Nuclear Physics B341 (1990) 50–100
North-Holland

# LOCAL DISCRETE SYMMETRY AND QUANTUM-MECHANICAL HAIR

John PRESKILL*,**

*California Institute of Technology, Pasadena, CA 91125, USA*

Lawrence M. KRAUSS**,***

*Center for Theoretical Physics and Department of Astronomy, Sloane Lab, Yale University, New Haven, CT 06517, USA*

Received 26 January 1990

A charge operator is constructed for a quantum field theory with an abelian discrete gauge symmetry, and a non-local order parameter is formulated that specifies how the gauge symmetry is realized. If the discrete gauge symmetry is manifest, then the charge inside a large region can be detected at the boundary of the region, even in a theory with no massless gauge fields. This long-range effect has no classical analog; it implies that a black hole can in principle carry "quantum-mechanical hair". If the gauge group is nonabelian, then a charged particle can transfer charge to a loop of cosmic string via the nonabelian Aharonov–Bohm effect. The string loop can carry charge even though there is no localized source of charge anywhere on the string or in its vicinity. The "total charge" in a closed universe must vanish, but, if the gauge group is nonabelian and the universe is not simply connected, then the "total charge" is not necessarily the same as the sum of all point charges contained in the universe.

## 1. Introduction

It was recently pointed out in ref. [1] that local gauge invariance may have interesting low-energy consequences even in the absence of light gauge fields. Imagine, for example, a gauge theory with continuous gauge group G that undergoes the Higgs mechanism at a symmetry-breaking mass scale $v$. If the only surviving manifest gauge symmetry is a *discrete*, but nontrivial, subgroup H of G, then all gauge fields acquire masses of order $ev$, where $e$ is the gauge coupling. The effective field theory that describes physics at energies well below the mass

* This work supported in part by the US Department of Energy under Contract no. DE-AC0381-ER40050.
** NSF Presidential Young Investigator.
*** Research supported in part by US Department of Energy.

scale $ev$ is a theory without gauge fields, but which respects the *local* H symmetry. Particles in this effective field theory may carry nontrivial H charges. Even though the classical electric field of a charged particle is screened by the Higgs condensate, the charge induces quantum-mechanical effects that *can* be detected in principle at arbitrarily long distances. In particular, an H charge exhibits a nontrivial Aharonov–Bohm effect upon circumnavigating a cosmic string of the underlying G gauge theory [2,3].

In fact, the notion of a local discrete H symmetry can be formulated even without appealing to an underlying theory with a continuous gauge group G that contains H [1,4]. The significance of local H symmetry can be appreciated directly if we consider a spacetime manifold M that is not simply connected. The fields of an H gauge theory defined on M need not be strictly periodic on the noncontractible closed loops of M; instead, the fields need only be periodic up to the action of an element of H. Such boundary conditions distinguish local H symmetry from global H symmetry, and allow the local symmetry to manifest itself through nontrivial Aharonov–Bohm phenomena.

While these observations are elementary, their consequences are potentially profound. For example, it is well known that among the quantum numbers that characterize a black hole are charges, like electric charge, that couple to massless gauge fields; the long-range field of a charged particle survives even as the particle crosses the horizon of the black hole. The possibility of discrete local symmetry suggests [1,4] that a black hole may also be endowed with other varieties of "hair" that, while invisible classically, can be detected through quantum interference effects. Such quantum-mechanical hair blurs the distinction between black holes and elementary particles, thus encouraging the speculation that this distinction may eventually fade away entirely.

Furthermore, the suggestion that discrete symmetries may be gauge symmetries provides a rationale for discrete symmetries in low-energy physics. Since no fundamental principle prevents global symmetries from being badly broken, one is naturally reluctant to impose any global symmetries, whether continuous or discrete, on a model field theory that purports to describe Nature. But a local symmetry, whether continuous or discrete, is intrinsically exact. Thus, the gauge principle may justify the discrete symmetries in various extensions of the standard model. For example, discrete symmetries are often invoked to ensure the phenomenological viability of models of low-energy supersymmetry. Such symmetries, if *global*, are jeopardized by the wormhole fluctuations in the topology of spacetime that may occur in quantum gravity [5,6]. But, if *local* these symmetries are invulnerable to wormhole dynamics [1].

Although the arguments contained there are persuasive, the discussion in ref. [1] did not provide a clearly stated definition of the charge in a theory with discrete local symmetry, or any precise mathematical criterion for whether the charge is detectable. In this paper we will fill these gaps in the previous discussion. We hope

to clarify the concept of discrete local symmetry, and to further explore some of its consequences.

The remainder of this paper is organized as follows. In sect. 2, we review the notion of a superselection rule, and observe that the central claim of ref. [1] is the existence of nontrivial superselection sectors labeled by the discrete gauge charge. We construct an operator realization of this charge in sect. 3 that is applicable when the discrete gauge symmetry is abelian, and we use this charge operator in sect. 4 to formulate a nonlocal order parameter that specifies how the discrete gauge symmetry is realized. The properties of the charge operator and the order parameter are illustrated in sect. 5 in the context of an explicit example – a $Z_2$ lattice gauge theory coupled to a $Z_2$ spin system. We describe in sect. 6 how a continuum field theory with local discrete symmetry can be formulated without introducing any gauge fields.

In sect. 7, we consider the case of nonabelian discrete gauge symmetry. We note a topological obstruction that prevents implementation of a discrete global gauge transformation when cosmic strings are present. One consequence of this obstruction is the nonabelian Aharonov–Bohm effect, whereby a charged particle can exchange charge with a loop of cosmic string. The string loop can carry charge, even though there is no localized source of charge anywhere on the string or in its vicinity. This phenomenon can also occur if the gauge group is continuous; for example, a string loop can carry electric charge $Q$ if the matching condition imposed by the string does not commute with $Q$. Similarly, in two spatial dimensions, charge can be carried by a pair of vortices. We observe in sect. 8 that, in the case of an abelian local discrete symmetry, the total charge contained in a closed universe must vanish. We also note that, even if cosmic strings are absent, a nonabelian Aharonov–Bohm effect can occur in a closed universe that is not simply connected; in this case, a charged particle can exchange charge with a "handle" that is attached to the universe. We consider the implications of this process in connection with wormhole physics. Sect. 9 contains our conclusions.

## 2. Superselection rules

In any discussion of charges in a gauge theory, the concept of a superselection rule plays a central role. We will now review this concept, emphasizing especially the fate of the superselection rule when a local symmetry undergoes the Higgs mechanism [7, 8].

A quantum field theory is said to respect a superselection rule if the physical Hilbert space $\mathscr{H}$ can be decomposed as a direct sum of distinct sectors

$$\mathscr{H} = \bigoplus_n \mathscr{H}_n, \tag{2.1}$$

such that every local observable $\mathcal{O}$ preserves the decomposition,

$$\langle m|\mathcal{O}|n\rangle = 0, \qquad m \neq n. \tag{2.2}$$

A familiar example is quantum electrodynamics in the Coulomb phase, which has superselection sectors labeled by the electric charge $Q$. The local observables of QED are gauge-invariant operators smeared in a compact region of spacetime, and *local* gauge-invariant operators always have charge $Q = 0$. More physically, a charged particle is accompanied by a long-range electric field; the particle cannot be annihilated unless its long-range field is annihilated as well, and no local observable can accomplish this task. (As in any discussion of superselection rules, we assume that the spatial volume is infinite.)

The superselection sectors of QED may also be characterized by their transformation properties under *global* gauge transformations. If $U(\omega)$ is the unitary operator acting on $\mathcal{H}$ that represents the global gauge transformation $e^{i\omega} \in U(1)$, then a state of charge $Q$ transforms according to

$$U(\omega)|Q\rangle = e^{i\omega Q}|Q\rangle. \tag{2.3}$$

Thus, states of definite charge are preserved by the global gauge transformations, as a state is identified with a ray in Hilbert space. A linear combination of two states from different charge sectors is not preserved, because $U(\omega)$ rotates the relative phase of the two states. However, because of the superselection rule, this relative phase is completely unobservable. It is therefore legitimate to confine our attention to a single sector, and regard all physical states as invariant under global gauge transformations; the property that physical states are invariant distinguishes global gauge symmetries from ordinary global symmetries. Even though local observables preserve each charge sector, we include all of the charge sectors in $\mathcal{H}$, since an isolated object can carry any amount of charge.

Another familiar example of a superselection rule arises if a global symmetry is spontaneously broken. In this case, there are degenerate vacua that transform one into another under the action of the global symmetry. Suppose, for example, that the global symmetry is $G = U(1)$, that there is an order parameter $\phi(x)$ transforming according to

$$U(\omega)\phi(x)U(\omega)^{-1} = e^{i\omega}\phi(x), \tag{2.4}$$

and that there are degenerate vacua labeled by $\alpha \in [0, 2\pi)$ such that

$$\langle\alpha|\phi(x)|\alpha\rangle = ve^{i\alpha}. \tag{2.5}$$

Then a distinct superselection sector can be constructed on each $\alpha$ vacuum; the $\alpha$-sector is spanned by smeared polynomials in local fields acting on the $\alpha$-vacuum.

Physically, the sector can change only if the phase of $\phi$ is rotated throughout an infinite spatial volume, and no local observable can accomplish this task. Because of the superselection rule, the degeneracy associated with a spontaneously broken symmetry cannot be directly detected by any local observer*.

Let us return now to the case of a local symmetry, and consider the fate of the charge superselection rule of QED when the theory is in the Higgs phase [7,8]. Suppose that the Higgs realization of the U(1) gauge symmetry is driven by the condensation of a scalar field $\phi$ that carries charge 1 (in units of $e$); that is, in the unitary gauge, $\phi$ has a nonvanishing vacuum expectation value. If all charged fields carry charge 1, then no nontrivial superselection rule is expected to survive in the Higgs phase. For example, let $\psi$ be another charge-1 scalar field. Then a $\psi$ excitation can be annihilated by the gauge-invariant operator $\phi^\dagger\psi$, which in unitary gauge becomes

$$\phi^\dagger\psi = v\psi + \ldots . \tag{2.6}$$

In the Higgs phase, since there is just one superselection sector, all states must transform trivially under global gauge transformations; in other words, all states have charge $Q = 0$. Physically, the electric charge of an arbitrary state is completely screened by the Higgs condensate. Charged particles no longer induce any effect that can be detected at spatial infinity. (One might prefer to say that the charge $Q$ is ill defined because of copious vacuum charge fluctuations. We find it more instructive to say that $Q$ vanishes, although defining $Q$ is a delicate matter. We elaborate on this viewpoint in sects. 3 and 4.)

More interesting than the case in which the gauge symmetry is completely broken is the case discussed in ref. [1], in which a discrete subgroup of the gauge group remains manifest. Consider now, for example, a U(1) gauge theory that contains a charge-$N$ field of $\eta$ and a charge-1 field $\phi$. Then, if $\eta$ condenses but $\phi$ does not, there is a surviving $Z_N$ gauge symmetry under which $\phi$ transforms as

$$U\left(\frac{2\pi k}{N}\right)\phi U\left(\frac{2\pi k}{N}\right)^{-1} = e^{2\pi i k/N}\phi , \qquad k = 0,1,\ldots,N-1 . \tag{2.7}$$

In this theory, charge screening is incomplete, because, as we will describe in more detail in sect. 3, the charge modulo $N$ of a state is not screened by the Higgs condensate. Thus, there is a nontrivial superselection rule. The Hilbert space decomposes into $N$ sectors labeled by the charge $Q$ mod $N$, with states of charge

* Loosely speaking, the $\alpha$-vacuum is degenerate with a zero-momentum Goldstone boson in the $\alpha$-sector. Technically, though, the ground state of the $\alpha$-sector is unique, since a momentum eigenstate is not normalizable.

$Q$ transforming under $Z_N$ gauge transformations according to

$$U\left(\frac{2\pi k}{N}\right)|Q\rangle = e^{2\pi i k Q/N}|Q\rangle . \tag{2.8}$$

Each sector is preserved by the gauge-invariant local observables.

Physically, the superselection rule arises because a $Z_N$ charge induces quantum-mechanical effects that can be detected at long range. In particular, this theory contains a stable cosmic string such that the U(1) gauge field far from the string satisfies

$$\exp\left(ie\oint_C A\cdot \mathrm{d}x\right) = e^{2\pi i/N}, \tag{2.9}$$

when integrated over a closed loop C that encloses the string. (The string encloses magnetic flux $\Phi_0/N$, where $\Phi_0 = 2\pi/e$ is the flux quantum and $e$ is the gauge coupling.) Therefore, the string may be used to measure the $Z_N$ charge of a low-energy projectile in a quantum interference experiment [2,3].

The examples described here illustrate that a superselection rule arises in a gauge theory whenever certain states are endowed with properties that cannot be destroyed by any process that is local in spacetime. In this sense, superselection rules provide a classification of the long-range "hair" that an excitation may carry.

The above discussion can be extended to the case in which the underlying gauge group G and the discrete subgroup H of G are nonabelian groups. If H is nonabelian, however, specifying the H charge involves subtleties that we will discuss in sects. 7 and 8.

## 3. The charge operator

We will now examine in more detail how the charge of a state may be defined in a gauge theory with a discrete gauge group. For the sake of concreteness, we continue to consider the model described in sect. 2, in which a U(1) gauge symmetry is broken to $Z_N$. This model is easily generalized.

Defining the $Z_N$ charge is delicate, because the $\eta$-condensate causes the electric field of a charge to decay as $e^{-mR}$, where $R$ is the distance from the charge and $m$ is the photon mass. Thus, the $Z_N$ charge cannot be detected classically; it is detectable at long range only through quantum-mechanical effects that survive in spite of the screening of the classical electric field.

It was suggested in ref. [1] that the $Z_N$ charge $Q_\Sigma$ contained within a closed surface $\Sigma$ can be expressed via the Gauss law in terms of a surface integral

$$\exp\left(\frac{2\pi i}{N}Q_\Sigma\right) = \exp\left(\frac{2\pi i}{Ne}\int_\Sigma E\cdot \mathrm{d}s\right) \equiv F(\Sigma) . \tag{3.1}$$

Heuristically, one might expect eq. (3.1) to make sense in spite of the exponential decay of the *classical* electric field, since the "flux operator" $F(\Sigma)$ is unable to see an object that carries charge $N$, and so should be unaffected by the $\eta$-condensate. The operator $U(2\pi/N)$ that represents a $Z_N$ gauge transformation could then be defined as the limit of $F(\Sigma)$ as the surface $\Sigma$ tends to spatial infinity. As we will see, this suggestion is nearly correct, but requires careful interpretation.

Eq. (3.1) is not precisely correct as it stands because quantum mechanical fluctuations cause the expectation value of $F(\Sigma)$ in a state $|\psi\rangle$ to decay rapidly as the size of $\Sigma$ increases. We have

$$\lim_{A(\Sigma)\to\infty} |\langle\psi|F(\Sigma)|\psi\rangle| \sim \exp[-\kappa A(\Sigma)], \tag{3.2}$$

where $A(\Sigma)$ denotes the area of the surface $\Sigma$, and $\kappa$ is a universal constant, independent of the state $|\psi\rangle$. This decay of $\langle F(\Sigma)\rangle$ is due to the quantum fluctuations of the charge-1 field $\phi$. Virtual pairs of charge-1 particles and antiparticles near the surface $\Sigma$ cause the charge contained inside $\Sigma$ to fluctuate. Furthermore, because the photon is massive, the charge fluctuations near two elements of $\Sigma$ become very weakly correlated when the elements are distantly separated. The finite correlation length thus gives rise to the characteristic "area-law" decay in eq. (3.2), and also ensures that the coefficient $\kappa$ does not depend on $|\psi\rangle$. The area-law behavior of $\langle F(\Sigma)\rangle$ occurs whether or not the local $Z_N$ symmetry is spontaneously broken (by condensation of $\phi$).

The cases of manifest and broken local $Z_N$ symmetry can be distinguished, however. Although the *modulus* of $\langle F(\Sigma)\rangle$ exhibits area-law decay, its *phase* does not get screened if the $Z_N$ symmetry remains manifest. Thus, eq. (3.1) can be salvaged by a mere multiplicative rescaling of $F(\Sigma)$. And since $\kappa$ is universal, we can isolate the phase of $F(\Sigma)$ by dividing by its vacuum expectation value. Thus, we may identify

$$\lim_{A(\Sigma)\to\infty} F(\Sigma)/\langle F(\Sigma)\rangle = \exp\left(\frac{2\pi i}{N}Q\right), \tag{3.3}$$

as the operator that represents a global $Z_N$ transformation*. In the remainder of this section and in the following two sections, we will discuss some of the properties of $F(\Sigma)$, in order to further elucidate the above arguments**.

We observe first of all that $F(\Sigma)$ may be regarded as a gauge transformation that has a discontinuity by the element $\exp(2\pi i/N) \in Z_N$ on the surface $\Sigma$. Since

* A similar definition of charge was proposed in a related context by Fröhlich and Marchetti [9].

