HUTP-82/A032

THE MAGNETIC MONOPOLE FIFTY YEARS LATER*

Sidney Coleman

Lyman Laboratory of Physics
Harvard University
Cambridge, Massachusetts 02138

These lectures were given at the 1981 International School of Subnuclear Physics, "Ettore Majorana". Versions of these lectures were also given at the VI Brazilian Symposium on Theoretical Physics, at Les Houches École d'été de Physique Théorique, and at the Banff Summer Institute on Particles and Fields.

*Research was supported in part by the U.S. National Science Foundation under Grant No. PHY77-22864.

6/82

TABLE OF CONTENTS

HUTP-82/A032

THE MAGNETIC MONOPOLE FIFTY YEARS LATER

Sidney Coleman

Lyman Laboratory of Physics
Harvard University
Cambridge, Massachusetts 02138

E R R A T A

The last paragraph on page 75 should be changed to:

The larger expectation value breaks SU(5) down to $[SU(3) \otimes SU(2) \otimes U(1)]/Z_6$. If we realize SU(5) as 5×5 matrices, SU(3) consists of transformations on the first three coordinates, SU(2) on the last two, and U(1) of diagonal matrices of the form

$$\text{diag}(e^{2i\theta}, e^{2i\theta}, e^{2i\theta}, e^{-3i\theta}, e^{-3i\theta}). \qquad (5.23)$$

If $e^{6i\theta} = 1$, this is in $SU(3) \otimes SU(2)$; this is why we must take the quotient of the direct product by Z_6. The SU(3) subgroup is ..... (continues as in original text).

On page 95, Footnote 10 should read:
This problem was solved very early on by I. Tamm, Z. Phys. 71, 141 (1931), and my results are the same as his, although my method is somewhat different. To my knowledge, the treatment in the literature closest to that given here is that of H. J. Lipkin, W. I. Weisberger, and M. Peshkin, Ann. of Phys. 53, 203 (1969).

THE MAGNETIC MONOPOLE FIFTY YEARS LATER

Sidney Coleman

Lyman Laboratory of Physics, Harvard University

Cambridge, Massachusetts 02138

1. INTRODUCTION

This is a jubilee year. In 1931, P. A. M. Dirac[1] founded the theory of magnetic monopoles. In the fifty years since, no one has observed a monopole; nevertheless, interest in the subject has never been higher than it is now.

There is good reason for this. For more than forty years, the magnetic monopole was an optional accessory; Dirac had shown how to build theories with monopoles, but you didn't have to use them if you didn't want to. Seven years ago, things changed; 't Hooft and Polyakov[2] showed that magnetic monopoles inevitably occur in certain gauge field theories. In particular, all grand unified theories necessarily contain monopoles. (In this context, a grand unified theory is one in which a semi-simple internal symmetry group spontaneously breaks down to electromagnetic U(1).) Many of us believe that grand unified theories describe nature, at least down to the Planck length. So where are the monopoles?

As we shall see, grand unified monopoles are very heavy; a typical mass is roughly a hundred times greater than the grand unified scale. Thus they are not likely to be made by contemporary accelerators or supernovae. However, energy was more abundant

shortly after the big bang. As Preskill[3] pointed out, naive estimates would lead one to believe that monopoles would have been produced so copiously in the very early universe and annihilated so inefficiently subsequently that they would at the current time form the dominant contribution to the mass of the universe.

Thus the absence of monopoles is significant. It tells us something about the extreme early universe or about extreme microphysics or about both. Since evidence on both these subjects is in short supply, monopoles are important.

This is the last I will say in these lectures about cosmology, or indeed about any reason for studying monopoles. We will have enough to occupy us just developing the basics of monopole theory.

I've tried in these lectures to go from the simple to the complex, from monopoles as seen at large distances to monopole internal structure, from classical physics to quantum mechanics. Of course, I've not been able to keep rigorously to this program; for example, elementary quantum mechanics will be with us from the very beginning, although we will not deal with the full complexities of quantum field theory until the very last lecture.

The organization of these lectures is as follows:

In Section 2 I shall discuss the theory of a classical magnetic monopole as seen from the outside. That is to say, I shall investigate only those questions that can be answered without looking at the monopole interior, without asking, for example, whether there is a real singularity at its core or only some complicated excitation of degrees of freedom that do not propagate out to large distances, like massive gauge fields.

In Section 3 I shall extend the analysis to a classical non-Abelian monopole. I emphasize that by this I do not mean an object that has massive non-Abelian fields in its core, but one which is surrounded by massless non-Abelian gauge fields that extend out to large distances. Since no such fields exist in nature, this may seem a silly exercise, but there are two good reasons for doing it.

Firstly, there is a pedagogical reason. In investigating monopoles in unbroken non-Abelian gauge theories, we will encounter mathematical structures that will be useful to us later, when we deal with the more complicated case of spontaneous symmetry breakdown. Secondly, there is a physical reason. Some of the monopoles that arise in grand unified theories have colored gauge fields surrounding them; they are color magnetic monopoles as well as ordinary electromagnetic monopoles. It is true that these colored fields are damped by confinement effects at distances greater than 10^{-13} cm. However, the cores of these monopoles are on the order of the grand unification scale, something like 10^{-28} cm. Thus we have fifteen orders of magnitude in which they look like non-Abelian monopoles, plenty of room for interesting physics.

In Section 4 we shall plunge into the belly of the beast, and see under what conditions the structures we have discovered at large distances can be continued to small distances without encountering singularities. This is a subject I discussed in my 1975 Erice lectures[4] (coming to it from a different starting point), and although I'll try to keep it to the minimum, there will be some unavoidable recycling of those lectures here.

Finally, in Section 5, I'll turn to quantum mechanics and discuss things like dyonic excitations and the effects of confinement.

There's a lot that's not in these lectures. I won't deal with cosmology, as I've said, nor will I have anything to say about the interactions of monopoles with ordinary matter, how they make tracks in emulsions and damp galactic magnetic fields. On a more theoretical level, I won't talk about exact solutions, index theorems, or fermion fractionalization. Finally, I've made no attempt to compose a full and fair bibliography. If I don't refer to a paper, please assume this is a result of ignorance and laziness, not of informed critical judgment.[5]

Most of what I know about this subject I've learned from conversations with Curtis Callan, Murray Gell-Mann, Jeffrey Goldstone, Roman Jackiw, Ken Johnson, David Olive, Gerard 't Hooft, and Erick Weinberg. It is a pleasure to acknowledge my debt.

Notational conventions: As usual, Greek indices run from 0 to 3, latin indices from the middle of the alphabet from 1 to 3. When doing ordinary vector analysis (as in all of Sec. 2), the signature of the three-dimensional metric is (+++); when working in four dimensions, the signature is (+---). I will usually set $\hbar$ and c equal to one, although occasionally I will restore explicit $\hbar$'s when discussing the approach to the classical limit. Rationalized units are used for electromagnetism; thus, the electromagnetic Lagrange density is $\frac{1}{2}(E^2-B^2)$ and there are no π's in Maxwell's equations.

2. ABELIAN MONOPOLES FROM AFAR

2.1 The Monopole Hoax and a First Look at Dirac's Quantization Condition

Consider some arrangement of stationary charges and currents restricted to a bounded region. Outside this region there are only stationary electromagnetic fields, vanishing at spatial infinity. Without detailed knowledge of the distribution of charges and currents, what can we say about the exterior fields?

The answer is known by anyone who has taken a course in electromagnetic theory. The exterior fields are a sum of electric monopole, magnetic dipole, electric quadrupole, etc. The successive terms in this series fall off faster and faster at infinity; the leading term at large distances is the electric monopole,

$$\vec{E} = \frac{e\,\vec{r}}{4\pi\, r^3} , \qquad \vec{B} = 0 , \tag{2.1}$$

where e is a real number, called the electric charge (of the system inside the region).

This analysis depends on the assumption that the only things inside the region are charges and currents, that is to say, that the empty-space Maxwell's equations,

$$\vec{\nabla}\cdot\vec{B} = 0 = \vec{\nabla}\times\vec{E} + \partial\vec{B}/\partial t , \tag{2.2}$$

remain true inside the region. If we drop this assumption, we must add to the series above a dual series, consisting of magnetic monopole,

electric dipole, etc. The leading term at large distances is the magnetic monopole,

$$\vec{E} = 0 , \qquad \vec{B} = \frac{g\vec{r}}{r^3} , \tag{2.3}$$

where g is a real number, called the magnetic charge (of the system within the region).[6] Systems carrying non-zero magnetic charge are somewhat confusingly also called magnetic monopoles.

As I stated in my introductory remarks, no one has ever observed a magnetic monopole. Let us imagine, though, that we attempt to hoax a gullible experimenter into believing that he has discovered a monopole. For this purpose we obtain a very long, very thin solenoid; it is best if it is many miles long and considerably thinner than a fermi. (This is very much a *gedanken* hoax.) We put one end of the solenoid in the experimenter's laboratory in Geneva and the other one here in Erice. We then turn on the current. The experimenter sees $4\pi g$ worth of magnetic flux emanating from his lab bench; he can not detect the fact that it is being fed in along the solenoid; he thinks he has a monopole.

Is there any way he can tell he has been hoaxed, that his monopole is a fake? Certainly not, if all he has access to are classical charged particles, for all these see are $\vec{E}$ and $\vec{B}$, and $\vec{E}$ and $\vec{B}$ are the same as they would be for a real monopole. (Except within the solenoid, but this is by assumption so thin as to be undetectable.)

However, the situation is very different if he has access to quantum charged particles. With these, a cunning experimenter can search for the solenoid via the Bohm-Aharonov effect.[7]

This is a variant of the famous two-slit diffraction experiment shown in Fig. 1. Charged particles emitted by the source A pass through the two slits in the screen B and are detected at the screen C. As anyone who has got six pages into a quantum mechanics text knows, the amplitudes for passage through the individual slits combine coherently; the probability density at C is

$$|\psi_1 + \psi_2|^2 , \tag{2.4}$$

where ψ_1 is the amplitude for passage through the first slit and ψ_2 that for passage through the second. If a solenoid, S, shown end on in the figure, is placed between the slits, the probability density at C is changed; it becomes

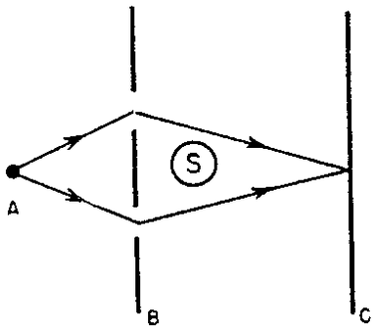

Figure 1

$$|\psi_1 + e^{ie\Phi}\psi_2|^2 \tag{2.5}$$

where e is the charge of the particle and Φ is the flux through the solenoid (in our case, $4\pi g$.)

Thus, by moving this apparatus about and observing the change in the interference pattern, our experimenter can detect the solenoid, unless eg is a half-integer,

$$eg = 0, \pm\tfrac{1}{2}, \pm 1, \ldots . \tag{2.6}$$

In this case, the expressions (2.4) and (2.5) are identical. The solenoid is undetected and our hoax has succeeded.

Equation (2.6) is the famous Dirac quantization condition.[1] As we shall see, if it is obeyed, the solenoid is not only undetectable by the Bohm-Aharonov effect, but undetectable by any conceivable method. It can be abstracted away to nothing, becoming a mathematical singularity of no physical significance, the Dirac string; all that is left behind is a genuine monopole. To demonstrate these assertions requires a closer analysis of the physics of a charged particle moving in a monopole field, to which I now turn.

2.2 Gauge Invariance and a Second Look at the Quantization Condition

Consider a spinless non-relativistic particle moving in some time-independent magnetic field (described in the usual way, by a vector potential, $\vec{A}$), and perhaps also in some scalar potential, V. The Schrödinger equation for this system is

$$i\frac{\partial\psi}{\partial t} = -\frac{1}{2m}(\vec{\nabla} - ie\vec{A})^2\psi + V\psi \ . \tag{2.7}$$

This system has a famous invariance, gauge invariance. For our purposes, we need only consider time-independent gauge transformations:

$$\psi \to e^{-ieX}\psi \ , \tag{2.8a}$$

$$\vec{A} \to \vec{A} - \vec{\nabla}X = \vec{A} - \frac{i}{e}e^{ieX}\vec{\nabla}e^{-ieX} \ . \tag{2.8b}$$

I have written the second of these equations in a somewhat unconventional form to emphasize that the only relevant quantity is $\exp(-ieX)$. Thus, for example, X and $X+2\pi/e$ are different functions, but they define the same gauge transformation.[8]

After these generalities, let us turn to the case of interest, a point monopole at the origin of coordinates. We can think of this as the field of a genuine point object, or as the field of some extended object as viewed from very large distances; it doesn't matter. Of course, it is impossible to find a vector potential whose curl is the monopole field, because the monopole field is not divergenceless. However, it is possible to find a potential that does the job except for a line extending from the monopole to infinity, "the Dirac string". Such a potential is simply the potential of our hoax, in the limit that the solenoid becomes infinitely long and infinitely thin.

For example, if I arrange the string along the negative z-axis, $\vec{A}$ is given by

$$\vec{A}\cdot d\vec{x} = g(1-\cos\theta)d\phi \ . \tag{2.9}$$

As promised, $\vec{A}$ is ill-defined on the string, $\theta = \pi$. Let's check that it has the right curl away from the string. The computation is done most simply using the methods of tensor analysis. The only non-vanishing covariant component of A is

$$A_\phi = g(1-\cos\theta) \ . \tag{2.10}$$

Thus, the only nonvanishing component of the field-strength tensor is

$$F_{\theta\phi} \equiv \partial_\theta A_\phi - \partial_\phi A_\theta = g\sin\theta \ . \tag{2.11}$$

This indeed corresponds to a radial magnetic field. To check that it has the right magnitude, we compute the flux through an infinitesimal element of solid angle:

$$F_{\theta\phi}\,d\theta\,d\phi = \frac{g}{r^2}\,r^2\sin\theta\,d\theta\,d\phi \ . \tag{2.12}$$

This is the desired result, g/r^2 times the infinitesimal element of area.

I emphasize that I have made no attempt to define $\vec{A}$ or compute $\vec{B}$ on the string. As we shall see, there is no need to inquire into these matters.

Of course, the choice of the negative z-axis is totally arbitrary. For example, I could just as well have put the string along the positive z-axis:

$$\vec{A}'\cdot d\vec{x} = -g(1+\cos\theta)d\phi \ . \tag{2.13}$$

It will be useful to us later to note that the two potentials we have introduced are simply related:

$$\vec{A}'(\vec{x}) = \vec{A}(-\vec{x}) \ . \tag{2.14}$$

In his classic monopole paper, Dirac showed that if the quantization condition was obeyed, the string was unobservable. The most straightforward way to demonstrate this is by showing that there is a gauge transformation that moves the string from the negative z-axis to any other desired location, that transforms the potential $\vec{A}$ into, for example, the potential $\vec{A}'$. However there are unpleasant features to this argument: the gauge transformation is necessarily singular at both the old and new locations of the string, and singular gauge transformations make people uneasy.

I will now give a refinement of this argument, due to Wu and Yang,[9] that avoids these difficulties. In the Wu-Yang construction, we never have to deal with singular gauge transformations, nor indeed with singular potentials (except, of course, at the origin, where $\vec{B}$

blows up and there is a real singularity). The price we pay for this is the necessity of using different vector potentials in different regions of space.

Let me define the upper region of space as the open set $\theta < 3\pi/4$ and the lower region as the open set $\theta > \pi/4$. The union of these two regions is all of space (except for the origin of coordinates, which is a singular point anyway). Let me define the monopole field by using the potential $\vec{A}$ in the upper region and the potential $\vec{A}'$ in the lower region. Thus neither potential has any singularity in its region of validity; in each case the string lies outside the region.

This is an adaptation of a mapmaker's stratagem. There is no way of mapping the spherical surface of the earth onto a portion of a plane without introducing singularities. For example, in the United Nations flag, a north polar projection, the south pole, a single point, is mapped into the circumference of the map. However, we can easily do the job with two maps. For example, we could use a north polar projection for the region north of the Tropic of Capricorn, and a south polar projection for that south of the Tropic of Cancer. The two maps together would form a singularity-free representation of the globe. However, we must be sure that the two maps fit together properly, that we have not by error received a map of the northern part of the earth and the southern part of Mars. That is to say, in their equatorial region of overlap, the two maps must describe the same geography.

Transposing this criterion from cartography to field theory, we must be sure that the two vector potentials describe the same physics in the overlap of the upper and lower regions, that is to say, that $\vec{A}'$ is a gauge transform of $\vec{A}$. Thus, in the overlap region, we must find χ such that

$$\vec{A} - \vec{A}' = \vec{\nabla}\chi . \tag{2.15}$$

This is easy:

$$(\vec{A} - \vec{A}') \cdot d\vec{x} = 2g\, d\phi . \tag{2.16}$$

Hence,

$$\chi = 2g\phi . \tag{2.17}$$

This is not a well-defined function in the overlap region; it is

multiple-valued. Fortunately, as I explained at the beginning of this section, the relevant object is not χ but $\exp(-ie\chi)$. This is single-valued if

$$eg = 0, \pm\tfrac{1}{2}, \pm 1 \ldots . \tag{2.6}$$

This is the Dirac quantization condition. This completes the argument.

In principle, the singularity-free Wu-Yang description of the monopole field is much cleaner than the Dirac description with its singular string. However, in practice, it's an awkward business to be continually changing gauges as one moves about. Thus, I'll mostly work with the Dirac description in the remainder of these lectures, and use due care with the string, just as, for example, one uses due care with the coordinate singularities when working with polar coordinates.

2.3 Remarks on the Quantization Condition

(1) All observed electric charges are integral multiples of a common unit; this is called charge quantization. There is no explanation of this striking phenomenon in either classical or quantum electrodynamics; nothing would go wrong if the proton charge were π times that of the electron, for example. One of the most attractive features of Dirac's monopole theory when it was first proposed was that it explained charge quantization. If there were monopoles in the universe (indeed, if there was just one monopole in the universe), all electric charges would be forced to be integral multiples of a common unit, the inverse of twice the minimum magnetic charge.

Of course, nowadays no one believes in pure electrodynamics. We believe the electrodynamic U(1) group is part of a larger group that suffers spontaneous symmetry breakdown. However, this explains charge quantization only if the larger group is semi-simple. (Actually, this is overstating the case. The larger group can be the product of a semi-simple group and a bunch of U(1) factors, as long as the electrodynamic group lies completely in the semi-simple factor.) As I said in the Introduction, and as I will demonstrate

later, in this case the theory inevitably contains monopoles.

Thus the connection between monopoles and charge quantization is firmer than ever, although the details are quite different from what was originally envisaged. Both are consequences of a common cause, spontaneous breakdown of a semi-simple symmetry.

(2) Although I won't have much to say about it in these lectures, considerable effort has been devoted over the years to developing quantum electrodynamics with monopoles, a relativistic field theory with both electrically and magnetically charged fundamental particles. A peculiar feature of this theory is that one can not investigate it using diagrammatic perturbation theory, the method that is so successful for ordinary QED. This is because perturbation theory of any kind is nonsense. The two coupling constants in the theory can not simultaneously be made arbitrarily small; because of the quantization condition, as e becomes small, g becomes large.

This does not mean that the effects of virtual monopoles are necessarily large, even for very small e. Large couplings enhance the effects of virtual particles, but large masses diminish them, both through large energy denominators and through the fall-off of form factors at large momentum transfers. Thus, even very strongly coupled particles can have small effects at low energies, if they are sufficiently massive. As we shall see, this is what happens for monopoles in gauge field theories. The masses of these particles are proportional to $1/e^2$; as e goes to zero, their effects are negligible at any fixed energy.

