbb{1}, \tag{2.4.27}$$

then G is spontaneously broken to $H=SO(3)$. The monopoles of this model are classified by

$$\pi_2(G/H)=\pi_1[SO(3)]=Z_2. \tag{2.4.28}$$

They are Z_2 monopoles carrying SO(3) magnetic charges. The monopole and antimonopole are indistinguishable.

It is interesting to examine the fate of these monopoles if there is a symmetry-breaking hierarchy of the form

$$G=SU(3)\overset{v_1}{\rightarrow}H_1=SO(3)\overset{v_2}{\rightarrow}H_2=U(1), \tag{2.4.29}$$

where $H_2=U(1)\subset SO(3)$ is generated by

$$Q=\mathrm{diag}(\tfrac{1}{2},-\tfrac{1}{2},0). \tag{2.4.30}$$

There will, of course, be $\pi_2(H_1/H_2)$ monopoles generated by the second stage of symmetry breakdown. These are light monopoles, with core radius of order $(ev_2)^{-1}$ and mass of order v_2/e, defined by topologically nontrivial loops in H_2 that can be contracted to a point in H_1.

But the light monopoles are not all the monopoles of this model; $\pi_2(G/H_2)$ is larger than $\pi_2(H_1/H_2)$, because there are topologically nontrivial loops in H_2 that cannot be contracted to a point in H_1, but are contractible in G. Thus, there are monopoles with half the magnetic charge of the minimal $\pi_1(H_1/H_2)$ monopole that are generated by the first stage of symmetry breakdown. These are heavy monopoles with a core radius of order $(ev_1)^{-1}$ and a mass of order v_1/e. They are just the Z_2 monopoles, which have been converted into Z monopoles with the Dirac magnetic charge by the physics of the second stage of symmetry breakdown. If we turn on v_2 smoothly, the Z_2 monopole, which is equivalent to its antiparticle, must choose the sign of its U(1) magnetic charge at random [49].

The heavy monopole has two cores, and most of its mass resides on its tiny inner core. But if two heavy monopoles are brought together, their inner cores can annihilate, and only the outer cores need survive. So the doubly charged light monopole can be regarded as a very tightly bound composite state of two singly charged heavy monopoles.

2.4.4. *The* SO(10) *model*

The SO(10) model is the next simplest realistic grand unified theory, after the SU(5) model. There are several possible choices for the symmetry-breaking hierarchy of the SO(10) model, and the properties of its monopoles depend on this choice. Rather than enumerate all the possibilities, let us focus on one particularly interesting case.

The group SO(10) is not simply connected, but has the simply connected covering group Spin(10). The 16-dimensional spinor representation of Spin(10) is a double-valued representation of SO(10) = Spin(10)/Z_2. All representations of Spin(10) can be constructed from direct products of 16's.

Let us suppose that the order parameter for the first stage of symmetry breakdown in the SO(10) model is a scalar field Φ that transforms as the 54-dimensional representation of SO(10): Φ can be written as a traceless symmetric 10×10 matrix transforming according to

$$\Phi(x) \to \Omega(x)\Phi(x)\Omega^{\mathrm{T}}(x), \tag{2.4.31}$$

where $\Omega(x) \in \mathrm{SO}(10)$. If Φ acquires the expectation value

$$\langle\Phi\rangle = \Phi_0 = v_1\,\mathrm{diag}(2,2,2,2,2,2,-3,-3,-3,-3), \tag{2.4.32}$$

then the unbroken subgroup H is locally isomorphic to SO(6) × SO(4). This group is, in turn, locally isomorphic to the direct product of SU(4), the covering group of SO(6), and SU(2) × SU(2), the covering group of SO(4).

To determine the global structure of the unbroken group, we check for nontrivial elements of SU(4) × SU(2) × SU(2) that act trivially in Spin(10). Since the fundamental spinor representation of Spin(10) transforms under SU(4) × SU(2) × SU(2) as

$$16 \to (4, 1, 2) + (\bar{4}, 2, 1), \tag{2.4.33}$$

we see that the element $(-\mathbb{1}_4, -\mathbb{1}_2, -\mathbb{1}_2)$ of SU(4) × SU(2) × SU(2) does act trivially on the spinor. Thus, the symmetry-breaking pattern is [50]

$$G = \text{Spin}(10) \xrightarrow{v_1} H_1 = [\text{SU}(4) \times \text{SU}(2) \times \text{SU}(2)]/\text{Z}_2. \tag{2.4.34}$$

The monopoles arising from this symmetry breakdown are Z_2 monopoles carrying SU(4) and SU(2) × SU(2) magnetic charges, classified by $\pi_2(G/H_1) = \pi_1(H_1) = \text{Z}_2$.

Now suppose that, at a lower mass scale v_2, the symmetry breakdown

$$\begin{aligned} H_1 &= [\text{SU}(4) \times \text{SU}(2) \times \text{SU}(2)]/\text{Z}_2 \\ &\xrightarrow{v_2} H_2 = [\text{SU}(3) \times \text{SU}(2) \times \text{U}(1)]/\text{Z}_6 \end{aligned} \tag{2.4.35}$$

occurs. (The order parameter could be a scalar field transforming as the 16-dimensional spinor representation of SO(10).) H_2 is exactly the same as the unbroken gauge group of the SU(5) model, and the monopole with the minimal U(1) magnetic charge in this SO(10) model also carries SU(3) and SU(2) magnetic charges, just like the monopole of the SU(5) model.

But, as in the example of section 2.4.3, the doubly charged monopole in this model is lighter than the monopole with minimal charge [51]. The minimal monopole defines a loop in H_2 that cannot be contracted to a point in H_1, but can be in G. So the core of this monopole has a radius of order $(ev_1)^{-1}$, and its mass is of order (v_1/e). The doubly charged monopole, however, has no SU(2) magnetic charge, and it defines a loop in H_2 that can be contracted to a point in H_1. It arises from the breakdown of H_1 to H_2, and has a core radius of order $(ev_2)^{-1}$ and a mass of order (v_2/e). Neither the minimal monopole nor the doubly charged monopole is much affected by the subsequent breakdown of H_2 to $H_3 = [\text{SU}(3) \times \text{U}(1)]/\text{Z}_3$.

In general, a grand unified theory with a complicated symmetry-breaking hierarchy may possess several stable monopoles with widely disparate masses, the monopole of minimal $\text{U}(1)_{\text{em}}$ charge being the heaviest. The SO(10) model described here is the simplest realistic example illustrating this possibility.

2.4.5. Monopoles and Alice strings

Let us consider again the model discussed in section (1.2). This model undergoes the symmetry breakdown

$$G = \mathrm{SO}(3) \rightarrow H = \mathrm{O}(2). \tag{2.4.36}$$

There are, of course, noncontractible loops in O(2) that can be contracted to a point in SO(3), so this model contains magnetic monopoles.

We noted earlier that the unbroken group O(2) contains a "charge conjugation" operator Ω_0 that flips the sign of the SO(2) generator Q,

$$\Omega_0 Q \Omega_0^{\dagger} = -Q. \tag{2.4.37}$$

Thus, there is a gauge transformation in H that changes the sign of an electric or magnetic charge. Apparently, there is no gauge-invariant way to distinguish a monopole from an antimonopole in this model. (A "hedgehog" is no different from an "antihedgehog", because the order parameter is a "headless" vector in three-dimensional space.) It seems, though, that one can distinguish a pair of monopoles (or antimonopoles) from a monopole–antimonopole pair; the ambiguity afflicts only the sign of the total charge, not the relative charge of two objects.

However, we must recall that, since O(2) is not connected, this model also has a string solution. Furthermore, an object that circles the string becomes gauge transformed by Ω_0. In particular, a monopole that winds once around the string becomes an antimonopole [18].

There is a local criterion for distinguishing between a pair of monopoles (or antimonopoles) and a monopole–antimonopole pair; we can bring the two objects together and see whether they will annihilate or not. But this criterion is not globally well defined if strings are present. Whether they annihilate or not depends on how many times the monopoles wind around the strings before they are brought together.

Magnetic charge is conserved, so the magentic charge lost by a monopole that winds around a string cannot disappear; it must be transferred to the string. If the string is open, the magnetic charge is transmitted to infinity along the string. But if the string is a closed loop, a finite magnetic charge density remains on the string, after it interacts with the monopole.

A cross section of a magnetically charged loop of string is sketched in fig. 19; the order parameter on a large sphere surrounding this loop is in a hedgehog configuration. On each cross section of the string, the order parameter winds through a path contained in a "plane" of G/H. The number of times this plane twists around as the loop of string is

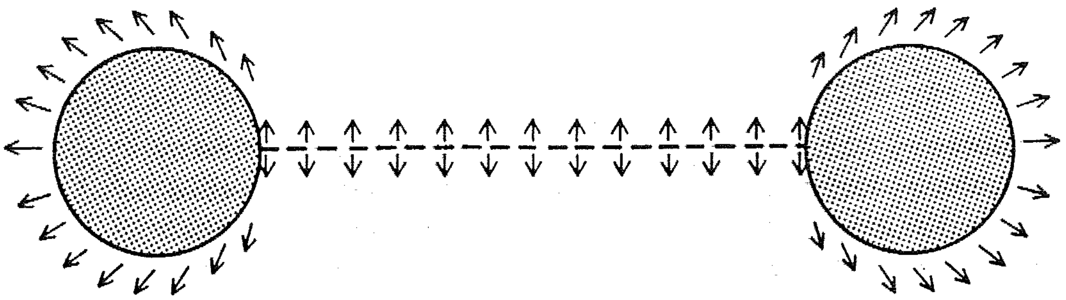


Fig. 19. The cross section of a magnetically charged loop of string.

traversed is a topological invariant of the string, and this topological invariant is the magnetic charge. The loop is a peculiar highly excited monopole, whose core has been distorted into a ring of radius R, and thickness $(ev)^{-1}$.

Exercise. The Pati–Salam model is an extension of the standard model with gauge group $\mathrm{SU}(4)_{\mathrm{color}} \times \mathrm{SU}(2)_{\mathrm{L}} \times \mathrm{SU}(2)_{\mathrm{R}}$ in which a single generation of fermions (plus a right-handed neutrino) transforms as the representation $(4,1,2)_{\mathrm{R}} + (4,2,1)_{\mathrm{L}}$. Describe the monopole of the Pati–Salam model. Specifically, find the monopole "charge" Q_{M}, both in a phase with unbroken gauge symmetry $\mathrm{SU}(3)_{\mathrm{color}} \times \mathrm{SU}(2)_{\mathrm{L}} \times \mathrm{U}(1)_{Y}$ and in a phase with unbroken gauge symmetry $\mathrm{SU}(3)_{\mathrm{color}} \times \mathrm{U}(1)_{\mathrm{em}}$.

2.5. Monopoles in other contexts

Topological considerations very similar to those that arise in the classification of monopoles can also be applied in quite different physical contexts. I will briefly describe two examples here. The first example concerns the topological properties of the phase acquired by the wavefunction of a quantum system under adiabatic transport [52]. The second example concerns a topological obstruction that makes it impossible to introduce spinors on certain manifolds [53].

2.5.1. Berry's phase

Consider a family of Hamiltonian systems parametrized by a set of variables $\boldsymbol{\lambda}$ that take values in a manifold M. And suppose that the Hamiltonian $H(\boldsymbol{\lambda})$ is a smooth function of $\boldsymbol{\lambda}$. If each $H(\boldsymbol{\lambda})$ has purely discrete spectrum and no degeneracies, then each (normalized) eigenfunction $\psi(\boldsymbol{\lambda})$ and its corresponding eigenvalue $E(\boldsymbol{\lambda})$ are also smooth functions of $\boldsymbol{\lambda}$, defined by

$$H(\boldsymbol{\lambda})\psi(\boldsymbol{\lambda}) = E(\boldsymbol{\lambda})\psi(\boldsymbol{\lambda}). \tag{2.5.1}$$

Actually, of course, eq. (2.5.1) leaves the phase of $\psi(\boldsymbol{\lambda})$ undetermined. It is surely possible to define the phase of $\psi(\boldsymbol{\lambda})$ *locally* on the manifold M so that it varies smoothly with $\boldsymbol{\lambda}$. But it is not necessarily possible to define the phase of $\psi(\boldsymbol{\lambda})$ *globally* on M. We would like to determine under what conditions a global definition of the phase *is* possible.

In order to compare the phases of $\psi(\boldsymbol{\lambda})$ at different values of $\boldsymbol{\lambda}$, it is useful to introduce a notion of parallel transport of the phase of ψ on the manifold M. A particularly natural notion of parallel transport is adiabatic transport. Given a path $\boldsymbol{\lambda}(t)$ in M parametrized by $t \in [0, 1]$, we may consider traversing this path infinitesimally slowly, allowing $\psi(\boldsymbol{\lambda}(t))$ to evolve according to the time-dependent Schrödinger equation. Then, since $H(\boldsymbol{\lambda}(t))$ has discrete spectrum and no degeneracies, the quantum adiabatic theorem ensures that the intial wave function $\psi(\boldsymbol{\lambda}(0))$ will evolve into the corresponding eigenstate $\psi(\boldsymbol{\lambda}(t))$, with a phase unambiguously related to that of $\psi(\boldsymbol{\lambda}(0))$. In other words, we may define

$$\psi(\boldsymbol{\lambda}(t)) = \lim_{T\to\infty} \exp\left[-\mathrm{i}\int_0^{Tt} \mathrm{d}s\, H(\boldsymbol{\lambda}(s/T))\right]\psi(\boldsymbol{\lambda}(0)). \tag{2.5.2}$$

The phase of $\psi(\boldsymbol{\lambda}(t))$ has the uninteresting component $\exp[-\mathrm{i}T\int_0^t \mathrm{d}s'\, E(\boldsymbol{\lambda}(s'))]$, which we will remove by making an additive redefinition of $H(\boldsymbol{\lambda})$, so that $E(\boldsymbol{\lambda}) = 0$. But the remaining phase has interesting properties that were first studied by Berry [52].