** In following this discussion, and in interpreting the figures, the reader may find it helpful to think about the case of three-dimensional euclidean space; in that case $\Sigma$ is a one-dimensional closed loop.

$F(\Sigma)$ may be expressed as

$$F(\Sigma) = \exp\left(\frac{2\pi i}{N}\int_{\Omega} \mathrm{d}^3 x J^0\right), \tag{3.4}$$

where $\Omega$ is the region enclosed by $\Sigma$, we see that $F(\Sigma)$ rotates the phase of a charge-$q$ matter field $\Phi_q$ according to

$$F(\Sigma)\Phi_q(x)F(\Sigma)^{-1} = \begin{cases} \mathrm{e}^{2\pi i q/N}\,\Phi_q(x), & x \in \Omega, \\ \Phi_q(x), & x \notin \Omega. \end{cases} \tag{3.5}$$

Except on $\Sigma = \partial\Omega$, then, $F(\Sigma)$ is a mere gauge transformation that acts trivially on physical states. But because the discontinuity on $\Sigma$ in eq. (3.5) is not accompanied by any corresponding twist in the gauge field, $F(\Sigma)$ is not a gauge transformation everywhere. On $\Sigma$, it creates a *physical* discontinuity in any matter fields that carry nontrivial $Z_N$ charge. As far as the charge-$N$ $\eta$ field is concerned, $F(\Sigma)$ is a smooth local gauge transformation, and so $F(\Sigma)$ is incapable of detecting the fluctuations of $\eta$. But the charge-1 $\phi$ field can see the discontinuity at $\Sigma$. $F(\Sigma)$ therefore detects the fluctuations of $\phi$, giving rise to the area-law decay of $\langle F(\Sigma)\rangle$ as described above.

(Now that we recognize $F(\Sigma)$ as a gauge transformation with a discontinuity on $\Sigma$, it is possible to generalize $F(\Sigma)$ to the case of an arbitrary underlying (nonabelian) gauge group G. However, $F(\Sigma)$ is a gauge-invariant operator only if the discontinuity takes values in the center of the group. The nonabelian case will be further discussed in sect. 7.)

It will also be useful to have a euclidean path integral representation of the operator $F(\Sigma)$. In path-integral language, a correlation function with an insertion of $F(\Sigma)$ is computed by summing over all field configurations such that the gauge potential $A_\mu$ has a "string" singularity on the surface $\Sigma$. The string carries "magnetic" flux $2\pi/Ne$; that is

$$\exp\left(ie\oint_{\mathrm{C}} A\cdot \mathrm{d}x\right) = \mathrm{e}^{2\pi i/N}, \tag{3.6}$$

where C is an infinitesimal closed loop that encloses the string. (Note that in four-dimensional euclidean space, a two-dimensional surface $\Sigma$ has co-dimension two, and so may be enclosed by a one-dimensional loop.)

By means of a singular gauge transformation, we may give an alternative interpretation to the restriction expressed in eq. (3.6) – when $F(\Sigma)$ is inserted, the matter field configurations included in the path integral are restricted to a class

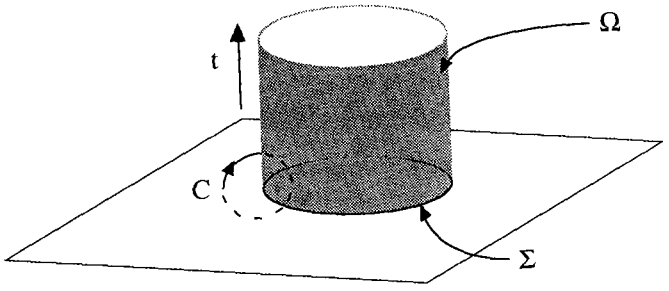


Fig. 1. An insertion of the flux operator $F(\Sigma)$ introduces a $Z_N$ discontinuity on a hypersurface $\Omega$ whose boundary is $\Sigma$. The gauge field has a $Z_N$ winding number on a closed loop C that links $\Sigma$.

that have a $Z_N$ discontinuity on a (co-dimension-one) hypersurface $\Omega$ that terminates on $\Sigma$. A (singular) gauge transformation can move this hypersurface, and $\Omega$ is therefore unphysical. But gauge transformations cannot move the boundary $\Sigma$ of $\Omega$, since $\Sigma$ may be identified by the gauge-invariant criterion equation (3.6).

By this means, we define $F(\Sigma)$ for an arbitrary closed two-surface $\Sigma$. But to establish the connection with the discontinuous gauge transformation considered previously, we suppose that $\Sigma$ lies in the time slice $t = 0$. Then, by a suitable gauge choice, we may choose the hypersurface $\Omega$ bounded by $\Sigma$ to lie in the region $t \geqslant 0$, as indicated in fig. 1. Because the $Z_N$ discontinuity on $\Omega$ is absent for $t < 0$ and present at $t > 0$, we can interpret $F(\Sigma)$ as an operator that creates a $Z_N$ discontinuity on $\Sigma$. Thus, for a two-surface $\Sigma$ lying in a time slice, our path integral formulation of $F(\Sigma)$ does indeed coincide with the operator $F(\Sigma)$ defined by eq. (3.1).

Of course, whether we use the operator language or the path-integral language, our definition of $F(\Sigma)$ is so far merely formal. To define it rigorously we will need a short-distance regulator that smooths the discontinuity on $\Sigma$. We will describe a lattice regularization of $F(\Sigma)$ in sect. 5.

If the surface $\Sigma$ does not lie on a time slice, then $F(\Sigma)$ has a different, and interesting, interpretation. A generic two-surface $\Sigma$ intersects a time slice on a one-dimensional closed loop, and eq. (3.6) tells us that this loop is the position of a magnetic flux tube that carries flux $\Phi = 2\pi/Ne$. An insertion of $F(\Sigma)$ in the path integral, then, is equivalent to introducing a cosmic string source that propagates on the euclidean worldsheet $\Sigma$ (fig. 2).

This observation allows us to connect the operator $F(\Sigma)$ with the thought experiment described in ref. [1]. It was emphasized there that the $Z_N$ charge of an object can be measured by scattering the object off a cosmic string that carries magnetic flux $\Phi = 2\pi/Ne$; at low energy, the scattering cross-section is dominated by the Aharonov–Bohm effect [2, 3]. Equivalently, we can imagine measuring the $Z_N$ charge contained inside a surface $\Sigma$ by adiabatically winding a cosmic string

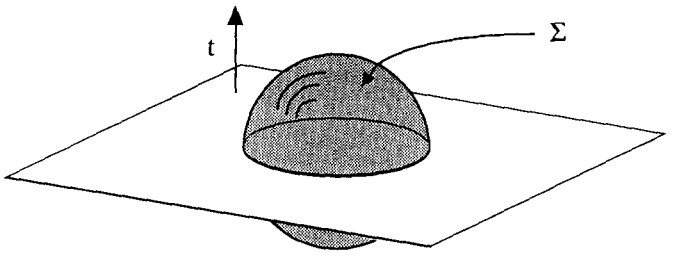


Fig. 2. $\Sigma$ may be regarded as the worldsheet of a loop of cosmic string.

around $\Sigma$ (fig. 3). The string acquires an Aharonov–Bohm phase $\exp(2\pi iQ/N)$ if $Q$ units of $Z_N$ charge are enclosed. The operator $F(\Sigma)$ embodies this thought experiment, since $F(\Sigma)$ picks up a phase $e^{2\pi i/N}$ from each $Z_N$ charge whose world line crosses the hypersurface $\Omega$ that is bounded by $\Sigma$ (fig. 4).

One may gain a greater appreciation of the interpretation of the operator $F(\Sigma)$ by contemplating the connection between $F(\Sigma)$ and the 't Hooft loop operator $B$(C) defined in ref. [10]. In the U(1) gauge theory that we have been considering here, the 't Hooft loop operator $B$(C), acting in a time slice, creates a cosmic string on the loop C that carries magnetic flux $\Phi = 2\pi/Ne$. Alternatively, we may think of this operator as a gauge transformation that has a discontinuity by $e^{2\pi i/N} \in Z_N$ on an open surface $\Sigma$ such that the boundary of $\Sigma$ is C. If the charge-$N$ field $\eta$ were the only charged field in the theory, then this discontinuity would be a pure gauge artifact, and the operator $B$(C) would be independent of the choice of the surface $\Sigma$ whose boundary is C. But if there is also a charge-1 field $\phi$, the discontinuity on $\Sigma$ is not merely a gauge artifact; thus $B$ depends on $\Sigma$ as well as C and should be denoted $B(C, \Sigma)$. (In the case studied in ref. [10], $B$(C) was dependent on C alone.) We thus recognize that the flux operator $F(\Sigma)$ is a degenerate case of the 't Hooft operator $B(C, \Sigma)$; $F(\Sigma)$ is obtained by shrinking the loop C to a point, so that the open surface $\Sigma$ becomes a closed surface (fig. 5).

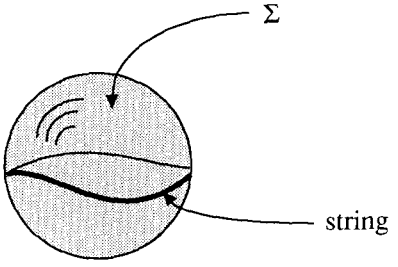


Fig. 3. A cosmic string that adiabatically winds around $\Sigma$ can detect the charge inside.

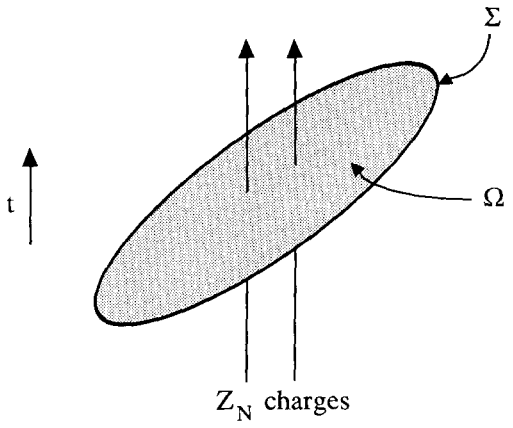


Fig. 4. Spacetime view of the detection procedure. A charge is detected if its worldline crosses the hypersurface $\Omega$ that is bounded by $\Sigma$.

In terms of the 't Hooft loop, our formal procedure for measuring the $Z_N$ charge enclosed by the surface $\Sigma$ may be described in an alternative language. For the sake of this discussion, let us imagine that all $Z_N$ charges are classical sources, so that there are no quantum fluctuations of the charge. In that event, it is not absolutely essential to introduce a surface $\Sigma$ in order to specify the operator $B(C, \Sigma)$, because the quantum fields do not see the discontinuity on $\Sigma$. However, unless we introduce such a surface, the 't Hooft loop operator $B(C)$ in the presence of classical $Z_N$ sources is typically an $N$-valued object; because of the Aharonov–Bohm effect, it acquires a phase $e^{2\pi i/N}$ upon winding around a unit $Z_N$ charge. Since multivalued objects are awkward to deal with, we may prefer to force $B(C)$ to be single-valued by arbitrarily restricting it to one of its $N$ branches. The price of single-valuedness is that $B(C)$ has a cut; we may, for example, select an arbitrary surface $\Sigma$ bounded by C and specify that the value of $B(C, \Sigma)$ jump

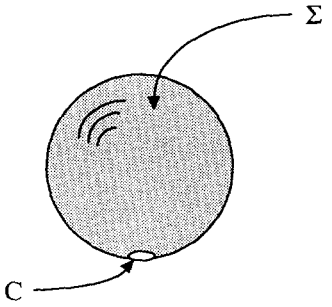


Fig. 5. The 't Hooft operator $B(C, \Sigma)$ becomes the flux operator $F(\Sigma)$ as the loop C shrinks to a point.

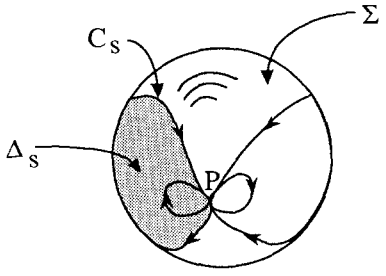


Fig. 6. A sequence of closed loops, all containing the common point P, that covers $\Sigma$.

discontinuously by the multiplicative factor $e^{2\pi i/N}$ whenever a unit $Z_N$ charge crosses $\Sigma$ (in the appropriate sense). Although $B(C, \Sigma)$ will then be single-valued, its phase will depend on the choice of the surface $\Sigma$.

Now, suppose that we wish to measure the $Z_N$ charge inside a closed surface $\Sigma$. If $\Sigma$ has the topology of a sphere, then we may construct a sequence of loops $C_s$, $s \in [0, 1]$, each containing a common point P, that sweep through the surface $\Sigma$. The sequence begins and ends with an infinitesimal loop at P (fig. 6). If $B(C_s)$ is the multivalued 't Hooft loop operator with no cuts, then $B(C_1)$ differs from $B(C_0)$ by the phase $\exp(2\pi i Q_\Sigma/N)$, where $Q_\Sigma$ is the enclosed $Z_N$ charge. Alternatively, we may consider the sequence of single-valued operators $B(C_s, \Delta_s)$ where $\Delta_s$ is a surface bounded by $C_s$ that lies on $\Sigma$. Since $\Delta_s$ never crosses any of the $Z_N$ charges enclosed by $\Sigma$, $B(C_1, \Delta_1 = \Sigma)$ and $B(C_0, \Delta_0 = 0)$ also differ by the phase $\exp(2\pi i Q_\Sigma/N)$. But it is evident that

$$B(C_0, \Sigma) = F(\Sigma) B(C_0, 0) . \tag{3.7}$$

As the loop C sweeps out the surface $\Sigma$, it drags behind it the surface on which the 't Hooft loop has its cut. When C shrinks back to a point, only the cut on $\Sigma$ is left behind. So we may think of $F(\Sigma)$ as a measure of the multivaluedness of the 't Hooft loop (at least when there are no quantum-mechanical charge fluctuations).

Expressed in this language, our procedure for measuring the $Z_N$ charge is strongly reminiscent of a closely analogous procedure that has been discussed previously. Coleman [11] and Srednicki and Susskind [12] considered the problem of measuring a $Z_N$ magnetic charge in the confining phase of an $SU(N)$ Yang–Mills theory. They noted that the confining phase has a mass gap, so that magnetic fields are screened and magnetic charge cannot be detected classically. But they claimed that the $Z_N$ magnetic charge induces quantum-mechanical effects that can be detected at long range. Specifically, they noted that the Wilson loop operator is multivalued in the presence of $Z_N$ magnetic charges, and that the $Z_N$ magnetic charge enclosed by a surface $\Sigma$ can be detected as the phase acquired by a Wilson loop that winds around $\Sigma$. (Monopoles were treated as classical sources in refs.

[11] and [12], and no attempt was made to take into account quantum-mechanical magnetic charge fluctuations.)

This analogy between the detection of $Z_N$ electric charge and the detection of $Z_N$ magnetic charge can be made precise. In 3 + 1 dimensions, a duality transformation can be formulated [10] that interchanges electric and magnetic charge, and hence also interchanges the Wilson loop operator and the 't Hooft loop operator. Duality thus relates confining behavior, in which the expectation value of the Wilson loop decays according to an area law and magnetic fields are screened, to Higgs behavior, in which the expectation value of the 't Hooft loop decays according to an area law and electric fields are screened. In a confining theory, the *phase* of a large Wilson loop is able to respond to the long-range field of a magnetic monopole, even though magnetic screening causes the field to decay exponentially, just because confinement also causes the *modulus* of the Wilson loop to decay rapidly [11]. Likewise, the rapid decay of the modulus of the 't Hooft loop in a Higgs theory enables the phase of the 't Hooft loop to respond to the weak long-range field of an electric charge.

We see, then, that the detection of a $Z_N$ electric charge by an 't Hooft loop and of a $Z_N$ magnetic charge by a Wilson loop involve essentially the same mathematics. In both cases, the crucial feature is that a unit $Z_N$ electric charge can see the Dirac string of a $Z_N$ magnetic charge through a nontrivial Aharonov–Bohm effect.

In fact, the argument of refs. [11] and [12] shows that black holes can in principle carry magnetic quantum-mechanical hair, as well as the electric quantum-mechanical hair proposed in ref. [1]. In an SU($N$) gauge theory, $Z_N$ magnetic monopoles are consistent with the Dirac quantization condition only if all matter fields are invariant under $Z_N$, the center of the group. These monopoles are confined by magnetic flux tubes in the Higgs phase, but in the confinement phase there is a nontrivial $Z_N$ magnetic charge superselection rule. The $Z_N$ magnetic charge of an object can be detected at long range, because a monopole has an Aharonov–Bohm interaction with an electric flux tube. (Were we to push the analogy with the electric superselection rule even further, we would distinguish two types of confinement phase. If $Z_N$ monopoles condense, then $Z_N$ magnetic charge is screened even quantum mechanically, and the electric flux tube becomes the boundary of a $Z_N$ domain wall [13].)