(3) Dyons are defined to be objects that carry both electric and magnetic charge. We can imagine constructing dyons by binding together electrically and magnetically charged particles, or we can imagine them as fundamental entities, not composed of more primitive objects. In either case, we can quickly work out the properties of dyons by exploiting the invariance of Maxwell's equations under duality rotations.

A duality rotation is defined by:

$$\vec{E} \to \vec{E}\cos\alpha + \vec{B}\sin\alpha \ ,$$
$$\vec{B} \to -\vec{E}\sin\alpha + \vec{B}\cos\alpha \ , \qquad (2.18)$$

where α is a real number. It is easy to check that this is an invariance of the empty-space Maxwell equations. We describe Eq. (2.18) by saying that $(\vec{E},\vec{B})$ is a duality vector. From Eqs. (2.3) and (2.4), $(e,4\pi g)$ is also a duality vector.

Given two dyons, with electric and magnetic charges e_i and g_i, $(i = 1,2)$, we can form two duality invariants bilinear in the charges:

the inner product, $\quad e_1e_2 + 16\pi^2 g_1g_2 \ , \qquad (2.19a)$

and the outer product, $\quad 4\pi(e_1g_2 - g_1e_2) \ . \qquad (2.19b)$

Any observable property of the two-dyon system must be expressible in terms of these invariants.

For example, the force a dyon fixed at the origin exerts on another, moving, dyon must be a linear function of these invariants. Thus, in the nonrelativistic limit, for example,

$$\vec{F} = (\frac{e_1e_2}{4\pi} + 4\pi g_1g_2)\vec{r}\,/\,r^3 + (e_1g_2 - g_1e_2)\vec{v}\times\vec{r}\,/\,r^3 \ , \qquad (2.20)$$

where "1" is the moving dyon and "2" the fixed one. The reason is that this is the only expression that agrees with the known results for both $g_1 = g_2 = 0$ and $g_1 = e_2 = 0$.

By the same reasoning, the Dirac quantization condition for dyons is

$$e_1g_2 - g_1e_2 = 0, \ \pm\tfrac{1}{2}, \ \pm 1 \ \ldots \ . \qquad (2.21)$$

This equation has some bizarre solutions. For example, it is satisfied by

$$e_i = n_ie + m_if \ ,$$
$$g_i = m_i/2e \ , \qquad (2.22)$$

where n_i and m_i are integers, e is an arbitrary real number, and so is f. That is to say, it is perfectly consistent with the quantization condition for all magnetically charged particles to have fractional electric charge. (We shall see eventually that not only is it consistent, there are circumstances in which it is inevitable.)

(4) When discussing the monopole hoax, I said that the solenoid

was invisible to classical charged particles but detectable (unless the quantization condition was satisfied) by quantum ones. However, the solenoid could also be detected by another kind of physical entity, a classical charged field. We have essentially already seen this. I have been talking about the Schrödinger equation as if it were the equation for a quantum probability amplitude (as it is), but all of our arguments would have been just as valid if ψ were just a classical field like any other (as Schrödinger briefly thought it was).

Thus, Dirac's quantization condition can be a consequence of classical physics, if classical physics contains charged fields. I make this point now because we will shortly be dealing with classical field theories that do contain charged fields, Yang-Mills field theories. There we will find the quantization condition again, and I don't want you to be disoriented by worrying about what a quantum effect is doing in a purely classical context.

(There is one difference between the classical-field and quantum-particle versions of the quantization condition that has been obscured by my use of units in which $\hbar$ is one. The electric charge of a classical field, the quantity that governs the strength of its electromagnetic interactions, is very different from the electric charge of a classical particle; the two even have different dimensions. They are linked together only by the quantum wave/particle duality; e_{field} is $e_{particle}/\hbar$. Thus, for fields, the condition says that ge is a half integer; for particles, that ge is a half-integral multiple of $\hbar$.)

2.4 Funny Business with Angular Momentum

Rotational invariance is a great simplifier of dynamical problems. For example, for a non-relativistic spinless particle moving in a central potential,

$$H = -\frac{1}{2m}\vec{\nabla}^2 + V(r) \ . \tag{2.23}$$

This operator commutes with angular momentum. On a subspace of states of given total angular momentum, H simplifies to:

$$H_\ell = -\frac{1}{2m}\left(\frac{\partial^2}{\partial r^2} + \frac{2}{r}\frac{\partial}{\partial r}\right) + \frac{\ell(\ell+1)}{2mr^2} + V \ , \tag{2.24}$$

where $\ell = 0, 1, 2 \ldots$.

In this subsection I will extend this result to a particle moving in the field of a monopole,[10]

$$H = -\frac{1}{2m}(\vec{\nabla} - ie\vec{A})^2 + V(r) \ , \tag{2.25}$$

where $\vec{A}$ is the monopole vector potential, Eq. (2.9). The first obstacle to the analysis is that the monopole potential spoils manifest rotational invariance. We will avoid this problem by introducing

$$\vec{D} \equiv \vec{\nabla} - ie\vec{A} \ . \tag{2.26}$$

This is a gauge-invariant operator. Further, H is expressible in terms of $\vec{D}$ and the position operator, $\vec{r}$,

$$H = -\frac{1}{2m}\vec{D}^2 + V(r) \ , \tag{2.27}$$

and these operators obey a rotationally-invariant set of commutation relations,[11]

$$[r_i, r_j] = 0 \tag{2.28}$$

$$[D_i, r_j] = \delta_{ij} \tag{2.29}$$

and

$$[D_i, D_j] = -ieg\epsilon_{ijk}r_k/r^3 \ . \tag{2.30}$$

Our method will be to work as much as possible with Eqs. (2.27) to (2.30) and avoid invoking the explicit form of $\vec{A}$.

Our first step is to construct an angular momentum operator, $\vec{L}$, a vector function of $\vec{D}$ and $\vec{r}$ obeying

$$[L_i, D_j] = i\epsilon_{ijk}D_k \tag{2.31}$$

$$[L_i, r_j] = i\epsilon_{ijk}r_k \ . \tag{2.32}$$

As a consequence of these, $\vec{L}$ will automatically obey

$$[L_i, L_j] = i\epsilon_{ijk}L_k \tag{2.33}$$

and

$$[L_i, H] = 0 \ . \tag{2.34}$$

The natural guess is

$$\vec{L} \stackrel{?}{=} -i\vec{r}\times\vec{D} \ . \tag{2.35}$$

However, this doesn't work; it has the right commutators with $\vec{r}$, but not with $\vec{D}$. The right answer turns out to be

$$\vec{L} = -i\vec{r}\times\vec{D} - eg\vec{r}/r \ . \qquad (2.36)$$

The second term looks very strange indeed; in Rabi's immortal words about something else altogether, "Who ordered that?"

I know of three ways to answer this question. I will sketch out two of them (leaving the details as an exercise), and give the third in full.

(1) I have already given the first answer: Eq. (2.36) gives the right commutation relations. I leave the verification of this statement as an exercise.

(2) From the Heisenberg equations of motion,

$$\dot{\vec{r}} = -i\vec{D}/m \ . \qquad (2.37)$$

Hence
$$\vec{L} = m\vec{r}\times\dot{\vec{r}} - eg\vec{r}/r \ . \qquad (2.38)$$

Thus, there is angular momentum in the system even if the particle is at rest! The only possible source of this angular momentum is the angular momentum of the electromagnetic field,

$$\vec{L}_{em} = \int d^3\vec{x}\,\vec{x}\times(\vec{E}\times\vec{B}) \ . \qquad 2.39)$$

It takes little labor to go far towards evaluating this integral. Firstly, it must be proportional to eg. Secondly, by dimensional analysis, it must be a homogeneous function of $\vec{r}$ of order zero. Thirdly, by rotational invariance, it must be proportional to $\vec{r}$, the only vector in the problem. Thus

$$\vec{L}_{em} = \beta eg\vec{r}/r \ , \qquad (2.40$$

where β is a numerical constant. The evaluation of β is left as an exercise.

(3) In my first physics course, I held a spinning bicycle wheel by its axis and attempted to rotate the axis. To my surprise, I felt a force orthogonal to the applied force; later I learned that this was because the system had angular momentum.

This is just what happens if we attempt to move a charged particle at rest in a monopole field. The Lorentz force moves the particle in a direction orthogonal to the initial impulse. This suggests

that this system also has angular momentum.

Let us attempt to compute this angular momentum by identifying its time rate of change with the external torque applied to the system. Inspired by the previous argument, let us make the *Ansatz*

$$\vec{L} = m\vec{r}\times\dot{\vec{r}} + \beta eg\vec{r}/r \ , \qquad (2.41)$$

where β is a constant we will fix in the course of the computation.

The equation of motion is

$$m\ddot{\vec{r}} = \vec{F}^{ext} + eg\dot{\vec{r}}\times\vec{r}/r^3 \ , \qquad (2.42)$$

where $\vec{F}^{ext}$ is the external force. Using this and the identity

$$\frac{d}{dt}(\vec{r}/r) = \dot{\vec{r}}/r - (\vec{r}\cdot\dot{\vec{r}})\vec{r}/r^3 \ , \qquad (2.43)$$

we find
$$\frac{d\vec{L}}{dt} = \vec{r}\times\vec{F}^{ext} + \vec{r}\times(eg\dot{\vec{r}}\times\vec{r})/r^3 + eg\beta[r^2\dot{\vec{r}} - (\vec{r}\cdot\dot{\vec{r}})\vec{r}]/r^3. \qquad (2.44)$$

Thus, if $\beta = -1$, $$d\vec{L}/dt = \vec{r}\times\vec{F}^{ext} \ . \qquad (2.45)$$

This completes the discussion of the extra term in the angular momentum. Now let us return to the analysis of the Hamiltonian (2.27).

We begin with the identity

$$\vec{D}\cdot\vec{D} = \vec{D}\cdot\vec{r}\,\frac{1}{r^2}\,\vec{r}\cdot\vec{D} - \vec{D}\times\vec{r}\cdot\frac{1}{r^2}\,\vec{r}\times\vec{D} \ . \qquad (2.46)$$

Here I have ordered the terms such that the identity is true whatever the commutators of $\vec{D}$ and $\vec{r}$. We will analyze the two terms in this expression separately.

Because $A_r = 0$, $$\vec{r}\cdot\vec{D} = \vec{r}\cdot\vec{\nabla} = r\frac{\partial}{\partial r} \ . \qquad (2.47)$$

Also, $$\vec{D}\cdot\vec{r} = \vec{r}\cdot\vec{D} + 3 \ . \qquad (2.48)$$

Thus, $$\vec{D}\cdot\vec{r}\,\frac{1}{r^2}\,\vec{r}\cdot\vec{D} = \left(r\frac{\partial}{\partial r} + 3\right)\frac{1}{r}\frac{\partial}{\partial r} = \frac{\partial^2}{\partial r^2} + \frac{2}{r}\frac{\partial}{\partial r} \ . \qquad (2.49)$$

From the commutators of $\vec{r}$ and $\vec{D}$, $\vec{r}\times\vec{D} = -\vec{D}\times\vec{r}$ commutes with r^2.

Thus, $$-\vec{D}\times\vec{r}\,\frac{1}{r^2}\,\vec{r}\times\vec{D} = \frac{1}{r^2}(\vec{r}\times\vec{D})^2 \ . \qquad (2.50)$$

If we square the expression for the angular momentum, Eq. (2.36), we

find $$\vec{L}\cdot\vec{L} = -(\vec{r}\times\vec{D})^2 + e^2g^2. \qquad (2.51)$$

Putting all this together, we find that on a subspace of states of given total angular momentum,

$$H_\ell = -\frac{1}{2m}\left(\frac{\partial^2}{\partial r^2} + \frac{2}{r}\frac{\partial}{\partial r}\right) + \frac{\ell(\ell+1) - e^2g^2}{2m\,r^2} + V\ , \qquad (2.52)$$

where, as usual, $\ell(\ell+1)$ is the eigenvalue of $\vec{L}\cdot\vec{L}$. Because $\vec{L}$ obeys the angular-momentum algebra, we know ℓ must be an integer or a half odd integer, but further analysis is required to determine which values of ℓ actually occur.

We all know the solution to this problem when $eg = 0$. Representations occur with $\ell = 0, 1, 2 \ldots$, and, at fixed r, each of them occurs only once. In elementary texts, this result is established by studying the solutions of the angular part of the wave equation. This method can certainly be extended to the case $eg \neq 0$, but it gets a bit sticky; one has to worry about singularities at the string, or, alternatively, about patching together solutions in the two regions. Therefore, I will solve the problem here by a slightly more abstract method that avoids these difficulties.

The general problem is this: Given a space of states that transform in some specified way under rotations, to find a set of basis vectors that transform according to the irreducible representations of the rotation group. Let me remind you of some standard formulas from the theory of rotations. If we label a general rotation in the standard way, with three Euler angles, α, β, and γ, then the states we are searching for obey

$$e^{-iL_z\alpha}\, e^{-iL_y\beta}\, e^{-iL_z\gamma}\, |\ell,m\rangle = \sum_{m'} D^{(\ell)}_{m'm}(\alpha,\beta,\gamma)\,|\ell,m'\rangle, \qquad (2.53)$$

where $$D^{(\ell)}_{m'm}(\alpha,\beta,\gamma) = e^{im'\alpha}\, d^{(\ell)}_{m'm}(\beta)\, e^{-im\gamma}\ , \qquad (2.54)$$

and $d^{(\ell)}$ is a matrix that can be found in any quantum mechanics text. (We shall not need its explicit form here.)

To warm up, let's do a case where we already know the answer, $eg = 0$. We are working at fixed r, so a complete set of basis vectors

are the eigenvectors of the angular position of the particle, which we describe in the usual way, by the two angles θ and ϕ. Any of these states can be obtained by applying an appropriate rotation to the state where the particle is at the north pole:

$$|\theta,\phi\rangle = e^{-iL_z\phi}\, e^{-iL_y\theta}\, |\theta = 0\rangle\ . \qquad (2.55)$$

(At $\theta = 0$, we don't need to specify ϕ.) We know the states we are searching for if we know their position-space wave-functions, $\langle\theta,\phi|\ell,m\rangle$. By the previous equations,

$$\begin{aligned}\langle\theta,\phi|\ell,m\rangle &= \langle\theta = 0|e^{iL_y\theta}\, e^{iL_z\phi}\, |\ell,m\rangle \\ &= \sum_{m'} e^{im\phi}\, d^{(\ell)}_{m'm}(\theta)\,\langle\theta = 0|\ell,m'\rangle\ . \end{aligned} \qquad (2.56)$$

Thus we know everything if we know $\langle\theta = 0|\ell,m'\rangle$. These coefficients obey an important consistency condition that is a consequence of

$$e^{-iL_z\alpha}\, |\theta = 0\rangle = |\theta = 0\rangle\ . \qquad (2.57)$$

This implies $$\langle\theta = 0|\ell,m'\rangle = 0,\ m' \neq 0\ . \qquad (2.58)$$

Thus we can construct $\langle\theta,\phi|\ell,m\rangle$ only for $\ell = 0, 1, 2, \ldots$, and for each of these values of ℓ, the solution is unique, save for an irrelevant normalization. Once we have constructed these functions, it is easy to check, using the multiplication rules for the D-matrices, that they do indeed transform in the desired way.

Now let us extend this analysis to the case $eg \neq 0$. The commutators of $\vec{L}$ and $\vec{r}$ are the same as before, so

$$|\theta,\phi\rangle = e^{-iL_z\phi}\, e^{-iL_y\theta}\, |\theta = 0\rangle \times (\text{phase factor})\ . \qquad (2.59)$$

The phase factor depends on what gauge we are working in. Fortunately, we don't need to know its explicit form; whatever it is, the main conclusion of the previous analysis is unchanged: we know everything if we know $\langle\theta = 0|\ell,m'\rangle$. The consistency condition, Eq. (2.58), is changed, though:

$$\begin{aligned} L_z|\theta = 0\rangle &= [-i\,\vec{r}\times\vec{D} - eg\,\vec{r}\,/\,r]_z\,|\theta = 0\rangle \\ &= -eg|\theta = 0\rangle\ . \end{aligned} \qquad (2.60)$$

Thus, $$\langle\theta = 0|\ell,m'\rangle = 0, \quad m' \neq -eg \ . \tag{2.61}$$

and the allowed values of the total angular momentum are

$$\ell = |eg|, \ |eg|+1, \ |eg|+2 \ \ldots \ . \tag{2.62}$$

As before, each of these occurs only once.

This completes the analysis.

Remarks: (1) The effect of the monopole is surprisingly simple. It merely changes slightly the centrifugal potential in the radial Schrödinger equation. If we can solve the Schrödinger equation for a given central potential without a monopole, we can solve it with a monopole. (2) Because ℓ is always greater than or equal to $|eg|$, the centrifugal potential is always positive, and the monopole by itself does not bind charged spinless particles. As one would expect, and as we shall see in a special case, the situation is very different when the particle has spin. (3) Dirac's quantization condition allows eg to be a half integer. Thus it is possible for two spinless particles, one carrying electric charge and the other carrying magnetic charge, to bind together to make a dyon with half-odd-integral angular momentum. This is puzzling from the viewpoint of the spin-statistics theorem. I will now explain the solution of this puzzle.

2.5 The Solution to the Spin-Statistics Puzzle

One might think that there is nothing to be said about the connection between spin and statistics within the framework of nonrelativistic quantum mechanics. This is not so; although relativistic field theory is indeed needed to show that spinless particles are bosons, nonrelativistic theory is then sufficient to deduce the statistics of composites made of these bosons. I will now show that a dyon made of a spinless electrically charged particle ("electron") and a spinless magnetically charged particle ("monopole") obeys Bose-Einstein statistics if eg is integral and Fermi-Dirac statistics if eg is half-odd-integral.[12]

We already know the Hamiltonian for an electron in the field of a monopole,

$$H = \frac{\left(\vec{p}_e - eg\,\vec{A}_D(\vec{r}_e - \vec{r}_m)\right)^2}{2m_e} + \ldots \tag{2.63}$$

Here the triple dots represent possible non-electromagnetic interactions, and A_D is the standard Dirac string potential,

$$\vec{A}_D(\vec{x}) \cdot d\vec{x} = (1 - \cos\theta)d\phi \ . \tag{2.64}$$

By a duality rotation, we thus know the Hamiltonian for a monopole in the field of an electron,

$$H = \frac{\left(\vec{p}_m + eg\,\vec{A}'_D(\vec{r}_m - \vec{r}_e)\right)^2}{2m_m} \ . \tag{2.65}$$

Here the sign has changed because the duality rotation that takes e into g takes g into minus e, and $\vec{A}'_D$ is some potential that is gauge-equivalent to $\vec{A}_D$.

To fix $\vec{A}'_D$, we consider a system made of one monopole and one electron. For any choice of $\vec{A}'_D$, $\vec{p}_e + \vec{p}_m$ is a constant of the motion, because H is translationally invariant. However, because

$$m_e \vec{v}_e = \vec{p}_e + eg\,\vec{A}_D(\vec{r}_e - \vec{r}_m) \ , \tag{2.66a}$$

and

$$m_m \vec{v}_m = \vec{p}_m - eg\,\vec{A}'_D(\vec{r}_m - \vec{r}_e) \ , \tag{2.66b}$$

$m_e\vec{v}_e + m_m\vec{v}_m$ is a constant of the motion (as the classical equations of motion say it should be) only if

$$\vec{A}'_D(\vec{x}) = \vec{A}_D(-\vec{x}) \ . \tag{2.67}$$

This is indeed gauge-equivalent to $\vec{A}_D$. (See Sec. 2.2.)