To discuss this phase, it is convenient to introduce reference wavefunctions $\phi(\boldsymbol{\lambda})$ which are eigenstates of $H(\boldsymbol{\lambda})$ with unit normalization,

$$H(\boldsymbol{\lambda})\phi(\boldsymbol{\lambda}) = 0, \qquad (\phi(\boldsymbol{\lambda}), \phi(\boldsymbol{\lambda})) = 1; \tag{2.5.3}$$

$\phi(\boldsymbol{\lambda})$ has a phase which is chosen arbitrarily but varies smoothly with $\boldsymbol{\lambda}$, at least locally. The phase of $\psi(\boldsymbol{\lambda}(t))$ can then be measured relative to that of $\phi(\boldsymbol{\lambda}(t))$; we may write

$$\psi(\boldsymbol{\lambda}(t)) = U(\boldsymbol{\lambda}(t))^{-1}\phi(\boldsymbol{\lambda}(t)), \tag{2.5.4}$$

where U is a pure phase. Because we have eliminated the uninteresting part of the phase of ψ, ψ changes only when the basis of eigenmodes of H rotates. We therefore have $(\psi, \mathrm{d}\psi/\mathrm{d}t) = 0$; the infinitesimal change in ψ is always orthogonal to ψ. It follows that

$$U^{-1}\,\mathrm{d}U = (\phi, \mathrm{d}\phi), \tag{2.5.5}$$

and U may be expressed as

$$U(\boldsymbol{\lambda}(1))=\exp\left(\mathrm{i}\int_C A\right)U(\boldsymbol{\lambda}(0)). \tag{2.5.6}$$

Here the one-form

$$A=-\mathrm{i}(\phi, \mathrm{d}\phi) \tag{2.5.7}$$

is real, since $(\phi, \phi)=1$, and it is integrated along the curve C defined by $\boldsymbol{\lambda}(t)$.

Both the one-form A and the phase U depend on our choice of reference wavefunctions $\phi(\boldsymbol{\lambda})$. If we make a redefinition of the phase of $\phi(\boldsymbol{\lambda})$ of the form

$$\phi(\boldsymbol{\lambda})\to\Omega(\boldsymbol{\lambda})\phi(\boldsymbol{\lambda}), \tag{2.5.8}$$

then U and A are transformed as

$$U(\boldsymbol{\lambda}(t))\to\Omega(\boldsymbol{\lambda}(t))U(\boldsymbol{\lambda}(t)), \tag{2.5.9}$$

$$A\to A-\mathrm{i}\Omega^{-1}\,\mathrm{d}\Omega. \tag{2.5.10}$$

We see that the freedom to redefine the phase of $\phi(\boldsymbol{\lambda})$ may be regarded as a U(1) gauge freedom, and A behaves just like an Abelian gauge field on the manifold M.

Since the change in the phase of ψ defined by adiabatic transport along an open path in M evidently depends on the coordinate system in which we express the phase, it tells us nothing about the intrinsic "geometry" of the $\psi(\boldsymbol{\lambda})$'s. What has geometrical meaning must be independent of coordinates. Such a quantity is the change in the phase of ψ defined by adiabatic transport around a closed path C, given by

$$\exp\left(\mathrm{i}\oint_C A\right), \tag{2.5.11}$$

which is invariant under the gauge transformation (2.5.8). By Stokes' Theorem, this can also be written as

$$\exp\left(\mathrm{i}\int_S F\right)\equiv\exp\left(\int_S (\mathrm{d}\phi, \mathrm{d}\phi)\right), \tag{2.5.12}$$

where S is a surface in M bounded by the closed path C.

Of course, the adiabatic phase, or "Berry phase" given by eq. (2.5.12) must not depend on the choice of the surface S that is bounded by C. The difference between any two surfaces is a closed surface, and we conclude that

$$\int_S F = 2\pi n, \tag{2.5.13}$$

where S is any closed surface in the manifold M, and n is an integer. Equation (2.5.13) is just the Dirac quantization condition, which we have found to arise in ordinary quantum mechanics in a surprising way.

When will the integer n in eq. (2.5.13) be nonzero? Obviously, it can be nonzero only if the surface S cannot be contracted to a point in M. But the existence of such noncontractible surfaces in M is not untypical. The Hamiltonions in a family $H(\boldsymbol{\lambda})$ generically have degeneracies on surfaces of codimension three. To understand this, one notes that as two energy levels closely approach each other, they are well described as an effective two-level system. A general two-by-two Hermitian matrix is specified by four parameters, while a two-by-two Hermitian matrix with two degenerate eigenvalues is specified by one parameter, so three conditions must be imposed to induce two levels to cross. We thus see that the values of $\boldsymbol{\lambda}$ for which $H(\boldsymbol{\lambda})$ has level crossings generically occur at isolated points in a three-dimensional parameter space, and that a two-dimensional surface in the parameter space may enclose one or more of these degenerate points. But adiabatic transport becomes ill-defined when level crossings occur (the adiabatic theorem breaks down), so we must exclude these degenerate points from our manifold M. Surfaces in the parameter space which enclose points at which level crossing occur are therefore noncontractible surfaces in M.

As an example, consider a spin-$\frac{1}{2}$ particle in a magnetic field, with Hamiltonian

$$H(\boldsymbol{x}) = \boldsymbol{x} \cdot \boldsymbol{\sigma}, \tag{2.5.14}$$

where the σ_i's are the Pauli matrices. If we fix $|\boldsymbol{x}| = 1$, the Hamiltonia of eq. (2.5.14) are parametrized by the points of a two-sphere, each representing a direction in which the magnetic field might point.

The calculation of the integral in eq. (2.5.13) is simplified by the rotational invariance of this problem. In the vicinity of the point $\boldsymbol{x} = (0, 0, 1)$, H may be written as

$$H(\varepsilon) = \begin{pmatrix} 1 & \varepsilon^* \\ \varepsilon & -1 \end{pmatrix}, \tag{2.5.15}$$

where $\varepsilon = x + \mathrm{i}y$, and terms of higher order in ε are dropped. $H(\varepsilon)$ has the eigenstate

$$\phi(\varepsilon)=\begin{pmatrix}1\\ \varepsilon/2\end{pmatrix}, \tag{2.5.16}$$

to order ε, and

$$F=-\mathrm{i}(\mathrm{d}\phi, \mathrm{d}\phi)=-\tfrac{1}{4}\mathrm{i}\, \mathrm{d}\varepsilon^* \wedge \mathrm{d}\varepsilon=\tfrac{1}{2}\,\mathrm{d}x \wedge \mathrm{d}y. \tag{2.5.17}$$

But $\mathrm{d}x \wedge \mathrm{d}y$ is just the area element on the sphere, and we evidently have

$$\int_S F=2\pi. \tag{2.5.18}$$

Adiabatic transport of this two-level system apparently defines a monopole of unit strength on the two-sphere. Note that the two-sphere is noncontractible, because there is a level crossing at $\boldsymbol{x}=0$.

This example has all the essential features of the general case. To integrate F over an arbitrary two-dimensional suface we exploit the fact that F is closed, $\mathrm{d}F=0$, to replace the surface by a set of infinitesimal spheres enclosing points in the parameter space where level crossings occur. Levels generically cross two at a time, so in the vicinity $\boldsymbol{\lambda}_0+\boldsymbol{x}$ of the level crossing at $\boldsymbol{\lambda}_0$, H can be replaced by the general two-level system

$$H(\boldsymbol{x})=(E_0+\boldsymbol{a}\cdot \boldsymbol{x})\Pi+\boldsymbol{\sigma}\cdot \boldsymbol{C}\boldsymbol{x}, \tag{2.5.19}$$

to linear order in $\boldsymbol{x}$. For this system the integral of F over the sphere $|\boldsymbol{x}|=\varepsilon$, which takes discrete values, must be independent of E_0, $\boldsymbol{a}$, and $\boldsymbol{C}$, as long as det $\boldsymbol{C}$ does not cross zero. (For det $\boldsymbol{C}=0$, the integral is ill-defined, because there are level crossings on the sphere.) Thus, the integral depends only on the sign of det $\boldsymbol{C}$, and we see that

$$\frac{1}{2\pi}\int_{\text{sphere}} F=1, \qquad \det \boldsymbol{C}>0,$$

$$\frac{1}{2\pi}\int_{\text{sphere}} F=-1, \quad \det \boldsymbol{C}<0. \tag{2.5.20}$$

(Making the reidentification $\hat{e}_y \to -\hat{e}_y$ changes the sign of both det $\boldsymbol{C}$ and the integral.) Finally, for the general surface S, we sum up the contributions of all the infinitesimal spheres, and obtain

$$\frac{1}{2\pi}\int_S F=\sum_{\substack{\text{level}\\ \text{crossings}}} \operatorname{sign}\det \boldsymbol{C}. \tag{2.5.21}$$

Equation (2.5.21) is Berry's Theorem [52]. It relates the "twist" of the adiabatic phase on a two-dimensional suface in the space of Hamiltonia to the level crossings enclosed by the surface.

This "twist" is a topological property of the "bundle" of wavefunctions $\psi(\boldsymbol{\lambda})$, and has nothing especially to do with adiabatic transport. It is an intrinsic topological property of the bundle that can be revealed by any smooth method of parallel transport. Berry's Theorem tells us that if the charge defined by the right-hand side of eq. (2.5.21) is nonzero for some surface S, then it is not possible for $\psi(\boldsymbol{\lambda})$ to be a smooth function defined globally on S.

Berry's Theorem has some interesting applications. For one thing, it is the basis for an illuminating discussion within the Hamiltonian framework of the origin of gauge anomalies [54]. But I will not go into that here.

2.5.2. Spin structures

Our second example of a "monopole" in an unusual context arises when we consider the problem of introducing spinors on a manifold. Before spinors are introduced on an n-dimensional manifold M, we first equip the manifold with a vielbein, an oriented orthonormal frame that varies smoothly on the manifold, and a connection that defines the notion of parallel transport of the vielbein along a path in M. The vielbein may be identified with the element of $\mathrm{SO}(n)$ that rotates it so that it coincides with a standard frame, and the connection may be regarded as an $\mathrm{SO}(n)$ gauge field.

If we are to introduce fermions, we must be able to associate with each point in M not just an element of $\mathrm{SO}(n)$, but an element of the covering group $\mathrm{Spin}(n)$. Only if this can be done consistently is it possible to introduce a spinor field on M with the crucial property that the field changes sign under a rotation through 2π. It turns out that, on some manifolds, there is a topological obstruction that prevents an $\mathrm{SO}(n)$ bundle from being covered twice by a corresponding $\mathrm{Spin}(n)$ bundle. Such a manifold will not admit spinors; it is said to lack a *spin structure* [53].

This topological obstruction can arise on the manifold M if M contains noncontractible two-spheres. To understand how it might arise, we may consider a sequence of closed loops containing a common point that sweep out one such noncontractible two-sphere; the sequence begins and ends with a trivial loop of vanishing length (fig. 20). With each closed loop, we may associate the element of $\mathrm{SO}(n)$ by which the vielbein is

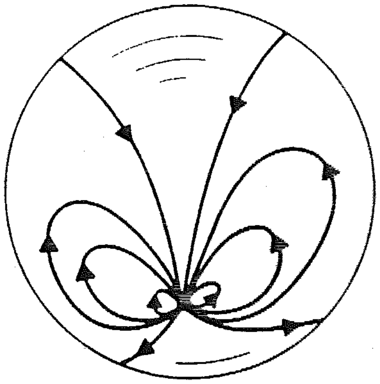


Fig. 20. A sequence of closed loops containing a common point that covers a two-sphere.

rotated under parallel transport around the loop. The two-sphere is a sequence of loops with which we may thus associate a closed path in SO(n). Suppose that this closed path is not contractible in SO(n). Then, when lifted to the covering group Spin(n), it is an open path, running from the element I to the element $-I$ in Spin(n). But this means that a spinor must change sign under parallel transport about an infinitesimal loop; it is impossible to introduce a smooth spinor field on the manifold.

Evidently, the manifold M will lack a spin structure if the loop in SO(n) associated with any noncontractible two-sphere in M is a noncontractible loop in SO(n). Whether the loop is contractible or not depends only on the topology of M. It is independent of the method by which the vielbein is continuously transported, because one method of continuous transport can be smoothly deformed into any other.

The relation of this discussion to the theory of magnetic monopoles is clear. For any noncontractible two-sphere in the manifold M, the loop in SO(n) described above is precisely the loop that classifies the topology of the SO(n) connection on the two-sphere. If this loop is noncontractible, the SO(n) connection is that of a Z_2 monopole.

The observation that, for a manifold without a spin structure, the SO(n) connection on some noncontractible two-sphere is a "monopole" connection suggests a cure for the problem. By introducing a Yang–Mills or Abelian gauge field on this two-sphere, we might arrange for the pathology in the transport of vielbeins on the two-sphere to be canceled by a corresponding pathology in the transport of the other gauge degree of freedom. For example, if we introduce an Abelian monopole field with strength g on the two-sphere, then spinors can be consistently defined on the two-sphere if they carry charges e satisfying the unusual

Dirac quantization condition

$$2eg = n + \tfrac{1}{2}, \qquad (2.5.22)$$

where n is an integer. If a monopole of appropriate strength can be introduced on each noncontractible two-sphere of M for which the transport of vielbeins is pathological, then it is possible to define spinors on M after all. M is then said to be endowed with a "generalized" spin structure [55].