However, in the realistic case of an $[SU(3)_{color} \times U(1)_{em}]/Z_3$ gauge theory, we may not legitimately speak of quantum-mechanical hair. Though magnetic monopoles carry a $Z_3$ color magnetic flux, there are no stable electric flux tubes, and hence no means of detecting quantum-mechanical hair. The $U(1)_{em}$ magnetic charge is the only magnetic quantum number that can be detected at long range, and the magnetic hair is entirely classical. If black holes can carry magnetic quantum-mechanical hair in Nature, then, this hair is not associated with the known strong interaction; rather, it must be associated with another, as yet unknown, confining gauge interaction that admits genuine $Z_N$ monopoles.

Incidentally, much as the $Z_N$ magnetic monopole number respects a superselection rule in a confining (3 + 1)-dimensional gauge theory, so the $Z_N$ vortex number respects a superselection rule in a (2 + 1)-dimensional theory with manifest local $Z_N$ symmetry. The $Z_N$ vortex number can be detected by Aharonov–Bohm scattering off a $Z_N$ electric charge. Although a gauge-invariant operator can be constructed that annihilates a vortex, this operator is not local; it has a semi-infinite string that can be seen by matter fields with nonvanishing $Z_N$ charge. This string is the (2 + 1)-dimensional analog of the surface that, in 3 + 1 dimensions, stretches across the 't Hooft loop. If the local $Z_N$ symmetry is spontaneously broken, then the vortices are confined. Hence, neither the vortex superselection rule nor the charge superselection rule survives.

We have now formulated the notion of quantum-mechanical hair in a reasonably precise language. To conclude this discussion, we wish to compare the quantum-mechanical hair associated with a local discrete symmetry to another exotic type of hair that was proposed recently. Bowick et al. [14] considered a theory in which a *global* U(1) symmetry is spontaneously broken. Such a theory contains an exactly massless Goldstone boson, the *axion*, and also a topological defect, the *axion string*. An axionic charge operator can be defined, and an object that carries this charge exhibits a nontrivial Aharonov–Bohm effect with an axion string. By means of this Aharonov–Bohm effect, axionic charge can in principle be detected at long range; thus, axionic charge is a type of hair.

Since axionic charge is detected via an Aharonov–Bohm interaction with a *global* string, much as $Z_N$ electric charge is detected via an Aharonov–Bohm interaction with a *gauge* string, these two types of hair appear to be related. Actually, axionic charge is more akin to the $Z_N$ vortex number that arises in a discrete gauge theory in 2 + 1 dimensions than to $Z_N$ electric charge. To see this connection more clearly, it is helpful to recognize that the massless axion field $\Theta$ is dual to a three-form field strength $H$ defined by

$$\frac{H}{2\pi} = f^2 * \mathrm{d}\Theta , \tag{3.8}$$

where $*$ denotes the Hodge dual. (We have normalized $\Theta$ so that it is a dimensionless periodic variable with period $2\pi$; $f$ is the mass scale that characterizes the spontaneous breakdown of the global U(1) symmetry.) This field strength $H$ can be expressed as the curl of a two-form potential $B$,

$$H = \mathrm{d}B , \tag{3.9}$$

and the axionic charge $q$ in a volume $\Omega$ that is enclosed by the surface $\Sigma$ can be

expressed as a surface integral

$$q = \frac{1}{2\pi}\int_{\Omega} H = \frac{1}{2\pi}\int_{\Sigma} B. \tag{3.10}$$

The quantity $q$ becomes more recognizable when we re-enact this duality transformation in 2 + 1 dimensions. Then the axion field is dual to an *electromagnetic* two-form $F$, and the axionic charge in a region $\Sigma$ that is enclosed by the loop C is

$$q = \frac{e}{2\pi}\int_{\Sigma} F = \frac{e}{2\pi}\int_{C} A; \tag{3.11}$$

it is just the magnetic flux in $\Sigma$, in units of the flux quantum. Furthermore, axionic charge is detected in 2 + 1 dimensions via an Aharanov–Bohm interaction with an axion *vortex*, and the vortex is transformed under duality into an electrically charged particle. Duality, then, maps the detection of axionic charge by an axion vortex, in 2 + 1 dimensions, to the detection, in (2 + 1)-dimensional electrodynamics, of magnetic flux with an electrically charged particle. (As in electrodynamics, the Aharonov–Bohm interaction is only sensitive to the axionic charge modulo an integer.)

It is also enlightening to consider the fate of the axionic hair when the dual electrodynamics is in a phase other than the Coulomb (massless) phase. For example, we may include finite-action magnetic monopole configurations in the euclidean path integral of (2 + 1)-dimensional electrodynamics [15]. Then the theory acquires a mass gap, and electric charges become confined by stable electric flux tubes. In the dual description in terms of the axion field, the magnetic monopoles break the global U(1) symmetry intrinsically [13]. Hence the axion acquires a nonzero mass and the axion vortex becomes the boundary of an axion domain wall; the domain wall is dual to the electric flux tube. In this confining phase, it is inappropriate to regard axionic charge as a type of hair that can be detected at long range. Instead, the axion vortex detects the axionic charge by dragging a domain wall across the charge.

If (2 + 1)-dimensional electrodynamics is in the Higgs phase, then there is a Meissner effect, and, consequently, magnetic flux is quantized. The axionic charge, or magnetic flux, contained in an isolated region cannot assume an arbitrary value, but must be an integer multiple of the flux quantum. If the electromagnetic gauge invariance is *completely* broken, then a quantum of magnetic flux cannot be detected via the Aharonov–Bohm effect. But if there is a surviving $Z_N$ local symmetry, then there is a $Z_N$ vortex number that *can* be detected at long range. In this sense, $Z_N$ vortex hair is the remnant of axionic hair that may survive the Higgs mechanism, in 2 + 1 dimensions.

In the dual description in terms of the axion field, the Higgs mechanism corresponds to the *restoration* of the global U(1) symmetry due to the condensation of axion vortices [13]. $Z_N$-valued axionic hair survives the Higgs mechanism if the vortex that condenses is not the minimal axion vortex with unit winding number, but rather a nonminimal vortex with winding number $N$.

This discussion of the $(2+1)$-dimensional cases is readily generalized to the detection of axionic charge by axion strings, in $3+1$ dimensions. If the axion is exactly massless, then the axionic charge $q$ may assume any value, and $q$ modulo an integer can be detected via the Aharonov–Bohm interaction of the charge with an axion string. If the U(1) global symmetry is intrinsically broken, then the axion has a nonzero mass and the axion string is the boundary of an axion domain wall. An axion string detects charge by dragging a domain wall across the region that contains the charge. If the U(1) global symmetry is manifest, then axion strings condense [16] and the axionic charge is quantized. If the strings that condense are nonminimal strings with winding number $N$, then axionic hair takes values in $Z_N$. The $Z_N$-valued axionic charge can be detected at long range by an axion string with winding number one.

It is convenient to describe axion physics in terms of the two-form potential $B$, because axionic hair can then be discussed in the language of classical field theory, whether or not the global U(1) symmetry is spontaneously broken. An object that carries axionic charge has a long-range $B$ field. Now, $B$ is not itself a gauge-invariant quantity, and the long-range $B$ field is actually a pure gauge locally, but the gauge-invariant axionic charge $q$ can be expressed as a surface integral of $B$, as in eq. (3.10). The statement that $B$ may be regarded as a "classical" field should not be misinterpreted. This statement means that quantum fluctuations of $B$ can be neglected. But $B$ is not a classical local observable, and the "topological" charge $q$ can be detected at long-range only via quantum-mechanical interference effects★.

This "classical" description of axionic hair is exploited by Bowick et al. [14] in their analysis of axionic black holes. They note that the classical field equations for $B$ coupled to Einstein gravity admit black hole solutions with nonzero axionic charge $q$. Furthermore, these axionic black holes obey a generalized uniqueness theorem. On a stationary black hole with a nonsingular event horizon, the potential $B$ is required to be a pure gauge locally, except at the singularity. Thus, the axionic charge $q$ is the sole physical attribute that characterizes the axion field of a stationary black hole [14].

A black hole that carries axionic charge contrasts sharply with a black hole that carries $Z_N$ electric charge. In classical field theory, black hole uniqueness theorems [17] require massive scalar and vector fields to vanish exactly on a stationary

★ In fact, a nonsingular gauge transformation on the surface $\Sigma$ can change $q$ by an integer; that is why only $q$ modulo an integer can be detected at long range.

black hole with a nonsingular event horizon. In a Higgs phase, then, a stationary black hole can carry no classical electric or scalar charge. Hence, the quantum-mechanical correlations that can be detected far from the black hole resist being encoded in any classical field-theoretic description. In this respect, $Z_N$ electric hair is a more subtle and elusive notion than axionic hair. The manifestation of these quantum-mechanical correlations during gravitational collapse, and their effect on the subsequent evaporation of a charged black hole, may be worthy of further study.

## 4. The order parameter

Our discussion in sect. 2 indicated that, in the U(1) gauge theory with charge-$N$ field $\eta$ and charge-1 field $\phi$, there are two distinct Higgs realizations of the gauge symmetry. These two realizations may be distinguished according to whether a $Z_N$ subgroup of U(1) remains manifest, or in other words, whether the theory respects a $Z_N$ superselection rule. If both realizations can be achieved for a suitable choice of the parameters in the model, then we expect that there are two distinct Higgs phases, separated by a phase boundary. In this section, we propose an order parameter that is sensitive to such a phase transition. Our order parameter is readily generalized to one that probes the realization of an arbitrary abelian discrete gauge symmetry.

Heuristically, whether the charge-1 field $\phi$ "condenses" determines whether the $Z_N$ symmetry is manifest or spontaneously broken. But Elitzur's theorem [18] cautions us that a gauge non-invariant local order parameter is unable to reveal the nontrivial phase structure. Our order parameter must instead be a gauge-invariant and nonlocal object.

It is familiar that such nonlocal order parameters can distinguish the Higgs realization from the confining realization of a gauge symmetry. Physically, the confining phase supports stable electric flux tubes and the Higgs phase supports stable magnetic flux tubes. (Both phases are distinguished from the Coulomb phase in that there is a mass gap, and, in the case of a Higgs phase in which the gauge symmetry is *completely* broken, no nontrivial charge superselection rule.) Mathematically, appropriate order parameters are the Wilson loop [19] and 't Hooft loop [10] operators.

Suppose, for example, that the gauge group is $G = SU(N)$ and that all of the fields in the theory transform trivially under the center $Z_N$ of $SU(N)$. The Wilson loop

$$W(C) = \mathrm{tr}\left[ P \exp\left( ig \oint_C A \cdot \mathrm{d}x \right) \right] \tag{4.1}$$

may be regarded as an insertion of a classical source, transforming as the defining

representation of SU($N$), that propagates along the worldline C. If electric flux is confined, then C becomes the boundary of the worldsheet of an electric flux tube; for sufficiently large loops, $W$(C) therefore exhibits the area-law behavior

$$\langle W(\mathrm{C})\rangle \sim \exp[-\kappa A(\mathrm{C})]\,, \tag{4.2}$$

where $A$(C) is the minimal area of a surface bounded by C, and $\kappa$ is the string tension. In the Higgs phase, electric flux is screened, and the Wilson loop has the perimeter-law behavior

$$\langle W(\mathrm{C})\rangle \sim \exp[-\mu P(\mathrm{C})]\,, \tag{4.3}$$

where $P$(C) is the length of C. Conversely, the 't Hooft loop $B$(C) may be regarded as an insertion of a classical $Z_N$ magnetic monopole source that propagates along the world line C. (That is, C is the boundary of a $Z_N$ Dirac string, as described in sect. 3.) In the Higgs phase, magnetic flux is confined and $B$(C) has area-law behavior, while in the confinement phase, magnetic flux is screened and $B$(C) has perimeter-law behavior.

But if an SU($N$) gauge theory is coupled to "quark" matter in the defining representation of SU($N$) (or in any representation that transforms faithfully under the center $Z_N$), then the confining and Higgs realizations can no longer be distinguished by the above criteria. Quark–antiquark pairs appear as quantum fluctuations, allowing the electric flux tube to break. The Wilson loop therefore always obeys the perimeter law. Because quarks transform nontrivially under $Z_N$, the 't Hooft loop becomes $B(\mathrm{C}, \Sigma)$; it depends on the choice of the surface $\Sigma$ bounded by C as discussed in sect. 3, and always obeys the area law.

With $W$(C) and $B$(C) failing to distinguish the Higgs and confining phases, one recognizes that no such distinction may be possible when "quarks" are present; there need be no sharp phase boundary that separates the Higgs and confining regions of the phase diagram [10, 20, 21]. However, we have also argued that two types of Higgs phases are possible, depending on the realization of the local $Z_N$ symmetry. It is presumably the Higgs phase with spontaneously broken $Z_N$ symmetry, and hence no nontrivial superselection rule, that is indistinguishable from the confining phase.

How, though, do we distinguish the two types of Higgs phases? Topological defects provide a potentially useful criterion. The Higgs phase with manifest $Z_N$ gauge symmetry supports cosmic strings that carry one unit of $Z_N$ magnetic flux. But when the local $Z_N$ symmetry is spontaneously broken, such a string becomes the boundary of a domain wall [22–24]. This observation might tempt one to propose that the behavior of $\langle F(\Sigma)\rangle$ distinguishes the two phases. We have seen that $F(\Sigma)$ may be regarded as an insertion of a classical $Z_N$ cosmic-string source propagating on a worldsheet $\Sigma$; if $\Sigma$ becomes the boundary of a domain wall, then

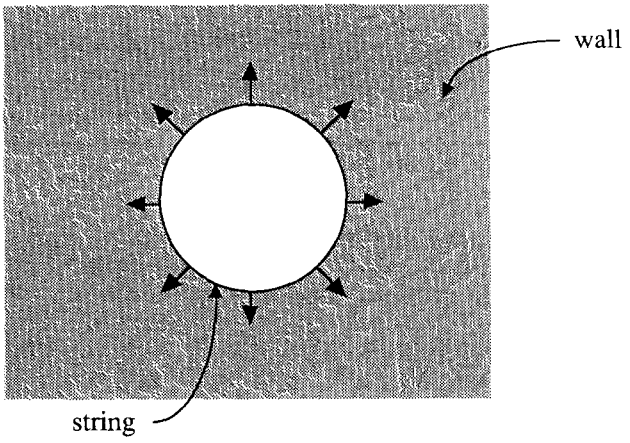


Fig. 7. Spontaneous nucleation and subsequent expansion of a loop of string causes a domain wall to decay.

we might expect $\langle F(\Sigma)\rangle$ to decay like exp[−Volume]. Unfortunately, this proposal does not quite work, because the domain wall bounded by $\Sigma$ is not absolutely stable. A hole in the wall, bounded by a $Z_N$ string, can arise as a quantum fluctuation. A sufficiently large hole will grow catastrophically, devouring the wall (fig. 7). Thus, even if the wall is very long-lived, $\langle F(\Sigma)\rangle$ will always decay like exp[−Area] for sufficiently large surfaces and there is no need for $\langle F(\Sigma)\rangle$ to behave non-analytically at the boundary between phases with manifest and broken local $Z_N$ symmetry.

The preferred way to distinguish the two phases is by means of the $Z_N$ charge superselection rule, as described in sect. 2. If the local $Z_N$ symmetry is manifest, then $Z_N$ electric charge is screened classically but not quantum mechanically; $Z_N$ charge can be detected at long range via the Aharonov–Bohm effect. If the local $Z_N$ symmetry is spontaneously broken, then $Z_N$ electric charge is completely screened. We may imagine introducing a classical source of $Z_N$ charge at the origin and then attempting to detect the charge at spatial infinity. The charge is detectable in principle if and only if the local $Z_N$ symmetry is manifest.

To restate this criterion mathematically, we recall that a Wilson loop operator $W(C)$ acts as a source of $Z_N$ charge, and that the flux operator $F(\Sigma)$ can detect $Z_N$ charge, as described in sect. 3. If we define

$$A(\Sigma, C) = \frac{\langle F(\Sigma)W(C)\rangle}{\langle F(\Sigma)\rangle\langle W(C)\rangle}, \tag{4.4}$$

then, if $Z_N$ electric charge is unscreened, we have

$$\lim A(\Sigma, C) = \exp\left[\frac{2\pi i}{N}k(\Sigma, C)\right]. \tag{4.5}$$

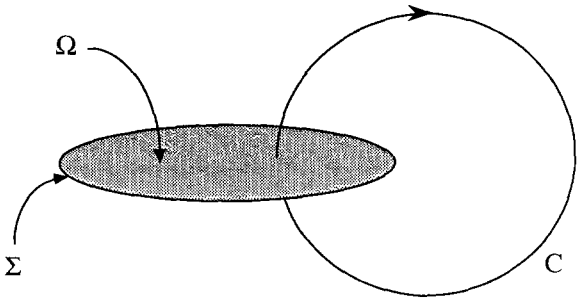


Fig. 8. A closed loop C and closed surface $\Sigma$ with linking number 1.