To summarize, the correct Hamiltonian for a monopole in the field of an electron is

$$H = \frac{\left(\vec{p}_m + eg\,\vec{A}_D(\vec{r}_e - \vec{r}_m)\right)^2}{2m_m} + \ldots \tag{2.68}$$

In exchanging monopoles and electrons, one changes the sign in front of the vector potential, but not the order of the terms within the vector potential.

We can now go on to the system of interest, two dyons, each made of a spinless monopole and a spinless electron. Just as when dealing

with two atoms, we describe the states of the system by a Schrödinger wave function,

$$\psi_{A_1 A_2}(\vec{r}_1, r_2)$$

Here the $\vec{r}$'s are the positions of the dyons, and the A's are discrete variables that give the internal states of the dyons, spins or excitation energies or whatever. Because our electrons and monopoles are bosons, standard arguments lead to

$$\psi_{A_1 A_2}(\vec{r}_1, \vec{r}_2) = \psi_{A_2 A_1}(\vec{r}_2, \vec{r}_1) \ . \tag{2.69}$$

We would normally say this implies that dyons are bosons. However, let us look more closely at the Hamiltonian for the two-dyon system:

$$H = \frac{(\vec{p}_1 - eg\vec{A}_D(\vec{r}_1 - \vec{r}_2) + eg\vec{A}_D(\vec{r}_2 - \vec{r}_1))^2}{2m} + \frac{(\vec{p}_2 - eg\vec{A}_D(\vec{r}_2 - \vec{r}_1) + eg\vec{A}_D(\vec{r}_1 - \vec{r}_2))^2}{2m} + \ldots \ , \tag{2.70}$$

where the triple dots represent Coulomb interactions as well as possible non-electromagnetic interactions. I hope the origin of the terms in this equation is clear; the electron in the first dyon sees the monopole in the second dyon, the monopole in the first dyon sees the electron in the second dyon, etc.

Equation (2.70) looks as if it describes the most horrible velocity-dependent forces, but there can be no such forces between two identical dyons, for there exists a duality rotation that makes their magnetic charges simultaneously vanish. Indeed, from Eq. (2.16),

$$A_D(\vec{x}) - A_D(-\vec{x}) = 2\vec{\nabla}\phi \ . \tag{2.71}$$

Thus, if we make a gauge transformation,

$$\psi \to \psi' = e^{2ieg\phi_{12}}\psi \ , \tag{2.72}$$

the horrible interactions disappear,

$$H \to H' = \frac{\vec{p}_1^{\,2}}{2m} + \frac{\vec{p}_2^{\,2}}{2m} + \ldots \ . \tag{2.73}$$

But this can change the symmetry of the wave function, for when $\vec{r}_1$ and $\vec{r}_2$ are interchanged, ϕ_{12} goes into $\phi_{12} + \pi$. Hence,

$$\psi'_{A_1 A_2}(\vec{r}_1, \vec{r}_2) = e^{2\pi ieg}\psi'_{A_2 A_1}(\vec{r}_2, \vec{r}_1) \ .$$

That is to say, ψ' is symmetric under interchange of the dyons if eg is integral and antisymmetric if eg is half-odd-integral.

We thus have two descriptions of the same system. One, given by ψ and H, says that our dyons are bosons whatever their spin, but that there are extraordinary long-range forces between them. The other, given by ψ' and H', says that there are no extraordinary forces, but that dyons obey the statistics appropriate to their spin. These two descriptions are connected by a gauge transformation, and therefore make the same prediction for all observable quantities. Nevertheless, there is no ambiguity here; it would be madness to choose the first description when the second is available. (After all, we could use the same gauge transformation to make any pair of fermions look like bosons, whatever their electric or magnetic charges.) Dyons unambiguously obey the spin-statistics theorem.

3. NON-ABELIAN MONOPOLES FROM AFAR

3.1 Gauge Field Theory - A Lightning Review

This subsection is a collection of definitions and formulas from classical gauge field theory, with occasional comments. I have inserted it here to establish notation and to remind you of some features of the subject that will be important to us later. It is far too compressed to be a pedagogical exposition; if you don't know basic gauge field theory already, you won't learn it here.

Gauge Transformations

The dynamical variables in gauge field theories fall into two classes, gauge fields and matter fields. We begin with the matter fields, which we assemble into a big vector, ϕ. A gauge transformation is labeled by a function, g(x), from space-time into some compact connected Lie group, G. Under such a transformation, the

matter fields transform according to some faithful unitary representation of G, D(g),

$$g(x) \in G\colon \Phi(x) \to D\big(g(x)\big)\Phi(x) \ . \tag{3.1}$$

For our purposes, it will be convenient to identify the abstract group element g with the matrix D(g), and write this equation as

$$g(x) \in G\colon \Phi(x) \to g(x)\Phi(x) \ . \tag{3.2}$$

In the neighborhood of the identity of G, a group element can be expanded in a power series,

$$g = 1 + \sum_{a=1}^{\dim G} \varepsilon^a T^a + O(\varepsilon^2) \ . \tag{3.3}$$

Here the ε's are some coordinates for the group, and the T's are a set of matrices called the infinitesimal generators of the group. The generators span a linear space called the Lie algebra of G. Because g is unitary, the T's are anti-Hermitian

$$T^a = -T^{a+} \ . \tag{3.4}$$

(We are following mathematicians' conventions here; physicists frequently insert a factor of i in the definition of the generators, to make them Hermitian.)

For a general group element, g,

$$g\,T^a\,g^{+} = D^{(adj)}_{ba}(g)T^b \ , \tag{3.5}$$

where $D^{(adj)}$ is a representation of the group, called the adjoint representation, and the sum on repeated indices is implied.

The commutator of any two generators is a linear combination of generators,

$$[T^a,T^b] = c^{abc}\,T^c \ , \tag{3.6}$$

where the c's are real coefficients called structure constants. They depend only on the abstract group G and not upon the particular representation we use to realize it. We will always choose the T's such that

$$\mathrm{Tr}\,T^a T^b = -N\delta^{ab} \ , \tag{3.7}$$

where N is a normalization constant. Thus,

$$c^{abc} = -N^{-1}\,\mathrm{Tr}[T^a,T^b]T^c \ . \tag{3.8}$$

from which it follows that c^{abc} is unchanged by even permutations of the indices and changes sign under odd permutations.

For example, if G is SU(2), the group of 2×2 unitary unimodular matrices, the standard choice is $T^a = -i\sigma^a/2$, where the σ's are the Pauli spin matrices. The adjoint representation is the vector representation, N is ½, and c^{abc} is ε^{abc}.

Gauge Fields

The gauge fields are a set of vector fields, A^a_μ, a = 1...dimG. It will be convenient for us to assemble these into a single matrix-valued vector field,

$$A_\mu = A^a_\mu T^a \ . \tag{3.9}$$

The gauge-transformation properties of this field are defined to be

$$g(x)\colon A_\mu \to g\,A_\mu\,g^{-1} + g\,\partial_\mu\,g^{-1} \ . \tag{3.10}$$

If we define the covariant derivative of the matter field by

$$D_\mu\,\Phi \equiv (\partial_\mu + A_\mu)\Phi \ , \tag{3.11}$$

then, under a gauge transformation,

$$g(x)\colon D_\mu\,\Phi \to g\,D_\mu\,\Phi \ . \tag{3.12}$$

The field-strength tensor, $F_{\mu\nu}$, is a matrix-valued field defined by

$$\begin{aligned}[D_\mu,D_\nu]\Phi &= (\partial_\mu A_\nu - \partial_\nu A_\mu + [A_\mu,A_\nu])\Phi \\ &\equiv F_{\mu\nu}\,\Phi \ .\end{aligned} \tag{3.13}$$

Under a gauge transformation, the field strength transforms according to the adjoint representation of G,

$$g(x)\colon F_{\mu\nu} \to g\,F_{\mu\nu}\,g^{-1} \ . \tag{3.14}$$

The covariant derivative of the field strength is defined by

$$D_\lambda\,F_{\mu\nu} = \partial_\lambda\,F_{\mu\nu} + [A_\lambda,F_{\mu\nu}] \ . \tag{3.15}$$

This transforms in the same way as F itself under gauge transformations.

Parallel to Eq. (3.9),

$$F_{\mu\nu} = F^a_{\mu\nu}\,T^a \ , \tag{3.16}$$

where $$F^a_{\mu\nu} = \partial_\mu A^a_\nu - \partial_\nu A^a_\mu + c^{abc} A^b_\mu A^c_\nu \ . \qquad (3.17)$$

Dynamics

The Lagrange density is

$$\mathscr{L} = \frac{1}{4Nf^2} \operatorname{Tr} F_{\mu\nu} F^{\mu\nu} + \mathscr{L}_m(\Phi, D_\mu \Phi) \ , \qquad (3.18)$$

where $\mathscr{L}_m$ is some invariant function,

$$\mathscr{L}_m(g\,\Phi, g\,D_\mu \Phi) = \mathscr{L}_m(\Phi, D_\mu \Phi) \ , \qquad (3.19)$$

and f is a real number, called the coupling constant. The gauge-field part of $\mathscr{L}$ can also be written as

$$\frac{1}{4Nf^2} \operatorname{Tr} F_{\mu\nu} F^{\mu\nu} = -\frac{1}{4f^2} F^a_{\mu\nu} F^{\mu\nu a} \ . \qquad (3.20)$$

If there are no matter fields, then the gauge fields obey

$$D_\mu F^{\mu\nu} = 0 \ . \qquad (3.21)$$

These are called the sourceless Yang-Mills equations.

If we define new fields, denoted by a prime, by

$$A^{a'}_\mu = f^{-1} A^a_\mu \ , \qquad (3.22)$$

and

$$F'^a_{\mu\nu} = f^{-1} F^a_{\mu\nu} = \partial_\mu A'^a_\nu - \partial_\nu A'^a_\mu + f\, c^{abc} A^{b'}_\mu A^{c'}_\nu \ , \qquad (3.23)$$

then

$$-\frac{1}{4f^2} F_{\mu\nu} F^{\mu\nu a} = -\frac{1}{4} F'^a_{\mu\nu} F^{\mu\nu' a} \qquad (3.24)$$

and

$$D_\mu \Phi = \partial_\mu \Phi + f A'^a_\mu T^a \Phi \ . \qquad (3.25)$$

From these we see that f is indeed a coupling constant, absent from the quadratic terms in the Lagrangian but present in the higher-order ones.

If the gauge group is a product of factors, simple groups and U(1) groups, then it is possible to have an independent coupling constant for each factor. The gauge-field part of the Lagrangian becomes

$$-\frac{1}{4} \sum_a \frac{1}{f_a^2} F^a_{\mu\nu} F^{\mu\nu a} \ , \qquad (3.26)$$

where f_a is the same for fields associated with the same factor group. Likewise, Eq. (3.25) becomes

$$D_\mu \Phi = \partial_\mu \Phi + \sum_a f_a A^{a'}_\mu T^a \Phi \ . \qquad (3.27)$$

Temporal Gauge

Temporal gauge is defined by $A_0 = 0$. To transform a gauge-field configuration into temporal gauge we need a function g(x) such that

$$g\,A_0\,g^{-1} + g\,\partial_0\,g^{-1} = 0 \ . \qquad (3.28)$$

A solution to this equation is

$$g^{-1}(x,t) = T \exp - \int_0^t dt'\, A_0(\vec{x},t') \ . \qquad (3.29$$

where T indicates the integral is time-ordered.

Temporal gauge is useful because the gauge-field part of the Lagrange density is $-\partial_0 A^a_i \partial^0 A^{ia} / 2f^2$, plus terms with no time derivatives. Thus the structure of the initial-value problem closely resembles that in a scalar field theory with non-derivative interactions, as does the form of Hamilton's equations.

The adaptation of temporal gauge does not destroy all gauge freedom; one can still make time-independent gauge transformations.

Group Elements Associated with Paths

A path in space-time is described by a function x(s), where s goes from 0 to 1. With every such path, we associate a group element, g, defined by

$$g = P \exp - \int_0^1 ds\, A_\mu \frac{dx^\mu}{ds} \ , \qquad (3.30)$$

where P indicates the integral is path-ordered. g can also be identified with g(1), where g(s) is the solution to

$$\frac{Dg}{Ds} \equiv \frac{dg}{ds} + A_\mu \frac{dx^\mu}{ds}\, g = 0 \ , \qquad (3.31)$$

with the boundary condition g(0) = 1. (Note the similarity to Eqs. (3.28) and (3.29).)

Under a gauge transformation, h(x),

$$h(x) \in G\colon\ g \to h\big(x(1)\big)\, g\, h\big(x(0)\big)^{-1} \ . \qquad (3.32)$$

This is even simpler if the path is closed, x(0) = x(1),

$$h(x) \in G\colon\ g \to h\big(x(0)\big)\, g\, h\big(x(0)\big)^{-1} \ . \qquad (3.33)$$

In this case, despite the fact that it is a non-local integral, g transforms just like $F_{\mu\nu}$, a local field.

For a closed path, Tr g is the famous Wilson loop factor. It is gauge invariant, and independent of where along the closed loop one chooses to start and end the path. It does depend on the particular matrix representation one uses; in the language of group theory, different representations have different characters.

For ordinary electromagnetism, the group element associated with a closed path is

$$g = \exp(ie\Phi) . \tag{3.34}$$

where Φ is the magnetic flux passing through a surface bounded by the path. Thus, the concept of group element associated with a closed path is one possible generalization to non-Abelian theory of the concept of magnetic flux. (The obvious alternative, defining magnetic flux by integrating the magnetic field, doesn't work; one gets nowhere adding together quantities that have different gauge-transformation properties, as do non-Abelian magnetic fields at different points.)

3.2 The Nature of the Classical Limit

We are going to spend some time and effort investigating the properties of magnetic monopoles in classical non-Abelian gauge theories. Thus it is reasonable to begin by asking when we can expect classical physics to describe quantum reality.

Let $S(\Phi)$ be the action functional for any classical field theory, with Φ now all the fields in the theory, gauge fields or whatever. The Euclidean form of Feynman's path integral formula says that the quantum version of the theory is defined by the functional integral

$$\int (d\Phi) e^{-S(\Phi)/\hbar} . \tag{3.35}$$

If we define new fields, $\Phi' = \Phi/\sqrt{\hbar}$, then, aside from an irrelevant normalization constant, this can be written as

$$\int (d\Phi') e^{-S(\Phi'\sqrt{\hbar})/\hbar} . \tag{3.36}$$

This trivial transformation reveals $\sqrt{\hbar}$ to be a coupling constant; it is absent from the quadratic terms in the argument of the exponential but present in the higher-order ones. The classical limit is a weak-coupling limit. Sending $\hbar$ to zero, with all coupling constants fixed, is the same as sending all coupling constants to zero (at appropriate rates), with $\hbar$ fixed.

Thus, classical gauge field theory (without symmetry breakdown), the subject of Sec. 3.1, should be a good guide to weakly-coupled quantum gauge field theory (again without symmetry breakdown). I know of two regimes in which such theories apply. One was mentioned in the introduction, quantum chromodynamics at distances less than the confinement length. Because the theory is asymptotically free, this is a regime of weak coupling. The other is the early universe. Weakly-coupled gauge fields which now have masses might well have been massless early on, when the structure of symmetry breakdown was different.

These then are the places where classical non-Abelian monopoles might be found: deep within hadrons or far in the past. Now let us see what they would look like if we found them.

3.3 Dynamical (GNO) Classification of Monopoles

At the beginning of Sec. 2, I described the standard series of stationary electromagnetic fields that could exist outside a black box of unknown content; the leading terms at large distances were the electric and magnetic monopole fields. The dynamical classification of monopoles, invented by Goddard, Nuyts, and Olive (GNO),[13] can be thought of as the extension of this analysis to non-Abelian gauge theories. I will now give a derivation of the GNO classification.

Just as in the electromagnetic case, the analysis will use only the equations that hold outside the black box, the sourceless field equations. Because these are difficult nonlinear equations, I will not attempt to construct a complete series of solutions, but just try to find the non-Abelian generalization of the magnetic monopole field.

I will exclude electric fields from the beginning by looking for solutions that are not just time-independent but time-reversal

invariant (in some appropriate gauge). Time reversal is the operation

$$T: A_0(\vec{x},t) \to -A_0(\vec{x},-t)$$
$$\vec{A}(\vec{x},t) \to A(\vec{x},-t) \tag{3.37}$$

(The alternative definition allowed in the Abelian case, this operation times minus one, is not an invariance of the non-Abelian theory.) Thus,

$$A_0 = 0, \qquad \partial_0 A_i = 0 \ , \tag{3.38}$$

and F_{0i} vanishes.

Equation (3.38) still allows us the freedom to make time-independent gauge transformations. I will use this freedom to make

$$A_r = 0 \ . \tag{3.39}$$

Just as one integrates along time lines to construct the gauge transformation that takes one to temporal gauge, Eq. (3.29), one here constructs the required gauge transformation by integrating along radial lines, starting, for example, at the unit sphere. Of course, this procedure could lead to trouble at the origin, where radial lines intersect, but this is of no concern to us; the origin is inside the black box, where we don't want to use the field equations anyway. Note that we still have the freedom to make gauge transformations that depend only on θ and ϕ, a freedom we shall use shortly.

We now assume that for large r, $\vec{A}$ can be expanded in powers of $1/r$,

$$\vec{A} = \frac{\vec{a}(\theta,\phi)}{r} + O\left(\frac{1}{r^2}\right) \ . \tag{3.40}$$

Because the Yang-Mills equations are non-linear, they will involve, in general, cross terms between the leading term in this expansion and the higher-order terms. However, only the leading term enters into the part of the equations proportional to $1/r^3$. Thus, if we are only interested in the leading term, as we are, it is legitimate to ignore the higher-order terms, as we shall.

Just as in the discussion of the Abelian case, it is convenient to write $\vec{A}$ in terms of its covariant components,

$$\vec{A} \cdot d\vec{x} = A_\theta(\theta,\phi)d\theta + A_\phi(\theta,\phi)d\phi \ . \tag{3.41}$$

If the field is to be nonsingular at the north and south poles, $A_\phi(0,\phi)$ and $A_\phi(\pi,\phi)$ must both vanish.[14]

I will now make one more gauge transformation to make

$$A_\theta = 0 \ . \tag{3.42}$$

This transformation can be constructed by integrating along lines of fixed ϕ, meridians, starting from the north pole. Of course, this choice of gauge can lead to an artificial singularity at the place where all the meridians intersect again, the south pole. In particular, it can lead to a non-zero (and ϕ dependent) $A_\phi(\pi,\phi)$, that is to say, to a Dirac-string singularity. Just as in the Abelian case, we'll live with this singularity for the time being. When we're done with solving the equations, we'll check that the singularity is indeed just a gauge artifact, that the string is undetectable.