The question whether a manifold M possesses a spin structure can be answered by explicit calculation in the case of a coset space $M = G/H$. There is a noncontractible two-sphere in M associated with each noncontractible loop in H that can be contracted to a point in G; that two-sphere is generated by the mapping that shrinks the loop. What must be checked is whether transport of the vielbein around the noncontractible loop in H is associated with a noncontractible loop in SO(n) [53].

Exercise. Show that the manifold

$$\mathrm{CP}^2 = \mathrm{SU}(3)/\mathrm{U}(2)$$

has no spin structure. (Consider the action of the minimal loop in U(2) on the "broken" SU(3) generators.) For what values of m does $\mathrm{CP}^m = \mathrm{SU}(m+1)/\mathrm{U}(m)$ have a spin structure?

2.6. Global color

In this section we will pursue a question that appears at first to be of merely mathematical interest: can a global gauge transformation be defined in the vicinity of a magnetic monopole? Rather surprisingly, we will find that the answer is no for a monopole with a non-Abelian long-range field, unless the gauge transformation acts trivially on the long-range field [56]. This result actually has some deep physical consequences, as we will better appreciate when we consider in the next section the semiclassical quantization of a classical monopole solution.

In order to address this question, we must define carefully what is meant by first a local gauge transformation, and then a global gauge transformation, in the vicinity of the monopole. We have seen that, to specify the gauge field on a sphere surrounding a monopole, we can split the sphere in half along the equator, and introduce smooth gauge potentials A_{U} and A_{L} on the upper and lower hemispheres. At the equator, the two potentials must be related as in eq. (2.1.10) by a single-valued gauge transformation $\Omega(\phi)$ that takes values in H, the unbroken gauge group. Now, if we wish to define a classical field or wavefunction for some "charged" object that transforms under H, we follow the same procedure again. Smooth fields f_{U} and f_{L} are specified on the two hemispheres that are required to satisfy the "matching condition"

$$f_{\mathrm{U}}(\theta=\pi/2, \phi)=\Omega(\phi)f_{\mathrm{L}}(\theta=\pi/2, \phi), \tag{2.6.1}$$

where $\Omega(\phi)$ is precisely the same gauge tranformation as appears in the matching condition, eq. (2.1.10), for the gauge field. It is necessary that the matching conditions for A and f are the same. The "function" f cannot be smoothly defined globally on the sphere, but its discontinuity at the equator is a mere gauge artifact, or coordinate singularity. (A mathematician would call f a "section" of a "nontrivial bundle".) If one performs a singular gauge tranformation that removes the discontinuity of A at the equator (while introducing a discontunuity somewhere else on the sphere), this gauge transformation must also make f continuous at the equator.

Now, a smooth local gauge transformation of f on the sphere is a pair of gauge transformations $\Omega_{\mathrm{U,L}}$ defined on the two hemispheres that preserves the matching condition (2.6.1):

$$\begin{aligned} &f_{\mathrm{U}}(\theta, \phi)\to\Omega_{\mathrm{U}}(\theta, \phi)f_{\mathrm{U}}(\theta, \phi), \quad \text{upper}\ (0\leqslant\theta\leqslant\pi/2),\\ &f_{\mathrm{L}}(\theta, \phi)\to\Omega_{\mathrm{L}}(\theta, \phi)f_{\mathrm{L}}(\theta, \phi), \quad \text{lower}\ (\pi/2\leqslant\theta\leqslant\pi),\\ &\Omega_{\mathrm{U}}(\theta=\pi/2, \phi)=\Omega(\phi)\Omega_{\mathrm{L}}(\theta=\pi/2, \phi)\Omega^{-1}(\phi). \end{aligned} \tag{2.6.2}$$

This gauge transformation is smooth in the sense that it produces no gauge-artifact singularities. A singular gauge transformation that removes the discontinuity in A along the equator also removes the discontinuity in Ω along the equator.

To define an infinitesimal *global* gauge transformation on the sphere, one must specify a set of generators $\{T^a\}$ of the unbroken gauge group H at each point (θ, ϕ) of the sphere. The statement that the transformation is global means that the commutation relations satisfied by the generators are independent of the position on the sphere. However, this condition does not remove the freedom to perform a local redefinition (depending on θ and ϕ) of the generators of the Lie algebra that preserves the structure constants, of the form

$$T^a(\theta, \phi)=\Sigma(\theta, \phi)T^a\Sigma^{-1}(\theta, \phi), \tag{2.6.3}$$

where $\Sigma\in H$ and $\{T^a\}$ is some standard choice of the generators. The redefinition of the generators determined by Σ is called an inner automorphism of the Lie algebra of H, and all redefinitions that preserve the Lie algebra and can be obtained by composing infinitesimal redefinitions have this form. The group of inner automorphisms is evidently isomorphic to H/K, where K is the center of H, since the elements of K, and only the elements of K, define trivial automorphisms.

A global gauge transformation of f on a sphere, like a local gauge transformation, must be consistent with the matching condition (2.6.1). To specify a global gauge transformation, we define inner automorphisms on the upper and lower hemispheres,

$$\begin{aligned} T^a_{\mathrm{U}}(\theta, \phi) &= \Sigma_{\mathrm{U}}(\theta, \phi) T^a \Sigma_{\mathrm{U}}^{-1}(\theta, \phi), \quad \text{upper } (0 \leq \theta \leq \pi/2), \\ T^a_{\mathrm{L}}(\theta, \phi) &= \Sigma_{\mathrm{L}}(\theta, \phi) T^a \Sigma_{\mathrm{L}}^{-1}(\theta, \phi), \quad \text{lower } (\pi/2 \leq \theta \leq \pi), \end{aligned} \tag{2.6.4}$$

that satisfy a matching condition

$$T^a_{\mathrm{U}}(\theta = \pi/2, \phi) = \Omega(\phi) T^a_{\mathrm{L}}(\theta = \pi/2, \phi) \Omega^{-1}(\phi), \tag{2.6.5}$$

or

$$\begin{aligned} T^a = {} & \Sigma_{\mathrm{U}}^{-1}(\pi/2, \phi) \Omega(\phi) \Sigma_{\mathrm{L}}(\pi/2; \phi) \\ & \times T^a \Sigma_{\mathrm{L}}^{-1}(\pi/2, \phi) \Omega^{-1}(\phi) \Sigma_{\mathrm{U}}(\pi/2, \phi). \end{aligned} \tag{2.6.6}$$

But eq. (2.6.6) says that $\Sigma_{\mathrm{U}}^{-1}(\pi/2, \phi)\Omega(\phi)\Sigma_{\mathrm{L}}(\pi/2, \phi)$ defines a trivial automorphism; it must be an element of the center of H.

Now, suppose that H is semisimple (is a compact Lie group with no U(1) factor). Then the center of H is discrete, and this element Ω_0 of the center must be a constant, independent of ϕ; we have

$$\Omega(\phi) = \Sigma_{\mathrm{U}}(\theta = \pi/2, \phi) \Omega_0 \Sigma_{\mathrm{L}}^{-1}(\theta = \pi/2, \phi). \tag{2.6.7}$$

If we now allow θ, the argument of Σ_{U} (Σ_{L}) to vary smoothly from $\theta = \pi/2$ to $\theta = 0$ ($\theta = \pi$) in eq. (2.6.7), we find that the loop $\Omega(\phi)$ can be continuously deformed to a point; it has winding number zero. We conclude, if H is semisimple, that a global H transformation can be performed on a sphere only if the sphere encloses no magnetic charge. In the vicintiy of a non-Abelian magnetic monopole, a global non-Abelian gauge transformation cannot be implemented [56].

If we try to implement global gauge transformations in some subgroup H' of H, eq. (2.6.6) tells us that we can succeed only if the loop $\Omega(\phi)$ in H can be chosen to commute with H'. In other words, there must be a gauge in which the H' rotations leave the long-range magnetic field of the monopole intact. For a monopole, like the SU(5) monopole, that has an $SU(3)_{\text{color}}$ magnetic field, only the gauge transformations in an $SU(2) \times U(1)$ subgroup of $SU(3)_{\text{color}}$ can be implemented globally.

2.7. *Semiclassical quantization and dyons*

Up to now, we have treated the monopole as a classical object. It is time to consider its quantum mechanical properties. Our analysis of these properties will be carried out in the semiclassical expansion, a systematic expansion in powers of $\hbar$.

To begin with, we consider the SO(3) gauge theory discussed in section 2.3. It is convenient to rescale the fields, so that the Lagrangian can be rewritten

$$L = \frac{1}{e^2}\left[-\tfrac{1}{4}F^a_{\mu\nu}F^{\mu\nu a} + \tfrac{1}{2}D_\mu \Phi^a D^\mu \Phi^a - \tfrac{1}{8}(m_S/m_V)(\Phi^a\Phi^a - m_V^2)^2\right], \tag{2.7.1}$$

where the gauge coupling e has been scaled out of $F_{\mu\nu}$ and $D_\mu\Phi$. The parameter $\hbar^{-1}$, normally set equal to one, multiplies the whole action. Thus, $\hbar$ can be absorbed into e^2, and we see that the semiclassical expansion is an expansion in e^2 with m_V and m_S fixed. In the classical limit $\hbar \to 0$, the size of the monopole core remains fixed while its mass diverges like $\hbar^{-1}$.

The lowest order in the semiclassical expansion is order e^{-2}; the classical core energy and Coulomb energy of the monopole are of this order. To go beyond lowest order, we fix the gauge somehow, and express each field as a sum of a classical background field (the monopole solution, a local minimum of the Hamiltonian) and a fluctuating quantum field; then we expand the Hamiltonian in powers of the quantum fields [39]. The first quantum corrections, of order e^0, are obtained by expanding to quadratic order in the quantum fields; this is equivalent to doing free field theory in the classical monopole background, and the eigenstates of the quadratic Hamiltonian can be interpreted as meson states in the vicinity of the monopole. The energies of these are the frequencies of small oscillations about the monopole solution. These frequencies can depend on m_V and m_S, but must be independent of e, which can be scaled out of the classical action. In higher order in e, the quantum fields interact, and the analysis becomes more complicated.

Another complication occurs if, in the expansion to quadratic order in the quantum fields, there are zero-frequency modes. Such zero-frequency modes should be anticipated if there are unbroken exact symmetries that act nontrivially on the classical monopole solution. Then the time-independent monopole solutions form a degenerate set, and the zero-frequency modes are infinitesimal displacements in the manifold of degenerate solutions.

For example, the monopole solution is not translationally invariant; therefore, it has translational zero modes. The translational modes are easily quantized. To obtain eigenstates of the Hamiltonian, we construct states that transform as irreducible representations of the translation group; these are plane waves carrying a momentum $\boldsymbol{p}$. For fixed $\boldsymbol{p}$, the energy of a monopole plane wave is $\mathrm{O}(e^2)$, because the monopole mass m is $\mathrm{O}(e^{-2})$:

$$E_p=\sqrt{m^2+p^2}=m+p^2/2m+\cdots=m+\mathrm{O}(e^2). \tag{2.7.2}$$

If the classical monopole solution were not rotationally invariant, it would have a moment of inertia of order e^{-2}, and rotational excitations with energy of order e^2. But, because the monopole solution is rotationally invariant, there are no such rotational excitations. (More precisely, the monopole is invariant under a spatial rotation combined with a global SO(3) gauge transformation, which is enough to ensure the absence of rotational excitations.)

A soliton can have zero-frequency modes associated with internal symmetries as well as space–time symmetries. In fact, the monopole of the SO(3) gauge theory is not invariant under a global $\mathrm{U}(1)_{\mathrm{em}}$ charge rotation, because the charged fields $W^{\pm}$ are excited inside the monopole core. (We can see that $W^{\pm}$ *must* be excited by an argument closely analogous to that used at the end of section 1.5 to show that charged fields are excited inside an Alice string.) To quantize the charge rotation degree of freedom, we diagonalize the Hamiltonian by constructing irreducible representations of $\mathrm{U}(1)_{\mathrm{em}}$; that is, states with definite electric charge Q. Thus, the quantum mechanical excitations of the fundamental monopole include *dyons*, particles that carry both magnetic and electric charge [57]. These dyons arise automatically upon the semiclassical quantization of the global charge rotation degree of freedom of the monopole.

Before describing the computation of the dyon spectrum, we should pause to reexamine the claim that the classical monopole solutions form a degenerate set related by $\mathrm{U}(1)_{\mathrm{em}}$ charge rotations. This claim sounds suspicious, because a $\mathrm{U}(1)_{\mathrm{em}}$ rotation is a gauge transformation. If we carry out canonical quantization in the gauge $A_0=0$, we ordinarily say that the physical states must be invariant under time-independent gauge transformations. But in fact we must distinguish between local gauge transformations, with finite support, and global gauge transformations, which act nontrivially at $r \to \infty$. The Gauss' law constraint requires physical states to be invariant under local gauge transformations, but under a global gauge rotation by an angle ε, a physical state with charge Q acquires the phase $e^{i\varepsilon Q}$. Therefore, a classical monopole solution can sensibly be regarded as a superposition of physical states of definite electric charge, and two monopole solutions related by a global charge rotation are distinct states in the physical Hilbert space, degenerate at the classical level.