Here the limit is taken with $\Sigma$ and C increasing to infinite size, and with the closest approach of $\Sigma$ to C also approaching infinity; $k(\Sigma, \mathrm{C})$ denotes the (integer-valued) linking number of the surface $\Sigma$ and loop C. That is, $k(\Sigma, \mathrm{C})$ is the (signed) number of times the world line C crosses a volume $\Omega$ that is bounded by $\Sigma$ (fig. 8). If, however, $Z_N$ electric charge is screened, then

$$\lim A(\Sigma, \mathrm{C}) = 1 . \tag{4.6}$$

The nonanalytic behavior of $A(\Sigma, \mathrm{C})$ guarantees that the two Higgs phases are separated by a well-defined phase boundary.

Our order parameter $A(\Sigma, \mathrm{C})$ can obviously be generalized to probe the realization of any abelian local discrete symmetry. In fact, we will see in sect. 6 that $A(\Sigma, \mathrm{C})$ can be constructed without any explicit reference to a continuous gauge group in which the discrete gauge group is embedded.

## 5. Example: Lattice theory with coupled $Z_2$ gauge and $Z_2$ spin variables

Since the above discussion is rather abstract, one desires an explicit model in which the behavior of the order parameter $A(\Sigma, \mathrm{C})$ can be studied analytically. We now present such an example. Our model is a $Z_2$ lattice gauge theory coupled to a $Z_2$ spin system; the spin system plays the role of "matter" that transforms nontrivially under the local discrete $Z_2$ symmetry [20, 25, 26]. The virtue of this model is that both its weak and strong coupling behavior can be analyzed using convergent expansions. Thus, the phase diagram of the model can be mapped out using perturbation theory. We will indeed find a phase boundary that separates a phase with screened $Z_2$ charge from a phase with unscreened $Z_2$ charge. In spite of the simplicity of the example, we believe that it captures all of the essential features of the general case.

The degrees of freedom of the model are gauge variables

$$U_l \in Z_2 \equiv \{1, -1\} , \tag{5.1}$$

residing on links (labeled by $l$) of a cubic four-dimensional spacetime lattice, and spin variables

$$\phi_i \in Z_2 \equiv \{1, -1\}, \tag{5.2}$$

residing on sites (labeled by $i$). The euclidean action is

$$S = S_{\text{gauge}} + S_{\text{spin}},$$

where

$$S_{\text{gauge}} = -\beta \sum_P U_P, \tag{5.4}$$

and

$$S_{\text{spin}} = -\gamma \sum_l (\phi U \phi)_l. \tag{5.5}$$

Here $U_P = \prod_{l \in P} U_l$ associates with each elementary plaquette (labeled by $P$), the product of the four $U_l$'s associated with the links of the plaquette, and $(\phi U \phi)_{ij} = \phi_i U_{ij} \phi_j$, for each pair $ij$ of nearest neighbor sites. The action is invariant under the $Z_2$ gauge transformation defined by

$$\eta_i \in Z_2 \equiv \{1, -1\}, \tag{5.6}$$

where the variables transform as

$$\phi_i \to \eta_i \phi_i, \qquad U_{ij} \to \eta_i U_{ij} \eta_j. \tag{5.7}$$

Expectation values of gauge-invariant quantities are computed using the normalized probability measure $Z^{-1} e^{-S}$, where

$$Z = \sum_{\{U\}\{\phi\}} e^{-S}. \tag{5.8}$$

Notice that, under a nontrivial *global* gauge transformation $\eta_i = -1$, the gauge variable $U_l$ is invariant but the spin variable $\phi_i$ is not. This model is therefore a prototype of the phenomenon that we are investigating; the "matter" field $\phi$ carries the local discrete $Z_2$ charge, and we wish to determine whether there is a nontrivial $Z_2$ superselection rule.

Our model is tractable because it can be analyzed by means of convergent perturbation expansions if both coupling parameters $\beta$ and $\gamma$ are either large or small. We will show by applying the order parameter $A(\Sigma, C)$ that $Z_2$ charge is unscreened if $\beta$ is large and $\gamma$ is small, and that $Z_2$ charge is screened if either $\beta$ is small or $\gamma$ is large. This nonanalytic behavior of $A(\Sigma, C)$ establishes the existence of a phase boundary.

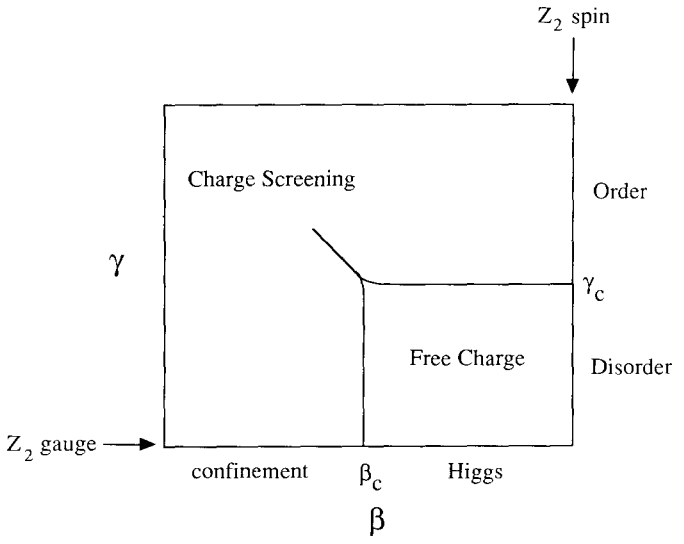


Fig. 9. Phase diagram of the $Z_2$ gauge-spin system.

It is easy to anticipate this conclusion if one contemplates the limiting behavior of the model on the boundaries of the phase diagram (fig. 9). The model becomes trivial in the limits $\beta = 0$ and $\gamma = \infty$. For $\beta = 0$, the $U_P$'s are completely unconstrained, and the $U_l$'s are therefore free to adjust so that $(\phi U \phi)_l = 1$ on every link, whatever the configuration of spins. For $\gamma = \infty$, $(\phi U \phi)_l = 1$ must be satisfied on every link, and hence $U_P = \prod_{l \in P} U_l = \prod_{l \in P} (\phi U \phi)_l = 1$ on every plaquette; there is no dependence on $\beta$.

More interesting are the limits $\beta = \infty$ and $\gamma = 0$. For $\beta = \infty$, the gauge variables are frozen at $U_l = 1$ (up to a gauge transformation) and the model reduces to a $Z_2$ Ising spin system. There is thus a second-order phase transition at a critical coupling $\gamma = \gamma_c$ on the $\beta = \infty$ axis; the spins are disordered for $\gamma < \gamma_c$ and ordered for $\gamma > \gamma_c$. For $\gamma = 0$, the spins decouple, and the model reduces to a $Z_2$ gauge system. There is thus a first-order phase transition (in four or more Euclidean dimensions) at a critical coupling $\beta = \beta_c$ on the $\gamma = 0$ axis; the gauge system is in the confining phase for $\beta < \beta_c$ and in the Higgs phase for $\beta > \beta_c$.

In the region of the phase diagram with $\beta$ large and $\gamma$ small, then, we expect a Higgs phase with disordered (uncondensed) spins. This region, in which $Z_2$ charge in unscreened, should be separated by a phase boundary from the large $\gamma$ and small $\beta$ regions. We therefore expect the phase diagram of the model to have the schematic form suggested in fig. 9.

This phase structure has been conjectured previously [20,25], and has been confirmed by Monte Carlo simulations [27]. Various attempts have been made to identify an order parameter that distinguishes the two phases. (These attempts are

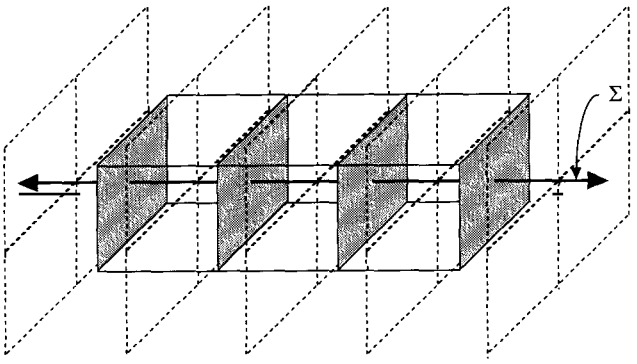


Fig. 10. The set of plaquettes (shaded) dual to a set $\Sigma$ of links (bold) of the dual lattice, in three euclidean dimensions. Dotted lines are links of the dual lattice, and solid lines are links of the original lattice. In four euclidean dimensions, the bold line should be interpreted as a slice through the surface $\Sigma$.

reviewed in ref. [7].) To our knowledge, though, no order parameter has been suggested before that probes *directly* as to whether $Z_2$ charge is screened, as $A(\Sigma, C)$ does. We will confirm the phase structure indicated in fig. 9 by computing $A(\Sigma, C)$ explicitly.

To begin, we must construct a lattice version of the flux operator $F(\Sigma)$. To do this, we consider $\Sigma$ to be a closed surface made up of plaquettes of the *dual* lattice; then each plaquette of $\Sigma$ is dual to a plaquette of the original lattice (fig. 10)*. To evaluate the path integral with an insertion of $F(\Sigma)$, we perform the transformation

$$U_P \to -U_P, \qquad P \in \Sigma^* \tag{5.9}$$

on these plaquettes that are dual to the plaquettes of $\Sigma$. That is, we flip the sign of $U_P$, or, equivalently, flip the sign of $\beta$ on these plaquettes, frustrating them. In effect, we place a unit of $Z_2$ "magnetic flux" on the surface $\Sigma$, so that $\Sigma$ can be regarded as the worldsheet of a $Z_2$ Dirac string**.

Before we proceed to the case of a $Z_2$ gauge system coupled to matter, consider first the pure gauge system. We noted that $F(\Sigma)$ is merely a gauge transformation in a pure gauge theory, and the surface $\Sigma$ is an artifact that can be moved by means of a singular gauge transformation. The lattice analog of a singular gauge transformation is the change of variable [28]

$$U_l \to -U_l. \tag{5.10}$$

* This construction is easier to visualize in three-dimensional euclidean space. Then $\Sigma$ is a closed *loop* made up of *links* of the dual lattice, and each link of $\Sigma$ is dual to a plaquette of the original lattice.

** This method of introducing a Dirac string on the lattice was used in ref. [28].

By performing this change of variable on a link contained in one of the plaquettes that is dual to $\Sigma$, we can move the surface $\Sigma$. In fact, $\Sigma$ is the boundary of a set of cubes of the dual lattice, and by performing $U_l \to -U_l$ on each of the links that are dual to the cubes of this set, we can shrink $\Sigma$ to a point*. So it is evident that $F(\Sigma)$ is a mere change of variable, and that

$$\langle F(\Sigma)\rangle = 1\,. \tag{5.10}$$

It is not the case, however, that $F(\Sigma)$ can be replaced by 1 when it is inserted in an arbitrary Green function. The Wilson loop on the lattice is defined as

$$W(\mathrm{C}) = \prod_{l\in \mathrm{C}} U_l\,, \tag{5.12}$$

where C is a closed loop of links. Now we easily see that

$$\langle F(\Sigma)W(\mathrm{C})\rangle = -\langle W(\mathrm{C})\rangle\,, \tag{5.13}$$

if the loop C and surface $\Sigma$ have linking number 1 (or any odd integer). This is because $F(\Sigma)$ is a change of variable that flips the sign of $U_l$ on an odd number of the links of C. Eq. (5.13) is the statement that $Z_2$ charge is not screened in the pure gauge system, or, equivalently, that a $Z_2$ cosmic string can be detected at long range by a $Z_2$ charge. In the pure gauge system, this statement is purely kinematic; it has nontrivial dynamical content only if fluctuating matter fields are introduced that carry the $Z_2$ charge.

It is somewhat enlightening nonetheless to consider how eq. (5.13) is fulfilled in the confining and Higgs phases of the pure gauge model. In the confining phase, we formulate a strong coupling expansion by re-expressing [26]

$$\mathrm{e}^{-S_{\mathrm{gauge}}} = N(\beta)\prod_P (1 + U_P \tanh\beta)\,, \tag{5.14}$$

and expanding in powers of $\tanh\beta$. Nonvanishing contributions to $\langle W(\mathrm{C})\rangle$ are associated with surfaces bounded by C that are "tiled" by extracting a factor of $U_P \tanh\beta$ from eq. (5.14) for each plaquette of the surface. Loosely speaking (because there are contributions from disconnected surfaces as well), a surface containing $A$ plaquettes makes a contribution of order $(\tanh\beta)^A$. The number of such surfaces grows sufficiently slowly with $A$ that the strong-coupling expansion has a finite radius of convergence [29]. Thus, for a planar loop, and sufficiently

* In three-dimensional euclidean space, $\Sigma$ is the boundary of a set of *plaquettes* of the dual lattice, and we shrink $\Sigma$ to a point by performing $U_l \to -U_l$ on each of the links dual to these plaquettes.

small $\tanh\beta$, we have

$$\langle W(\mathrm{C})\rangle = (\tanh\beta)^A + \ldots , \tag{5.15}$$

where $A$ is the area of the *minimal* surface bounded by C, and the remainder is negligibly small. Equation (5.15) shows that the $Z_2$ gauge theory exhibits confinement at strong coupling.

If $\Sigma$ and C have an odd linking number, then

$$\langle F(\Sigma)W(\mathrm{C})\rangle = -(\tanh\beta)^A + \ldots , \tag{5.16}$$

because the minimal surface bounded by C contains an odd number of plaquettes on which $\tanh\beta$ has flipped sign. Heuristically, a source of $Z_2$ charge can "see" the Dirac string because it drags along an electric flux tube (represented by the minimal-area surface) that crosses the string at some point [12].

In the Higgs phase, a weak-coupling expansion in $e^{-2\beta}$ can be carried out, where the order in the expansion is determined by the number of "frustrated" plaquettes with $U_P = -1$. Since there is a duality transformation that interchanges strong and weak coupling in this model [26, 28], the weak-coupling expansion also has a finite radius of convergence. At weak coupling, the leading nontrivial contribution to $\langle W(\mathrm{C})\rangle$ arises when one of the links $l$ on C assumes the value $U_l = -1$; this configuration has $W(\mathrm{C}) = -1$ and six frustrated plaquettes. Thus,

$$\langle W(\mathrm{C})\rangle \sim \frac{\exp\left(-L(e^{-2\beta})^6 + \ldots\right)}{\exp\left(L(e^{-2\beta})^6 + \ldots\right)} = \exp\left(-2L(e^{-2\beta})^6 + \ldots\right), \tag{5.17}$$

where $L$ is the number of links on C. The exponential in the numerator of eq. (5.17) results from summing over the $L^N/N!$ ways of flipping the sign on $N$ links contained in C; the denominator is the contribution from these configurations to the partition function $Z = \Sigma e^{-S}$. Eq. (5.17) shows that the $Z_2$ gauge theory is not confining at weak coupling.

The leading contribution to $\langle F(\Sigma)W(\mathrm{C})\rangle$ at weak coupling arises from a configuration of the link variables such that none of the plaquettes dual to $\Sigma$ are frustrated★. In order to avoid frustrating any plaquettes, we must choose $U_l = -1$ on all of the links that are dual to the cubes enclosed by $\Sigma$★★. This is the sense in which, in the Higgs phase, the cosmic string represented by $\Sigma$ has "hair" that can

★ We say that a plaquette is frustrated if the plaquette action is not at its minimum. Thus, when $F(\Sigma)$ is inserted in a Green function, a plaquette $P$ dual to $\Sigma$ is frustrated if $U_P = +1$, while a plaquette $P$ not dual to $\Sigma$ is frustrated if $U_P = -1$.

★★ In three-dimensional euclidean space, we choose $U_l = -1$ on all of the links that are dual to the *plaquettes* enclosed by $\Sigma$.

be detected at long range; $F(\Sigma)$ evidently flips the sign of $W(C)$ if $\Sigma$ and C have an odd linking number.

Let us now turn to the less trivial case of coupled $Z_2$ gauge and spin systems. As for the gauge coupling $\beta$, the dependence on the spin coupling $\gamma$ can be studied by means of strong-coupling and weak-coupling expansions. At strong coupling, we write

$$e^{-S_{spin}} = N(\gamma) \prod_l [1 + (\phi U \phi)_l \tanh \gamma] \tag{5.18}$$

and expand in powers of $\tanh \gamma$. At weak coupling, we expand in powers of $e^{-2\gamma}$, with a factor of $e^{-2\gamma}$ arising from each frustrated link with $(\phi U \phi)_l = -1$. The strong-coupling expansion has a finite radius of convergence because the number of closed loops of length $L$ grows sufficiently slowly with $L$; the weak-coupling expansion can be seen to have a finite radius of convergence by a duality argument. We now distinguish four cases.

*(i)* $\beta, \gamma \ll 1$

In this region, the Wilson loop exhibits perimeter law behavior. For a sufficiently large loop C, the leading contribution to $\langle W(C) \rangle$ arises when a factor of $(\phi U \phi)_l \tanh \gamma$ is extracted from eq. (5.18) for each link of C. Thus

$$\langle W(C) \rangle = (\tanh \gamma)^L + \ldots , \tag{5.19}$$

where $L$ is the length of C. The interpretation is clear: the theory confines $Z_2$ charge, but the electric flux tube can break. $W(C)$ creates a $Z_2$-invariant "hadron", rather than an isolated source of $Z_2$ charge.