We are now in a position to (finally) use the field equations. The field strength tensor has only one non-vanishing component,

$$F_{\theta\phi} = \partial_\theta A_\phi \ . \tag{3.43}$$

In curvilinear coordinates, the sourceless Yang-Mills equations take the form

$$\partial_\mu \sqrt{g}\, F^{\mu\nu} + [A_\mu, \sqrt{g}\, F^{\mu\nu}] = 0 \ . \tag{3.44}$$

In our case,

$$\sqrt{g}\, F^{\theta\phi} = \frac{1}{r^2 \sin\theta} \partial_\theta A_\phi \ . \tag{3.45}$$

Thus there are two non-trivial field equations. One is

$$\partial_\theta \frac{1}{\sin\theta} \partial_\theta A_\phi = 0 \ . \tag{3.46}$$

The general solution of this, consistent with the vanishing of $A_\phi(0,\phi)$, is

$$A_\phi = Q(1-\cos\theta) \ , \tag{3.47}$$

where Q is an arbitrary (matrix-valued) function of ϕ. However, the other field equation

$$\partial_\phi \sqrt{g}\, F^{\phi\theta} + [A_\phi, \sqrt{g}\, F^{\phi\theta}] = -\partial_\phi Q = 0 \ , \tag{3.48}$$

tells us Q must be a constant.

Thus, up to a gauge transformation, the non-Abelian monopole field is constructed by multiplying the Abelian monopole field by a constant matrix. That this procedure leads to non-Abelian monopoles is trivial; what is non-trivial is that it leads to all of them.

Because the non-Abelian monopole is so simply related to the Abelian one, it is easy to work out the quantization condition. We just follow the arguments of Sec. 2.2, with trivial changes. The potential with the north-pointing string is defined by

$$A'_\phi = -Q(1+\cos\theta) \ . \tag{3.49}$$

$\vec{A}$ and $\vec{A}'$ are transformed into each other by the gauge transformation,

$$g = \exp 2Q\phi \ , \tag{3.50}$$

which is single-valued if

$$\exp 4\pi Q = 1 \ . \tag{3.51}$$

This is the quantization condition.

As a small consistency check, let us verify that this reduces to the correct condition when we specialize to the Abelian theory of Sec. 2. There has been a slight notational change; what we now call $\vec{A}$ we then called $-ie\vec{A}$. (See the expression for the covariant derivative, Eq. (3.11).) Thus, Q becomes $-ieg$, and Eq. (3.51) becomes the Dirac quantization condition, as it should.

In high-energy theory, we tend to focus on the Lie algebra of a group and ignore its global structure; for example, we indiscriminately refer to the isospin group as SU(2) or SO(3). This is an especially bad habit in monopole theory, because the quantization condition is sensitive to the global structure of G; the allowed set of monopoles is different for SU(2) and SO(3). For example, let us suppose that Q is proportional to I_3, the generator of rotations about the third axis in isospin space. If G is SO(3), Q can be any half-integral multiple of I_3, because a rotation by 2π is the identity. If G is SU(2), though, only integral multiples are allowed, because a rotation by 4π is needed to get back to the identity.

There is nothing deep in this; it can all be understood in the

elementary terms of Sec. 2.1. To say G is SO(3) is to say that all particles have integral isospin, and therefore integral values of I_3. To say G is SU(2) is to allow half-odd integral values. With a richer set of test particles one can do a richer set of Bohm-Aharonov experiments and detect solenoids formerly undetectable.

To have some examples to use later on, let me work out the explicit form of Q for some representative groups.

If G is SU(n), Q must be a traceless antihermitian $n \times n$ matrix. Such a matrix can always be diagonalized by a unitary transformation, that is to say, by a constant gauge transformation. Thus

$$Q = -\frac{i}{2}\,\mathrm{diag}\,(q_1, q_2, \ldots q_n) \ . \tag{3.52}$$

Tracelessness implies that the sum of the q's is zero, while the quantization condition implies that each q is an integer. Because we can arbitrarily permute the q's by a gauge transformation, only the set of q's is relevant, not their order.

This becomes expecially simple for SU(2). Here,

$$Q = -\frac{i}{2}\,\mathrm{diag}\,(q,-q) = -\frac{i}{2} q\sigma_3 \ . \tag{3.53}$$

We can use our freedom to permute the eigenvalues of Q to insure that q is always non-negative, $q = 0, 1, \ldots$.

Some gauge field theories based upon the Lie algebra of SU(n) have for their global group, G, not SU(n), but $SU(n)/Z_n$. (Z_n is the n-element finite group consisting of the nth roots of unity, that is to say, the integral powers of $\exp(2\pi i/n)$.) One example is the theory of gauge fields only, without matter fields, gluons without quarks. Gauge fields transform according to the adjoint representation of the group; two SU(n) matrices that differ only by a factor belonging to Z_n will be represented by the same matrix in the adjoint representation.

To treat this case without my equations growing too long, let me introduce some *ad hoc* notation. I will denote by Q_{fund} the matrix that represents a given abstract group generator in the n-dimensional fundamental representation of the group, and by Q_{adj} the matrix that

represents the same generator in the (n^2-1)-dimensional adjoint representation. The quantization condition in the case at hand is $\exp(4\pi Q_{adj}) = 1$. This is true if and only if

$$\exp(4\pi Q_{fund}) \in Z_n \ . \tag{3.54}$$

Thus, as before,

$$Q_{fund} = -\frac{i}{2}\,\mathrm{diag}\,(q_1 \cdots q_n) \tag{3.55}$$

but now

$$q_m = \frac{r}{n} + \text{integer} \ , \tag{3.56}$$

where r is an integer, independent of m. As before, the q's must sum to zero, and only the set of q's is relevant, not their order.

This becomes especially simple for $SU(2)/Z_2 = SO(3)$. Just as in Eq. (3.53), Q_{fund} is $-\frac{1}{2}iq\sigma_3$, with $q \geq 0$. However, now q can be a half-integer.

3.4 Topological (Lubkin) Classification of Monopoles

The topological classification of monopole fields was developed almost twenty years ago by Elihu Lubkin in an important and unjustly ignored paper.[15] We begin by investigating a situation in which a black box of unknown content is surrounded by gauge fields, the same situation we studied in the dynamical classification. However, now we shall not assume the gauge fields are stationary solutions of the sourceless Yang-Mills equations, or indeed any kind of solutions to any dynamical equations. Instead, we shall associate with the gauge-field configuration a topological charge, a quantity that is unchanged by arbitrary continuous deformation of the configuration, and thus in particular is unchanged by time evolution in accordance with any sensible equation of motion.

I shall first describe Lubkin's construction and develop some of its important properties. This will involve a (very) short course in (very little) homotopy theory. No matter how low your tolerance for abstract mathematics, please pay close attention here; otherwise, you will understand nothing of what follows. After this, I shall investigate the connection of the topological classification with the dynamical classification.

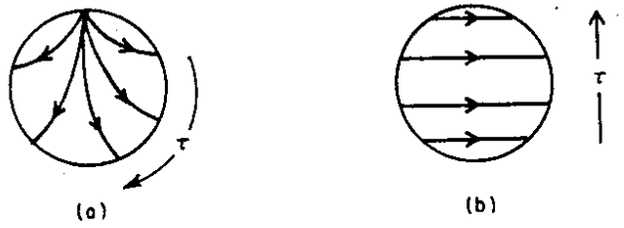

Figure 2

Figure 2a shows a family of closed paths lying on a sphere surrounding the black box. Each path begins and ends at the north pole; the return portions of the paths are on the back side of the sphere and can't be seen in the figure. These paths are labeled by a parameter τ that ranges from 0 to 1; the first and last paths in the family, $\tau = 0$ and 1, are trivial paths that never leave the pole. As τ goes through its range, the paths sweep around the black box, like a magician's hoop sweeping around a floating lady.

In case you have trouble visualizing this from Fig. 2a, I've drawn the identical structure in a different way in Fig. 2b. This is a south polar projection of the sphere; the north pole is represented by the circumference of the disc. The paths are the horizontal lines.

Along each path we may integrate the gauge field to obtain the group element associated with the path. In this way, from our family of paths, we obtain a path in group space, $g(\tau)$, beginning at the identity and ending at the identity. (We need not worry about encountering gauge-dependent singularities, like Dirac strings, as we sweep around the sphere. We can always make a gauge transformation to get the singularity out of our way; by Eq. (3.33), this will not affect $g(\tau)$, as long as we take care to insure that the gauge transformation is the identity at the north pole.)

The path $g(\tau)$ depends on practically everything in the problem: the sphere we choose, the details of how we construct the family of closed paths on the sphere, the time at which we do the computation

(if our gauge field is changing with time), the gauge in which we are working. However, it depends continuously on all of these things, and therefore the associated element of the first homotopy group of G, $\pi_1(G)$, is unchanged.

To explain the preceding sentence requires the promised short course in homotopy theory.[16]

For any topological space, X, a path, $x(t)$, is a continuous function from the interval $[0,1]$ into X. Let $x(t)$ and $x'(t)$ be two paths with the same endpoints, that is to say, $x(0) = x'(0)$ and $x(1) = x'(1)$. Then we say these paths are "homotopic" or "in the same homotopy class" if one can be continuously distorted into the other, keeping the endpoints fixed. Phrased in equations, $x(t)$ and $x'(t)$ are homotopic if there exists a continuous function of two variables, $F(s,t)$, $0 \le s,t \le 1$, such that

$$F(0,t) = x(t) ,$$

$$F(1,t) = x'(t) ,$$

and

$$F(s,0) = x(0) = x'(0) ,$$

$$F(s,1) = x(1) = x'(1) . \qquad (3.57)$$

An example is shown in Fig. 3. The topological space, X, is a portion of the plane with a disc (the shaded region) removed. Four paths are shown; all have the same initial and final point, x_0. Path C is homotopic to path D and also to the trivial path, $x(t) = x_0$ for all t. Otherwise, no two paths shown are homotopic. (I hope these statements are obvious to you, because they are rather difficult to prove analytically.)

Given two paths such that one ends where the other begins, the product of the paths is defined by first going along the first path and then going along the second. In equations,

$$x \cdot x'(t) = x(2t) , \qquad 0 \le t \le \tfrac{1}{2} ,$$

$$= x'(2t-1) , \qquad \tfrac{1}{2} \le t \le 1 . \qquad (3.58)$$

The products of homotopic paths are clearly homotopic.

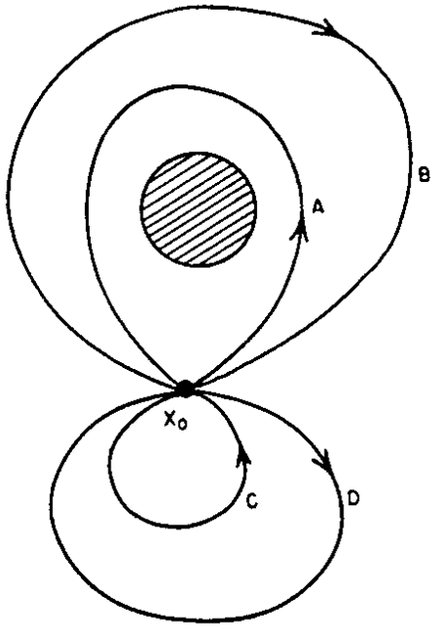

Figure 3

The inverse of a path is defined by going along the path in the opposite direction. In equations,

$$x(t)^{-1} = x(1-t) . \qquad (3.59)$$

The product of a path and its inverse (in either order) is clearly homotopic to a trivial (*i.e.*, constant) path. These concepts are also exemplified in Fig. 3; A is homotopic to the inverse of B.

If we consider all homotopy classes of closed paths, with fixed initial-final point, x_0, these operations of multiplication and inversion define a group structure. The group obtained in this way is called the first homotopy group of X, and is denoted by $\pi_1(X)$. We do not have to specify x_0, because $\pi_1(X)$ is independent of x_0, if X is connected. Proof: Let y be some path going from x_0 to some other point, x_1. Then the mapping, $x \to y \cdot x \cdot y^{-1}$, maps every closed path

beginning and ending at x_1 into one beginning and ending at x_0 in such a way that the group operations are preserved.

For the example of Fig. 3, the homotopy classes are labeled by an integer, the winding number, the net number of times the path winds around the shaded disc. For example, path A has winding number 1, path B, -1, paths C and D, 0. The winding number of a product of paths is the sum of the winding numbers of the factors. Thus, π_1 is the additive group of the integers, sometimes called Z.

If X is connected and $\pi_1(X)$ is trivial, we say X is simply connected. It is easy to show that in this case any two paths connecting the same two points are homotopic.

This completes the first part of the short course on homotopy theory. There will be a second part to the course, on the computation of π_1 for compact Lie groups. However, even at this stage, we can understand some things about monopoles:

(1) You should now understand my cryptic statement of a few paragraphs back, that Lubkin's loop construction associates with the gauge fields outside the black box an element of $\pi_1(G)$. This element is the topological charge I referred to at the beginning of this section.

(2) We can see that the topological charge is gauge invariant. If we make a gauge transformation equal to h at the north pole, then $g(\tau)$ is transformed into $hg(\tau)h^{-1}$. Because G is connected, h can be continuously deformed to 1, and the transformed path is homotopic to the original one.

(3) Let us consider a world that contains two black boxes. We can compute the topological charge of each of the boxes, by surrounding it by a sphere that does not contain the other, or we can compute the topological charge of the total system, by surrounding both boxes by a large sphere. In the latter case, we can continuously distort the sphere into an hourglass, two spheres, each surrounding a box, connected by a tube pinched down to a point at the middle. We choose this midpoint to be the "north pole" of the distorted sphere.

Now, when we sweep loops about the distorted sphere, they pass first around one sphere and then around the other; the path in group space is the product of the paths for the individual spheres. Thus the topological charge of the combined system is the (group theoretical) product of the topological charges of the individual components. As a byproduct, we have produced a very indirect argument that $\pi_1(G)$ is Abelian, since it obviously doesn't matter in what order we do things. We'll shortly have a more direct demonstration of the same result.

(4) We see that the forbidden region, the black box, is essential if we are to obtain a non-trivial structure. For, if the box were not there, we could shrink the surrounding sphere to a point. In this limit, the associated element of π_1 would be the identity, because all paths on the sphere would be trivial. But, since it is an invariant under continuous transformations, if it is the identity in the limit, it must have been the identity to begin with.

If we are to obtain a non-trivial topological charge, there must be something inside the box other than just non-singular gauge fields. In Sec. 4 we shall see what that something is.

I now return to the course. As promised, I will tell you how to compute π_1 for any compact connected Lie group. The best approach to this problem is an indirect one; I'll begin by classifying all Lie groups with a given Lie algebra. Once this is done, the computation of π_1 will turn out to be trivial.

I assume you know that the Lie algebra of any compact Lie group is the direct sum of copies of certain fundamental Lie algebras. These are the Lie algebras of U(1), of the three infinite families of classical groups, SO(n), SU(n), and Sp(n), and of the five exceptional Lie groups.

For each of these, there exists a simply connected group with the given Lie algebra. For the algebra of U(1), it is R, the additive group of real numbers. For the algebras of SU(n) and Sp(2n), it is these very groups. For the algebra of SO(n), it is the double

covering of SO(n), sometimes called Spin(n). I won't worry here about the five exceptional groups; we'll never use them in these lectures.

Thus, by taking direct products of these groups, we can construct a simply connected group whose Lie algebra is isomorphic to that of a given compact connected Lie group, G. I will denote this group by $\bar{G}$. $\bar{G}$ is called the covering group of G; the reason is that it covers G in the same way SU(2) covers $SO(3) = SU(2)/Z_2$. To be precise, it is possible to show that G is isomorphic to the quotient group $\bar{G}/K$, where K is some discrete subgroup of the center of $\bar{G}$. (I remind you that the center of a group is the subgroup consisting of all elements that commute with every element of the group.) This is a standard theorem of Lie-group theory and I ask you to take it on trust.

Thus, to classify all groups with a given algebra, we have to find all the discrete subgroups of the center of $\bar{G}$. In all the cases we shall encounter, finding the subgroups will be trivial once we find the center.

The center of $\bar{G}$ is the product of the centers of its factors. All of these are easy to compute. R is Abelian, and thus all center. The center of SU(n) is Z_n. The center of Sp(n) consists of 1 and -1. The center of Spin(n), for even n, consists of the two elements that are mapped into 1 in SO(n) and the two elements that are mapped into -1. For odd n, -1 is not in SO(n), so the center of Spin(n) consists only of the two elements mapped into 1.

As an example, let me use this apparatus to work out all the groups with Lie algebras isomorphic to that of $\bar{G} = SU(2) \otimes SU(2)$. To emphasize the differences between these groups, as I construct the groups I will also describe their irreducible representations. If I describe the elements of $\bar{G}$ in the standard way, by (g_1,g_2), the center consists of (1,1), (1,-1), (-1,1) and (-1,-1). This has five subgroups, and thus we have five groups with the given algebra.

(a) K is (1,1). G is $SU(2) \otimes SU(2)$. The representations of G are of the form $D^{(s_1)}(g_1) \otimes D^{(s_2)}(g_2)$, with the D's representations of SU(2) and s_1 and s_2 half-integers.

(b) K consists of (1,1) and (1,-1). G is $SU(2) \otimes SO(3)$. The representations are as before; s_1 is a half-integer and s_2 an integer.

(c) K consists of (1,1) and (-1,1). This is the same as the previous case with the two factors transposed.

(d) K consists of all four elements of the center. G is $SO(3) \otimes SO(3)$. Both s_1 and s_2 are integers.

(e) K consists of (1,1) and (-1,-1). This is in many ways the most interesting case. G is not a direct product. Either s_1 and s_2 are both integers or they are both half-odd-integers. This is reminiscent of the kind of structures we find in realistic theories, where the unbroken gauge group has the algebra of SU(3) (color) $\otimes$ U(1) (electromagnetism), but it is not a direct product, because only particles of non-zero triality have fractional charge.

Now that we have classified the Lie groups, it is easy to compute their first homotopy groups. We reason as follows. In general, the mapping of $\bar{G}$ into G is many-to-one; there is no unique element of $\bar{G}$ mapped into a given element of G. However, because K is discrete, for every continuous path in G beginning at the identity, there is a unique <u>continuous</u> path in $\bar{G}$ beginning at the identity. If the path in G is closed, that is to say, if it ends at the identity, the path in $\bar{G}$ must end at some group element that is mapped into the identity, that is to say, at some element of K. Let us consider two closed paths in G such that the corresponding paths in $\bar{G}$ end at the same element of K. Because $\bar{G}$ is simply connected, the paths in $\bar{G}$ can be continuously deformed into each other; thus, so can the corresponding paths in G. Hence, the homotopy classes of closed paths in G are in one-to-one correspondence with the elements of K. It is easy to show that the product of two closed paths corresponds to the products of the two group elements; that is to say,

that $\pi_1(G)$ is isomorphic to K.

This concludes the short course in homotopy theory. Now let us return to magnetic monopoles.

The Lubkin construction is totally gauge invariant, but carrying it through can be tedious; we have to solve an independent differential equation (or, equivalently, evaluate an independent path-ordered integral) for each value of τ. Things can be simplified considerably by going to a special gauge, "string gauge".

This is very much like the gauge we used in the dynamical classification of monopoles. On each sphere of fixed r, we start at the north pole and gauge transform along meridians to make $A_\theta(\theta,\phi)$ vanish. This may induce a non-zero $A_\phi(\pi,\phi)$, a Dirac string along the south polar axis. When we did this in Sec. 2.3, we assumed the vector potential was proportional to 1/r; thus it sufficed to make the gauge transformation on some one sphere. Here, we make no such assumption, so we might have to make an independent gauge transformation on each sphere. This can induce a non-zero A_r, but this is no problem; A_r never enters into the Lubkin construction.

We now choose our family of closed paths as shown in Fig. 4. This is a north polar projection of the sphere; the circumference of the disc is the south pole, $\theta = \pi$. The path labeled by τ goes from the north pole to the south pole along the meridian $\phi = 0$ and returns along $\phi = 2\pi\tau$.