There is another, less formal, way of explaining why a monopole carries a $U(1)_{em}$ collective coordinate [58]. The configuration space for the classical dynamics of a gauge theory is the space of field configurations modulo gauge transformations; therefore two monopole configurations related by a global $U(1)_{em}$ rotation ought not to be considered as separate objects. Suppose we consider, though , not a single monopole, but a static monopole–antimonopole pair. This static configuration is an approximate classical solution if the monopole and antimonopole are very heavy and widely separated. What collective coordinates are needed to characterize the space of monopole–antimonopole "solutions"? We can construct such a solution by patching together a monopole solution and an antimonopole solution. Both the scalar field (order parameter) and gauge field must match where we do the patching. But if we perform a *relative* gauge rotation of the monopole and antimonopole by an element of the unbroken gauge group H, this rotation does not disturb the scalar field far from the poles, and the fields will still match *provided that the gauge rotation acts trivially on the long-range gauge field of the monopole.* Given one monopole–antimonopole pair, we can cut it in two, rotate the monopole relative to the antimonopole, and glue it together again, thus obtaining a new pair that is not merely a gauge transformation of the original pair. Therefore, a relative global H rotation *is* a proper collective coordinate of a monopole–antimonopole pair. Furthermore, if the monopole has a non-Abelian long-range field, the only H rotations that are sensible in this connection are those that leave the long-range field undisturbed. And, as we saw in the last section, these are precisely the global gauge rotations of an isolated monopole that can be implemented.

We can construct a spectrum of excitations of the monopole–antimonopole pair by projecting out states that transform as irreducible representations of the group of global gauge transformations that serve as collective coordinates for the pair. Since the excitation energy is localized near the pole, we may as well forget about the antimonopole and construct the excitations of an isolated monopole.

Having agreed that a charge rotation is a sensible collective coordinate, we must now, in order to find the energies of the dyon excitations, compute the moment of inertia associated with such a rotation. This computation involves subtleties associated with gauge invariance that must be dealt with carefully [59-61], but let us at first ignore these subtleties, to get a feel for how the computation works.

For purposes of illustration, consider a field theory involving a real scalar field ϕ that has a static soliton solution $\phi = \phi_0(x)$, and suppose that there is a compact "isospin" symmetry of the theory that acts nontrivially on the soliton, so that there is a compact manifold of degenerate soliton solutions. We wish to quantize the "motion" of the soliton on this compact manifold. An infinitesimal motion on this manifold has the form

$$\phi(x, t) = (1 + \varepsilon_a(t) T^a)\phi_0(x), \tag{2.7.3}$$

where the T^a's are (antihermitian) generators of the isospin symmetry group. Since $\phi(x, t)$ is the soliton solution at each fixed t, only the kinetic terms in the action have a nontrivial dependence on the trajectories $\phi(x, t)$. If the kinetic term is the conventional one for a real scalar field, the Lagrangian, after a suitable field rescaling, is

$$L = \int \mathrm{d}^3x \frac{1}{2e^2}(\partial_0\phi)^2 = \tfrac{1}{2} I^{ab}\dot{\varepsilon}_a\dot{\varepsilon}_b, \quad I^{ab} = \frac{1}{e^2}\int \mathrm{d}^3x\,(T^a\phi_0)(T^b\phi_0). \tag{2.7.4}$$

This is the effective Lagrangian that describes the dynamics of the isospin rotator degree of freedom of the soliton. It is the Lagrangian of a rigid body in isospin space.

To canonically quantize, we construct the Hamiltonian

$$H = \tfrac{1}{2} p^a (I^{-1})_{ab} p^b, \quad p^a = I^{ab}\dot{\varepsilon}_b. \tag{2.7.5}$$

Since the isospin group is compact, the ε's are periodic variables, and the p's have discrete eigenvalues. H can be expressed in terms of the Casimirs of the isospin group.

In the case of the monopole of the SO(3) model with unbroken group $\mathrm{U}(1)_{\mathrm{em}}$, the Hamiltonian is

$$H = \frac{1}{2I} Q^2, \tag{2.7.6}$$

where Q is the electric charge operator (in units of e); this is the Hamiltonian of a planar rotor. The group $U(1)_{\mathrm{em}}$ is compact and the eigenvalues of Q are integers. (Q is actually the generator $\mathrm{diag}(\frac{1}{2}, -\frac{1}{2})$ of SU(2), but dyons with half-odd integer Q do not occur, because the monopole solution is invariant under the center of SU(2); thus, $\exp(\mathrm{i}2\pi Q) = 1$ acting on the monopole.) Applying dimensional analysis to the rescaled Lagrangian (eq. (2.7.1)), we see that I is of order $1/e^2 m_{\mathrm{V}}$. Thus, the dyon excitations are split from the monopole ground state by an amount of order $(eQ)^2 m_{\mathrm{V}}$, the Coulomb self-energy of an electric charge eQ localized on the monopole core of radius m_{V}^{-1}.

For a monopole with a non-Abelian long-range field, the gauge rotations in the subgroup H' that can be globally implemented rotate the core of the monopole without disturbing the long-range field, and the associated semiclassical excitations are charge excitations localized on the core. There is no spectrum of dyon excitations associated with the rotations in the gauge group H that act nontrivially on the long-range field because these excitations cannot be supported by the monopole core. They are carried out to large distances by the non-Abelian magnetic field, and are lost in the gluon continuum. They do appear explicitly, however, in the excitation spectrum of a widely separated monopole-antimonopole pair, with energy splittings inversely proportional to the separation of the pair [58].

Having now understood the basic procedure for quantizing the collective coordinates of a soliton, let us consider more carefully how the inertia tensor is computed in a gauge theory. We will work in the temporal gauge $A_0 = 0$, and suppose for definiteness that the scalar field Φ is in the adjoint representation of the gauge group. The static classical monopole solutions form a degenerate set; we denote one representative of this set by

$$\boldsymbol{A} = \boldsymbol{A}_{\mathrm{mon}}(\boldsymbol{x}), \qquad \Phi = \Phi_{\mathrm{mon}}(\boldsymbol{x}). \tag{2.7.7}$$

The motions in the manifold of degenerate monopole configurations that we want to quantize are (time-dependent) global gauge transformations of the form

$$\begin{aligned} \boldsymbol{A}(\boldsymbol{x}, t) &= \Omega^{-1}(\boldsymbol{x}, t)\boldsymbol{A}_{\mathrm{mon}}(\boldsymbol{x})\Omega(\boldsymbol{x}, t) - \mathrm{i}\Omega^{-1}(\boldsymbol{x}, t)\boldsymbol{\nabla}\Omega(\boldsymbol{x}, t), \\ A_0(\boldsymbol{x}, t) &= 0, \qquad \Phi(\boldsymbol{x}, t) = \Omega^{-1}(\boldsymbol{x}, t)\Phi_{\mathrm{mon}}(\boldsymbol{x})\Omega(\boldsymbol{x}, t). \end{aligned} \tag{2.7.8}$$

The statement that $\Omega(\boldsymbol{x}, t)$ is a "global" gauge transformation does not mean that Ω is independent of $\boldsymbol{x}$; it merely means that Ω acts nontrivially at spatial infinity, that

$$\lim_{r\to\infty} \Omega(x, t) = \Omega(t) \tag{2.7.9}$$

is a nontrivial function of t.

What makes the semiclassical quantization of solitons subtle in a gauge theory is that we must consider only those motions on the manifold of soliton solutions that preserve the physical subspace of the Hilbert space. In other words, the motion eq. (2.7.8) is required to be consistent with the Gauss' law constraint. This condition determines the function $\Omega(x, t)$, given its asymptotic behavior, eq. (2.7.9) [60, 61].

For the purpose of finding $\Omega(x, t)$, and of calculating the effective Lagrangian for the global gauge rotations, it is quite convenient to observe that the motion eq. (2.7.8) is gauge equivalent to

$$\boldsymbol{A}(x, t) = \boldsymbol{A}_{\text{mon}}(x), \qquad A_0(x, t) = \mathrm{i}\left(\frac{\partial}{\partial t}\,\Omega(x, t)\right)\Omega^{-1}(x, t),$$

$$\Phi(x, t) = \Phi_{\text{mon}}(x). \tag{2.7.10}$$

(Equation (2.7.8) is *not* a gauge transformation of the static monopole solution, if Ω depends on t.) Now all the time dependence is in A_0, and an effective Lagrangian for the motion is found by plugging into eq. (2.7.1); we obtain

$$L = \frac{1}{e^2}\int \mathrm{d}^3x\, \mathrm{tr}\Big[(\boldsymbol{D}_{\text{mon}}A_0)^2 - [A_0, \Phi]^2\Big] + \cdots, \tag{2.7.11}$$

the remaining terms being time independent.

The Gauss' law constraint may be written

$$D^iF_{0i} = -\boldsymbol{D}^2_{\text{mon}}A_0 = J_0 = -[\Phi, [\Phi, A_0]]. \tag{2.7.12}$$

This equation tells us how to extend the function $\Omega(x, t)$ from its value $\Omega(t)$ at spatial infinity down into the core of the monopole. Furthermore, integrating eq. (2.7.11) by parts, and invoking eq. (2.7.12), our effective Lagrangian may be rewritten as a surface integral,

$$L = \frac{1}{e^2}\int_{r=\infty} \mathrm{d}\Omega\, r^2\, \mathrm{tr}\left(A_0 \frac{\partial}{\partial r} A_0\right). \tag{2.7.13}$$

(To obtain eq. (2.7.13), we have used the fact that the long-range tail of the monopole vector potential satisfies $\hat{\boldsymbol{r}}\cdot \boldsymbol{A}_{\text{mon}} = 0$.)

The asymptotic behavior of A_0 at spatial infinity is

$$A_0^a(x,t) \xrightarrow[r\to\infty]{} 2\mathrm{i}\,\mathrm{tr}\left[\lambda^a\left(\frac{\mathrm{d}}{\mathrm{d}t}\Omega(t)\right)\Omega^{-1}(t)\right] = \omega^a, \tag{2.7.14}$$

where λ^a is a generator of the gauge group, and ω^a is the corresponding "angular velocity". Corrections to this asymptotic form can be expanded in powers of $1/r$; the leading correction in $1/r$ is needed to compute eq. (2.7.13). Since eq. (2.7.12) for A_0 is linear, its solution is linear in ω^a, and must have the form

$$A_0^a = \left(\delta^{ab} - \frac{c^{ab}}{r} + \cdots\right)\omega_b. \tag{2.7.15}$$

Finally, the effective Lagrangian takes the form [60, 61]

$$L = \tfrac{1}{2} I^{ab}\omega_a\omega_b, \quad I^{ab} = \frac{4\pi}{e^2} c^{ab}; \tag{2.7.16}$$

the inertia tensor I^{ab} may be determined by solving the Gauss' law condition eq. (2.7.12). The quantity c^{ab} has the dimension of length, and will turn out to be a tensor of order one multiplying m_V^{-1}, the size of the monopole core.

As we have seen, the gauge rotations of the monopole are restricted to the subgroup H' of the unbroken gauge group that is globally implementable. The Hamiltonian can be expressed in terms of the Casimirs of H', and can be diagonalized by constructing states in irreducible representations of H'. I will not explicitly calculate the Hamiltonian here; we now know in principle how it can be done. But it is helpful to consider the qualitative features of the dyon spectrum for a few examples.

2.7.1. SU(3) *monopole*

This is the example section 2.4.1, with the pattern of symmetry breakdown

$$G = \mathrm{SU}(3) \to H = [\mathrm{SU}(2)\times\mathrm{U}(1)]/\mathrm{Z}_2. \tag{2.7.17}$$

The minimal monopole has magnetic charge

$$Q_\mathrm{M} = \mathrm{diag}(1, 0, -1). \tag{2.7.18}$$

H' is the subgroup of H that commutes with Q_M; it is

$$H' = \mathrm{U}(1)\times\mathrm{U}(1). \tag{2.7.19}$$

This monopole has a long-range SU(2) gauge field, and only a U(1) subgroup of SU(2) leaves the long-range field intact.

One might naively expect the monopole to have two collective coordinates, but in fact only one of the two U(1)'s in H' acts nontrivially on the monopole. The minimal monopole of this SU(3) model is just the 't Hooft–Polyakov SO(3) monopole embedded in an SU(2) subgroup of SU(3), and the U(1) that commutes with this SU(2) subgroup leaves the monopole solution invariant. Therefore, the only charge carried by the dyons is the U(1) charge generated by

$$Q' = \mathrm{diag}(1, 0, -1), \tag{2.7.20}$$

and the dyon excitation energies have the form

$$E_{\mathrm{dyon}} = a e^2 m_{\mathrm{V}} Q'^2, \tag{2.7.21}$$

where Q' is an even integer and a is a numerical constant of order one. The first dyon excitation has the H' quantum numbers of the heavy W-boson. In particular, under the U(1) generated by

$$Q = \mathrm{diag}(\tfrac{1}{2}, \tfrac{1}{2}, -1), \tag{2.7.22}$$

it carries charge $\frac{3}{2}$. All dyons have Q an integer multiple of $\frac{3}{2}$ because $\exp(\frac{4}{3}\pi \mathrm{i} Q)$ is an element of the center of SU(3), which leaves the monopole solution invariant. Each dyon is constructed as a coherent superposition of classical solutions with different charge orientations, and hence $\exp(\frac{4}{3}\pi \mathrm{i} Q)$ must be 1 acting on any dyon state.

2.7.2. Nonminimal monopole

Consider, in the same model as before, the monopole with magnetic charge

$$Q_{\mathrm{M}} = \mathrm{diag}(1, 1, -2). \tag{2.7.23}$$

This is a "nonminimal monopole" associated with a loop in G/H that is homotopic to the composition of two minimal loops. Supposing that this monopole is stable against decay into two minimal monopoles, what are its dyon excitations?