To determine the leading nontrivial behavior of $\langle F(\Sigma) \rangle$, consider the effect on the partition function $Z$ of changing the sign of $\beta$ on the plaquettes dual to $\Sigma$. Since a contribution to $Z$ of order $(\tanh \gamma)^4 \tanh \beta$ arises from tiling a single plaquette with $U_p \tanh \beta$ and covering each link of the plaquette with $(\phi U \phi)_l$, we find that

$$\langle F(\Sigma) \rangle = \frac{\exp[-A(\tanh \gamma)^4 \tanh \beta + \ldots]}{\exp[A(\tanh \gamma)^4 \tanh \beta + \ldots]}$$

$$= \exp[-2A(\tanh \gamma)^4 \tanh \beta + \ldots] , \tag{5.20}$$

where $A$ is the number of plaquettes of $\Sigma$. (The exponentiation results from summing over the $A^N/N!$ ways of tiling $N$ of the plaquettes dual to $\Sigma$.) The interpretation is again clear: $\langle F(\Sigma) \rangle$ has area-law behavior because virtual pairs of $Z_2$ charges cause uncorrelated fluctuations in the total $Z_2$ charge enclosed by $\Sigma$, as we discussed in sect. 3.

It is quite evident that

$$\lim A(\Sigma, \mathrm{C}) = \lim \frac{\langle F(\Sigma) W(\mathrm{C}) \rangle}{\langle F(\Sigma) \rangle \langle W(\mathrm{C}) \rangle} = 1 . \tag{5.21}$$

Since the leading behavior of $W(\mathrm{C})$ is zeroth-order in $\tanh \beta$, it is completely unaffected by changing the sign of $\beta$ on the plaquettes dual to $\Sigma$. Only nonleading contributions to $\langle W(\mathrm{C}) \rangle$ that decay like exp[ − Area] are affected by $F(\Sigma)$ if $\Sigma$ and C are far apart. Because of confinement, there are no free $Z_2$ charges, and it is not possible to detect $Z_2$ charge at long range.

*(ii)* $\beta, \gamma \gg 1$

Much as in the pure gauge theory, the leading nontrivial contribution to $\langle W(\mathrm{C}) \rangle$ at weak coupling arises when one of the links on C has $U_l = -1$, except that now flipping $U_l$ frustrates the spins on the link as well as the six plaquettes that contain the link. Thus, we find

$$\langle W(\mathrm{C}) \rangle = \exp\left[ -2L(\mathrm{e}^{-2\beta})^6 \mathrm{e}^{-2\gamma} + \ldots \right], \tag{5.22}$$

where $L$ is the length of the loop C.

The flux operator $F(\Sigma)$ frustrates the plaquettes dual to $\Sigma$, and so its leading behavior is

$$\langle F(\Sigma) \rangle = (\mathrm{e}^{-2\beta})^A + \ldots , \tag{5.23}$$

where $A$ is the area of $\Sigma$. Equation (5.23) is actually the correct leading weak-coupling behavior only for sufficiently large surfaces, where sufficiently large means, roughly speaking,

$$(\mathrm{e}^{-2\gamma})^{A^{1/2}} \lesssim \mathrm{e}^{-2\beta} . \tag{5.24}$$

If eq. (5.24) is not satisfied, then we can do better than eq. (5.23) by flipping a set of links so as to shrink the surface $\Sigma$ dual to the frustrated plaquettes to a surface $\tilde{\Sigma}$ of smaller area $\tilde{A} < A$. Shrinking the surface saves some factors of $\mathrm{e}^{-2\beta}$, but at the price of new factors of $\mathrm{e}^{-2\gamma}$ that result from frustrating the spins on the links that are dual to the volume $V$ enclosed between $\Sigma$ and $\tilde{\Sigma}$. Since shrinking the surface by $\delta A$ requires that the spins be frustrated in a volume $\delta V \sim A^{1/2} \delta A$, it becomes disadvantageous to shrink $\Sigma$ when eq. (5.24) is satisfied.

We can see again that

$$\lim A(\Sigma, \mathrm{C}) = 1 , \tag{5.25}$$

because, for $\Sigma$ sufficiently large, the leading contribution to $\langle F(\Sigma) \rangle$ does not require $U_l$ to flip sign on links that are deep inside the volume bounded by $\Sigma$. A

cosmic string has no hair because hair is too costly; the action due to the hair scales like the volume enclosed by the worldsheet of the string. This is just the phenomenon noted in sect. 3 – condensation of the matter field causes the cosmic string to become the boundary of a domain wall, but the wall is unstable and decays by nucleation of a loop of string.

*(iii)* $\beta \ll 1$, $\gamma \gg 1$

The leading nontrivial contribution to $\langle W(\mathrm{C})\rangle$ is zeroth order in $\tanh\beta$ and arises, again, when one link on C is flipped in sign. This frustrates the spins on that link, so that we find

$$\langle W(\mathrm{C})\rangle = \exp(-2Le^{-2\gamma} + \dots)\,. \tag{5.26}$$

where $L$ is the length of C.

The leading contribution to $\langle F(\Sigma)\rangle$ is zeroth-order in $e^{-2\gamma}$. In the $\gamma \to 0$ limit all plaquette variables are frozen at $U_P = 1$. By considering the effect on the partition function of changing the sign of $\beta$ on the plaquettes dual to $\Sigma$, we therefore find

$$\langle F(\Sigma)\rangle = \exp(-2A\tanh\beta + \dots)\,, \tag{5.27}$$

where $A$ is the area of $\Sigma$.

It is obvious again that

$$\lim A(\Sigma, \mathrm{C}) = 1\,. \tag{5.28}$$

$Z_2$ charge is both confined and screened by the spin condensate.

*(iv)* $\beta \gg 1$, $\gamma \ll 1$

The leading nontrivial contribution to $\langle W(\mathrm{C})\rangle$ arises in zeroth order in $\tanh\gamma$, and we therefore have

$$\langle W(\mathrm{C})\rangle = \exp\left[-2L(e^{-2\beta})^6 + \dots\right], \tag{5.29}$$

just as in the weakly-coupled pure gauge theory.

There is a contribution to $\langle F(\Sigma)\rangle$ of the form $(e^{-2\beta})^A$ that arises when all of the plaquettes dual to $\Sigma$ are frustrated, just as in case (ii) above. But now a much larger contribution is obtained by flipping all of the links dual to the volume enclosed by $\Sigma$. Then $U_P = -1$ on the plaquettes dual to $\Sigma$ and $U_P = 1$ elsewhere, so that no plaquette variables are frustrated. By expanding the spin partition function with the plaquette variables frozen at these values, we find

$$\langle F(\Sigma)\rangle = \exp\left[-2A(\tanh\gamma)^4 + \dots\right]. \tag{5.30}$$

The crucial feature is that the configurations that dominate $\langle F(\Sigma)\rangle$ have the

gauge variables $U_l$ flipped in a volume enclosed by $\Sigma$. This happens because the gauge variables are ordered, and it is therefore costly to frustrate them, while the spins are disordered, and are therefore nearly indifferent to a flip in the sign of their nearest-neighbor couplings inside $\Sigma$. Thus, a cosmic string has hair, and we have shown that

$$\lim A(\Sigma, C) = -1, \tag{5.31}$$

if $\Sigma$ and C have an odd linking number.

We have therefore established that the model is in a phase with unscreened $Z_2$ charge for $\beta \gg 1$ and $\gamma \ll 1$, and that $A(\Sigma, C)$ serves as an order parameter for the phase transition.

It is obvious that this order parameter $A(\Sigma, C)$ can be generalized to a lattice gauge theory with arbitrary gauge group G, when the discrete symmetry whose realization is to be probed is in the center of G.

## 6. Local discrete symmetry without gauge fields

We have seen how a field theory that respects a local discrete symmetry can arise by means of the Higgs mechanism as the low-energy limit of an underlying theory with a continuous gauge group. We wish to explain in this section how a theory with a discrete gauge symmetry can be formulated directly, without ever introducing any gauge fields. We will see that the order parameter defined in sect. 4 can still be constructed in order to probe whether a nontrivial superselection rule is respected by the theory.

The discussion in this section will be somewhat formal, however, in that, although we will use notation appropriate for a continuum field theory, a finite ultraviolet cutoff will be implicit, and a nontrivial continuum limit of the discrete gauge theory need not necessarily exist. Actually, it is obvious that a discrete gauge theory without gauge fields can be constructed, for we can integrate out any gauge fields that acquire mass by the Higgs mechanism, thus obtaining an effective field theory with a cutoff [30]. But it is convenient and rather enlightening to describe this effective field theory directly, without any reference to the physics at mass scales above the cutoff.

To keep the discussion concrete, we consider again the U(1) gauge theory described previously, with a charge-$N$ scalar field $\eta$, and a charge-1 scalar field $\phi$. (As usual, the generalization is straightforward.) We imagine that $\eta$ condenses at the mass scale $v$, so that the photon acquires mass $\mu = Nev$. We may integrate out $\eta$ and the photon to obtain an effective field theory for the surviving field $\phi$, with cutoff $\Lambda \sim \mu$.

For the purpose of writing the lagrangian of this theory, it is convenient to retain the pure gauge Goldstone degree of freedom, the phase of $\eta$, that is eaten by the

photon. We denote this phase by $\Theta$, where

$$\eta = \rho\, e^{-i\Theta}; \tag{6.1}$$

thus $\Theta$ is a periodic variable defined modulo $2\pi$. Under a U(1) gauge transformation parametrized by $\omega$, the fields $\phi$ and $\Theta$ transform according to

$$\phi \to e^{i\omega}\phi\,, \qquad \Theta \to \Theta - N\omega\,. \tag{6.2}$$

The effective lagrangian with (nonlinearly realized) local U(1) invariance is [30]

$$\begin{aligned}\mathscr{L}(\phi,\Theta) &= \mathscr{L}\left(\phi\, e^{i\Theta/N}\right)\\ &= \partial_\mu\left(\phi\, e^{i\Theta/N}\right)^\dagger \partial^\mu\left(\phi\, e^{i\Theta/N}\right)\\ &\quad - V\left(\phi\, e^{i\Theta/N}\right) + \ldots\,,\end{aligned} \tag{6.3}$$

where the ellipsis indicates terms higher order in derivatives. That is, the action is a local functional of the U(1) invariant field

$$\Phi = \phi\, e^{i\Theta/N}. \tag{6.4}$$

Furthermore, since the $Z_N$ transformation

$$\Phi \to e^{2\pi i/N}\Phi \tag{6.5}$$

is merely a rotation of $\Theta$ by $2\pi$, and $\Theta$ is a periodic variable, the action must be invariant under eq. (6.5).

Although the lagrangian $\mathscr{L}(\phi,\Theta)$ respects a local U(1) symmetry, it really describes (redundantly) a $Z_N$ gauge theory. The point is that $\Theta$ is purely a gauge degree of freedom; locally at least, we are free to rotate $\Theta$ to zero by means of eq. (6.2). The physical content of $\Theta$ therefore resides entirely in any topological obstructions that prevent us from rotating $\Theta$ to zero globally. In particular, there are field configurations such that $\Theta$ is ill defined on a codimension-two "string" and has unit winding number around the string. For such a configuration, $\Theta$ can be gauged away only at the cost of making $\Phi$ $N$-valued in the vicinity of the string.

If spacetime is simply connected, then the sole purpose of $\Theta$ is to identify where the strings are. Thus, a theory of a complex scalar field $\Phi$ with a local $Z_N$ symmetry differs from a theory with a global $Z_N$ symmetry in that the dynamical variables include both $\Phi$ and a $Z_N$ string degree of freedom. In principle, the action of the theory could include a Nambu–Goto term, or a more complicated dependence on the string worldsheet, but for the minimal version given by eq. (6.3), the classical string tension vanishes*. The only effect of the string, then, is to impose a nontrivial boundary condition on $\Phi$ – that around a loop C that links the

* Of course, the effective theory that describes the low-energy limit of the Higgs phase of a U(1) gauge theory would have a positive string tension.

string surface once, $\Phi$ is not strictly periodic but is instead periodic up to the element $e^{2\pi i/N}$ of $Z_N$ [1,4]. (Because of this boundary condition, the quantum fluctuations of $\Phi$ generate an *effective* string tension.)

If the spacetime manifold M is not simply connected, then another type of obstruction arises that prevents $\Theta$ from being completely gauged away. If $\gamma$ is a noncontractible loop in M, then $\Theta$ may have a winding number $k_\gamma$ about $\gamma$

$$(\delta\Theta)_\gamma = 2\pi k_\gamma . \tag{6.6}$$

This winding number $k_\gamma$ (modulo $N$) determines the boundary condition satisfied by $\Phi$ on the loop $\gamma$; $\Phi$ is periodic up to $\exp[i(\delta\Theta)_\gamma/N] \in Z_N$. A mod $N$ integer $k$ may thus be associated with each homology cycle of M, and the $k$'s should also be regarded as dynamical variables – they are to be summed over in the path integral.

Another way to describe this distinction between global and local $Z_N$ symmetry is to note that, in the case of a local symmetry, $\Phi$ is actually to be *identified* with $e^{2\pi i/N}\Phi$. Hence, the field $\Phi$ takes values not in a smooth manifold, but in an *orbifold* [1,4,30] with a conical singularity at $\Phi = 0$. That $\Phi$ takes values in $\mathbf{C}/Z_N$ rather than $\mathbf{C}$ is of no consequence, however, aside from the topological considerations noted above.

A Wilson loop operator $W(C)$ can be expressed in terms of $\Theta$ as

$$W(C) = \exp\left[\frac{i}{N}\oint_C dx\cdot\partial\Theta\right] \in Z_N . \tag{6.7}$$

This operator merely counts the (signed) number of strings that link the closed loop C; that is, for a given configuration of strings (and given values of the mod $N$ integers $k$ associated with the homology cycles of the spacetime manifold), $W(C)$ identifies the boundary conditions that are satisfied by $\Phi$ on C. The flux operator $F(\Sigma)$ may also be defined as before; an insertion of $F(\Sigma)$ constrains $\Phi$ to twist by $e^{2\pi i/N}$ on a loop that links $\Sigma$ once. Thus, the order parameter $A(\Sigma, C)$ can be constructed as in eq. (4.4). This order parameter provides a criterion that specifies whether the $Z_N$ charge is screened. If $A(\Sigma, C)$ behaves as in eq. (4.5), then $Z_N$ charges induce quantum-mechanical effects that can be detected at long range via the Aharonov–Bohm effect, and there is a nontrivial $Z_N$ superselection rule.

If the $Z_N$ symmetry is spontaneously broken due to the condensation of $\Phi$, then $Z_N$ charge is screened. But one may gain a further appreciation of the difference between global and local discrete symmetry by considering the topological defects that result from the symmetry breakdown. When a discrete global symmetry is spontaneously broken, there are topologically stable domain walls. If the spontaneously broken discrete symmetry is a local symmetry, however, then stable domain walls do not exist [22–24]. A domain wall can end on a string, and a wall may therefore decay by means of the spontaneous nucleation of a closed loop of string that creates a hole in the wall (fig. 7).

If spacetime is simply connected, then, as we have seen, the difference between a local and global discrete symmetry rests on the existence of strings (either as

dynamical objects or as classical sources). Even if the strings have zero tension classically, quantum fluctuations induce a string tension; this is the interpretation of $\kappa$ in eq. (3.2). Thus, strings tend to decouple at low energy, obscuring the distinction between local and global discrete symmetries.

Because of this renormalization of the string tension, the distinction between a local and global discrete symmetry might not survive in the continuum limit of a scalar field theory, at least if spacetime is simply connected. A continuum theory that contains strings can be constructed only if the string tension, as well as the $\Phi$ mass, can be chosen to be arbitrarily small in units of the cutoff $\Lambda$. Conventional wisdom holds that this is impossible in four dimensions; dynamical strings would induce nontrivial $\Phi$ interactions, and we could thus construct a nontrivial continuum limit of a self-coupled scalar field theory, without introducing any gauge fields.

Finally, we conclude this section with a brief comment about discrete gauge theories that contain chiral fermions. In a gauge theory with a continuous gauge group, the fermion content is restricted by the requirement that perturbative [31] and nonperturbative [32] gauge anomalies must cancel. It is natural to wonder whether similar restrictions apply to a gauge theory with a discrete gauge group.

If we insist that a theory with manifest local $Z_N$ symmetry, for example, be obtained as the low-energy limit of an underlying U(1) gauge theory, then the $Z_N$ quantum numbers of the light fermions are restricted by the requirement that the gauge anomalies must cancel in the underlying theory. But if the $Z_N$ gauge theory is formulated directly, without ever introducing any gauge fields, then we claim that there are no such restrictions on the fermion content. Regardless of the charges of the fermions, the fermionic effective action is manifestly gauge-invariant; the only smooth $Z_N$ gauge transformations are constant, and these have no effect on the boundary conditions satisfied by the fermions. Nor do any cancellations occur, when we sum over all possible boundary conditions, that render gauge-invariant Green functions ill defined. We also note that the nontrivial boundary conditions that distinguish a theory with a local discrete symmetry from one with a global discrete symmetry have no impact on the short-distance behavior of the theory. Therefore, gauge invariance appears to have no bearing on renormalizability*.