Because A_θ vanishes,

$$g(\tau) = P \exp - \int_0^{2\pi\tau} A_\phi(\pi,\phi)\, d\phi \ . \tag{3.60}$$

Thus we need to evaluate the integral for only one path, an infinitesimal loop circling the Dirac string. This expression becomes especially simple for one of the Goddard-Nuyts-Olive solutions, Eq. (3.47),

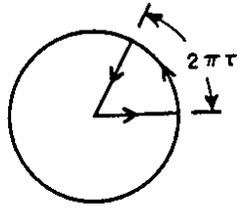

Figure 4

$$g(\tau) = \exp(-4\pi\tau Q) \ . \tag{3.61}$$

This equation enables us to readily compute the topological class into which a given GNO monopole field falls. Let us begin with ordinary electromagnetism, for which G is U(1) and Q = -ieg. For eg = n/2,

$$g(\tau) = \exp(i2\pi n\tau) \ . \tag{3.62}$$

This winds n times around U(1) as τ goes from 0 to 1. Thus, for each element of π_1, there is one and only one monopole field. Topological charge is magnetic charge. The situation is quite different for other groups, though. For example, for SU(n), we found an infinite number of GNO monopole fields, but SU(n) is simply connected; π_1 has only a single element. Likewise, for $SU(n)/Z_n$ there are an infinite number of GNO monopole fields but only n elements in $\pi_1 = Z_n$. (As an exercise, you might want to work out which element of Z_n is associated with each of the fields.)

Thus, the topological classification is in general coarser than the dynamical one. This is what one might expect; the topological classification is based on fewer assumptions than the dynamical one, and therefore should contain less information.

I've been misleading you. Not because anything I've said in the preceding paragraph is a lie, but because I have been withholding an important piece of information: most GNO monopole fields are unstable.

3.5 The Collapse of the Dynamical Classification

Stability analysis is trivial for an Abelian monopole, because the field equations are linear. This is not so for a non-Abelian monopole. Even at large distances, the field equations do not linearize; if the gauge fields fall off like 1/r, derivatives and commutators are of comparable magnitude, both $O(1/r^2)$.

I will now investigate small vibrations about an arbitrary SO(3) monopole field.[17] (Of course, this will also take care of SU(2) monopole fields, a subset of the SO(3) ones.) After the computation is

done, I will discuss the extension to an arbitrary GNO monopole field for an arbitrary gauge group. The calculation is lengthy and I will present it in outline only; you should have no trouble filling in the missing steps, if you're interested.

We work in temporal gauge, $A_0 = 0$, and write the gauge field as an SO(3) monopole plus a small perturbation,

$$\vec{A} = -\tfrac{1}{2} i q \sigma_3 \vec{A}_D + \delta\vec{A} \ , \tag{3.63}$$

where, as explained in Sec. 3.3, $q = 0, \tfrac{1}{2}, \ldots$, and A_D is the Dirac string potential, Eq. (2.64). $\pi_1(SO(3))$ is Z_2, so there are only two topological classes. It is easy to check that the monopoles with integer q's are all in one class and the ones with half-odd-integer q's are all in the other.

We write the perturbation as an explicit 2×2 matrix,

$$\delta\vec{A} = -\tfrac{1}{2} i \begin{pmatrix} \vec{\phi} & \vec{\psi} \\ \vec{\psi}^* & -\vec{\phi} \end{pmatrix} \tag{3.64}$$

where $\vec{\phi}$ is a real vector field and $\vec{\psi}$ is a complex one. If we linearize the field equations in the perturbation, an easy computation shows that $\vec{\phi}$ obeys a free wave equation,

$$-\partial_0^2 \vec{\phi} = \vec{\nabla} \times (\vec{\nabla} \times \vec{\phi}) \ , \tag{3.65}$$

while $\vec{\psi}$ obeys a more complicated equation

$$\begin{aligned} -\partial_0^2 \vec{\psi} &= \vec{D} \times (\vec{D} \times \vec{\psi}) + \frac{iq\vec{r}}{r^3} \times \vec{\psi} \\ &\equiv H\vec{\psi} \ , \end{aligned} \tag{3.66}$$

where

$$\vec{D} = \vec{\nabla} - iq\vec{A}_D \ . \tag{3.67}$$

(Note that q here plays the role of e_g in Sec. 2.)

The stability of the system under small perturbations is determined by the eigenvalue spectrum of the differential operator H. If H has a negative eigenvalue, the associated eigenmode has exponential time behavior, and the monopole field is unstable.

H has a large set of eigenmodes with eigenvalue zero. This is a consequence of the invariance of the temporal-gauge equations of motion under time-independent gauge transformations. We write such a transformation as

$$g(\vec{x}) = 1 - \tfrac{1}{2} i \begin{pmatrix} \lambda(\vec{x}) & \chi(\vec{x}) \\ \chi^*(\vec{x}) & -\lambda(\vec{x}) \end{pmatrix} + \cdots \ , \tag{3.68}$$

where the triple dots indicate quadratic and higher-order terms in λ and χ. An easy computation shows that, under this transformation,

$$\vec{\psi} \to \vec{\psi} + \vec{D}\chi + \cdots \ . \tag{3.69}$$

Since $\vec{\psi} = 0$ is certainly a solution of Eq. (3.66), so must be its gauge transform, $\vec{\psi} = D\chi$. Thus,

$$H\vec{D}\chi = 0 \tag{3.70}$$

for arbitrary χ.

We call these trivial eigenmodes gauge modes. All of the interesting physics is in the physical modes, the modes that are orthogonal to the gauge modes. For a physical mode,

$$\int d^3x \, \vec{\psi}^* \cdot \vec{D}\chi = 0 \ , \tag{3.71}$$

for any χ. Equivalently,

$$\vec{D} \cdot \vec{\psi} = 0 \ . \tag{3.72}$$

As always, rotational invariance is a great simplifier. H commutes with

$$\vec{J} = \vec{L} + \vec{S} \ , \tag{3.73}$$

where, just as in Sec. 3.4,

$$\vec{L} = -i\vec{r} \times \vec{D} - q\frac{\vec{r}}{r} \ , \tag{3.74}$$

and $\vec{S}$ is the usual spin operator for vector fields, defined by

$$(\vec{a} \cdot \vec{S})\vec{b} = i\vec{a} \times \vec{b} \ , \tag{3.75}$$

for any two vectors $\vec{a}$ and $\vec{b}$. As explained in Sec. 2.4, the orbital angular momentum takes on the values,

$$\ell = q, q+1, q+2 \ \ldots \ . \tag{3.76}$$

Thus, by the usual rules for adding angular momenta, the total angular momentum takes on the values

$$\begin{aligned} j &= q-1,\ q,\ q+1\ \ldots\ , \qquad && q \geq 1\ , \\ &= q,\ q+1\ \ldots\ , && q = 0 \text{ or } \tfrac{1}{2}\ . \end{aligned} \tag{3.77}$$

(I won't bother here to keep track of how often each value of j occurs.)

We can use Eq. (3.75) to purge H of cross products:

$$H = -(\vec{S}\cdot\vec{D})^2 + q\,\frac{\vec{S}\cdot\vec{r}}{r^2}\ . \tag{3.78}$$

It is then straightforward to use the methods of Sec. 2.4 to show that, acting on functions of definite j,

$$\begin{aligned} H\vec{\psi} = &\left[-\frac{\partial^2}{\partial r^2} - \frac{2}{r}\frac{\partial}{\partial r} + \frac{j(j+1)-q^2}{r^2}\right]\vec{\psi} \\ &+ \vec{X}\vec{D}\cdot\vec{\psi} + \vec{D}(\vec{Y}\cdot\vec{\psi})\ . \end{aligned} \tag{3.79}$$

Here $\vec{X}$ and $\vec{Y}$ are two ugly objects whose explicit form is of no interest to us, since the terms in which they occur make no contribution to the matrix element of H between physical modes. (The $\vec{X}$ term annihilates the mode on the right and the $\vec{Y}$ term the mode on the left.)

We are now home. For $q \geq 1$, j can be $q-1$, and

$$j(j+1) - q^2 = -q\ . \tag{3.80}$$

That is to say, the centrifugal potential is attractive. This is bad news. Just how bad can be seen by computing the expectation value of H for the radial function

$$\begin{aligned} \psi &= 0\ , && r < R\ , \\ &= \frac{1}{r}\left(\sqrt{r} - \sqrt{R}\right)e^{-r/a}, && r \geq R\ , \end{aligned} \tag{3.81}$$

where R and a are positive numbers. The expectation value is given by

$$\begin{aligned} \langle H\rangle &= \int_0^\infty r^2\,dr\,\psi^*\left(-\frac{d^2}{dr^2} - \frac{2}{r}\frac{d}{dr} - \frac{q}{r^2}\right)\psi \\ &= \int_0^\infty dr\,\left[r^2(d\psi/dr)^2 - q\,\psi^2\right] \\ &= (\tfrac{1}{4} - q)\ln a + \ldots\ , \end{aligned} \tag{3.82}$$

where the triple dots denote terms that have a finite limit as a goes to infinity. For any fixed R, this expression becomes negative for sufficiently large a.

Thus, not only does H have negative eigenvalues, the existence of these eigenvalues is totally insensitive to the form of the gauge field at short distances (or, indeed, at any finite distance). This is good to know, because we don't trust our expressions at short distances, inside the black box. All the GNO monopole fields with $q \geq 1$ are unstable under arbitrarily small perturbations at arbitrarily large distances; that is to say, they decay by the emission of non-Abelian radiation.

(In case you've been worrying about it, the modes we have been studying are indeed physical modes. Taking the covariant divergence is a rotationally invariant operation. Thus if $\vec{\psi}$ has $j = q-1$, so does $\vec{D}\cdot\vec{\psi}$. But we showed in Sec. 2.4 that every non-zero scalar function has $j \geq q$. Thus $\vec{D}\cdot\vec{\psi} = 0$.)

For SO(3), there are only two stable GNO monopole fields, $q = 0$ (no monopole at all) and $q = \frac{1}{2}$, one for each of the two topological classes. It's reasonable that there should be at least one stable monopole in each topological class: We can't radiate away topological charge; the passage of radiation through a Lubkin sphere is just another continuous deformation. What's surprising is that there is only one stable monopole for each topological class. For SU(2), it's also true that there's only one stable monopole in each topological class. SU(2) is simply connected, so there's only one topological class, and only $q = 0$ is a SU(2) GNO solution.

Once we have gone through the hard work of the stability analysis for SO(3), there's no need to get involved with differential equations again for more general groups. For example, let me do the analysis for $SU(n)/Z_n$.

As in Eq. (3.55), we write the gauge field in terms of the $n \times n$ matrices of the fundamental representation. Writing these out in components

$$\vec{A}_{ij} = -\tfrac{1}{2} i [q_i \delta_{ij} \vec{A}_D(\vec{x}) + \delta A_{ij}] \ , \qquad (3.83)$$

where there is no sum on i. In this expression

$$q_i = \frac{r}{n} + \text{integer} \ , \qquad (3.56)$$

and the q's sum to zero. It's easy to check that r = 0, 1 ... n-1 labels the topological class of the monopole. It's also easy to check that δA_{ij} obeys a differential equation identical in form to that obeyed by $\vec{\psi}$, Eq. (3.66), with the substitution

$$q \to q_i - q_j \ . \qquad (3.84)$$

Thus, the condition for infinitesimal stability is

$$q_i - q_j = 0, \pm 1 \ , \qquad \text{all } i, j \ . \qquad (3.85)$$

The only way this condition can hold is if the q's assume only two values. Because the q's sum to zero, one of these values must be non-negative. I will rearrange the q's such that the first s of them assume this value,

$$q_i = q_1 \geq 0 \ , \quad 1 \leq i \leq s \ , \qquad (3.86a)$$

Hence, by Eq. (3.86),

$$q_i = q_1 - 1 \ , \quad s+1 \leq i \leq n \ . \qquad (3.86b)$$

These sum to zero only if

$$q_1 = (n-s)/n \ . \qquad (3.87)$$

Thus, r is n-s; once again there is one and only one stable monopole in each topological class.

It's straightforward to extend this analysis to all of the classical groups, and, if you can remember their definitions, to the five exceptional Lie groups. Alternatively, if you're skilled at lifting weights and digging up roots, you can smash the general problem in one blow using the structure theory of Lie groups.[19] I won't do it either way here, but just tell you that no matter how you do it, the answer is the same: there is only one stable GNO monopole field for each topological class; the only stability is topological stability.

You may find it helpful in thinking about this to consider a simple mechanical problem that has the same property, an elastic loop constrained to lie on the surface of a sphere. For purposes of this problem, an elastic loop is a system whose potential energy is proportional to its length. Thus to find the time-independent solutions of the equations of motion is to find the closed geodesics on a sphere. These are the trivial geodesic (the loop bunched up at a single point), one transit of a great circle, two transits of a great circle, etc. Because a sphere is simply connected, there is only one topological class. Here it is easy to see that the only stability is topological stability. If the loop is wound around the sphere, we have but to move it ever so slightly, and it will spring away, scrunching itself up into a single point.

3.6 An Application

Topology is power. If we understand a dynamical system in topological terms, we can often deduce its qualitative features without messing around with detailed quantitative computations. As an example, I will here discuss the force between distantly separated non-Abelian monopoles. ("Distantly separated" only because we don't yet know what monopoles look like at short distances.)

For distantly separated Abelian monopoles, the force can be either repulsive or attractive, depending on whether the magnetic charges are of like or unlike sign. As we shall see, the situation is different for non-Abelian monopoles; if G is semi-simple, the force is always attractive. ("Semi-simple" means without U(1) factors; if U(1) factors are present in G, they can produce an Abelian repulsion that can overwhelm the attraction from the other factors.)

I will begin with SO(3) and afterwards generalize the result. As we have just seen, for SO(3) there is only one stable monopole,

$$\vec{A} = -\tfrac{1}{2} i \sigma_3 \vec{A}_D(\vec{x}) \ . \qquad (3.88)$$

Of course, there are many gauge-equivalent ways of writing this; in particular, minus this expression is just the same monopole in another gauge.

Because of this, there are two ways of writing the field of two

monopoles $$\vec{A} = -\tfrac{1}{2} i \sigma_3 [\vec{A}_D(\vec{x}-\vec{r}_1) \pm \vec{A}_D(\vec{x}-\vec{r}_2)] \ . \qquad (3.89)$$

This superposition of two solutions is a solution because everything lies in a single Abelian subgroup; thus the nonlinear commutator terms in the Yang-Mills equations never appear. If we had attempted to add a monopole pointing in the 3-direction to its gauge transform pointing in the 2-direction, for example, the nonlinear terms would have made trouble. There may be other ways, less trivial than these, of putting together two monopoles, but I have been unable to find them; for purposes of this discussion, I will assume these two are the only ones.

In the Abelian case, an expression like Eq. (3.89) would be interpreted as either the superposition of two monopoles (plus sign) or of a monopole and an antimonopole (minus sign). At the risk of being repetitious, I emphasize that this is not the case here. The minus sign is simply a gauge transform of the plus sign; the two signs correspond to two different ways of putting together the same two monopoles, just as spin one and spin zero are two different ways of putting together the same pair of spin-$\frac{1}{2}$ objects.

Despite the difference in interpretation, the computation of the interaction energy stored in the magnetic field is the same as in Abelian gauge theory,

$$E_{int} \propto \pm 1/|\vec{r}_1-\vec{r}_2| \ , \qquad (3.90$$

repulsive in the plus case, attractive in the minus. The difference from the Abelian case appears when we study the field at large distances,

$$\vec{A} = -\tfrac{1}{2} i \sigma_3 \vec{A}_D(\vec{x}) \, [1 \pm 1] + O(1/r^2) \ . \qquad (3.91)$$

Both of these are GNO monopole fields; they have to be, because they are time-independent solutions of the Yang-Mills equations. Both are in the same topological class; they have to be, because the topological charge of a two-monopole system is always the product of the topological charges of the individual monopoles, no matter how we weave the fields together. But only one of them is stable, because there is only one stable GNO field in each topological class, and it is the minus field.

Thus, even if we were able to put the two monopoles together in the repulsive plus configuration, they wouldn't stay there for a minute; they would emit non-Abelian radiation and settle down to the attractive minus configuration. This is much like the situation for two bar magnets, each free to pivot about its center of mass. It doesn't matter what the initial orientation of the magnets is; they will realign themselves until they are in the maximally attractive (anti-aligned) configuration. Here the realignment takes place in an internal rather than a geometrical space, but the physics is much the same.

Now let us go on to $SU(n)/Z_n$. Here we have n stable monopole fields, so the two monopoles may be gauge-inequivalent. Nevertheless, as we shall see, the force is still attractive. We write the field of the two monopoles as

$$\vec{A} = Q_1 \vec{A}_D(\vec{x}-\vec{r}_1) + Q_2^P \vec{A}_D(\vec{x}-\vec{r}_2) \ . \qquad (3.92)$$

My notation here needs some explanation. Q_1 and Q_2 are the two $n \times n$ matrices that appear in the single-monopole fields. As before, we wish to avoid the nonlinear terms in the Yang-Mills equations, so we choose Q_1 and Q_2 to commute; this implies that they can be simultaneously diagonalized. This still leaves us the freedom to permute the eigenvalues of Q_2, say while keeping those of Q_1 fixed. We pick one arrangement of eigenvalues as the standard one, and denote the others by Q_2^P, where P is one of the n! permutations on n objects. Different ways of putting together the two monopoles correspond to different choices of the permutation P, but, just as before, the total topological charge is independent of P.

Because there is only one stable GNO field for a given topological charge, most choices of P will lead to unstable fields at large distances. We wish to find the energy in the unique stable configuration. For any P, just as before,

$$E_{int} \propto -\mathrm{Tr}\, Q_1 Q_2^P / |\vec{r}_1-\vec{r}_2| \ . \qquad (3.93)$$

(The minus sign is there because the Q's are anti-Hermitian.) If we sum over permutations,

$$\sum_P Q_2^P = (n-1)!\,\mathrm{Tr}\,Q_2 = 0 \ . \tag{3.94}$$

Thus the energy averaged over all configurations vanishes, and therefore the energy in the unique stable configuration, the configuration of minimum energy, must be negative. Once again, the force is attractive.

If you know anything about the general theory of Lie groups, you can see that it is trivial to generalize this argument. The only property of the Q's we needed was their tracelessness, and this holds for an arbitrary semi-simple group. (For group mavens, the precise statement required is that only the zero element of the Cartan subalgebra is invariant under the Weyl group.) Thus, even in this case, any two monopoles attract each other.

4. INSIDE THE MONOPOLE

4.1 Spontaneous Symmetry Breakdown - A Lightning Review

Up to now, we have kept our distance from the monopole. We are now going to open the black box, to see whether the structures we have found at large distances can be continued down to r = 0.

We will work in the context of gauge field theories with spontaneous symmetry breakdown. These theories are the daily bread of contemporary high-energy physics; nevertheless, I'll give a lightning review of them here, much like the review of Sec. 3.1, but even more compressed. My purposes are the same as before, to establish common notation and emphasize salient points.

In Sec. 3.1, we divided the fields in our theory into gauge fields and matter fields; it will now be convenient to divide the matter fields into scalar fields (assumed, for convenience, to be all real) and other(typically, spinor) fields. We will change notation slightly and use ϕ to denote a big vector made up of the scalar fields only.

We will restrict ourselves to theories for which the matter Lagrange density is of the form

$$\mathscr{L}_m = \tfrac{1}{2} D_\mu \phi \cdot D^\mu \phi - U(\phi) + \ldots \ , \tag{4.1}$$

where U is some G-invariant function and the triple dots indicate terms involving the non-scalar matter fields.