Now the long-range field of the monopole is preserved by the unbroken SU(2), and $H' = H$. The dyon excitations can be assembled into irreducible representations of H, and have excitation energies [61]

$$E_{\mathrm{dyon}} = e^2 m_{\mathrm{V}} (aI(I+1) + bQ^2), \tag{2.7.24}$$

where a and b are numbers of order one. By the same reasoning as above, Q is an integer multiple of 3/2, and $I + Q$ must be an integer if the multiplet is to provide a single-valued representation of [SU(2) ×

$\mathrm{U}(1)]/\mathrm{Z}_2$. Of course, the dyon spectrum of this nonminimal monopole is qualitatively quite distinct from the dyon spectrum of the minimal monopole. For one thing, the excitations in the nonminimal case are $(2I+1)$-fold degenerate, while all excitations in the minimal case are nondegenerate.

2.7.3. SU(5) *model*

This is the example of section 2.4.2, which closely resembles the example of section 2.4.1. The pattern of symmetry breakdown is

$$G=\mathrm{SU}(5)\to H=[\mathrm{SU}(3)_{\mathrm{c}}\times \mathrm{U}(1)_{\mathrm{em}}]/\mathrm{Z}_3, \tag{2.7.25}$$

with $\mathrm{U}(1)_{\mathrm{em}}$ generated by

$$Q_{\mathrm{em}}=\mathrm{diag}(\tfrac{1}{3},\tfrac{1}{3},\tfrac{1}{3},0,-2). \tag{2.7.26}$$

The minimal monopole carries the magnetic charge

$$Q_{\mathrm{M}}=Q'=\mathrm{diag}(0,0,1,0,-1). \tag{2.7.27}$$

The subgroup of H that commutes with Q_{M} is

$$H'=\mathrm{SU}(2)\times\mathrm{U}(1)\times\mathrm{U}(1), \tag{2.7.28}$$

but just as in the previous example, only a single $\mathrm{U}(1)\in H'$, that generated by Q', acts nontrivially on the monopole. The dyon spectrum is again given by eq. (2.7.21). Now $\exp(\frac{3}{2}\pi \mathrm{i} Q_{\mathrm{em}})$ leaves the monopole solution invariant, and all dyons therefore carry electric charge Q_{em} that is an integer multiple of $\frac{4}{3}$.

2.8. Catalysis

We have seen that dyons emerge when we carry out the semiclassical quantization of a classical monopole solution. The existence of this tower of dyonic excitations of the monopole has very important consequences when we consider the scattering of fermions off monopoles. Indeed, we will see that monopole-fermion scattering has a truly spectacular property. When a monopole and a fermion collide at an energy much less than the inverse size m_{V} of the monopole core, the outcome is strongly dependent on the structure of the core. In particular, in a typical grand unified theory, there are heavy gauge bosons with masses of order m_{V} and couplings that violate conservation of baryon number; in such a theory the cross section for baryon number changing scattering of a fermion off a monopole at low energy is large, and independent of m_V [62, 63].

This result seems to violate a cherished principle of quantum field theory, the decoupling principle, which asserts that the effects of very-short-distance physics must be suppressed at low energy by a power of the short-distance scale. In this respect, monopole–fermion scattering appears to be a unique phenomenon.

We can begin to understand some of the peculiar features of monopole-fermion scattering by considering the classical motion of a charged particle with electric charge e and mass m in the background of a magnetic monopole with magnetic charge g. If the monopole is fixed at the origin, and $\boldsymbol{r}$ is the position of the charged particle, then the classical equation of motion is

$$m\ddot{\boldsymbol{r}} = eg\frac{\dot{\boldsymbol{r}}\times\hat{\boldsymbol{r}}}{r^2}. \tag{2.8.1}$$

Using the equation of motion, one easily verifies that the quantity

$$\boldsymbol{J} = m\boldsymbol{r}\times\dot{\boldsymbol{r}} - eg\hat{\boldsymbol{r}} \tag{2.8.2}$$

is a constant of the motion. $\boldsymbol{J}$ is just the usual angular momentum, except for the peculiar extra term $-eg\hat{\boldsymbol{r}}$, which can be interpreted as the angular momentum stored in the electromagnetic field.

Since the "usual" angular momentum is perpendicular to $\boldsymbol{r}$, we have

$$\boldsymbol{J}\cdot\hat{\boldsymbol{r}} = -eg. \tag{2.8.3}$$

From the conservation of angular momentum $\boldsymbol{J}$ we conclude that the trajectory of the charged particle is confined, not to a plane as in a typical central force problem, but to a cone with its apex at the monopole and opening half angle θ such that $\cos\theta = |eg|/J$. The magnitude v of the velocity $\dot{\boldsymbol{r}}$ of the monopole is also a constant of the motion, because the magnetic force is always perpendicular to $\dot{\boldsymbol{r}}$. The square of $\boldsymbol{J}$ can be written as

$$J^2 = m^2v^2b^2 + e^2g^2, \tag{2.8.4}$$

in terms of the "instantaneous" impact parameter b. We conclude that b is a constant of the motion too; in particular, the initial impact parameter is the same as the distance of closest approach of the charged particle to the monopole.

In the limit of small b, J approaches $|eg|$, and the cone on which the trajectory lies becomes very narrow; the scattering angle approaches π, and the trajectory winds many times around the axis of the cone [64]. But if b is *exactly* zero, then the charged particle experiences no force at all, and the scattering angle is zero. The particle passes through $\boldsymbol{r}=0$, where $\boldsymbol{J}$ is ill-defined, and conservation of angular momentum breaks down. Thus, the limit of zero impact parameter is very singular. One might expect to see a remnant of this odd classical behavior in the quantum theory; if so, it seems that only the lowest partial wave is likely to be afflicted.

In order to perform a partial wave analysis of monopole-fermion scattering, we must find the eigenstates of the operators $\boldsymbol{J}^2$ and J_z, where $\boldsymbol{J}$ is the operator

$$\boldsymbol{J}=\boldsymbol{r}\times(\boldsymbol{p}-e\boldsymbol{A})-eg\hat{\boldsymbol{r}}, \tag{2.8.5}$$

and $\boldsymbol{A}$ is the vector potential of the monopole. This task is delicate, because $\boldsymbol{A}$ cannot be expressed as a nonsingular function. It is convenient to adopt the strategy of sections 2.1 and 2.5 and introduce potentials $\boldsymbol{A}_{\mathrm{U}}$ and $\boldsymbol{A}_{\mathrm{L}}$ defined on the upper and lower hemispheres that obey a nontrivial matching condition. With $\boldsymbol{A}$ given by eq. (2.1.12), one finds [37]

$$\begin{aligned} J_z^{\mathrm{U}} &= -\mathrm{i}\frac{\partial}{\partial\phi}-q, \quad \text{upper } (0\leq\theta\leq\pi/2),\\ J_z^{\mathrm{L}} &= -\mathrm{i}\frac{\partial}{\partial\phi}+q, \quad \text{lower } (\pi/2\leq\theta\leq\pi), \end{aligned} \tag{2.8.6}$$

where the notation $q=eg$ has been introduced. A wave"function" that is consistent with the matching condition and is an eigenstate of J_z with eigenvalue m takes the form

$$Y^q_{j,m}=f^q_{j,m}(\theta)\begin{cases}\mathrm{e}^{\mathrm{i}(m+q)\phi}, & \text{upper } (0\leq\theta\leq\pi/2),\\ \mathrm{e}^{\mathrm{i}(m-q)\phi}, & \text{lower } (\pi/2\leq\theta\leq\pi).\end{cases} \tag{2.8.7}$$

Here $f^q_{j,m}(\theta)$ is a nonsingular function on the sphere, to be determined by the requirement that $Y^q_{j,m}$ is an eigenstate of $\boldsymbol{J}^2$ with eigenvalue $j(j+1)$.

The Dirac quantization condition requires that q is a half-integer, and single-valuedness on each hemisphere requires that $m-q$ is an integer. Since the components of $\boldsymbol{J}$ obey the usual angular momentum algebra (check this!), we know that, for given j, m takes the values $j, j-1, \dots, -j$. Thus, $j-q$ is an integer, and we conclude that the angular momentum j of the "monopole harmonic" $Y^q_{j,m}$ is an integer or half-odd-integer depending on whether q is an integer or half-odd-integer. Moreover, because

$$J^2=[\boldsymbol{r}\times(\boldsymbol{p}-e\boldsymbol{A})]^2+q^2 \tag{2.8.8}$$

is the sum of two positive pieces, it is manifest that $j(j+1)>q^2$. Indeed, by solving the eigenvalue problem for $f^q_{j,m}$, one finds that j can take the values [37]

$$j=|q|,|q|+1,|q|+2,\ldots. \tag{2.8.9}$$

The quantum mechanics of a charged spin-0 boson interacting with a magnetic monopole turns out not to be very exciting; a centrifugal barrier prevents the charged particle from penetrating to the magnetic pole [39]. Much more interesting is the case of a spin-$\frac{1}{2}$ fermion. For a spin-$\frac{1}{2}$ fermion, the angular momentum becomes

$$\boldsymbol{J}=\boldsymbol{r}\times(\boldsymbol{p}-e\boldsymbol{A})-q\hat{\boldsymbol{r}}+\tfrac{1}{2}\boldsymbol{\sigma}, \tag{2.8.10}$$

and the eigenstates of $\boldsymbol{J}^2$ and J_z are easily constructed by addition of angular momentum. For example, if $q=\frac{1}{2}$, the $j=0$ angular wavefunction is

$$\eta^{j=0}=\frac{1}{\sqrt{2}}\begin{pmatrix} Y^{1/2}_{1/2,-1/2} \\ -Y^{1/2}_{1/2,1/2}\end{pmatrix}, \tag{2.8.11}$$

where the Y's are monopole harmonics.

The Dirac Hamiltonian for a massless spin-$\frac{1}{2}$ fermion in the field of the monopole is

$$H=\boldsymbol{\alpha}\cdot(-\mathrm{i}\boldsymbol{\nabla}-e\boldsymbol{A}). \tag{2.8.12}$$

If we use the representation

$$\boldsymbol{\alpha}=\begin{pmatrix}0 & \boldsymbol{\sigma}\\ \boldsymbol{\sigma} & 0\end{pmatrix},\qquad \beta=\begin{pmatrix}1 & 0\\ 0 & -1\end{pmatrix},\qquad \gamma_5=\begin{pmatrix}0 & 1\\ 1 & 0\end{pmatrix}, \tag{2.8.13}$$

then the eigenstates of H in the lowest partial wave $j=|q|-\frac{1}{2}$ have the form

$$\psi(r,\theta,\phi)=\frac{1}{r}\begin{pmatrix}\chi_1(r)\eta^j_1(\theta,\phi)\\ \chi_1(r)\eta^j_2(\theta,\phi)\\ \chi_2(r)\eta^j_1(\theta,\phi)\\ \chi_2(r)\eta^j_2(\theta,\phi)\end{pmatrix}, \tag{2.8.14}$$

and the Dirac equation $H\psi=E\psi$ reduces to the two-component radial equation [65]

$$H\chi(r) = -\frac{\mathrm{i}q}{|q|}\gamma_5 \frac{\mathrm{d}}{\mathrm{d}r}\chi(r) = E\chi(r). \tag{2.8.15}$$

Remarkably, the monopole vector potential has disappeared from the problem; the radial equation describes a *free* (1+1)-dimensional spinor propagating radially.

The solutions to eq. (2.8.15) can be chosen to be eigenstates of γ_5, and the eigenvalues ± 1 of γ_5 can be identified with the (3+1)-dimensional helicity of the fermion. Strangely, the helicity of a solution is correlated with whether it is an incoming or outgoing wave. If $q > 0$, we have

$$\begin{aligned} &\gamma_5 = +1, \quad \chi(r) \propto \mathrm{e}^{\mathrm{i}Er}, \quad \text{outgoing}, \\ &\gamma_5 = -1, \quad \chi(r) \propto \mathrm{e}^{-\mathrm{i}Er}, \quad \text{incoming}. \end{aligned} \tag{2.8.16}$$

(The helicities of the solutions are reversed if $q < 0$.) The cause of this peculiar asymmetry of the helicity states in the lowest partial wave can be traced back directly to the peculiar extra term $-q\hat{\boldsymbol{r}}$ in the angular momentum of a charged particle in the field of the monopole (fig. 21). An incoming (outgoing) fermion must have negative (positive) helicity to be in the lowest partial wave, with $j = q - \frac{1}{2}$.

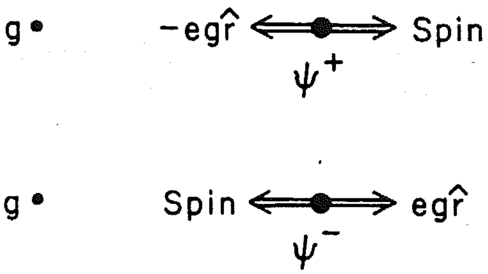


Fig. 21. In the field of a magnetic monopole, whether the spin of a fermion with *minimal* angular momentum points toward or away from the monopole is correlated with the charge of the fermion.

To study scattering, we must determine how the incoming and outgoing waves match up at the origin. However, both solutions are singular at the origin, the location of the pole, and the Dirac equation provides no criterion for matching up the incoming and outgoing waves. Therefore, the Dirac Hamiltonian is not self-adjoint; probability fails to be conserved unless the Hamiltonian is suitably augmented by a boundary condition at the pole that relates the incoming and outgoing waves [66].