## 7. Nonabelian local discrete symmetry

We will now extend the previous discussion to the case of a nonabelian discrete unbroken gauge group H. Our objective is to identify the nontrivial superselection sectors in a theory with manifest H symmetry, and hence to classify the possible

* These conclusions about gauge anomalies are further elaborated in ref. [33].

types of quantum-mechanical hair. This generalization involves some new subtleties.

In the case of an abelian gauge group, we argued that there is a distinct superselection sector associated with each (one-dimensional) irreducible representation of the gauge group. It is natural to expect that this classification will continue to hold in the nonabelian case. Indeed, we will argue that the irreducible representation of H according to which a projectile transforms can be determined in principle by scattering the projectile off of the various cosmic strings of the H gauge theory. Equivalently, this information can be extracted by winding loops of the various strings around the projectile, as described in sect. 3.

However, in contrast to the discussion of the abelian case in sect. 3, we have not succeeded in constructing a gauge-invariant operator whose eigenvalues label the distinct superselection sectors. The strategy of seeking an operator realization of winding a cosmic string around a region fails, in part because we are unable to find a gauge-invariant operator that creates a cosmic string. In fact, we find that there is a topological obstruction to implementing a global H gauge transformation in the presence of strings (including virtual strings that can arise as quantum fluctuations). A similar obstruction can occur, even if strings are absent, in an H gauge theory that satisfies nontrivial boundary conditions on a multiply-connected spacetime (see sect. 8.)

These mathematical statements have a physical counterpart in the remarkable properties of the Aharonov–Bohm effect in the nonabelian case [34]. Scattering of a projectile by a string can change the charge of the projectile. However, since a local process cannot affect long-range hair, the irreducible representation of H according to which projectile plus string transform must remain unchanged. Thus, a closed loop of string must be capable of carrying H-charge. We find that this is indeed so, but that, oddly, the charge "carried" by the string cannot be localized anywhere on the string or in its vicinity.

There is no obstruction to defining a global gauge transformation that is in the *center* of H; the corresponding charges are precisely those charges that cannot be changed by Aharonov–Bohm scattering. Charge operators in the center can be constructed exactly as in the abelian case. But the operator realization of *nonabelian* gauge charge is very elusive. We believe nonetheless that the available evidence supports the conclusion that hair in an H gauge theory is classified by the irreducible representations of H.

We will now consider the nonabelian Aharonov–Bohm effect in greater detail. For most of this discussion, we will find it convenient to take spacetime to be $2+1$-dimensional; rather than strings, then, the theory contains vortices, at least some of which are stable particles. Our analysis applies to strings in $3+1$ dimensions with only minor modifications.

As we noted in sect. 3, a $(2+1)$-dimensional discrete gauge theory respects a topological superselection rule; states carry a topologically conserved vortex num-

ber. Physically, vortex number is a type of hair because it can be detected at long range via the Aharonov–Bohm effect, just as charge can be detected.

A vortex has hair if fields that transform nontrivially under the manifest discrete local H symmetry are not strictly periodic on closed loops that enclose the vortex, but are instead periodic only up to a nontrivial element $\Omega_0$ of H. (We will refer to this boundary condition satisfied by the fields on the vortex background as the "matching condition" imposed by the vortex.) This element $\Omega_0$ may be regarded as the "magnetic flux" of the vortex. That is, if H is embedded in a continuous gauge group G that has undergone the Higgs mechanism, then we may write

$$\Omega_0 = P \exp\left( i \oint_{C,x} \boldsymbol{A} \cdot \mathrm{d}\boldsymbol{x} \right) \in \mathrm{H}. \tag{7.1}$$

Here $\boldsymbol{A}$ is the G gauge field, and the path-ordered integration is carried out on a large oriented closed path C that encloses the vortex, beginning and ending at the point $\boldsymbol{x}$. (H is the subgroup of G that leaves invariant the symmetry-breaking order parameter at $\boldsymbol{x}$). While a finite-energy vortex configuration can be constructed with flux $\Omega_0$ for each element $\Omega_0 \in \mathrm{H}$, two distinct elements of H do not necessarily correspond to physically distinguishable boundary conditions. This is because $\Omega_0$ is not gauge invariant; under a gauge transformation that preserves the order parameter at $\boldsymbol{x}$, it transforms as

$$\Omega_0 \to h \Omega_0 h^{-1}, \qquad h \in \mathrm{H}. \tag{7.2}$$

Thus, it is the conjugacy classes of H that classify the boundary conditions, and hence the distinct types of vortex hair. (The corresponding topological statement, if G is simply connected, is that while $\pi_1(\mathrm{G}/\mathrm{H}) = \mathrm{H}$ classifies the closed paths in G/H that begin and end at a specified point, the closed loops in G/H with no specified endpoint are classified by the conjugacy classes of H [35, 36].)

For the purpose of describing physics on the vortex background, it is often convenient to choose a (singular) gauge such that $\boldsymbol{A}$ vanishes far from the vortex core. In this gauge, fields that transform nontrivially under H are typically multivalued on the vortex background. As in the abelian case, this multivaluedness results in a nontrivial Aharonov–Bohm effect. But the vortices of a theory with nonabelian local discrete H symmetry have a characteristic feature that is not shared by the abelian case: H gauge transformations are also multivalued on the vortex background. This multivaluedness of the discrete gauge transformations has remarkable consequences.

These consequences are best appreciated within the context of particular examples. We will describe two examples here. In our first example, the unbroken gauge group H is not actually a discrete group at all; instead it is a continuous nonabelian group with two distinct connected components. We present this example in some detail because it is quite simple and yet nicely illustrates some of our main points.

In the second example, H is a discrete nonabelian group. This example is a bit more complicated than the first, but it is more representative of the general case.

In our first example [24, 37], the gauge group is SU(2) and the order parameter $\Phi$ is in the 5-dimensional representation of SU(2). We may express $\Phi$ as a symmetric traceless $3 \times 3$ matrix that transforms according to

$$\Phi \to \Omega \Phi \Omega^{\mathrm{T}}, \qquad \Omega \in \mathrm{SO}(3). \tag{7.3}$$

If the expectation value of $\Phi$ in unitary gauge is

$$\langle \Phi \rangle = v \,\mathrm{diag}(1, 1, -2), \tag{7.4}$$

then SO(3) is broken to a subgroup that is isomorphic to O(2). The unbroken group has two connected components. The component connected to the identity is SO(2), containing all rotations about the $z$-axis. The other component contains each 180° rotation about an axis in the $x$–$y$ plane.

When SO(3) is lifted to SU(2), O(2) is covered twice by a group called Pin(2). The elements of the two connected components of Pin(2) may be parametrized as

$$\left\{\exp\left(i\frac{\theta}{2}\sigma_z\right)\right\}, \left\{i\sigma_y \exp\left(i\frac{\theta}{2}\sigma_z\right)\right\}, \quad \theta \in [0, 4\pi). \tag{7.5}$$

Because H = Pin(2) is disconnected, there is a topologically stable vortex whose "magnetic flux" is in the disconnected component of Pin(2); for a particular choice of gauge the flux is

$$\Omega_0 = P \exp\left(i \int \boldsymbol{A} \cdot \mathrm{d}\boldsymbol{x}\right) = i\sigma_y. \tag{7.6}$$

Thus, $\Omega_0$ does not commute with the charge operator $Q = \frac{1}{2}\sigma_z$ that generates the unbroken $\mathrm{U}(1) \subset \mathrm{Pin}(2)$; instead we have

$$\Omega_0 Q \Omega_0^{-1} = -Q. \tag{7.7}$$

Therefore, a charged particle that voyages around the vortex returns to its starting point with its charge flipped in sign. For this reason, the vortex of this model was dubbed the "Alice vortex" in ref. [37] – whoever circumnavigates the vortex is reflected in the charge-conjugation looking-glass.

Eq. (7.7) tells us that electric charge, and hence also electric or magnetic field, are necessarily two-valued in the background of an Alice vortex. For example, whether two point charges have the same sign or opposite sign can be determined locally if the charges are brought together in a region where there are no vortices nearby. (The charges either repel or attract one another.) But whether two charges

have the same sign is not well defined globally. If the charges are initially distantly separated, we must bring them together to measure the relative sign of their charges. But the outcome of the experiment depends on the path that we choose in reuniting them; in particular, on how many times each charge winds around the vortex before they meet. Similarly, the sign of the electric field is ambiguous because we measure the field locally by observing the response to the field of a positive test charge, but the sign of the test charge is not globally well defined.

Because of the double valuedness of the electric field, there is no sensible way to define the electric flux through a surface that contains an odd number of vortices. In particular, then, the total electric charge of a state that contains an odd number of vortices is ill defined*. Note that we are unable to take refuge in the observation that the charge of a state can be extracted from the transformation properties of the state under global gauge transformations, as in eq. (2.3). The difficulty is that there is a topological obstruction to defining a global gauge transformation on the vortex background, precisely *because* the charge $Q$ is double-valued. Since

$$\Omega_0 \, e^{i\omega Q} \Omega_0^{-1} = e^{-i\omega Q}, \tag{7.8}$$

a constant U(1) transformation on a loop enclosing a single vortex is consistent with the matching condition imposed by the vortex only if $\omega = 2\pi n$, where $n$ is an integer. Thus, $e^{2\pi i Q}$, the nontrivial element of the *center* of $H = \mathrm{Pin}(2)$, is a globally-defined quantity, or the electric charge $Q$ is well defined modulo an integer. The globally-defined charge, then, only tells us whether the number of fundamental charges with $|Q| = \frac{1}{2}$ is odd or even.

(This phenomenon, that a global gauge transformation can be defined on a surface enclosing a vortex only if the transformation commutes with the matching condition imposed by the vortex, is strongly reminiscent of an observation due to Nelson and Manohar [38] and Balachandran et al. [38]. They found that a global gauge transformation can be implemented on a surface enclosing a magnetic monopole only if the transformation leaves the long-range field of the monopole unchanged.)

Although the total electric charge is a meaningless notion in the background of a single vortex, this is not so in the background of *two* vortices. (The two vortex sector is topologically trivial in that the Alice vortex and antivortex are gauge

* Strictly speaking, the total charge vanishes for any state of finite energy in the infinite volume limit, because charge is (logarithmically) confined in (2 + 1)-dimensional QED. We ignore this technicality here. The reader who is concerned about this point may prefer to promote the discussion below of charge in a two-vortex background to the case of charge in the background of a string in 3 + 1 dimensions.

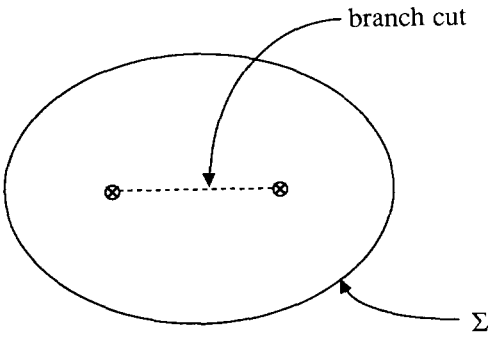


Fig. 11. Two Alice vortices connected by a branch cut. The electric field is single-valued and smooth on $\Sigma$.

equivalent.) While the charge and electric field $\boldsymbol{E}$ remain globally two-valued, we can consistently restrict $\boldsymbol{E}$ to one of its two branches on a loop that encloses both vortices, because the matching conditions are trivial on this loop. In fact, we can restrict $\boldsymbol{E}$ to a single branch on the whole plane at the cost of introducing a branch cut connecting the two vortices across which $\boldsymbol{E}$ is required to change sign (fig. 11). Of course, this branch cut is a completely unphysical gauge artifact. The electric flux through a surface that encloses both vortices is well defined up to an overall sign, and so the absolute value of the total charge contained in the surface can in principle be measured. Correspondingly, a "global" gauge transformation can be implemented on a region that contains two vortices; it is a constant transformation on the boundary of the region, and is deformed in the interior of the region so as to vanish on the vortices and the string connecting them. (This is analogous to the observation by Coleman and Nelson [39], that a global gauge transformation can always be implemented on the background on a monopole–antimonopole pair, but that the transformations that change the long-range monopole field are required to vanish on the monopole core.)

Although the electric charge, and hence the classical "hair", are well defined for a region that contains two vortices, this charge is not the same as the sum of the charges of all the particles contained in the region. Since the sign of a charge cannot be globally defined, there is no unambiguous way to add charges together.

This observation gives rise to a puzzle. As depicted in fig. 12, we can imagine a region that contains two vortices and two point charges, such that the total charge vanishes as measured on a distant boundary. Suppose that the two point charges are united, and found to be equal and opposite in sign. Call the charges $\pm Q$. Now let the particle with charge $+Q$ voyage around one of the vortices, while the other charge stays home. When the charges are reunited, both have charge $-Q$. Yet the total charge, as measured at arbitrarily long-range, must not have changed. It seems that $2Q$ units of charge are unaccounted for. Where did this charge go?

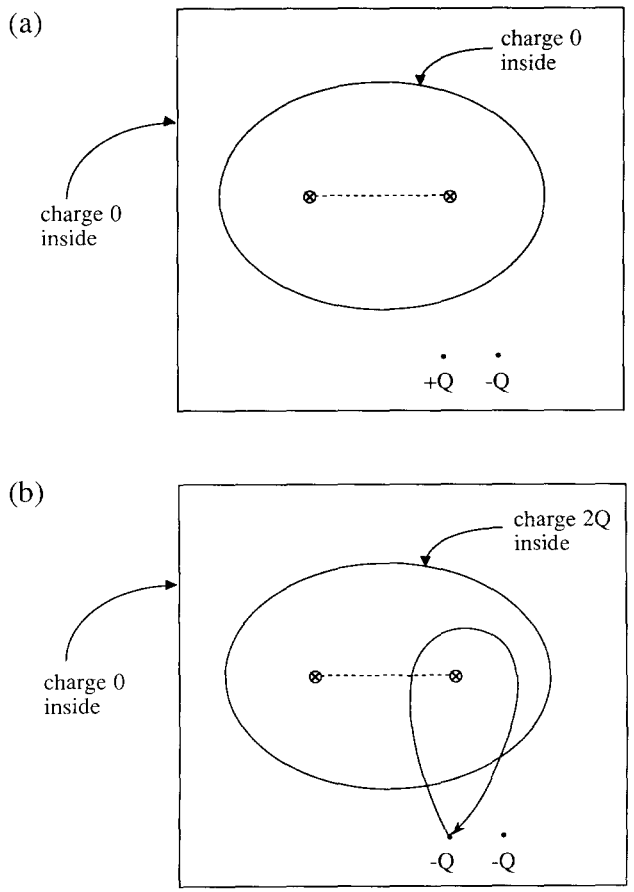


Fig. 12. A charged particle voyages around a vortex and returns with the sign of its charge changed. $2Q$ units of charge have been deposited inside $\Sigma$.

It is natural to suspect that the charge has been transferred to the vortex, but one quickly sees that this is an unlikely explanation. If there does exist an excitation of the vortex with $2Q$ units of charge bound to the vortex core, we expect that excitation to be split from the vortex ground state by a finite amount of Coulomb energy*. But we may imagine that the charge-$Q$ particle circles the vortex adiabatically**; then the charged vortex could not get excited. Moreover,

* More correctly, since the Coulomb energy of an isolated charged particle is infinite in two dimensions, we should compare two configurations with vanishing total charge. The configuration with nonzero charge localized on the vortex should be split from the configuration with zero charge on the vortex by a finite gap.

** We need to do some work to move the charged particle in the fields of the other particles, but this work may be made as small as desired.

there is no indication within the semiclassical approximation that the vortex *has* charged excitations. Charged excitations of solitons arise semiclassically from quantizing the global charge rotor degree of freedom of the soliton. But we have seen that global charge rotations of the Alice vortex do not exist.

Even more telling, we may imagine that the charged particle circumnavigates the vortex while distantly separated from it. Since all charged fields have finite mass, there seems to be no mechanism by which the particle could transmit charge to the vortex at arbitrarily long range.

Even if charge cannot be carried by an isolated vortex, conservation of the electric flux at spatial infinity requires that charge can be carried by a vortex pair. We seem, though, to require a miracle, for the charged particle must be able to excite a charged state of the vortex pair without actually transferring any charge to the pair. It is the two-valuedness of the electric field that makes it possible to perform this miracle.