Up to gauge transformations, the ground states of the theory are states for which all non-scalar fields vanish, and the scalar fields are space-time independent and at an absolute minimum of U. We will use quantum language for this classical situation, and refer to the ground states as "vacua", and the ground-state value of ϕ as "the vacuum expectation value of ϕ". We will always assume a constant has been added to U such that the ground-state energy is zero.

Let $\langle\phi\rangle$ be some (arbitrarily selected) minimum of U; then $g\langle\phi\rangle$ is also a minimum, for any g in G. We will assume all minima of U are of this form. Then all the ground states are physically equivalent, and with no loss of generality we can restrict ourselves to the case in which the vacuum expectation value of ϕ is $\langle\phi\rangle$. (This assumption excludes interesting phenomena such as accidental degeneracy, in which no symmetry connects the ground states, and Goldstone bosons, in which a symmetry connects them but not a gauge symmetry. We make the assumption here just to keep our arguments as simple as possible; it's not difficult to extend the analysis to the more general case.)

We define H to be the subgroup of G that leaves $\langle\phi\rangle$ invariant,

$$h \in H \text{ iff } \quad h\langle\phi\rangle = \langle\phi\rangle \ . \tag{4.2}$$

We say the symmetry group G has spontaneously broken down to H. The gauge fields associated with H remain massless; the others combine with some of the scalar fields to form massive vector fields. (This is the famous Higgs mechanism.) The vector mass matrix is given by

$$\mu^2_{ab} = f_a T_a \langle\phi\rangle \cdot f_b T_b \langle\phi\rangle \ , \tag{4.3}$$

where there is no sum on repeated indices. The scalar fields absorbed into the massive vector fields are those that correspond to perturbations of the vacuum of the form $\delta\phi = T_a\langle\phi\rangle$.

Had we made a different arbitrary choice of $\langle\phi\rangle$, H would have been a different subgroup of G, although one isomorphic to our original H. Occasionally we will find ourselves working in a gauge such that the vacuum expectation value of Φ varies in space; it will be important then to remember that H varies with it.

As an example of these ideas, let G be SO(n), Φ an n-vector, and

$$U = \frac{\lambda}{4}\,(\Phi\cdot\Phi - c^2)^2 \ , \qquad (4.4)$$

where λ and c are positive numbers. The possible choices for $\langle\Phi\rangle$ are all vectors of length c. If we make the choice

$$\langle\Phi^a\rangle = c\,\delta^{an} \ , \qquad a = 1\ldots n \qquad (4.5)$$

then H is the subgroup of rotations on the first n-1 coordinates, SO(n-1). Of the original $\frac{1}{2}n(n-1)$ gauge fields, $\frac{1}{2}(n-1)(n-2)$ remain massless; the other n-1 combine with n-1 scalar fields to become massive; one scalar field remains untouched.

In what follows, we shall make much use of this theory with n = 3. This case faintly resembles reality in that there is only one massless gauge meson, which we can identify with the photon; for this reason it was incorporated by Georgi and Glashow into an ingenious but erroneous alternative to the Weinberg-Salam model, and is sometimes called the Georgi-Glashow model.

4.2 Making Monopoles[19]

I will now explain how magnetic monopoles arise in spontaneously broken gauge field theories. This phenomenon was first discovered by 't Hooft and Polyakov,[2] working in the SO(3) model I've just described. It is one of the most dazzling effects in field theory; magnetic monopoles miraculously appear in theories that contain no fundamental magnetically charged fields at all.

It turns out that the non-scalar matter fields play no role in building monopoles, so, for simplicity, I will assume that we are working in a theory of gauge fields and scalar fields only. In this theory, we will study finite-energy nonsingular field configurations

at some fixed time. Later on, we'll worry about how these configurations evolve in time.

The energy density is

$$\mathcal{E}^{\infty} = \tfrac{1}{2}\,\vec{D\Phi}\cdot\vec{D\Phi} + U(\Phi) + \text{other positive terms} \ . \qquad (4.6)$$

For the energy integral to converge, each of the two terms displayed must vanish at large r. To keep the argument simple, I will assume here that they are strictly zero outside some radius R,

$$U = \vec{D\Phi} = 0 \ , \qquad r \geq R \ . \qquad (4.7)$$

I stress that this is a totally bananas assumption. It is not a consequence of finiteness of the energy and it is not even true of any of the known finite-energy solutions to the field equations. I make it here for pedagogical reasons only, so I can construct an argument in which the underlying structure is not buried under analysis of how rapidly and uniformly limits are attained. I invite those of you skilled at real analysis to generalize the argument to weaker and more sensible assumptions.

Equation (4.7) implies that Φ must be at a minimum of U for $r \geq R$, but it may be at different minima at different points. However, we can always gauge transform such that

$$\Phi = \langle\Phi\rangle \ , \qquad r \geq R \ . \qquad (4.8)$$

We do this in two steps. First we make a gauge transformation depending only on r such that Φ is transformed to $\langle\Phi\rangle$ along the north polar axis, for $r \geq R$. Then, for each fixed r, we start at the north pole and transform along meridians to make $\Phi = \langle\Phi\rangle$ everywhere. This second step may introduce a Dirac string singularity along the south polar axis, where all meridians intersect again.

Equations (4.7) and (4.8) imply that

$$\vec{D\Phi} = \vec{A}\langle\Phi\rangle = 0 \ , \qquad r \geq R \ . \qquad (4.9)$$

Thus, after we have made our gauge transformations, the only gauge fields that exist for $r \geq R$ are those associated with the unbroken subgroup H. Of course, this is just what we would have expected;

only massless fields can extend to large distances.

Thus we have the same structure as in Sec. 3; outside of a black box, the sphere of radius R, there is nothing but massless gauge fields. The only change is a mild notational one; the group we called G in Sec. 3 we now call H. Thus, outside the sphere, we have the usual classification of field configurations by their topological charges, elements of $\pi_1(H)$.

If all there were inside the sphere were H gauge fields, then a nonsingular field would always have trivial topological charge. I gave the argument for this in Sec. 3.4, but let me give it again here: With every sphere, we associate a path in group space. If we take a sphere around a monopole and shrink it to a point, the associated path must become the constant path. Thus there are only two possibilities. Either the path changes continuously, in which case it was in the trivial homotopy class to begin with, or it changes discontinuously, in which case we have encountered a singularity. (Note that since the topological charge is gauge-invariant, singularities that are mere gauge artifacts, like Dirac strings, will not do the job; a genuine gauge-invariant singularity is needed.)

We now see what is different in the case at hand. Inside the black box, we have G gauge fields, not just H ones, and the path may move out of H and into the larger group G. It is quite possible for a path that is homotopically nontrivial in H to be homotopically trivial in G; a topological knot that can not be disentangled in the smaller space may fall apart in the larger. In this way we can have a nontrivial topological charge without a singularity.

This can be phrased in somewhat more abstract language. Because H is a subgroup of G, every path in H is a path in G. This induces a mapping of $\pi_1(H)$ into $\pi_1(G)$. The kernel of this mapping is defined, as always, as the subgroup that is mapped into the identity. Thus our result can be stated as follows:

> The condition for a nonsingular monopole is that the topological charge is in the kernel of the mapping $\pi_1(H) \to \pi_1(G)$.

We have seen that this condition is necessary. I will now demonstrate that it is sufficient by constructing a nonsingular finite-energy field configuration for every topological charge in the kernel.

For any topological charge, there is a field at large distances with that charge, the appropriate GNO field,

$$\vec{A}_{GNO} \cdot d\vec{x} = Q(1 - \cos\theta)d\phi \ . \tag{4.10}$$

Here, Q is the Lie algebra of H,

$$Q\langle\Phi\rangle = 0 \ , \tag{4.11}$$

and the path

$$g(\tau) = e^{4\pi Q\tau} \ , \qquad 0 \le \tau \le 1 \ , \tag{4.12}$$

is in the specified homotopy class.

In G, $g(\tau)$ is homotopic to the trivial path. Thus, there exists a continuous function of two variables, $g(\theta,\phi) \in G$, $\theta \in [0,\pi]$, $\phi \in [0,2\pi]$, such that

$$g(0,\phi) = g(\theta,0) = g(\theta,2\pi) = 1 \tag{4.13a}$$

and

$$g(\pi,\phi) = e^{2Q\phi} \ . \tag{4.13b}$$

We will use this to define our vector and scalar fields for $r \ge R$,

$$\Phi = g\langle\Phi\rangle \ , \tag{4.14a}$$

$$\vec{A} = g\vec{A}_{GNO}g^{-1} + g\vec{\nabla}g^{-1} \ . \tag{4.14b}$$

This is simply a gauge transform of the GNO field, so it is still a finite-energy field configuration. However, the Dirac string has completely disappeared; all our fields are now manifestly nonsingular at all angles. The price we have paid is that we have made the vacuum expectation value of Φ angle-dependent.

It is now trivial to continue this to $r \le R$,

$$\Phi(r,\theta,\phi) = \frac{r^2}{R^2}\Phi(R,\theta,\phi) \ , \tag{4.15a}$$

$$\vec{A}(r,\theta,\phi) = \frac{r^2}{R^2}\vec{A}(R,\theta,\phi) \ . \tag{4.15b}$$

This is manifestly nonsingular and of finite energy all the way down to $r = 0$. Note that we could not have continued things into the

interior in this way if we had not first removed the string; if we had attempted to simply scale down the GNO field, we would have violated the quantization condition.

This completes the proof of sufficiency.

The fact that the condition is both necessary and sufficient makes it the jewel of monopole theory, well worthy of being honored with a box. It enables us to tell instantly, without solving a single differential equation, whether a given theory admits nonsingular monopoles, and what kinds it admits. I will give three examples:

(1) G is SO(3) and H is SO(2). This is the theory described at the end of Sec. 4.1, the Georgi-Glashow model. If we identify the unbroken subgroup with the electromagnetic group, then large distance analysis allows eg to be 0, $\pm\frac{1}{2}$, ± 1, etc. Topologically, these correspond to paths that go 2eg times around SO(2). Only paths that go an even number of times around SO(2) can be deformed into the trivial path in SO(3); thus the allowed values of eg are the alternate terms in the series, eg = 0, ± 1, etc.

(2) G is U(2) and H is U(1), embedded as the subgroup of U(2) that leaves the first basis vector in the two-dimensional unitary space invariant. This is the Weinberg-Salam model. G is locally isomorphic to U(1) ⊗ SU(2), and any path that winds around H also winds around the U(1) factor in G. Thus only trivial topological charge is allowed, eg = 0.

(3) G is any semisimple group and H is any group with a U(1) factor. These are the grand unified models referred to in Sec. 1. $\pi_1(G)$ is finite and $\pi_1(H)$ is infinite, so the kernel of the mapping must be infinite and the theory must contain monopoles, as I asserted in Sec. 1. Of course, we can't tell precisely what magnetic charges are allowed until we know the details of the theory.

We have concerned ourselves here with finite-energy nonsingular field configurations at fixed time. If we have a well-posed initial-value problem, any such configuration will evolve into some solution of the equations of motion, but it need not be time-independent. Thus, none of our analysis demonstrates the possibility of building time-independent monopoles. This is not important. We theorists like time-independent solutions, but that is because they are easy to analyze and we are lazy. If an experimenter finds a black box surrounded by a monopole field, this is of interest whether the box contains time-independent fields or oscillating, rotating, ergodically quivering fields.

We do know one thing about the time evolution of configurations with nontrivial topological charge. Whatever they do, they can not dissipate utterly, simply leak out of the box in the form of ordinary radiation fields, massive or massless. This is because radiation does not carry topological charge, and topological charge is conserved. Topology is power.

4.3 The 't Hooft-Polyakov Object

Despite these sour words about time-independent solutions, I will devote some time here to discussing the famous time-independent monopole found in the Georgi-Glashow model, Eq. (4.4), by 't Hooft and Polyakov.[2] Searching for general time-independent solutions is difficult because one has to deal with many functions of three variables; the trick is to simplify things by looking for solutions that are symmetric under some astutely chosen subgroup of the symmetry group of the theory.

It turns out that a fruitful subgroup for this purpose is the SO(3) subgroup consisting of simultaneous spatial and internal rotations. Scalar fields invariant under this are necessarily of the form

$$\phi^a = f(r^2)\,\frac{r^a}{r}\,, \tag{4.16}$$

where the internal index a runs from 1 to 3. To keep the energy finite at infinity, $f(\infty)$ must equal c. To keep the field nonsingular at r = 0, f(0) must vanish.

This is not only invariant under the stated SO(3) group, it is also invariant under parity, if we define ϕ to be pseudoscalar. The only gauge fields invariant under both these symmetries are of the form

$$A^{ia} = h(r^2)\epsilon^{iak} r_k \ . \tag{4.17}$$

Finiteness of the energy puts restrictions on the large-r behavior of h, but I won't bother to work them out here.

Thus finding time-independent solutions becomes a simple problem in the calculus of variations; we must minimize the energy as a functional of two functions of a single variable. This problem is not beyond the reach of either functional or numerical analysis; I hope you will find it plausible when I tell you that it is possible both to prove a solution exists and to compute this solution with good accuracy on a pocket calculator.

We have a time-independent solution, but is it a monopole? The easiest way to answer this question is to transform the solution into string gauge. From Eq. (4.16),

$$\phi(r,\theta,\phi) = g(\theta,\phi)\phi(r,\theta=0) \ , \tag{4.18}$$

where

$$g(\theta,\phi) = e^{T_3\phi}\, e^{T_2\theta}\, e^{-T_3\phi} \ , \tag{4.19}$$

and the T's are the generators of SO(3). Equation (4.18) would be true even if I left the last factor out of Eq. (4.19); I put it in so g would be well defined at $\theta = 0$.

Let us now make a gauge transformation, using g^{-1}:

$$\phi(r,\theta,\phi) \to g^{-1}\phi = \phi(r,\theta=0) \ . \tag{4.20}$$

This aligns the vacuum expectation value throughout space; in all directions, the unbroken group H is the SO(2) subgroup generated by T_3. Under the same gauge transformation,

$$\vec{A} \to g^{-1}\vec{A}g + g^{-1}\vec{\nabla}g \ . \tag{4.21}$$

Only the second term in this expression produces a Dirac-string singularity on the south polar axis. Here,

$$g(\pi,\phi) = e^{T_2\pi}\, e^{-2T_3\phi} \ , \tag{4.22}$$

and

$$A_\phi(\pi,\phi) \to -2T_3 \ . \tag{4.23}$$

Integrating this, we see that we circle the SO(2) subgroup twice when we go once around the string. This is a monopole with $eg = 1$, the minimal value allowed by the topological considerations of Sec. 4.2.

4.4 Why Monopoles are Heavy

I said in the introduction that monopoles are typically very massive. Now that we understand their topological structure we can see why this is so.

Let me begin by considering a theory in which all the heavy gauge fields have similar masses, on the order of some typical mass, μ, and all of the gauge couplings are on the order of some typical coupling, e. In any non-radiant configuration, the heavy gauge fields fall off with distance like $\exp(-\mu r)$. Thus we expect the core of the monopole, the region in which the heavy fields are signigicant, to have a size on the order of $1/\mu$. Outside the core, only massless fields should be significant, and the monopole should look like a magnetic Coulomb field (in the Abelian case) or a GNO field (in the general case).

It is easy to estimate the energy stored outside the core. Just to be definite, let me do the computation for an Abelian monopole of magnetic charge g,

$$\begin{aligned} E_{\text{magnetic}} &= \tfrac{1}{2} \int_{r>O(1/\mu)} d^3\vec{x}\,|\vec{B}|^2 \\ &= 2\pi g^2 \int_{O(1/\mu)}^{\infty} dr/r^2 \\ &= 2\pi g^2\, O(\mu) \ . \end{aligned} \tag{4.24}$$

Because the energy density of the theory is positive, the energy inside the core can only add to this. Thus,

$$m \geq O(\mu/e^2) \ . \tag{4.25}$$

In this order-of-magnitude form, the bound is clearly also valid

for non-Abelian monopoles. Monopoles are heavy because the integral for the electromagnetic energy of a Coulomb field is divergent at short distances.

If we're willing to make some plausible guesses, we can replace this inequality with an equality. If we imagine increasing the core radius, the magnetic energy decreases. Since the monopole is in equilibrium, this must be compensated for by an increase of the internal energy. (If the monopole is time-dependent, this is still true, if we average over time.) Thus, it is plausible to assume that the core energy is of the same magnitude as the magnetic energy.

Now let us go on to a theory which has many mass scales. This typically occurs when there is a hierarchy of symmetry breakdown,

$$H \subset G_1 \subset G_2 \subset G_3 \ \ldots \ , \tag{4.26}$$

with an associated hierarchy of masses

$$\mu_1 \ll \mu_2 \ll \mu_3 \ \ldots \ , \tag{4.27}$$

where μ_i is the typical mass of a gauge field in G_i (but not G_{i-1}). For example, in the grand unification model of Georgi and Glashow,[20] SU(5) breaks down to SU(3) (color) $\otimes$ U(2) (electroweak); the gauge fields associated with this get masses on the order of 10^{14} GeV. U(2) in turn breaks down to U(1), as in the Weinberg-Salam model; the gauge fields here get masses on the order of 10^2 GeV.

Associated with the sequence (4.26), there is a sequence of mappings

$$\pi_1(H) \to \pi_1(G_1) \to \pi_1(G_2) \ \ldots \ . \tag{4.28}$$

We can build a monopole with any topological charge that is eventually mapped into the identity. Let us suppose this first happens at the group G_i. Then the monopole field remains Coulombic down to distances of the order of $1/\mu_i$, and the monopole mass obeys

$$m \simeq O(\mu_i/e^2) \ . \tag{4.29}$$

As before, if we're willing to make some plausible guesses, this inequality can be replaced by an equality.

As an example, in the Georgi-Glashow grand unified theory, we have to go to SU(5), as we saw in Sec. 4.3; thus the monopole mass is on the order of 10^{16} GeV, or 10^{-8} grams.

Of course, it's possible for different elements of $\pi_1(H)$ to be mapped into the identity at different stages of the hierarchy; this gives the possibility of a wide variety of monopole mass scales. For example, consider a theory in which SU(3) breaks down to its real subgroup, SO(3), at some large mass scale; at a much smaller scale, SO(3) breaks down to SO(2) in the familiar way. This has monopoles with both integral and half-odd-integral values of eg, the latter much heavier than the former.

4.5 The Bogomol'nyi Bound and The Prasad-Sommerfield Limit

We have derived rough estimates of monopole masses. For certain theories, it is possible to derive a rigorous lower bound, the Bogomol'nyi bound.[21] I will give the derivation here for the Georgi-Glashow model; it can readily be extended to any theory in which the scalar fields transform according to the adjoint representation of the gauge group.

We write the energy of the theory as

$$E = \tfrac{1}{2}\int d^3\vec{x}\left[\vec{E}^a\cdot\vec{E}^a + \vec{B}^a\cdot\vec{B}^a + (D_0\phi^a)^2 + \vec{D}\phi^a\cdot\vec{D}\phi^a + \frac{\lambda}{2}(\phi^a\phi^a - c^2)^2\right] . \tag{4.30}$$

Here the vectors indicate spatial transformation properties and the indices internal symmetry ones. Also, E^a and B^a are defined as in electromagnetism,

$$E_i^a = F_{0i}^a \ , \qquad B_i^a = \tfrac{1}{2}\varepsilon_{ijk}F_{jk}^a \ , \tag{4.31}$$

and we have rescaled the fields as in Eq. (3.23), so there is no coupling constant in the expression for the energy.