In other words, the Dirac Hamiltonian actually describes a family of quantum mechanical systems, distinguished by different boundary conditions satisfied by the fermions at the magnetic monopole. The boundary condition must be specified before the outcome of a scattering event can be determined. For example, consider a four-component Dirac fermion, the (massless) electron, scattering from a monopole, with $q=-\frac{1}{2}$. In the lowest partial wave, there are two possible incoming states and two possible outgoing states, namely

$$\begin{aligned} &\text{incoming:} \quad e_L^-, e_R^+, \\ &\text{outgoing:} \quad e_R^-, e_L^+. \end{aligned} \tag{2.8.17}$$

An incoming left-handed electron e_L^- will emerge from the collision as a right-handed electron e_R^- or a left-handed positron e_L^+, in some linear combination determined by the boundary condition. No choice of the boundary condition can conserve both electric charge and angular momentum, although these are both good symmetries of the Hamiltonian. The monopole transfers either charge or angular momentum to the monopole.

The need for a boundary condition to determine the final state of an electron scattering from a point monopole is the crucial feature of monopole-fermion scattering that results in a violation of the decoupling principle. The decoupling principle leads one to expect that the amplitude for monopole-fermion scattering at energies much less than the inverse size of the monopole core does not depend on the structure of the core, except for power corrections that vanish as the size of the core goes to zero. Up to power corrections, the amplitude should be calculable in a low-energy "effective theory" in which the core is regarded as pointlike, and the properties of the core need not be specified. This expectation fails because monopole-fermion scattering is inherently ambiguous when the monopole is pointlike. Information about the core of the monopole survives in the low-energy effective theory as a boundary condition needed to specify the outcome of a scattering event. A low-energy fermion in the lowest partial wave can penetrate to the monopole core, and be strongly influenced by its structure. In particular, the boundary condition may violate a symmetry (like baryon number) that would otherwise be a good symmetry of the low-energy effective theory.

One sees from the above discussion that the analysis of the scattering of a low-energy fermion by a nonsingular 't Hooft–Polyakov monopole divides naturally into two steps. In the first step, we determine, by considering in the semiclassical approximation the interaction of a fermion with a monopole of finite core size, the appropriate boundary conditions to impose as the core effectively shrinks to zero radius. In the second step, we study the interactions of fermions satisfying the boundary conditions with a point monopole, taking into account as fully as possible the effects of fermion pair creation. But to justify and carry out this procedure in detail, we must consider carefully how the semiclassical expansion is formulated.

In the first nontrivial order of the semiclassical expansion, order e^0, the monopole field may be treated as a classical background field. To this order, it should be legitimate to treat the scattering of a fermion from a nonsingular monopole by solving the Dirac equation in the monopole background. By solving the Dirac equation in the background of an 't Hooft–Polyakov monopole, one finds that an incoming SU(2) doublet fermion in the lowest partial wave emerges from the collision with its helicity preserved, but with its electric charge flipped [67]. One might have guessed this result without doing a detailed computation; electric charge becomes ill-defined at the center of the classical hedgehog solution, and the sign of the charge operator $\boldsymbol{T}\cdot\boldsymbol{\phi}/|\boldsymbol{\phi}|$ flips as the core is traversed. Furthermore, we have already found that the excitations of the monopole that emerge upon semiclassical quantization are electrically charged dyons, so it is plausible that the fermion transfers charge to the monopole, exciting the dyon degree of freedom. It would have been much more puzzling if we found that the fermion transfers angular momentum to the monopole, since no rotational excitations of the monopole were revealed by the semiclassical analysis.

On the other hand, there is something suspicious about the conclusion that the fermion transfers charge to the monopole, producing a dyon excitation. We wish to consider the limit of very low fermion energy E, and when E is much less than the dyon excitation energy E_{dyon} of order $e^2 m_{\text{V}}$, it is absurd to say that the dyon degree of freedom is excited. The problem is that, if we are interested in monopole–fermion scattering for $E \ll e^2 m_{\text{V}}$, the semiclassical limit is not the proper limit to study. In the semiclassical expansion, e^2 is regarded as small and m_{V} and E are order one; thus, in this expansion E is always formally much larger than E_{dyon}, which is why we were able to conclude in order e^0 of the semiclassical expansion that the dyon degree of freedom is excited.

If we are interested in monopole–fermion scattering below the threshold for exciting the dyon, we should study, not the usual semiclassical limit, but a different limit [68],

$$e^2 \to 0, \qquad E_{\text{dyon}} \sim e^2 m_V \to \infty, \tag{2.8.18}$$

with the fermion energy E fixed. In other words, we must *not* neglect the Coulomb energy of the dyon, even though it is formally small in the semiclassical expansion. Fortunately, though, all other effects which are formally of order e^2 *can* be safely neglected. The quantum fields are localized on a scale of order E^{-1}, rather than m_V^{-1}; therefore the Coulomb interactions of quantum fields with the monopole core and with each other are small and can be ignored. Indeed, the *only* unconventional effect which must be retained in the lowest-order semiclassical investigation of the limit eq. (2.8.18) is the dyon self-energy. This is the crucial observation that makes the analysis tractable.

For the purpose of the leading semiclassical approximation, the monopole–fermion system can thus be described by a Lagrangian containing three terms. The first term specifies how the fermions propagate in the classical background field of the monopole. We will confine our attention to the lowest partial wave, the only partial wave with a nontrivial coupling to the monopole core. Thus, the fermion fields solve the free radial Dirac equation

$$\left(\frac{\partial}{\partial t}+\frac{q}{|q|}\gamma_5\frac{\partial}{\partial r}\right)\chi(r, t)=0. \tag{2.8.19}$$

For a definite sign of $(q/|q|)\gamma_5, \chi$ behaves like a free chiral fermion in $(1+1)$ dimensions; it is a right-mover for $(q/|q|)\gamma_5=+1$ and a left-mover for $(q/|q|)\gamma_5=-1$. In the lowest partial wave, a pair of four-dimensional Weyl fermions with $\gamma_5=1$ and $q/|q|=\pm1$ may be described by a pair of two-dimensional chiral fermions propagating on $r\in[0,\infty)$; the positive-charge fermion becomes a right-moving (outgoing) fermion $\psi^+(r)$, and the negative-charge fermion becomes a left-moving (incoming) fermion $\psi^-(r)$. This pair of chiral fermion fields on the half line can be mapped to a single right-moving fermion field $\psi_R(x)$ on the full line $x\in(-\infty,\infty)$ through the identification

$$\psi_R(x)=\psi^+(x), \quad x>0, \tag{2.8.20}$$

$$\psi_R(x)=\psi^-(-x), \quad x<0. \tag{2.8.21}$$

Thus, the first term in our Lagrangian becomes simply

$$L_1=\int \mathrm{d}x\, \mathrm{i}\psi_R^\dagger\left(\frac{\partial}{\partial t}+\frac{\partial}{\partial x}\right)\psi_R. \tag{2.8.22}$$

The second term in our Lagrangian is the self-energy of the charge rotor collective coordinate of the monopole. Denoting the charge orientation of the monopole by a periodic variable θ with period 2π, this term has the form

$$L_2 = \tfrac{1}{2} I \dot{\theta}^2, \tag{2.8.23}$$

where I is a "moment of inertia" of order $(e^2 m_V)^{-1}$.

The third term in the Lagrangian is a coupling between the fermions and the charge rotor θ. This coupling arises because the boundary condition relating ψ^+ to ψ^- at $r = 0$ depends on the charge orientation of the monopole. For a standard orientation of the charge rotor ($\theta = 0$), we can choose a phase convention for the fields so that the charge exchange boundary condition is written

$$\psi^+_{\text{out}}(r=0) = \psi^-_{\text{in}}(r=0), \quad \theta = 0. \tag{2.8.24}$$

But if a global charge rotation by the angle θ is now performed, the phases of ψ^+ and ψ^- are rotated on opposite directions, and the boundary condition becomes

$$\psi^+_{\text{out}}(r=0) = \mathrm{e}^{-\mathrm{i}\theta} \psi^-_{\text{in}}(r=0). \tag{2.8.25}$$

This boundary condition induces a coupling between the charge rotor θ and the fermions in the lowest partial wave. Note that the structure of the monopole core is essential to the derivation of this coupling, for it is the structure of the core that determines that the charge exchange boundary condition is appropriate.

Through the identification eqs. (2.8.20, 21), the boundary condition eq. (2.8.25) becomes

$$\psi(x=0^+, t) = \mathrm{e}^{-\mathrm{i}\theta(t)} \psi(x=0^-, t). \tag{2.8.26}$$

To incorporate this boundary condition into our (1+1)-dimensional field theory, we add to the Lagrangian the coupling [68]

$$L_3 = -\theta(t) \int \mathrm{d}x\, \psi^\dagger_{\mathrm{R}}(x,t) f(x) \psi_{\mathrm{R}}(x,t), \tag{2.8.27}$$

where f is a function with support in the small interval $(-\varepsilon, \varepsilon)$ that integrates to one,

$$\int_{-\varepsilon}^{\varepsilon} f(x)\, \mathrm{d}x = 1. \tag{2.8.28}$$

This coupling simulates the effect of the tiny monopole core. At low energy, all frequencies are small compared to ε^{-1}, and $\theta(t)$ can be regarded as nearly constant as the fermion traverses the interval $(-\varepsilon, \varepsilon)$. Therefore, the equation of motion for ψ derived from L_1+L_3,

$$\left[\frac{\partial}{\partial t}+\frac{\partial}{\partial x}+\mathrm{i}\theta(t)f(x)\right]\psi_{\mathrm{R}}(x, t)=0, \tag{2.8.29}$$

has the approximate solution

$$\psi_{\mathrm{R}}(x, t)=\exp\left[-\mathrm{i}\theta(t)\int_{-\varepsilon}^{x}\mathrm{d}x f(x)\right]\psi_0(x-t), \tag{2.8.30}$$

from which follows

$$\psi_{\mathrm{R}}(\varepsilon, t)=\mathrm{e}^{-\mathrm{i}\theta(t)}\psi_{\mathrm{R}}(-\varepsilon, t-2\varepsilon). \tag{2.8.31}$$

Equation (2.8.31) reduces to the boundary condition eq. (2.8.26) in the limit of small ε.

Having constructed the Lagrangian

$$\begin{aligned} L &= L_1+L_2+L_3 \\ &= \int \mathrm{d}x\, \mathrm{i}\psi_{\mathrm{R}}^{\dagger}(x, t)\left(\frac{\partial}{\partial t}+\frac{\partial}{\partial x}+\mathrm{i}\theta(t)f(x)\right)\psi_{\mathrm{R}}(x, t)+\tfrac{1}{2}I\dot{\theta}^2, \end{aligned} \tag{2.8.32}$$

we are now almost ready to study the monopole–fermion system in the limit eq. (2.8.18). But first we need to tinker with the Lagrangian a little. At the classical level, our Lagrangian has two desirable properties: the dynamical variable θ that represents a charge rotation is a periodic variable with period 2π, and the Noether charge associated with a θ rotation is a conserved quantity. Unfortunately, both properties are spoiled by quantum effects in our (1+1)-dimensional field theory. To restore these properties, we must add a suitable counterterm to the Lagrangian.

To see that θ is a periodic variable at the classical level, we perform the change of variable

$$\psi_{\mathrm{R}}\to \mathrm{e}^{\mathrm{i}\phi g(x)}\psi_{\mathrm{R}}, \tag{2.8.33}$$

which is evidently equivalent to

$$\theta f(x)\to \theta f(x)+\phi\frac{\mathrm{d}}{\mathrm{d}x}g(x). \tag{2.8.34}$$

If we define

$$g(x)=\int_{-\varepsilon}^{x} dx'\,f(x')-\tfrac{1}{2}, \tag{2.8.35}$$

then we have

$$\theta\to\theta+\phi. \tag{2.8.36}$$

Furthermore, since

$$g(x)=-\tfrac{1}{2}, \quad x<-\varepsilon, \tag{2.8.37}$$

$$g(x)=+\tfrac{1}{2}, \quad x>\varepsilon, \tag{2.8.38}$$

we see that the change of variable, eq. (2.8.33), acts trivially on ψ outside the interval $(-\varepsilon,\varepsilon)$ if ϕ is an integral multiple of 2π; in particular, the boundary condition is unmodified. Thus, the (classical) physics of this model is left invariant by a 2π rotation of θ.

The Lagrangian eq. (2.8.32) is invariant under the transformation eq. (2.8.33), accompanied by $\theta\to\theta-\phi$. There is an associated (classically) conserved Noether charge

$$Q_{\text{tot}}=\int dx\,\psi_{\text{R}}^{\dagger}(x,t)g(x)\psi_{\text{R}}(x,t)+Q, \quad Q=I\dot{\theta}(t). \tag{2.8.39}$$

Q_{tot} is just the sum of the fermionic electric charge and the electric charge Q carried by the dyon. The conservation law says that charge lost by the fermions is transferred to the monopole.