Fig. 13 shows one branch of the static two-valued electric field of a point charge $Q$ in the vicinity of a vortex pair, for a sequence of positions of the point charge. The total electric charge as measured at spatial infinity (on this branch) is also assumed to be $Q$. When the charge reaches the cut that connects the two vortices, it disappears behind the cut just as its image charge $-Q$ emerges from behind the cut. The electric flux $-2\pi Q$ emanating from the image charge returns to the second sheet through the cut, and the flux $2\pi Q$ emanating from the original charge likewise returns to the first sheet through the cut. After the charge has passed the cut, then, an observer on a closed surface that encloses the two vortices, but not the point charge, measures electric flux $4\pi Q$ through the surface, and infers that $2Q$ units of charge are inside. In fact, though, this charge is not localized anywhere. The electric flux through any closed box on the two-sheeted surface vanishes, if the box does not contain the point charge, its image, or either vortex. Nevertheless, the vortex pair, initially in its zero-charge ground state, has been excited to a charge-$2Q$ state after the point charge has passed through. The vortex pair now has hair. The electric field lines trapped by the pair cause the vortices to repel each other*.

On a background with an even number of vortices at specified positions, the classical electric field on a single sheet is not uniquely determined by the positions and values of all point charges on that sheet. One must also know the charge "carried" by each cut. As the point charges move with respect to this background, the values of the point charges may change sign, but the "charge" on the cuts also changes, in such a way that the total charge, the "hair", stays constant. If $|Q| = \frac{1}{2}$ is the quantum of charge, then the sum of the values of the point charges can

* That a pair of Alice vortices or a loop of Alice string can carry unlocalized electric charge has also been noted by Alford et al. [44]. They propose the apt name "Cheshire charge" for this phenomenon.

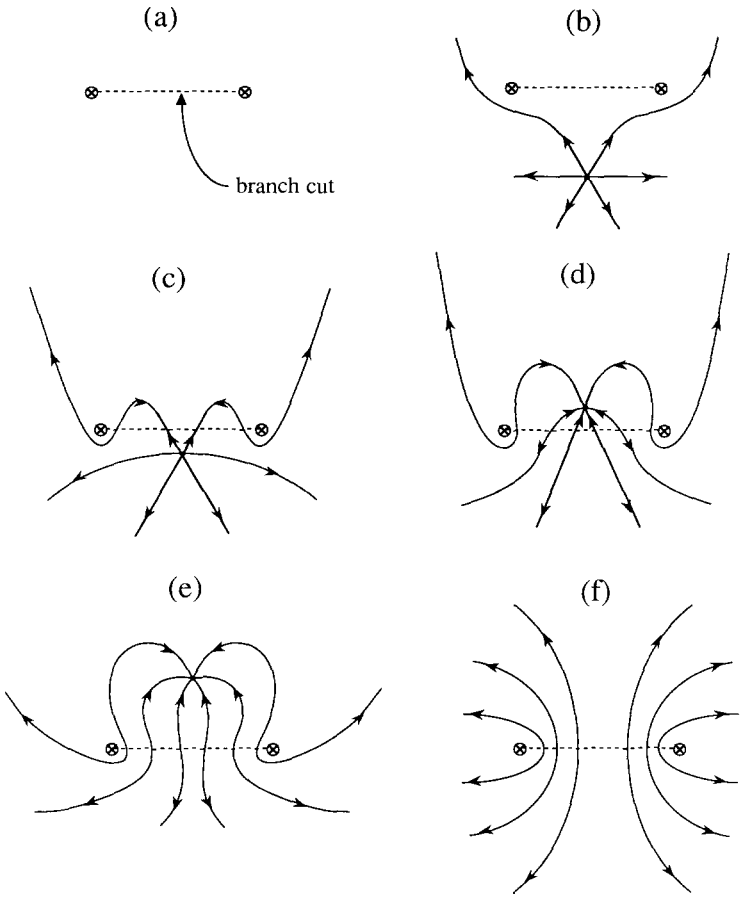


Fig. 13. One branch of the static electric field of a point charge in the vicinity of a vortex pair, for a sequence of positions of the point charge.

change by only an integer. This is in accord with our earlier observation that $e^{2\pi iQ}$ is a well-defined quantum number, even when the number of vortices is odd.

(A monopole–antimonopole pair, where the monopole has a nonabelian magnetic field, also has charged excitations [39]. In that case, however, there are light charged fields, the nonabelian gauge fields, that carry the charge of the excitation. The charged excitations of a pair of Alice vortices are quite different, as we have just seen.)

The above discussion of a pair of Alice vortices in two spatial dimensions generalizes immediately to a loop of Alice string in three spatial dimensions. From the nonabelian Aharonov–Bohm effect, we infer that a loop of Alice string can have classical hair. A loop can carry any integer amount of electric charge, and a

loop with charge $Q$ and length $L$ has a Coulomb energy of order $Q^2e^2/L$. A sufficiently large charged loop is therefore metastable; all charged particles are massive, so the loop cannot reduce its charge by emitting charged particles.

The oscillations of the loop cause it to emit both gravitational and electromagnetic radiation. In order of magnitude the gravitational and electromagnetic power are

$$P_{\text{grav}} \sim GM^2/L^2, \qquad P_{\text{em}} \sim Q^2e^2/L^2, \tag{7.9}$$

where $M$ is the mass of the loop. The loop loses energy and shrinks. It may eventually reach a stable configuration in which the string tension is balanced by the electric flux trapped by the loop. However, the mass of this configuration is of order

$$M \sim Qe\sqrt{\mu}\,, \tag{7.10}$$

where $\mu$ is the string tension, and so is comparable to the mass of $Q$ charged vector bosons. Thus, the decay of this minimal charged loop to vector bosons may be kinematically allowed.

Having unraveled the interplay of the nonabelian Aharonov–Bohm effect and classical hair in the case of the Alice vortex, we are now prepared to return to our main interest – *quantum-mechanical* hair. Again, it is useful to consider a particular example in some detail.

Our example [36, 40, 41] will be the same model as before, but with a different pattern of symmetry breakdown. We now suppose that the order parameter $\Phi$ has, in unitary gauge, the expectation value

$$\langle \Phi \rangle = v\,\mathrm{diag}(1+\delta, 1-\delta, -2)\,; \tag{7.11}$$

this reduces to eq. (7.4) in the limit $\delta \to 0$. For $\delta \neq 0$, the unbroken subgroup of SO(3) is the four-element subgroup $D_2 \sim Z_2 \times Z_2$, which contains the identity and 180° rotations about each of the $x$, $y$ and $z$ axes. When SO(3) is lifted to SU(2), $D_2$ is covered twice by the eight-element quaternionic group

$$\mathbf{Q} = \{\pm\mathbf{1}, \pm i\sigma_x, \pm i\sigma_y, \pm\sigma_z\}\,. \tag{7.12}$$

Thus, $H = \mathbf{Q}$ is a discrete nonabelian group.

The model with unbroken $\mathbf{Q}$ symmetry, unlike the model with unbroken Pin(2) symmetry, has no massless gauge fields. It also differs from the Pin(2) model in another significant respect that concerns the properties of vortices. The conjugacy classes of $\mathbf{Q}$ are

$$\{\mathbf{1}\}, \{-\mathbf{1}\}, \{\pm i\sigma_x\}, \{\pm i\sigma_y\}, \{\pm i\sigma_z\}\,. \tag{7.13}$$

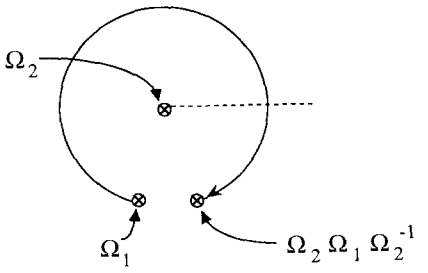


Fig. 14. When a vortex with flux $\Omega_1$ is transported around a vortex with flux $\Omega_2$, its flux becomes conjugated.

As we have noted, the conjugacy classes of a discrete gauge group H classify the values of the "magnetic flux" or topological charge. Hence, the **Q** model contains four distinct types of vortices.

Multiplication of conjugacy classes is in general ambiguous, and there is a corresponding ambiguity when two vortices are patched together to make a vortex pair [35,36]. For example, the product of two elements of the conjugacy class $\{\pm i\sigma_x\}$ can be either **1** or $-\mathbf{1}$, which are two distinct classes. When two vortices that both represent the class $\{\pm i\sigma_x\}$ are brought together, then, they may or may not be able to annihilate each other, depending on how the patching has been performed. This phenomenon is actually closely analogous to the property we observed in the Pin(2) model, that whether two charged particles have equal or opposite charges is dependent on the choice of a path that connects the two particles. In general, if a vortex with magnetic flux $\Omega_1$ is transported around a vortex with magnetic flux $\Omega_2$, its flux becomes conjugated

$$\Omega_1 \to \Omega_2 \Omega_1 \Omega_2^{-1}. \tag{7.14}$$

So if a topologically trivial pair of vortices is produced, say with flux $i\sigma_x$ and $-i\sigma_x$, and one of the vortices voyages around a $\{\pm i\sigma_y\}$ vortex, then the magnetic flux of the pair has become $-\mathbf{1}$ when the vortices are reunited (fig. 14). This example illustrates why magnetic flux and topological charge can be globally defined only up to conjugacy. (The corresponding phenomenon in three spatial dimensions is that $\{\pm i\sigma_x\}$ and $\{\pm i\sigma_y\}$ strings cannot pass through each other without becoming *entangled* [36,40].) Since topological quantum numbers do not prevent the $\{-\mathbf{1}\}$ vortex from decaying to a pair of $\{\pm i\sigma_a\}$ vortices, this vortex may be unstable, at least for a range of values of the parameters of the model.

The number of conjugacy classes of a finite group H is also the number of inequivalent irreducible representations of H. The irreducible representations of **Q** include, aside from the defining two-dimensional representation and the trivial

TABLE 1

| Representation | $\{\pm i\sigma_x\}$ | $\{\pm i\sigma_y\}$ | $\{\pm i\sigma_z\}$ |
|---|---|---|---|
| $1_0$ | 1 | 1 | 1 |
| $1_x$ | 1 | −1 | −1 |
| $1_y$ | −1 | 1 | −1 |
| $1_z$ | −1 | −1 | 1 |

one-dimensional representation, three other nontrivial one-dimensional representations. The one-dimensional representations all represent $\pm\mathbf{1}$ by the identity and represent the other elements of **Q** as shown in table 1. The five irreducible representations are the five distinct values that the "charge" can assume in the **Q**-model.

The charge of a state in an H gauge theory is ill defined unless the magnetic flux of the state has a value in the center of H. For **Q**, the center is $\{\mathbf{1}, -\mathbf{1}\}$. In particular, then, **Q**-charge is ill-defined in the background of a single $\{\pm i\sigma_x\}$ vortex but well-defined in the background of a pair of $\{\pm i\sigma_x\}$ vortices. This charge, as well as the vortex number, is a type of quantum-mechanical hair that can be detected at long-range by means of the Aharonov–Bohm effect.

We say that Aharonov–Bohm scattering of a charged projectile by one member of a vortex pair is *nonabelian* if charge is transferred to the pair and is *abelian* if no charge transfer occurs. As for the case of the Alice vortex, we can see that nonabelian Aharonov–Bohm scattering must be possible in the **Q**-model. Before turning to the nonabelian effect, however, let us first note that the abelian Aharonov–Bohm effect can in principle be used to unambiguously determine the charge of a projectile in the **Q**-model.

First, charges in the center of a discrete gauge group H can always be detected by vortices with magnetic flux in the center of H; the Aharonov–Bohm scattering involving these vortices is always abelian. In the case of **Q**, the vortex with flux $\Omega_0 = -\mathbf{1}$ distinguishes the doublet representation, with which it has a nontrivial Aharonov–Bohm effect, from the four singlet representations, with which it has no Aharonov–Bohm effect.

Aharonov–Bohm scattering of the singlet charges is necessarily abelian. We may distinguish among the four distinct one-dimensional representations by scattering off the three remaining types of vortices. We can read off from table 1 that the $\{\pm i\sigma_x\}$ vortex, for example, scatters projectiles with charge $1_y$ or $1_z$, but does not scatter projectiles with charge $1_0$ or $1_x$, while the $\{\pm i\sigma_y\}$ vortex scatters $1_x$ and $1_z$, but not $1_0$ or $1_y$. So scattering off these two vortices evidently distinguishes the four singlet representations of **Q**.

We can also invoke the abelian Aharonov–Bohm effect to measure the charge in a large region, if the magnetic flux in the region is in the center of **Q**. As described

in sect. 3, the charge can be detected as the Aharonov–Bohm phase acquired by a vortex that circumnavigates the region.

The nonabelian Aharonov–Bohm effect occurs in the **Q**-model when a projectile in the doublet representation of **Q** scatters off a vortex whose magnetic flux is not in the center of **Q**. Consider, for example, a pair of vortices, each with magnetic flux $\{\pm i\sigma_a\}$. If the pair initially has the trivial $1_0$ charge, and a projectile in the doublet representation scatters off one member of the pair, then the pair becomes excited to the $1_a$ representation. The charge carried by a vortex pair is restricted to the singlet representations of **Q**, as the center charge of **Q** must be well defined even in the background of a single vortex.

We can gain some insight into the mechanism of nonabelian Aharonov–Bohm scattering in the **Q**-model by contemplating a model with a symmetry-breaking hierarchy. If $\delta$ in eq. (7.11) satisfies $\delta \ll 1$, then the model has three different energy regimes. At short distances SU(2) is effectively restored; at intermediate distances, Pin(2) is a good symmetry; and at long distances only the **Q**-symmetry survives. Then our earlier description of a charged particle interacting with an Alice vortex applies to Aharonov–Bohm scattering at intermediate energy of a doublet off a pair of $\{\pm i\sigma_y\}$ vortices with intermediate separation. A projectile in the doublet representation of **Q** must be in a $|Q| = n + \frac{1}{2}$ representation of Pin(2), where $n$ is a non-negative integer, and excites an initially uncharged state of the vortex pair to a state with charge $|Q| = 2n + 1$. A pair of $\{\pm i\sigma_y\}$ vortices in the $|Q| = 2n + 1$ representation of Pin(2) transforms as the $1_y$ representation of **Q**. At long distances, the Pin(2) charge of the vortex pair becomes screened by the condensate, but the **Q**-charge is a type of hair that cannot be screened. Nor can the **Q**-charge become screened as the separation of the vortex pair smoothly varies from an intermediate distance to a long distance.

We see, now, that the **Q**-model lends support to our general claims. Quantum-mechanical hair in the model is classified by magnetic flux, and in the sectors with magnetic flux in the center of **Q**, by an irreducible representation of **Q**. Furthermore, **Q**-charge can be carried by an isolated vortex pair.

In general, if H is a discrete gauge group, then a vortex with magnetic flux in the conjugacy class $\{h_0\}$ can be patched together with a vortex with flux in the class $\{h_0^{-1}\}$ to make a configuration with trivial flux. The possible values of H charge that this pair can carry are determined as follows: H can act on a representative $h_0$ of a conjugacy class according to

$$h\colon h_0 \to h h_0 h^{-1}, \qquad h \in \mathrm{H}. \tag{7.15}$$

The representatives of a class thus transform as a (in general reducible) representation of H with dimension equal to the order of the class. The irreducible representations of H contained in this representation are the allowed values of the charge of the vortex pair.

For example, the class $\{\pm i\sigma_y\}$ of **Q** transforms as the representation $1_0 + 1_y$ of **Q**. The class

$$\left\{P(\theta) = i\sigma_y \exp\left(i\frac{\theta}{2}\sigma_z\right)\right\} \tag{7.16}$$

of Pin(2) transforms as

$$\exp\left(i\frac{\omega}{2}\sigma_z\right)\colon P(\theta) \to P(\theta + 2\omega)\,; \tag{7.17}$$

this is an infinite-dimensional representation of Pin(2) that contains all the integer-$|Q|$ irreducible representations.

A gauge-invariant operator can be constructed corresponding to each element in the center $\overline{\mathrm{H}}$ of the unbroken gauge group H. Each such operator may be defined as in eq. (3.3), and an order parameter $A(\Sigma, \mathrm{C})$ can be formulated as in eq. (4.4) that probes how the local $\overline{\mathrm{H}}$ symmetry is realized. However, these statements require a qualification. If H is embedded in a continuous gauge group G that undergoes the Higgs phenomenon, then the operator $F(\Sigma)$ corresponding to the $\overline{\mathrm{H}}$ charges is invariant under local G transformations only if $\overline{\mathrm{H}}$ is contained in the center of G. If this is not the case, then the $\overline{\mathrm{H}}$ charges must be constructed in a low-energy effective theory with local H symmetry, as we described in sect. 6.

If the unbroken discrete gauge group H is nonabelian, though, we have argued that there is a charge superselection sector associated with each irreducible representation of H. This is a richer classification than can be probed by the $\overline{\mathrm{H}}$ charge operators alone. But we have been unable to construct gauge-invariant operators whose eigenvalues distinguish the various sectors. Our difficulties result from the peculiar properties of the nonabelian Aharonov–Bohm effect.

In particular, in two spatial dimensions, the H-charge contained in a region is in general ill defined, unless the "magnetic flux" contained in the region is in the center $\overline{\mathrm{H}}$ of H. Still, one can hope that an H-charge superselection rule can be formulated in the sector of the theory with magnetic flux in $\overline{\mathrm{H}}$. Similarly, in three dimensions, H-charge should be well defined in a region that contains cosmic strings as long as no "branch cuts" intersect the boundary of the region.