To derive the bound, we shall need two properties of the magnetic field. One is

$$\vec{D}\cdot\vec{B}^a = 0 \ . \tag{4.32}$$

This is a consequence of the Jacobi identity for covariant differentiation. The other is

$$\lim_{r\to\infty}\int d^2S\,\vec{n}\cdot\vec{B}^a\phi^a = 4\pi g c \ , \tag{4.33}$$

where d^2S is the usual element of surface area and $\vec{n}$ is the outward-pointing normal. This is most easily seen to be true by going to the gauge introduced in Sec. 4.2, where, at large distances,

$$\phi^a = \langle\phi^a\rangle = \delta^{a3} c \ , \tag{4.34}$$

and the only surviving component of $\vec{B}^a$ is $\vec{B}^3$, a magnetic monopole field.

We are now ready to go.

$$\begin{aligned} E &\geq \tfrac{1}{2}\int d^3\vec{x}(\vec{B}^a\cdot\vec{B}^a + \vec{D}\phi^a\cdot\vec{D}\phi^a) \\ &= \tfrac{1}{2}\int d^3\vec{x}(\vec{B}^a \pm \vec{D}\phi^a)^2 \\ &\quad \mp 4\pi g c \ , \end{aligned} \tag{4.35}$$

by integration by parts and Eqs. (4.32) and (4.33). Thus

$$E \geq |4\pi g c| \ , \tag{4.36}$$

the desired result.

Because g is O(1/e) and μ is O(ce), the right-hand side of this equation is $O(\mu/e^2)$, consistent with the arguments of Sec. 4.4. Actually, this is evidence for nothing but dimensional analysis; once we have discarded the scalar coupling constant, λ, the only quantity we can build with the dimensions of energy is μ/e^2.

It is actually possible to saturate the bound in an appropriate limit, first studied by Prasad and Sommerfield.[22] In deriving the bound, we discarded three of the five terms in Eq. (4.30). Two of them, the E^2 term and the $(D_0\Phi)^2$ term, automatically vanish if we assume time-reversal invariance. We make the third one, the scalar potential, vanish by going to the limit $\lambda \to 0^+$. Phrased less formally, we eliminate all λ-dependent terms from the equations of motion, but retain the boundary condition that $\phi^a\phi^a = c^2$ at large distances.

In this limit, Eq. (4.35) is an equality, not an inequality, and we can saturate the bound if we can find solutions of

$$\vec{B}^a \pm \vec{D}\phi^a = 0 \ , \tag{4.37}$$

where the sign in this equation depends on the sign of g. As a bonus, any solution of Eq. (4.36) is also a time-independent solution of the equations of motion; a minimum of the energy functional is *a fortiori* a stationary point.

Equation (4.37) is a first-order differential equation, and considerably easier to analyze than the second-order equations of motion. I don't have time to give the analysis here, or even to describe its main results in any detail. However, it turns out that the equation does have solutions, and, delightfully, not just single-monopole solutions but also many-monopole ones. This is not so incredible as it may seem. In the Prasad-Sommerfield limit, the scalar field becomes massless, and therefore it is possible for many monopoles to exist in static equilibrium, their scalar attraction balancing their magnetic repulsion.

5. QUANTUM THEORY

5.1 Quantum Monopoles and Isorotational Excitations[19]

In Sec. 3.2, I argued that a quantum field theory should most closely resemble the corresponding classical theory in the limit of weak coupling. In this section, I will build on this observation to develop a quantitative approximation scheme for theories whose classical limits possess time-independent monopole solutions.[23]

For definiteness, we'll work with the Georgi-Glashow model,

$$\begin{aligned} \mathscr{L} = &-\frac{1}{4f^2}F^a_{\mu\nu}F^{\mu\nu a} + \frac{1}{2}D_\mu\phi^a\cdot D^\mu\phi^a \\ &-\frac{\lambda}{4}(\phi^a\cdot\phi^a - c^2)^2 \ . \end{aligned} \tag{5.1}$$

We define new variables by $\phi' = \phi f$, $c' = cf$, and $\lambda' = \lambda/f^2$. In terms of these,

$$\begin{aligned} \mathscr{L} = \frac{1}{f^2}\Big[&-\frac{1}{4}F^a_{\mu\nu}F^{\mu\nu a} + \frac{1}{2}D_\mu\phi'^a D^\mu\phi'^a \\ &-\frac{\lambda'}{4}(\phi'^a\cdot\phi'^a - c'^2)\Big] \ . \end{aligned} \tag{5.2}$$

I propose to study this theory in the limit of small f, with e' and λ' held fixed. This is the classical limit, as discussed in

Sec. 3.2; the relevant quantity in the quantum theory is $\mathscr{L}/\hbar$, so the limit of small f with fixed $\hbar$ is the same as the limit of small $\hbar$ with fixed f. Also, when things are written in this way, the classical monopole solution is independent of f; a factor that multiplies the total Lagrangian divides out of the equations of motion. (We will use these redefined parameters for the rest of this discussion, so I'll drop the primes from now on.)

The theory is especially easy to work with in temporal gauge, $A_0 = 0$,

$$\mathscr{L} = \frac{1}{f^2}\left[\frac{1}{2}\partial_0 \vec{A}^a \cdot \partial_0 \vec{A}^a + \frac{1}{2}\partial_0 \phi^a \partial_0 \phi^a + \text{terms without time derivatives}\right]. \tag{5.3}$$

Thus the canonical momentum density conjugate to ϕ^a is

$$\pi^a = f^{-2}\,\partial_0 \phi^a \ , \tag{5.4}$$

that conjugate to A^a is

$$\vec{\pi}^a = f^{-2}\,\partial_0 \vec{A}^a \ , \tag{5.5}$$

and the Hamiltonian is given by

$$f^2 H = \frac{1}{2} f^4 \int [\vec{\pi}^a \cdot \vec{\pi}^a + \pi^a \pi^a] d^3\vec{x} + V \ , \tag{5.6}$$

where V is minus the integral of the terms without time derivatives in Eq. (5.3) The classical monopole solution is a stationary point of the functional V; indeed, if the monopole is stable (as we shall assume it is), it is a local minimum of V. I emphasize that all powers of f are explicitly displayed in Eq. (5.6); V is independent of f.

Equation (5.6) defines a peculiar Hamiltonian from the viewpoint of ordinary perturbation theory. Firstly, there is an explicit f^2 on the left-hand side of the equation. Of course, this is a trivial peculiarity; if we can find an expansion for the energy eigenfunctions and eigenvalues of $f^2 H$, we can find one for those of H. Secondly, the small parameter multiplies the kinetic energy, the term quadratic in canonical momenta, rather than the potential energy, the term independent of canonical momenta. This is very strange; have we ever encountered such a system before?

Yes, we have. For this is the situation for a diatomic molecule:

$$H = \frac{\vec{p}^2}{2M} + V(r) \ , \tag{5.7}$$

where M is the reduced nuclear mass. The standard expansion in the study of the spectra of diatomic molecules uses as the small parameter 1/M, the coefficient of the kinetic energy. Of course, our system is not exactly a diatomic molecule. What it is exactly is a polyatomic molecule,

$$H = \sum_{i=1}^{N} \frac{\vec{p}_i^2}{2M_i} + \vec{V}(\vec{r}_1 \ldots \vec{r}_N) \ . \tag{5.8}$$

To be precise, it is an infinitely polyatomic molecule, where all the nuclei have mass $1/f^4$. Thus the problem of constructing quantum monopoles is one that was solved completely fifty years ago.

I will now explain this solution, first by reminding you of the familiar results for a diatomic molecule, then by telling you the trivial extension to a polyatomic molecule, and, finally, by making the even more trivial transcription of this extension into the language of field theory.

For the diatomic molecule, we assume the interatomic potential is as shown in Fig. 5. The minimum of V is at $r = r_0$, and $V(r) = E_0$. The first three approximations to the low-lying energy eigenstates and their eigenvalues are shown in Table 1.

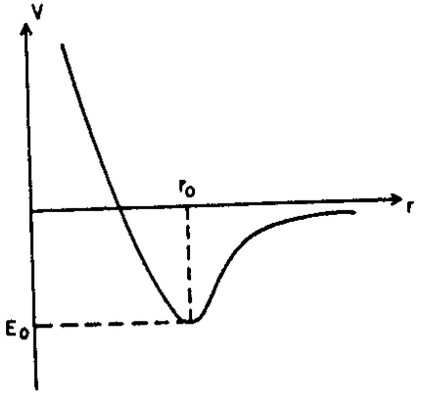

Figure 5

Table 1

The Diatomic Molecule

Order of Approximation	Energy Eigenstate	Energy Eigenvalue
0	$\|r_0,\theta,\phi\rangle$	E_0
1	$\|n,\theta,\phi\rangle$	$+(n+\frac{1}{2})\sqrt{V''(r_0)/M}$
2	$\|n,\ell,m\rangle$	$+\ell(\ell+1)/2M r_0^2+\ldots$

As we see from the table, the proper expansion parameter for energy eigenvalues is $1/\sqrt{M}$; this will become f^2 in the field-theory problem. (The right-hand side of the table is cumulative; that is to say, the energy in first order is the sum of the first two entries, etc.) I will now explain the origin of the table.

In zeroth order, we neglect the kinetic energy altogether. The particle sits at the bottom of the potential well, in an eigenstate of the position operator, $\vec{r}$. The magnitude of the position is fixed at r_0, but the angular position is arbitrary. This is not much like the real spectrum revealed by molecular spectroscopy; in particular, there is a totally spurious infinite angular degeneracy. As we shall see, this degeneracy is lifted only in second order.

In first order, we begin to see the effects of the vibration of the particle about its equilibrium position. Since M is very large, the particle does not vibrate very far, and, to first order, we can replace the potential near equilibrium by a harmonic potential,

$$V(r) = E_0 + \tfrac{1}{2} V''(r_0)(r-r_0)^2 \ . \qquad (5.9)$$

The energy eigenfunctions are now harmonic-oscillator wave functions in r, but still angular delta-functions. They are labeled by the usual oscillator excitation number, n, and have the usual oscillator energy. These are the famous vibrational levels of molecular spectroscopy.

It is only in second order that we begin to see the effects of rotation; this is because the zeroth-order moment of inertia of the molecule is $M r_0^2$. The degeneracy in angle is removed; angular eigenstates are replaced by angular-momentum eigenstates,

$$|n,\ell,m\rangle = \int d\Omega\, Y_{\ell m}(\theta,\phi)|n,\theta,\phi\rangle \ , \qquad (5.10)$$

and a rigid-rotator term is added to the energy. These are the famous rotational levels of molecular spectroscopy. Note that the rotational structure involves no properties of V that have not entered earlier approximations. In addition, we begin to see the effects of departures from the harmonic approximation, Eq. (5.9). I have indicated these terms (vibrational-vibrational coupling) in the table by triple dots. They depend on the detailed form of V (in particular, on its third and fourth derivatives at r_0). Unlike the rotational term, they do not affect the qualitative features of the problem, nor do the higher terms in the expansion.

The extension of all this to a polyatomic molecule is trivial. Unless the equilibrium configuration of the molecule is one in which all the nuclei are aligned, the equilibrium configurations are labeled, like the positions of a rigid body, by three Euler angles rather than two polar angles. As a consequence of this, the rotational spectrum, once it appears in second order, will be that of a rigid body, rather than a rigid rotator. Also, there are many ways to vibrate about equilibrium, and the single integer n is replaced by a string of integers n_i, one for each normal mode.

It is straightforward to transcribe this to field theory; the only question is what replaces the rotational spectrum. The zeroth-order molecular energy eigenstates were rotationally degenerate because the classical equilibrium state was not invariant under rotations. We will have similar phenomena in field theory whenever the classical equilibrium state, the monopole, is not invariant under the unbroken symmetry group of the theory. How rich a "rotational spectrum" we have depends on how assymetric the monopole is, how many solutions we can generate by application of the symmetry group.

At the very minimum, the monopole solution is not translationally invariant, so we have at least a three-parameter family of time-

Table 2

The Quantum Monopole

Order of Approximation	Energy Eigenstate	Energy Eigenvalue
0	$\|\vec{r}\rangle$	V_0/f^2
1	$\|n_1,n_2\ldots\vec{r}\rangle$	$+\sum_i \omega_i(n_i+\frac{1}{2})$
2	$\|n_1,n_2\ldots\vec{P}\rangle$	$+\frac{f^2\vec{P}^2}{2V_0}+\ldots$

independent solutions, labeled by the position of the center of the monopole, $\vec{r}$. I have constructed Table 2 on the assumption this is the only degeneracy.

This is very much a transcription of Table 1. The small parameter 1/M, the coefficient of the kinetic energy in Eq. (5.7), has been replaced by the small parameter f^4, the coefficient of the kinetic energy in Eq. (5.6), and all the eigenvalues have been divided by f^2, because of the f^2 on the left in Eq. (5.6), but otherwise the right-hand columns in the two tables are almost identical.

Now let's go through the table in detail.

To zeroth order, the energy eigenstates are eigenstates of the field operators, with eigenvalues given by the classical solutions to the field equations. In equations,

$$\phi^a_{op}(\vec{x})|\vec{r}\rangle = \phi^a_{cl}(\vec{x}-\vec{r})|\vec{r}\rangle \ , \qquad (5.11)$$

where "op" indicates an operator and "cl" the classical solution. Of course, a similar equation holds for $\vec{A}^a$. V_0 is the value of the functional V at the monopole solution. We see that to leading order the monopole mass is proportional to $1/f^2$, something we already knew on other grounds.

In first order, we have a sum over normal modes, the eigenmodes of classical small vibrations about the monopole solution. Of course, we have a system with an infinite number of degrees of freedom, so we can have continuum eigenmodes as well as discrete ones; for these

the sum should be replaced by an integral. As usual when passing from particle mechanics to field theory, we reinterpret harmonic-oscillator excitation numbers as meson normal-mode occupation numbers. The state where all n's vanish is an isolated monopole; states with nonvanishing n's correspond to one or more mesons bound to the monopole (discrete eigenmode) or passing by the monopole (continuum eigenmode). To this order, there is no sign of meson-meson interactions in the energy because meson-meson interactions are of order f^2; Their effects are analogous to the vibrational-vibrational coupling in the molecule, and, like it, they are lurking behind the three dots in the second-order energy.

There is a first-order correction to the mass of the monopole, $\Sigma\frac{1}{2}\omega_i$. This sum is divergent, but, at least in a renormalizable theory, the difference between it and the corresponding sum for the vacuum state is finite and is a genuine correction to the classical monopole mass.

In second order, the degeneracy is lifted. Because the degeneracy is caused by translational invariance, not rotational invariance, the energy eigenstates are not angular-momentum eigenstates, as in Eq. (5.10), but linear-momentum eigenstates,

$$|n_1\ldots\vec{P}\rangle = \int \frac{d^3\vec{r}}{(2\pi)^{3/2}} e^{i\vec{P}\cdot\vec{r}} |n_1\ldots\vec{r}\rangle \ . \qquad (5.12)$$

By a fluke, we know the form of the second-order energy without having to do any computations; the expression in the table just comes from

$$\sqrt{\vec{P}^2+M^2} = M + \frac{f^2\vec{P}^2}{2V_0} + O(f^4) \ . \qquad (5.13)$$

Now let us relax our assumption that the only degeneracy is translational. There can be either further geometrical degeneracy or internal-symmetry degeneracy.

Geometrical degeneracy leads to a spectrum resembling that of molecular physics. If the monopole solution is not spherically symmetric, but does have an axis of rotational symmetry, the classical solutions are labeled by two polar angles, and in second order the

system develops a rigid-rotator spectrum, like the diatomic molecule. If the system has no axis of symmetry at all, three Euler angles are required, and the spectrum is a rigid-body spectrum, like that of the polyatomic molecule. Interesting variations on these phenomena can appear if the classical solution is invariant under some discrete subgroup of O(3). For example, if there is both an axis of rotational symmetry and an orthogonal plane of reflection symmetry, the odd angular momenta do not appear in the rotator spectrum.

Internal-symmetry degeneracy occurs if the classical solution is not invariant under H, the unbroken subgroup of G. As always, the second-order energy eigenstates are linear combinations of the eigenstates of the field operators. However, now the linear combinations transform as irreducible representations of H, rather than the translation or rotation group. The "rotational spectrum" lies in internal-symmetry space, and "rotational levels" carry H quantum numbers. I will call these states "isorotational levels". (Spin is to isospin as rotation is to isorotation.)

To compute the energies of the isorotational levels requires knowing the form of the kinetic energy operator acting on wave-functions restricted to the surface of minima of V. This is a generalization of the most straightforward way of solving the rigid rotator, by analyzing the angular part of the Laplace operator. We don't need to do the computation to see that the H-singlet state is always the lowest level; this state has a constant wave-function, and is always annihilated by any generalized Laplace operator, whatever its detailed form.[24]

As an example, let us consider the 't Hooft-Polyakov monopole. To avoid confusion caused by a spatially-dependent H, we will work in the string gauge defined in Eqs. (4.19-21); Φ points everywhere in the 3-direction, and H is everywhere the SO(2) subgroup of rotations about the 3-axis. As usual, we identify this with the group of electromagnetism. The fields in the theory consist of a neutral scalar, a neutral massless vector (the photon), and positively and

negatively charged massive vectors. The monopole solution is invariant under H only if the charged fields vanish everywhere. But they can not; if all charged fields vanish, the electromagnetic field obeys the free Maxwell equations, and a monopole field at large distances implies a gauge-invariant singularity at the origin. Note that this argument is independent of the details of the model; it is true for any magnetic monopole in any spontaneously broken gauge field theory.[25]

Thus the zeroth order eigenstates are labeled not just by the position of the center, r, but also by a complex number of modulus one, $\exp(i\alpha)$, the value of the charged field at some standard point. Under SO(2) rotations,

$$e^{-iQ_{op}\lambda}\,|e^{i\alpha},\vec{r}\rangle = |e^{i(\alpha+e\lambda)},\vec{r}\rangle\ , \qquad (5.14)$$

where I have normalized the electric charge operator, Q_{op}, such that the charged field has charge e. The isorotational levels are

$$|m,\vec{P}\rangle = \int_0^{2\pi}\frac{d\alpha}{\sqrt{2\pi}}\,e^{-im\alpha}\int\frac{d^3\vec{r}}{(2\pi)^{3/2}}\,e^{-i\vec{P}\cdot\vec{r}}\,|e^{i\alpha},\vec{r}\rangle\ , \qquad (5.15)$$

where m is an integer and I have suppressed the vibrational quantum numbers. These are charge eigenstates as well as momentum eigenstates,

$$Q_{op}|m,\vec{P}\rangle = m\,e\,|m,\vec{P}\rangle\ . \qquad (5.16)$$

The monopole has brought forth dyons.[26]

I emphasize that these are not fundamental dyons; they are inevitable, not optional, and their properties are computable, not adjustable. Neither are they bound states of a monopole and a charged meson; their excitation energies are proportional to f^2, while the meson mass is O(1). For the same reason, the dyons are stable, at least until very high m; they can not de-excite by emitting a charged meson.

Only the last of these statements is not true in general. In a theory with a hierarchy of symmetry breakdown and mass scales, f^2 times a large mass scale may still be much greater than the masses

of the light particles in the theory. If these light particles carry the proper quantum numbers, the dyons may all decay into the ground-state monopole.