However, as we saw back in section 1.5, conservation of Q_{tot} is spoiled by an anomaly. Equation (2.8.32) is the Lagrangian of a chiral fermion in $(1+1)$ dimensions coupled to the background electric field

$$eE(x,t)=-\theta(t)\frac{d}{dx}f(x). \tag{2.8.40}$$

According to the anomaly equation, the fermionic electric charge density satisfies

$$\left(\frac{\partial}{\partial t}+\frac{\partial}{\partial x}\right)\psi_{\text{R}}^{\dagger}\psi_{\text{R}}=\frac{eE}{2\pi}=-\frac{\theta}{2\pi}\frac{df}{dx} \tag{2.8.41}$$

(cf. eq. (1.5.13); eq. (2.8.41) holds if we define the composite operator $\psi_{\text{R}}^{\dagger}\psi_{\text{R}}$ by, for example, covariant point splitting). We conclude that Q_{tot}

changes at the rate

$$\dot{Q}_{\rm tot}=-\frac{\theta(t)}{2\pi}\int_{-\varepsilon}^{\varepsilon}\mathrm{d}x\, g(x)\frac{\mathrm{d}}{\mathrm{d}x}f(x)=\frac{\theta(t)}{2\pi}\int_{-\varepsilon}^{\varepsilon}\mathrm{d}x\,[f(x)]^2, \tag{2.8.42}$$

where an integration by parts has been performed in the last step. To restore conservation of electric charge, we add to our model Lagrangian the final term

$$L_4=\tfrac{1}{2}C\theta^2,\quad C=\frac{1}{2\pi}\int_{-\varepsilon}^{\varepsilon}\mathrm{d}x\,[f(x)]^2. \tag{2.8.43}$$

This term modifies the equation of motion for θ, increasing $I\ddot{\theta}$ so that $\dot{Q}_{\rm tot}$ now vanishes.

The new term L_4 may appear to spoil the periodicity of θ, but in fact it does not. The change of variable eq. (2.8.33) actually has a nontrivial Jacobian associated with the anomaly, so that we *need* the counterterm L_4 to ensure that physics at the quantum level is really unchanged by a 2π rotation of θ.

Finally, we have a (1+1)-dimensional model field theory that correctly represents the interactions of a doublet of massless Weyl fermions at low energy with an 't Hooft-Polyakov monopole. Trivially generalizing to a model with N identical fermion flavors, we have the Lagrangian [68]

$$L=\int \mathrm{d}x\,\mathrm{i}\sum_{k=1}^{N}\psi_k^\dagger(x,t)\left(\frac{\partial}{\partial t}+\frac{\partial}{\partial x}+\mathrm{i}\theta(t)f(x)\right)\psi_k(x)+\tfrac{1}{2}I\dot{\theta}^2-\tfrac{1}{2}C\theta^2,$$
$$C=\frac{N}{2\pi}\int_{-\varepsilon}^{\varepsilon}\mathrm{d}x\,[f(x)]^2. \tag{2.8.44}$$

It only remains to solve the model.

The θ equation of motion derived from the Lagrangian eq. (2.8.44) is

$$\dot{Q}=-\int_{-\varepsilon}^{\varepsilon}\mathrm{d}x\sum_{k=1}^{N}f(x)\psi_k^\dagger(x)\psi_k(x)-C\theta, \tag{2.8.45}$$

where $Q=I\dot{\theta}$ is the dyon charge, the momentum conjugate to θ. This equation, together with the local anomaly equations

$$\left(\frac{\partial}{\partial t}+\frac{\partial}{\partial x}\right)\psi_k^\dagger\psi_k=-\frac{\theta}{2\pi}\frac{\mathrm{d}f}{\mathrm{d}x}, \tag{2.8.46}$$

can be solved simultaneously for the $\psi_k^\dagger\psi_k$'s and Q. It is convenient to introduce a variable

$$j_k(x,t)=\psi_k^\dagger\psi_k(x,t)+\frac{\theta(t)}{2\pi}f(x), \tag{2.8.47}$$

in terms of which these equations become

$$\dot{Q}(t)=-\int_{-\varepsilon}^{\varepsilon}\mathrm{d}x\sum_{k=1}^{N}f(x)j_k(x,t),$$

$$\left(\frac{\partial}{\partial t}+\frac{\partial}{\partial x}\right)j_k(x,t)=\frac{1}{2\pi I}f(x)Q(t). \tag{2.8.48}$$

Since j_k and $\psi_k^\dagger\psi_k$ agree outside the core region $-\varepsilon<x<\varepsilon$, we may just as well study the time evolution of j_k as $\psi_k^\dagger\psi_k$ to determine the flow of fermionic charge into and out of the monopole.

Outside the core, j_k propagates freely to the right. To solve for j_k inside the core, we may treat $Q(t)$ as constant for the time that it takes to traverse the core, obtaining

$$j_k(x,t)=j_k^0(x-t)+\frac{1}{2\pi I}Q(t)\int_{-\varepsilon}^{x}\mathrm{d}x'f(x'). \tag{2.8.49}$$

Substituting back into eq. (2.8.48), we find

$$\dot{Q}(t)=-\sum_{k=1}^{N}j_k(-\varepsilon,t)-\frac{N}{4\pi I}Q(t), \tag{2.8.50}$$

using eq. (2.8.28).

In eq. (2.8.50), the rate of change of the dyon charge is expressed as a sum of two terms. The first term is the rate at which fermionic charge flows into the monopole core. The second term arises solely because of the anomaly, and can be viewed as the rate at which charge on the monopole is disposed of by anomalous fermion production. If there is no incoming flow of charge, the expectation value of eq. (2.8.50) becomes

$$\frac{\mathrm{d}}{\mathrm{d}t}\langle Q\rangle=-\frac{N}{4\pi I}\langle Q\rangle. \tag{2.8.51}$$

Hence a dyon excitation decays at a rate [69] $\Gamma=N/4\pi I$, that is comparable to the splitting $1/2I$ between dyon energy levels; the tower of dyonic excitations is wiped out by fermion emission. (The factor N occurs because the dyon emits all fermion flavors democratically.) From a (3+1)-dimensional point of view, there is a strong expectation value of

$\boldsymbol{E}\cdot\boldsymbol{B}$ near the core of a dyon that, due to the chiral anomaly, induces fermion emission and rapidly causes the dyon to deexcite.

Since the monopole disposes of the fermionic charge deposited on it on a time scale much shorter than E^{-1}, where E is the energy of an incoming fermion, charge cannot accumulate on the monopole, and $\dot{Q}$ is effectively zero. Therefore, the solution to eq. (2.8.50) is

$$Q(t) = -\frac{4\pi I}{N}\sum_{k=1}^{N} j_k(-\varepsilon, t). \tag{2.8.52}$$

Substituting back into eq. (2.8.49), we obtain

$$j_k(\varepsilon, t) = j_k(-\varepsilon, t) - \frac{2}{N}\sum_{k=1}^{N} j_k(-\varepsilon, t). \tag{2.8.53}$$

In any scattering process, the total fermionic charge that flows into or out of the monopole can be found by integrating the currents,

$$n_k^{\text{in}} = \int_{-\infty}^{\infty} \mathrm{d}t\, j_k(-\varepsilon, t), \tag{2.8.54}$$

$$n_k^{\text{out}} = \int_{-\infty}^{\infty} \mathrm{d}t\, j_k(\varepsilon, t), \tag{2.8.55}$$

and so we obtain [68]

$$n_k^{\text{out}} = n_k^{\text{in}} - \frac{2}{N}\sum_{l=1}^{N} n_l^{\text{in}}. \tag{2.8.56}$$

Equation (2.8.56) is the main result of our analysis. It describes the relation between the flavor quantum numbers of the initial and final states in monopole-fermion scattering. This relation is determined by a subtle interplay of the chiral anomaly and the boundary conditions satisfied by the fermions at the monopole core.

If a single fermion of flavor $k=1$ is incident on the monopole in the lowest partial wave, then the final state must satisfy

$$n_k^{\text{out}} = \delta_{k,1} - 2/N. \tag{2.8.57}$$

The nature of the final state evidently depends on N, the total number of flavors. Let us consider a few special cases.

(i) $N=1$; $n_1^{\text{out}} = -1$. In this case, an incoming ψ_{R}^{-} emerges as the $(\psi_{\mathrm{R}}^{+})^{\mathrm{c}}$ antiparticle of ψ_{R}^{+}. In effect, the helicity of the fermion is flipped, which is the only possibility consistent with conservation of electric charge.

(Actually, an SU(2) model with any odd number of chiral fermion doublets is known to be rendered inconsistent by a global anomaly [70], so this $N=1$ model does not really exist.)

(ii) $N=2$, $n_1^{\text{out}}=0$, $n_2^{\text{out}}=-1$. An incoming $\psi_{\text{R},1}^-$ emerges as an outgoing $(\psi_{\text{R},2}^+)^{\text{c}}$. If we think of the two flavors of Weyl doublets as the left-handed and right-handed components of a Dirac doublet, we may describe the process as $\psi_{\text{R}}^- \rightarrow \psi_{\text{L}}^-$; this is helicity flip again [62, 63]. (There may also be, in this case as in those considered below, an indefinite number of flavor-neutral fermion pairs in the final state.)

(iii) $N=4$; $n_1^{\text{out}}=\frac{1}{2}$, $n_{2,3,4}^{\text{out}}=-\frac{1}{2}$. Now we encounter something strange and unexpected. What emerges from the collision of an incident fermion with the monopole are not final state fermions of the usual sort, but pulses of fermionic vacuum polarization with fractional fermion number in each flavor channel [71, 72]. These excitations, called "semitons", are allowed final states because the fermion masses have been neglected. If the fermions really have masses, or are confined in hadrons, integer fermion numbers are detected in the final state - an electron is either there or it is not. Thus, the semitons must be destabilized by explicit fermion mass terms (or confining interactions); they must be able to evolve into states with integer fermion number over distance scales comparable to the fermion Compton wavelength (or the confinement scale).

The decay of a semiton into genuine final state fermions is a long-distance process having little to do with the physics of monopole-fermion interactions, and it is not yet understood in any detail. But we can get a better idea of what types of processes are possible in monopole-fermion scattering by considering appropriate initial states such that the final states are conventional states with integer fermion number. For example, if $n_1^{\text{in}}=n_2^{\text{in}}=1$, $n_3^{\text{in}}=n_4^{\text{in}}=0$, then eq. (2.8.56) gives $n_3^{\text{out}}=n_4^{\text{out}}=-1$, $n_1^{\text{out}}=n_2^{\text{out}}=0$; the process

$$\psi_{\text{R},1}^-\psi_{\text{R},2}^- \rightarrow (\psi_{\text{R},3}^+)^{\text{c}}(\psi_{\text{R},4}^+)^{\text{c}} \tag{2.8.58}$$

can occur. Obviously, this process fails to conserve the flavor quantum numbers.

It is evident that the chiral anomaly must play an essential role in the above processes. Were it not for the chiral anomaly, the chirality and flavor of a massless fermion would be preserved. The physics of the monopole core is also essential; it determines what charge is transferred by the incoming fermion to the monopole core. But the anomaly determines how the charge deposited on the core is subsequently disposed of. And only the anomaly, not the boundary condition at the core, changes chirality and flavor.

The language of the previous paragraph implies that monopole-fermion scattering can be described as a two-stage process. First, the incoming fermion scatters off the monopole, producing a dyon excitation. Second, the dyon decays to the ground state monopole plus some final state fermions. However, we have already stressed that this language is quite misleading. A low-energy incoming fermion is surely unable to excite the very energetic dyon. The key role of the anomaly in the scattering process nonetheless suggests a picture in which charge temporarily resides on the monopole, and generates a radial electric field in the vicinity of the core. The parallel electric and magnetic fields near the core can then drive the anomalous production of fermions. How can this picture be reconciled with the obvious fact that excitation of the dyon is not kinematically allowed?

The answer is implicit in our analysis of the model eq. (2.8.44). We derived in eq. (2.8.52) the relation

$$Q=\frac{4\pi I}{N}J\sim\frac{r_0}{Ne^2}J, \tag{2.8.59}$$

between the charge Q on the monopole and the incoming fermion current J, where r_0 is the size of the monopole core. The electric field near the surface of the core is of order eQ/r_0^2, and the anomaly equation predicts that chiral charge is produced at the rate

$$\langle\dot{Q}_5\rangle\sim e^2N\int \mathrm{d}^3r\,\langle \boldsymbol{E}\cdot\boldsymbol{B}\rangle\sim J. \tag{2.8.60}$$

The change in Q_5 integrated over time is therefore of order one. Since J is of order the momentum p of the incoming fermion, we see that the expectation value of Q is very small when p is far below the dyon excitation energy I^{-1}, reflecting the inaccessibility of the dyon excitation. But the small charge Q persists for a long time p^{-1}, which allows the cumulative effect of the anomaly to be sizable [73].

The analysis described above applies to processes catalyzed by the 't Hooft–Polyakov monopole of an SU(2) gauge theory, and only to processes involving fermions in the doublet representation of SU(2). We may be interested in fermions in other representations, or in models with larger gauge groups. When the analysis is suitably generalized, some qualitatively new features emerge. To appreciate one such feature, consider the nonminimal SU(3) monopole of section 2.7, with magnetic charge

$$Q_{\mathrm{M}}=\mathrm{diag}(1,1,-2).$$

If a right-handed fermion triplet interacts with the monopole, two members of the triplet have $q=\frac{1}{2}$ and one member has $q=-1$ (with q defined as in eq. (2.8.6)). Therefore, the modes that penetrate to the monopole core are an incoming fermion with $j=\frac{1}{2}$ and two outgoing fermions with $j=0$. Evidently, the boundary conditions satisfied by the fermions at the monpole core *must* require that angular momentum is transferred to the monopole; the dyon excitations in this model carry *spin* [73]. Indeed, careful consideration of the semiclassical quantization of this monopole shows that, although the static monopole solution is spherically symmetric, the time-dependent configurations that arise in the quantization procedure are not spherically symmetric, and the excitations therefore carry angular momentum [61]. Actually, if one wishes to demonstrate that the dyons carry spin, it is much simpler to analyze the catalysis process, instead of carrying out the semiclassical quantization procedure in detail.