The task of assigning a definite H-charge to a bounded region is complicated, however, by quantum fluctuations. Virtual vortex pairs near the boundary of the region cause the enclosed magnetic flux to fluctuate. Therefore, computing the expectation value of the charge necessarily involves averaging over values of the enclosed flux for which the charge is ill-defined. Because this is merely a surface effect, though, we expect that a well-defined H-charge can be extracted in the limit of infinite volume.

A related problem is that a global H-gauge transformation cannot be defined when vortices are present. This problem is even more acute for a discrete gauge

group like **Q**, than for a continuous nonabelian group like Pin(2). In the case of Pin(2), a global U(1) gauge transformation can be defined acting on a state that contains an even number of vortices. This transformation is nontrivial at spatial infinity, but can be smoothly deformed to the identity on all vortices and branch cuts. If H is discrete, no such smooth deformation is possible. Only the gauge transformations in the center $\overline{H}$ of H can be performed on a general background, for the other transformations fail to preserve the nontrivial matching conditions on the branch cuts.

A mathematical pursuit might conclude, then, that only the $\overline{H}$ charges can be defined in the presence of vortices; even virtual vortex pairs deep inside a region obstruct the global H transformations acting on the region. We are reluctant to accept so drastic a conclusion. As we have argued, the Aharonov–Bohm phases acquired by various vortices or cosmic strings that wind around a region provide us in principle with sufficient information to determine the H charge contained in the region.

We have been unable to find an operator realization of this experiment, because we do not know how to construct any gauge-invariant operator that creates a cosmic string or vortex pair, if the string or vortex exhibit the nonabelian Aharonov–Bohm effect. This in itself is rather curious, especially in the case of the vortex, which can be a stable particle in a (2 + 1)-dimensional field theory. There should be *S*-matrix elements with vortices in the asymptotic in and out states, but we do not know how to obtain these *S*-matrix elements by applying the LSZ procedure to gauge-invariant Green functions.

We note in passing that a vortex bound to a charged particle provides a realization of nonabelian statistics in 2 + 1 dimensions, if the vortex and particle exhibit the nonabelian Aharonov–Bohm effect. A pair of gauge-equivalent "identical" particles undergoes a nonabelian transformation when the particles are interchanged. The physical interpretation of this construction is rather obscure, since an isolated particle has ill-defined charge, and a pair of particles can carry unlocalized charge.

## 8. The charge of a closed universe

An important distinction between gauge and global symmetries is emphasized in ref. [1]. Because a gauge symmetry arises when the variables used to describe a system are redundant, gauge symmetries, unlike global symmetries, are intrinsically exact. A gauge symmetry that is weakly broken by a small perturbation is an oxymoron.

This distinction has a familiar consequence in connection with the topology-changing interactions that may occur in quantum gravity [5]. Fig. 15 depicts a wormhole in euclidean spacetime, a process in which a parent universe gives birth to a tiny closed baby universe that is disconnected from the parent, and the baby

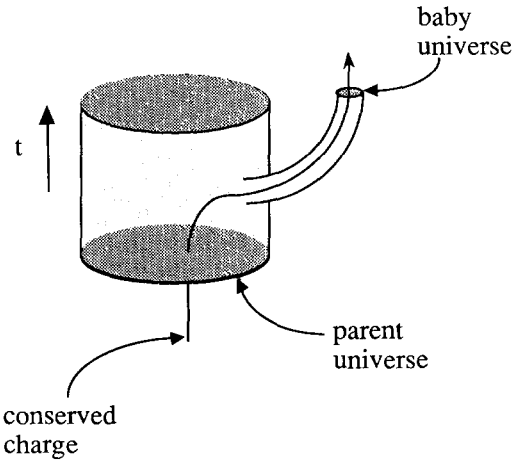


Fig. 15. A globally conserved charge disappears down a wormhole.

universe carries away a unit of a globally conserved charge. An observer in the parent universe sees the charge disappear, and interprets this event as a violation of the global conservation law.

The corresponding process in which a nonvanishing amount of a *locally* conserved charge is swallowed by a wormhole must be forbidden. An observer in the parent universe would interpret this event as a violation of a local conservation law, which we know to be impossible. It is familiar how wormhole physics is reconciled with gauge invariance in the case of a locally conserved charge, like electric charge, that couples to a massless gauge field. The baby closed universe is required to carry vanishing total charge. This requirement is a consequence of the Gauss law – charge can be expressed as a surface integral, but the baby universe has no boundary. From the perspective of the parent universe, a region contained in the universe can pinch off and become a disconnected baby universe only if the region has no hair that can be detected far outside the region.

This heuristic understanding of why wormholes respect gauge symmetries extends to the case of local discrete symmetries; discrete gauge charges, although screened classically, have quantum mechanical hair that can be detected at the boundary of a manifold by means of the Aharonov–Bohm effect. In a manifold without boundary, such hair is absent, and so the total gauge charge must be trivial.

In the case of an abelian discrete gauge symmetry, the charge operator defined in sect. 3 explicitly expresses the charge enclosed by a surface in terms of fields on that surface. Eq. (3.1) is the analog of the Gauss law for a discrete charge, and suffices to ensure that a closed manifold contains zero total charge. That is, a closed manifold can be obtained from a manifold with boundary $\Sigma$ in the limit in

which $\Sigma$ shrinks to a point; in this limit $F(\Sigma)$ approaches one, and the charge enclosed by $\Sigma$ vanishes. In the nonabelian case, although we are unable to construct a gauge-invariant operator realization of the Gauss law, we have seen that the hair can in principle be detected by cosmic strings at the boundary.

In the nonabelian case, however, the "total charge" is not necessarily the same as the sum of all the point charges contained in the universe. If the universe is not a simply connected manifold, and the gauge group is nonabelian, then addition of charges may be ambiguous, just as we found in the background of a string or vortex pair. To illustrate this phenomenon, consider the model with unbroken gauge group H = Pin(2) that we described in sect. 7. A universe with a handle attached to it might be an "Alice universe" on which the electric field $\boldsymbol{E}$ and electric charge $Q$ are double-valued; if we restrict $\boldsymbol{E}$ and $Q$ to a single branch, then, there must be a branch cut on a closed surface contained in the handle, where $\boldsymbol{E}$ and $Q$ change sign. In such an Alice universe, a charged particle that voyages through the handle returns to its starting point with its charge flipped in sign [42].

Of course, this process in which a charge changes sign by traversing a handle must not modify the electric field as measured far away from the handle. Indeed, just as we saw that a charged particle that passes through a loop of Alice string must transfer charge to the loop, a charged particle that traverses an "Alice handle" transfers charge to the handle. The charge of the handle can be defined by means of the electric flux through a surface that is the boundary of a region that contains the handle, as long as this surface does not intersect the branch cut. While the handle itself has a well-defined electric charge, this charge is not localized anywhere inside the handle. The branch cut appears to be the source of the flux, but the cut is an unphysical gauge artifact, and no actual charge resides there.

In general, if $\boldsymbol{E}$ is restricted to a single branch, then the "total charge" is the sum of the point charges and of the "charges" on all of the branch cuts; this is the quantity that must vanish in a closed universe. Fig. 16 depicts the electric field of a single charged particle in an Alice universe; the charge $Q$ of the particle is compensated by the charge $-Q$ of the handle.

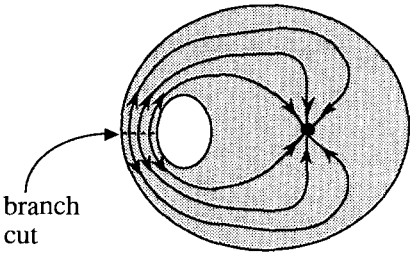


Fig. 16. One branch of the electric field of a point charge in an Alice universe.

Since the element $e^{2\pi iQ}$ is in the center of Pin(2), it commutes with the nontrivial transition function on the branch cut, and a handle is therefore not permitted to carry this charge. This means that $Q$ modulo an integer is a well-defined additive quantum number that can be computed by summing the values of the point charges. In particular, then, the number of elementary point charges with $|Q| = \frac{1}{2}$ must be even in a closed universe.

This discussion of charge in the Pin(2) model is also applicable to the case of nonabelian *discrete* gauge symmetry, as we argued in sect. 7.

## 9. Conclusions

In a gauge theory, a superselection rule arises whenever there are states that are endowed with properties that can be detected at arbitrarily long range. In this sense, the superselection rules provide a classification of the types of "hair" that a localized object can carry. This paper has aimed to clarify the nature of the superselection rules in theories with local symmetry. In particular, we have examined the notion of "quantum-mechanical hair" that is invisible classically, but can be detected via the Aharonov–Bohm effect. Such quantum-mechanical hair can in principle be carried by a black hole.

In the case of quantum-mechanical hair that is associated with an abelian discrete gauge symmetry, our discussion has been reasonably complete. We constructed a charge operator whose eigenvalues distinguish the various superselection sectors of the theory, and formulated a nonlocal order parameter that probes the realization of the local symmetry. Our construction translates into operator language the heuristic idea that the charge inside a region can be measured as the Aharonov–Bohm phase that is acquired by a cosmic string that winds around the region. The charge operator provides, in particular, an operator description of the quantum-mechanical hair on a black hole.

In the case of a nonabelian discrete gauge symmetry, our grasp of the superselection rules is in a less satisfactory state. We proposed that, as in the abelian case, the charge superselection sectors are classified by the irreducible representations of the gauge group, and we described how the representation content can be inferred from measurable Aharonov–Bohm phases. However, we succeeded in constructing gauge-invariant charge operators only for the charges in the center of the gauge group. We could not translate the procedure for measuring Aharonov–Bohm phases into operator language because, curiously, we could not find a gauge-invariant operator that creates the desired cosmic string. This difficulty is closely related to a remarkable feature of the nonabelian Aharonov–Bohm effect – a loop of cosmic string can carry charge, even though the charge cannot be localized anywhere on the string or in its vicinity.

The implications of the Aharonov–Bohm effect, and of local discrete symmetry, are surprisingly deep; we feel that these implications have not yet been fully

explored. We expect, in particular, that further investigation of the nonabelian Aharonov–Bohm effect will prove to be highly rewarding.

Alford et al. [43, 44] have independently investigated the properties of quantum-mechanical hair and the nonabelian Aharonov–Bohm effect.

J.P. gratefully acknowledges a helpful discussion with Sidney Coleman concerning the Alice string and Cheshire charge. L.K. acknowledges useful discussions with Frank Wilczek about discrete gauge charges and local discrete symmetry. We have also benefited from conversations with Frank Accetta, Jim Hughes, Joe Polchinski, David Politzer, Soo-Jong Rey, Opher Shapira, and Lenny Susskind. This work was initiated at the Aspen Center for Physics.

## Note added in proof

That a general global gauge transformation cannot be implemented in the background of a nonabelian vortex (as described in sect. 7) was previously pointed out by Balachandran, Lizzi and Rogers [45]. The effect of an axion mass on axionic charge (discussed in sect. 3) has been addressed independently by Allen and Bowick [46].

## References

[1] L.M. Krauss and F. Wilczek, Phys. Rev. Lett. 62 (1989) 1221
[2] Y. Aharonov and D. Bohm, Phys. Rev. 115 (1959) 485
[3] R. Rohm, Princeton University Ph.D. Thesis (1985) unpublished;
M.G. Alford and F. Wilczek, Phys. Rev. Lett. 62 (1989) 1071
[4] L.M. Krauss, Gen. Rel. Grav. 22 (1990) 253
[5] S.W. Hawking, Phys. Lett. B195 (1987) 337; Phys. Rev. D37 (1988) 904;
G.V. Lavrelashvili, V. Rubakov and P.G. Tinyakov, JETP Lett. 46 (1987) 167;
S.B. Giddings and A. Strominger, Nucl. Phys. B306 (1988) 890; B307 (1988) 854;
S. Coleman, Nucl. Phys. B307 (1988) 867; B310 (1988) 643
[6] G. Gilbert, Nucl. Phys. B328 (1989) 159
[7] K. Fredenhagen, *in* Fundamental Problems of Gauge Field Theory, ed. G. Velo and A.S. Wightman (Plenum, New York, 1986)
[8] G. Morchio and F. Strocchi, *in* Fundamental Problems of Gauge Field Theory, ed. G. Velo and A.S. Wightman (Plenum, New York, 1986)
[9] J. Fröhlich and P.A. Marchetti, Commun. Math. Phys. 121 (1989) 177
[10] G. 't Hooft, Nucl. Phys. B138 (1978); B153 (1979) 141
[11] S. Coleman, *in* The Unity of the Fundamental Interactions, ed. A. Zichichi (Plenum, New York, 1983)
[12] M. Srednicki and L. Susskind, Nucl. Phys. B179 (1981) 239
[13] A.K. Gupta, J. Hughes, J. Preskill and M.B. Wise, Nucl. Phys. B333 (1990) 195
[14] M.J. Bowick, S.B. Giddings, J.A. Harvey, G.T. Horowitz, and A. Strominger, Phys. Rev. Lett. 61 (1988) 2823
[15] A.M. Polyakov, Phys. Lett. B59 (1975) 82; Nucl. Phys. B120 (1978) 477
[16] S.-J. Rey, Phys. Rev. D40 (1989) 3396
[17] J.D. Bekenstein, Phys. Rev. D5 (1972) 1239, 2403;
C. Teitelboim, Phys. Rev. D5 (1972) 2941

[18] S. Elitzur, Phys. Rev. D12 (1975) 3978
[19] K.G. Wilson, Phys. Rev. D10 (1974) 2445
[20] E. Fradkin and S. Shenker, Phys. Rev. D19 (1979) 3682;
T. Banks and E. Rabinovici, Nucl. Phys. B160 (1979) 349
[21] S. Dimopoulos, S. Raby, and L. Susskind, Nucl. Phys. B173 (1980) 208
[22] T.W.B. Kibble, G. Lazarides, and Q. Shafi, Phys. Rev. D26 (1982) 435
[23] A. Vilenkin, Phys. Rep. 121 (1985) 263
[24] J. Preskill, *in* Architecture of the Fundamental Interactions at Short Distances, ed. P. Ramond and R. Stora (North-Holland, Amsterdam, 1987)
[25] F. Wegner, J. Math. Phys. 12 (1971) 2259
[26] R. Balian, J.M. Drouffe, and C. Itzykson, Phys. Rev. D11 (1975) 2098; 2104
[27] M. Creutz, Phys. Rev. D21 (1980) 1006;
G.A. Jongeward, J.D. Stack, and J. Jayaprakash, Phys. Rev. D21 (1980) 3360
[28] A. Ukawa, P. Windey, and A. Guth, Phys. Rev. D21 (1980) 1013
[29] K. Osterwalder and E. Seiler, Ann. Phys. (NY) 110 (1978) 440;
E. Seiler, Gauge Theories as a Problem of Constructive Quantum Field Theory and Statistical Mechanics (Springer, Berlin, 1982)
[30] T. Banks, Nucl. Phys. B323 (1989) 90
[31] D.J. Gross and R. Jackiw, Phys. Rev. D6 (1972) 477
C. Bouchiat, J. Iliopoulos, and Ph. Meyer, Phys. Lett. B38 (1972) 519
[32] E. Witten, Phys. Lett. B117 (1982) 432
[33] J. Preskill, Gauge anomalies in an effective field theory, Caltech preprint CALT-68-1493 (1990)
[34] T.T. Wu and C.N. Yang, Phys. Rev. D12 (1975) 3845
P.A. Horváthy, Phys. Rev. D33 (1986) 407;
R. Sundrum and L.J. Tassie, J. Math. Phys. 27 (1986) 1566
[35] S. Coleman, *in*: New Phenomena in Subnuclear Physics, ed. A. Zichichi (Plenum, New York, 1977)
[36] N.D. Mermin, Rev. Mod. Phys. 51 (1979) 591
[37] A.S. Schwarz, Nucl. Phys. B208 (1982) 141
[38] P. Nelson and A. Manohar, Phys. Rev. Lett. 50 (1983) 943;
A. Balachandran, G. Marmo, M. Mukunda, J. Nilsson, E. Sudarshan, and F. Zaccaria, Phys. Rev. Lett. 50 (1983) 1553
[39] S. Coleman and P. Nelson, Nucl. Phys. B237 (1984) 1
[40] V. Poénaru and G. Toulouse, J. Phys. (Paris) 38 (1977) 887
[41] T.W.B. Kibble, Phys. Rep. 67 (1980) 183
[42] J. Kiskis, Phys. Rev. D17 (1978) 3196
[43] M.G. Alford, J. March-Russell and F. Wilczek, Nucl. Phys. B337 (1990) 695
[44] M.G. Alford, K. Benson, S. Coleman, J. March-Russell, and F. Wilczek, Phys. Rev. Lett. 64 (1990) 1632
[45] A.P. Balachandran, F. Lizzi and V.G.J. Rodgers, Phys. Rev. Lett. 52 (1984) 1818
[46] T.J. Allen, M.J. Bowick and A. Lahiri, Axionic black holes from massive axions, Syracuse preprint SU-HEP-4238-405 (1989)