Let us apply our newly found understanding of monopole–fermion scattering to a particularly interesting example, the SU(5) model. The classical monopole solution lives in a particular SU(2) subgroup of SU(5), and the appropriate boundary conditions for the fermions can be inferred from our earlier discussion of the 't Hooft–Polyakov monopole. The diagonal SU(2) generator is $\frac{1}{2}Q'$ of eq. (2.7.27), and the SU(2) representation content of a single generation of fermions (the representation $\overline{10}+5$ of SU(5)) is readily seen to be the four doublets

$$\begin{pmatrix}\bar{d}_3\\ e^-\end{pmatrix}_{\mathrm{L}},\quad \begin{pmatrix}u_1\\ \bar{u}_2\end{pmatrix}_{\mathrm{L}},\quad \begin{pmatrix}u_2\\ \bar{u}_1\end{pmatrix}_{\mathrm{L}},\quad \begin{pmatrix}e^+\\ d_3\end{pmatrix}_{\mathrm{L}}, \tag{2.8.61}$$

plus singlets, where 1, 2, 3 are color indices. For $2q=+1$, the top member of each doublet (or the antiparticle of the bottom member) is an incoming state with access to the monopole core. The boundary condition couples each incoming fermion to the outgoing fermion that is the bottom member of the same doublet. The $\mathrm{U}(1)_{Q'}$ and $\mathrm{U}(1)_{\mathrm{em}}$ charges transferred to the monopole are precisely the charges of the minimal dyon found in section 2.7.

For the purpose of studying the monopole–fermion interactions, this model reduces to an SU(2) model with four doublets, and the process eq. (2.8.58) found earlier becomes:

$$u_{1\mathrm{L}}u_{2\mathrm{L}} \to e^{+}_{\mathrm{R}}\bar{d}_{3\mathrm{R}}. \tag{2.8.62}$$

That a monopole can catalyze this baryon number nonconserving process was discovered by Rubakov [62] and Callan [63].

Accurate calculations of the rates of such baryon number changing reactions are not easily performed, in part because the evolution of semitons into "final state" quarks and leptons is not yet understood in quantitative detail. But in the leading semiclassical approximation, the reaction

$$u_{1\mathrm{L}} \to \tfrac{1}{2}(\bar{u}_{2\mathrm{L}} \quad u_{1\mathrm{R}} \quad \bar{d}_{3\mathrm{R}} \quad e^{+}_{\mathrm{R}}) + \text{flavor-neutral pairs} \tag{2.8.63}$$

saturates the unitarity limit in the lowest partial wave. It is reasonable to expect that the semiton intermediate state can evolve with probability of order one into a final state with a baryon number different from the initial state [74]. Thus, the baryon number changing cross section for a quark of energy E scattering from a monopole is of order E^{-2}, if E^{-1} is much greater than the radius of the monopole core, and much less than both the Compton wavelength of the quark and the size of a hadron. It is also to be expected that adding more generations of fermions will have little qualitative effect on the baryon number changing processes. The main new feature in the many-generation case is that the boundary conditions and hence the scattering amplitudes depend on generalized Cabibbo-like mixing angles [75].

The baryon number nonconservation catalyzed by an SU(5) monopole cannot be regarded as a consequence of the chiral anomaly; tr Q'^2B actually vanishes, where Q' is the charge carried by the monopole and B is baryon number. The violation of baryon number really arises from the boundary conditions satisfied by the fermions at the monopole core. The boundary condition allows the baryon number transferred to the monopole to be either $-\frac{2}{3}$ or $+\frac{1}{3}$. Thus, the dyon does not have definite baryon number, and it is capable of mediating baryon number changing processes.

There *is* a baryon number violating anomaly in the standard model, but it is effective only if the electric or magnetic field has a component in the Z^0 direction. (This anomaly generated baryon number nonconservation on the superconducting string of section 1.5.) It is possible to embed the standard model in a grand unified theory such that the minimal monopole carries a Z^0 charge, at least at distances from the monopole core less than M_Z^{-1}; an example is the Pati–Salam model. (See the exercise at the end of section 2.4.) Indeed, in the Pati–Salam model, baryon number is a good classical symmetry, so the anomaly is the only possible source of nonconservation of B. The B-changing processes catalyzed by Pati–Salam monopoles that arise from the anomaly, like the processes catalyzed by SU(5) monopoles that arise from B-violation at the monopole core, have large rates completely unsuppressed by the tiny size of the core [73, 76, 77].

Exercise. By considering the boundary conditions satisfied by the fermions at the core of a Pati–Salam monopole, infer enough about dyon quantum numbers to show that anomalous production of baryon number occurs in the vicinity of a Pati–Salam dyon.

References

[1] H. Nielsen and P. Olesen, Nucl. Phys. B 61 (1973) 45.

[2] A.S. Schwarz, Nucl. Phys. B 208 (1982) 141;
A.S. Schwarz and Yu.S. Tyupkin, Nucl. Phys. B 209 (1982) 427.

[3] M. Hindmarsh and T. Kibble, Phys. Rev. Lett. 55 (1985) 2398.

[4] S. Coleman, Classical lumps and their quantum descendents, in: New Phenomena in Subnuclear Physics, ed. A. Zichichi (Plenum, London, 1977).

[5] Ya. B. Zeldovich, I.Yu. Kobzarev and L.B. Okun, Sov. Phys. JETP 40 (1975) 1.

[6] A.H. Guth, Phys. Rev. D 23 (1981) 347.

[7] T.W.B. Kibble, G. Lazarides and Q. Shafi, Phys. Rev. D 26 (1982) 435.

[8] A. Vilenkin and A.E. Everett, Phys. Rev. Lett. 48 (1982) 1867.

[9] R. Peccei and H. Quinn, Phys. Rev. Lett. 38 (1977) 1440.

[10] P. Sikivie, Phys, Rev. Lett. 48 (1982) 1156.

[11] S. Weinberg, Phys. Rev. Lett. 40 (1978) 223;
F. Wilczek, Phys, Rev. Lett. 40 (1978) 279.

[12] H. Georgi and M.B. Wise, Phys. Lett. B 116 (1982) 123.

[13] G. Lazarides and Q. Shafi, Phys. Lett. B 115 (1982) 21.

[14] F. Wilczek, Phys. Rev. Lett. 49 (1982) 1549;
D.B. Reiss, Phys. Lett. B 115 (1982) 217;
S.M. Barr, D.B. Reiss, and A. Zee, Phys. Lett. B 116 (1982) 227.

[15] E. Witten, Nucl. Phys. B 249 (1985) 557.

[16] E. Weinberg, Phys. Rev. D 24 (1981) 2669;
R. Jackiw and P. Rossi, Nucl. Phys. B 190 (1981) 681.

[17] C. Callan and J. Harvey, Nucl. Phys. B 250 (1985) 427;
G. Lazarides and Q. Shafi, Phys. Lett. B 151 (1985) 123.

[18] S. Coleman and P. Ginsparg (1982), unpublished.

[19] A. Vilenkin, Phys. Rev. Lett. 46 (1981) 1169, 1494 (E).

[20] A. Vilenkin, Phys. Rep. 121 (1985) 263.

[21] N. Turok, Cosmic strings and the correlations of Abell clusters, UC Santa Barbara preprint 85-0535 (1985).

[22] R. Brandenberger and N. Turok, Phys. Rev. D 33 (1986) 2175.

[23] T.W.B. Kibble, J. Phys. A 9 (1976) 1387.

[24] P.G. de Gennes, Scaling Concepts in Polymer Physics (Cornell University Press, Ithaca, 1979).

[25] R. Scherrer and J. Frieman, Cosmic strings as random walks, Enrico Fermi Institute preprint EFI-86-07 (1986).

[26] T. Vachaspati and A. Vilenkin, Phys, Rev. D 30 (1984) 2036.

[27] A. Vilenkin, Phys. Rev. D 24 (1981) 2082.

[28] A. Albrecht and N. Turok, Phys. Rev. Lett. 54 (1985) 1868.

[29] J. Preskill and M.B. Wise (1983), unpublished.

[30] Ya.B. Zeldovich, Mon. Not. R. Astron. Soc. 192 (1980) 663.

[31] R. Brandenberger and N. Turok, Phys. Rev. D 33 (1986) 2182.

[32] T.W.B. Kibble and N. Turok, Phys. Lett. B 116 (1982) 141.

[33] A. Vilenkin, Phys. Lett. B 107 (1981) 47;
T. Vachaspati and A. Vilenkin, Phys. Rev. D 31 (1985) 3052.

[34] C.J. Hogan and M.J. Rees, Nature 311 (1984) 109;
E. Witten, Phys. Rev. D 30 (1984) 272.

[35] P. Sikivie, Axions in cosmology, in: Gif-sur-Yvette Proceedings (1983).

[36] P.A.M. Dirac, Proc. R. Soc. London A 133 (1931) 60.

[37] T.T. Wu and C.N. Yang, Phys. Rev. D 12 (1975) 3845; Nucl. Phys. B 107 (1976) 365.

[38] E. Lubkin, Ann. Phys. 23 (1963) 233.

[39] S. Coleman, The magnetic monopole fifty years later, in: The Unity of the Fundamental Interactions, ed. A. Zichichi (Plenum, London, 1983).

[40] J. Preskill, Ann. Rev. Nucl. Part. Sci. 34 (1984) 461.

[41] R. Brandt and F. Neri, Nucl. Phys. B 161 (1979) 253.

[42] G. 't Hooft, Nucl. Phys. B 138 (1978) 1.

[43] G. 't Hooft, Nucl. Phys. B 79 (1974) 276.

[44] A.M. Polyakov, JETP Lett. 20 (1974) 194.

[45] P. Goddard, J. Nuyts and D. Olive, Nucl. Phys. B 125 (1977) 1.

[46] G. 't Hooft, Nucl. Phys. B 105 (1976) 538;
E. Corrigan, D. Olive, D. Fairlie and J. Nuyts, Nucl. Phys. B 106 (1976) 475.

[47] F.A. Bais, Phys. Lett. B 98 (1981) 437.

[48] C. Gardner and J. Harvey, Phys. Rev. Lett. 52 (1984) 879.

[49] E. Weinberg, D. London and J. Rosner, Nucl. Phys. B 236 (1984) 90.

[50] G. Lazarides, M. Magg and Q. Shafi, Phys. Lett. B 97 (1980) 87.

[51] S. Dawson and A.N. Schellekens, Phys. Rev. D 27 (1983) 2119.

[52] M. Berry, Proc. R. Soc. London A 392 (1984) 45;
B. Simon, Phys. Rev. Lett. 51 (1983) 2167.

[53] R. Geroch. J. Math. Phys. 9 (1968) 1739; 11 (1970) 343;
L. Castellani, L.J. Romans and N.P. Warner, Nucl. Phys. B 241 (1984) 429.

[54] L. Alvarez-Gaumé and P. Nelson, Commun. Math. Phys. 99 (1985) 103.

[55] S.W. Hawking and C.N. Pope, Phys. Lett. B 73 (1978) 42;
M.F. Atiyah, R. Bott and V.K. Patodi, Inv. Math. 19 (1973) 279.

[56] P. Nelson and A. Manohar, Phys. Rev. lett. 50 (1983) 943;
A. Balachandran et al., Phys. Rev. Lett. 50 (1983) 1553.

[57] B. Julia and A. Zee, Phys. Rev. D 11 (1975) 2227.

[58] S. Coleman and P. Nelson, Nucl. Phys. B 237 (1984) 1.

[59] E. Tomboulis and G. Woo, Nucl. Phys. B 107 (1976) 221.

[60] A. Aboulsaoud, Nucl. Phys. B 226 (1983) 309.
[61] L. Dixon, Nucl. Phys. B 248 (1984) 90.
[62] V. Rubakov, JETP Lett. 33 (1981) 644; Nucl. Phys. B 203 (1982) 311.
[63] C.G. Callan, Phys. Rev. D 25 (1982) 2141; D 26 (1982) 2058; Nucl. Phys. B 212 (1983) 391.
[64] D. Boulware et al., Phys. Rev. D 14 (1976) 2708.
[65] Y. Kazama, C.N. Yang and A.S. Goldhaber, Phys. Rev. D 15 (1977) 2287.
[66] A.S. Goldhaber, Phys. Rev. D 16 (1977) 1815.
[67] R. Jackiw and C. Rebbi, Phys. Rev. D 13 (1976) 3398;
C. Besson, Ph.D. Thesis, Princeton University (1982), unpublished.
[68] J. Polchinski, Nucl. Phys. B 242 (1984) 345.
[69] A. Blaer, N. Christ and J.F. Tang, Phys. Rev. Lett. 49 (1981) 1364; Phys. Rev. D 25 (1982) 2128.
[70] E. Witten, Phys. Lett. B 117 (1982) 324.
[71] C.G. Callan, The monopole catalysis *S*-matrix, in: La Jolla Proceedings (1983).
[72] A. Sen, Phys. Rev. D 28 (1983) 876.
[73] C.G. Callan, Monopole update, in: Fifth Workshop on Grand Unification (1984).
[74] S. Dawson and A.N. Schellekens, Phys. Rev. D 28 (1983) 3125.
[75] J. Ellis, D.V. Nanopoulos and K.A. Olive, Phys. Lett. B 116 (1982) 127;
F.A. Bais et al., Nucl. Phys. B 219 (1983) 189.
[76] F. Wilczek, Phys. Rev. Lett. 48 (1982) 1146.
[77] A.S. Goldhaber, Monopoles, gauge fields, and anomalies, in: Fourth Workshop on Grand Unification, eds P. Langacker, P. Steinhardt and A. Weldon (Birkhauser, Boston, 1983);
A. Sen, Nucl. Phys. B 250 (1985) 1; Phys. Lett. B 153 (1985) 55;
A.N. Schellekens, Phys. Rev. D 29 (1984) 2378.