5.2 The Witten Effect

There is a famous CP-violating term that can be added to the Lagrangian of a gauge field theory, the θ-term,

$$\mathscr{L}' = \frac{\theta}{32\pi^2}\,\epsilon^{\mu\nu\lambda\sigma}\,F^a_{\mu\nu}F^a_{\lambda\sigma}\ , \tag{5.17}$$

where θ is a real number, and the fields are normalized as in Sec. 3.1. This term is a total derivative, so it has no effect on the equations of motion; nevertheless, it has profound effects on the physics of the theory. How this comes about is part of the theory of instantons, which I have no intention of reviewing here. However, I will need one result of this theory: θ is an angle, that is to say, all physical phenomena are periodic functions of θ with period 2π.

Witten showed that for Abelian monopoles the θ-term has a striking influence on the dyon spectrum.[27] I will give here a derivation of the Witten effect which shows that it is independent of the dynamics of the heavy fields inside the monopole. The heavy fields are important only in that they insure that a monopole exists in the first place; everything that happens afterwards involves only the fields outside the monopole, the ordinary electromagnetic fields. Had I the courage, I could have discussed the Witten effect in Sec. 2.

Let me write out the θ-term, totally ignoring the heavy fields. That is to say, we replace $F^a_{\mu\nu}$ by a single field $F_{\mu\nu}$, where

$$F_{0i} = e\,E_i\ , \qquad F_{ij} = e\,\epsilon_{ijk}B_k\ , \tag{5.18}$$

and $\vec{E}$ and $\vec{B}$ are the conventionally normalized electric and magnetic fields. In terms of these,

$$\mathscr{L}' = \frac{\theta e^2}{4\pi^2}\,\vec{E}\cdot\vec{B}\ . \tag{5.19}$$

We now write these fields as ordinary electromagnetic fields (for simplicity assumed static) plus a monopole background:

$$\vec{E} = \vec{\nabla}\phi\ ,$$
$$\vec{B} = \vec{\nabla}\times\vec{A} + g\vec{r}/r^3\ . \tag{5.20}$$

where $\vec{A}$ and ϕ are the conventional vector and scalar potentials. (We will generalize to dyon backgrounds shortly.) We find

$$L' = \int d^3\vec{r}\,\mathscr{L}'$$
$$= \frac{e^2 g\theta}{\pi}\int d^3\vec{r}\,\phi(\vec{r})\delta^{(3)}(\vec{r})\ , \tag{5.21}$$

by integration by parts.

This is the standard coupling of the scalar potential to an electric charge of magnitude $e^2g\theta/\pi$ localized at the monopole. In the presence of the θ-term, magnetic charge induces electric charge.

We can understand this in a rough way. If we had added an $\vec{E}\cdot\vec{E}$ term to the Lagrange density, it would have represented a dielectric constant, a term that would make a given electric charge induce a (typically shielding) additional electric charge. A $\vec{B}\cdot\vec{B}$ term would have a similar effect on magnetic charge. Thus it is not surprising that an $\vec{E}\cdot\vec{B}$ term causes magnetic charge to induce electric charge.

The problem with this argument is that the effect is not reciprocal; if we add an electric monopole background, its contribution to L' vanishes upon integration by parts. However, this makes it easy to extend things to dyons. In the presence of the θ-term, the dyon charges are given by

$$Q = e(m + eg\,\theta/\pi)\ , \tag{5.22}$$

with m an integer. This violates CP, but not the quantization condition. (See the discussion after Eq. (2.22).)

Because eg is always a half-integer, when θ increases by 2π, the original series of charges is recreated, with each term moving 2eg places forward. With the methods used here, we can't say anything about the θ dependence of dyon energies, but Witten has looked into the interior of the monopole, and he reports that by the time a dyon has replaced another, its energy has become that appropriate to its new charge. In a word, θ is an angle; physics is a periodic function

of θ with period 2π. This is a charming result, considering that instantons never entered the argument.

5.3 A Little More About SU(5) Monopoles

In the earlier parts of these lectures, I've made some comments about the monopoles that occur in the SU(5) theory of Georgi and Glashow. We'll now look a little more closely at these objects.[28] There are two reasons for doing this. Firstly, the SU(5) theory is in itself interesting. It is the simplest of a family of grand unified theories, one of which might well describe reality, at least at energies below the Planck mass. Secondly, the theory is sufficiently rich in its structure to serve as a good example of the interplay of many of the fundamental ideas of monopole theory.

I'll begin by summarizing the relevant features of the theory. The theory is a gauge field theory with gauge group SU(5). This is a simple group, so there is only one coupling constant, f; f is on the order of e, the ordinary electromagnetic coupling. (It is not precisely e both because of Clebsch-Gordon coefficients and because of renormalization effects in going from large distances, where e is defined, to 10^{-28} cm, where monopoles live.)

All we will need to know about the scalar fields in the theory is that their interactions are such that there are two mass scales, or, equivalently, two scales of scalar vacuum expectation value.

The larger expectation value breaks SU(5) down to SU(3) $\otimes$ SU(2) $\otimes$ U(1). If we realize SU(5) as 5×5 matrices, SU(3) consists of transformations on the first three coordinates, SU(2) on the last two, and U(1) of diagonal matrices of the form

$$\mathrm{diag}(e^{2i\theta}, e^{2i\theta}, e^{2i\theta}, e^{3i\theta}, e^{-3i\theta}) \ . \tag{5.23}$$

None of these (except the identity) are in SU(3) $\otimes$ SU(2), so the group really is the direct product, not just something locally isomorphic to it. The SU(3) subgroup is identified with color; the SU(2) $\otimes$ U(1) subgroup with the group of the Weinberg-Salam model. As a result of this symmetry breakdown, twelve of the original twenty-four gauge

fields acquire masses on the order of 10^{14} GeV. (It is only at these high mass scales that the color gauge coupling is comparable in strength to the electroweak ones.) These superheavy mesons transform as the $(3,\bar{2}) \oplus (\bar{3},2)$ representation of SU(3) $\otimes$ SU(2).

The smaller vacuum expectation value breaks the group down further, to SU(3) (color) $\otimes$ U(1) (electromagnetism); as a result of this symmetry breakdown, three gauge fields acquire masses on the order of 10^2 GeV. The electromagnetic group is generated by

$$Q_{em} = -i\,\mathrm{diag}(\tfrac{1}{3}, \tfrac{1}{3}, \tfrac{1}{3}, -1, 0) \ , \tag{5.24}$$

where the generator has been normalized such that the proton has unit charge. The factors of $\frac{1}{3}$ are a sign that the theory contains fractionally charged particles: quarks, of course, but also the superheavy vector mesons. Because of these factors,

$$\exp(2\pi Q_{em}) \neq 1 \ . \tag{5.25}$$

We have to go three times as far,

$$\exp(6\pi Q_{em}) = 1 \ . \tag{5.26}$$

$\exp(2\pi Q_{em})$ is an SU(3) matrix though, so really the group H is not SU(3) $\otimes$ U(1) but SU(3)$\otimes$ U(1)/Z_3. This is just a fancy way of saying that only particles of non-zero triality have fractional charge. (*cf.* example (e) at the end of Sec. 3.4.)

In addition to these gauge symmetries, the theory possesses a continuous global internal symmetry which, when all the dust of symmetry breakdown has settled, leads to the conservation of the difference of baryon and lepton number, B - L. (B and L are not separately conserved; the theory famously predicts proton decay.)

We already know some things about this theory. We know that it has monopoles (Sec. 4.2), that the masses of these monopoles are of order 10^{16} GeV and their cores of order 10^{-28} cm in size (Sec. 4.4), and that they have a spectrum of isorotational excitations, dyons, with a level spacing on the order of 10^{12} GeV (Sec. 5.1).

Now let us look at the monopoles in more detail. At large

distances, the monopole field must be of the standard form,

$$\vec{A} = Q \vec{A}_D \ , \tag{5.27}$$

where Q is a generator of SU(3) ⊗ U(1). One's first thought is that because the theory contains fractionally charged particles, $g = \frac{1}{2} e$ is not allowed; $g = \frac{3}{2} e$ is needed. Indeed, $Q = Q_{em}/2$ does not satisfy the quantization condition;

$$Q = 3Q_{em} / 2 \tag{5.28a}$$

is needed, as we see from Eqs. (5.25) and (5.26).

(Let me raise and lay a spectre. All fractionally charged particles are confined; for example, we can not separate the components of quark-antiquark pair by more than roughly 10^{-13} cm. In using confined particles to derive the quantization condition, are we not making an error? No, we are not. The Dirac string is infinitely thin and infinitely long. We can imagine bringing the pair to within 10^{-14} cm of the string, a thousand light years from the monopole, and diffracting the quark around the string while holding the antiquark fixed. Confinement is irrelevant to this point.)

However, there is a second possibility,[29]

$$Q = (Q_{em} + Q_Y) / 2 \tag{5.28b}$$

where Q_Y is the generator of color hypercharge,

$$Q_Y = - i \, \mathrm{diag}(\tfrac{2}{3}, -\tfrac{1}{3}, -\tfrac{1}{3}, 0, 0) \ . \tag{5.29}$$

Thus,

$$Q = - i \, \mathrm{diag}(1, 0, 0, -1, 0) / 2 \ , \tag{5.30}$$

and the quantization condition is satisfied.

This is a combination of an electromagnetic monopole and a chromomagnetic monopole. Colorless particles, like leptons and hadrons, see only the electromagnetic field, and find $g = \frac{1}{2} e$. However, because all fractionally charged particles are colored, a fractionally charged particle sees both the electromagnetic and the chromomagnetic field, and the combined effect of these two fields renders the string undetectable.

Of course (5.28a) and (5.28b) are not the only monopole fields there are. Up to a gauge transformation, the solution of the quantization condition and the non-Abelian stability condition is

$$Q = - i \, \mathrm{diag}(r, s, s, -r-2s, 0) / 2 \ , \tag{5.31}$$

where r and s are integers such that

$$r-s = 0, \pm 1 \ . \tag{5.32}$$

Because SU(3) is simply connected, the only topological charge is the Abelian magnetic charge, r+2s.

We can estimate the energies of these configurations by the method of Sec. 4.4. The energy stored in the field outside the core is proportional to

$$- \mathrm{Tr}\, Q^2 = s^2 + \tfrac{1}{2}(r+s)^2 \ . \tag{5.33}$$

The nontrivial monopole of lowest energy is r = 1, s = 0, the combination monopole, Eq. (5.28b). If we take Eq. (5.33) dead seriously as an estimate of energy, it is easy to show that every other nontrivial monopole has more than enough energy to decay into an appropriate number of combination monopoles. For example, the pure electromagnetic monopole, Eq. (5.28a), with r = s = 1, has six times the energy of a combination monopole, and is thus unstable to decay into three such objects, the minimum number needed to conserve topological charge.

(Another spectre: In our energy estimate we neglected factors of coupling constants. By treating the color coupling constant as if it were the same as the electromagnetic coupling constant, are we not making an error? No, we are not. The two couplings are indeed very different at large distances, but they are the same at 10^{-28} cm, near the monopole core, and this is the region that dominates the energy integral.)

Dokos and Tomaras[28] have constructed a combination monopole that is a time-independent solution of the field equations; their method is a transposition of the 't Hooft-Polyakov construction to an appropriate SU(2) subgroup of SU(5).

For our purposes, all we need to know about this solution is

its degeneracy, for this suffices to determine the spectrum of isorotational levels. It turns out that one of the superheavy vector fields, X^a (a = 1, 2, 3) is nonzero in the Dokos-Tomares solution. This field is a color 3, with electric charge $-^4/_3$ and B - L = $-^2/_3$. Further, once the value of this field at some standard point is given, the solution is uniquely determined.

Since any complex 3-vector can be turned into any other of the same magnitude by an SU(3) matrix, the manifold of solutions is in one-to-one correspondence with the set of all unit complex three-vectors,

$$X^a \bar{X}_a = 1 \ , \tag{5.34}$$

and the problem of constructing the isorotational levels is the problem of constructing the functions of these vectors that transform as irreducible representations of the symmetry group.

This is easily done. A complete set of functions on the manifold consists of all monomials in X and $\bar{X}$. We can write these as tensors,

$$X^{a_1 \cdots a_n}_{b_1 \cdots b_m} = X^{a_1} \cdots X^{a_n} \bar{X}_{b_1} \cdots \bar{X}_{b_n} \ . \tag{5.35}$$

These tensors are almost the objects that form the basis space for the irreducible representation of SU(3) called (n,m). The only difference is that the irreducible tensors are traceless. However, it is straightforward to subtract the trace from these expressions. Because of Eq. (5.34), the tensors of lower rank we obtain by taking the traces are not new functions, just objects already constructed as monomials of lower degree. To compute the electric charge and (B-L) assignments of these tensors is trivial; we just get n-m times the contribution of a single X.

Thus: The isorotational levels transform as the representation (n,m) of color SU(3), with each representation occurring once and only once. These levels have an electric charge of 4(m-n)/3 and a B-L of 2(m-n)/3. Note the coupling of charge excitation to color excitation. These are chromodyons.

Let us look more closely at the physics of chromodyons. To

begin with, I will ignore the existence of fermions.

Given two levels with the same value of m-n, the higher can always decay into the lower by the emission of massless color gluons. (The fact that the gluons are confined is irrelevant. They are confined at 10^{-13} cm and the decay takes place at 10^{-28} cm; worrying about the effects of gluon confinement on the decay of a chromodyon is like worrying about the effects of the walls of the laboratory on the decay of a radioactive nucleus.) Thus we expect only one stable level for a given m-n. These levels can not decay by the emission of superheavy vector mesons, because their excitation energy is too low; they can not decay by the emission of color gluons because these do not carry charge; they can not decay by the emission of Weinberg-Salam gauge mesons because these do not carry color; they are stable.

Although they are stable, they are colored (except for the ground state monopole). Thus an (n,m) chromodyon and an (m,n) chromodyon will form a colorless pair, bound together by the confining force. The fate of this pair depends on the interaction between its components at short distances. (Although not so short that the cores overlap.)

This is dominated by the magnetic interaction. If the components have opposite Abelian magnetic charge, that is to say, if they are excitations of monopole and antimonopole, the magnetic force is attractive and the components will annihilate each other.

A short computation is needed if the components have the same Abelian charge; we have to worry about the competition between Abelian repulsion and non-Abelian attraction, as explained in Sec. 3.6. If we represent the field of one monopole by a matrix Q_1 and the other by a matrix Q_2, the magnetic interaction energy is proportional to $-\mathrm{Tr}\, Q_1 Q_2 / r_{12}$. We can always choose our gauge so

$$Q_1 = -i\,\mathrm{diag}(1, 0, 0, -1, 0)/2 \ . \tag{5.36}$$

Q_2 can be any matrix obtained from this by permutation on the first three entries (the color entries). We minimize the interaction energy by choosing

$$Q_2 = -i\,\mathrm{diag}(0, 1, 0, -1, 0)/2 \ , \qquad (5.37)$$

but even at the minimum the energy is still positive. Abelian repulsion wins over non-Abelian attraction.

Thus we have a system composed of two particles with an attractive interaction between them at large distances and a repulsive interaction at short distances. This is the diatomic molecule again, not as an analogy this time but as the real thing, with the full panoply of vibrational and rotational levels, though on a somewhat larger energy scale than is usual in molecular physics.

But all of this is mere fantasy. In the real world there are light fermions, quarks and leptons, and all the chromodyons can decay to the ground-state monopole by emitting quark-lepton pairs. Pity.

5.4 Renormalization of Abelian Magnetic Charge

Some years ago, in a famous series of papers, Julian Schwinger generalized quantum electrodynamics to include fundamental magnetic monopoles.[30] In the course of this work, he investigated the renormalization of magnetic charge. Even though fundamental monopoles are not our primary interest, I'll review Schwinger's results here for later comparison with ours.

Schwinger found that a necessary condition for consistent quantization of the theory was

$$e_0 g_0 = n_0/2 \ , \qquad (5.38)$$

where e_0 and g_0 are the bare coupling constants, the parameters that appear in the Lagrangian of the theory, and n_0 is an integer. (In fact, at first Schwinger argued that n_0 had to be even, but this point is irrelevant to the issues at hand, and I'll ignore it here.) The standard arguments for the Dirac quantization condition have nothing to do with the fundamental or composite character of the monopoles; they tell us that

$$eg = n/2 \ , \qquad (5.39)$$

where e and g are the physical coupling constants and n is an integer.

The analysis of electric charge renormalization proceeds much as in ordinary electrodynamics. Because there are magnetic currents as well as electric ones, the propagator for the unrenormalized electromagnetic field is expressed as an integral over three spectral weight functions rather than one; however, there is still only one photon, and only one single-photon pole term. Thus there is no problem in following the usual path, defining Z_3 to be the coefficient of this term and showing that

$$e = Z_3^{1/2} e_0 \ . \qquad (5.40)$$

As usual, the spectral representation implies that if the theory is nontrivial, Z_3 is strictly less than one.

The analysis of magnetic charge renormalization is trivially obtained by performing the duality rotation that exchanges E and B. The spectral weight functions change places but the photon remains the photon. Thus Z_3 is unchanged and

$$g = Z_3^{1/2} g_0 \ . \qquad (5.41)$$

Hence,

$$eg = Z_3 e_0 g_0 \ . \qquad (5.42)$$

This is the origin of Schwinger's statement that Z_3 is a rational number.

This is for fundamental monopoles. Our interest, though, is in composite monopoles, like the 't Hooft-Polyakov object. Is Eq. (5.42) still true for these? Clearly not, but neither is it false; it is meaningless. In a theory of composite monopoles, e_0 still appears as a parameter in the Lagrangian, but g_0 is nowhere to be found.

Nevertheless, all is not lost. If we define e_λ as the quantity that gives the strength of the interaction between an electron and a tiny test charge at distance λ, then e_0 is the limit of e_λ as λ goes to zero. This suggests that we define g_λ in a similar way, as the parameter that gives the strength of the interaction between a monopole and a tiny test current loop at distance λ. In this case, though, we can not send λ to zero; once λ is smaller than the size of the monopole core there is no unambiguous way to separate electromagnetism from the larger non-Abelian structure in which it is em-

bedded, no unambiguous way to define the current loop. However, until we get to the core there is no problem. Thus we can ask how e_λ and g_λ, for λ larger than the core radius, are related to e and g, their large-distance limits. This is not the same question as that addressed by Schwinger, but it is an interesting one; there is plenty of room for large renormalization effects between infinite distance and the monopole core.

However, there is no room for renormalization effects caused by virtual monopoles. For weak coupling, monopole Compton wavelengths are much smaller than monopole geometrical sizes. (That is to say, monopoles are much heavier than massive vector mesons.) Thus, at the distance scales we are working at, between λ and infinity, the effects of virtual monopoles are negligible. (I emphasize that "weak coupling" here does not mean merely "to lowest order in e". If we think in terms of an expansion in powers of $\hbar$, the effects of virtual monopoles are barrier-penetration effects and are exponentially suppressed.) Thus, for our purposes, it is completely legitimate to replace the full theory by a simplified one in which the only dynamical variables are the electromagnetic field and the fields of ordinary charged particles.

For notational simplicity, we will restrict ourselves to the case in which the only charged field is a Dirac electron; the generalization is trivial. The theory we will study is defined by

$$\begin{aligned}\mathscr{L} = &-\frac{1}{4}(\partial_\mu A_\nu - \partial_\nu A_\mu)^2 + \frac{1}{2\alpha}(\partial_\lambda A^\lambda)^2 \\ &+ \bar{\psi}(i\not{\partial} - e_\lambda \not{A} - e_\lambda g_\lambda \not{A}_D - m_0)\psi \\ &+ e_\lambda J_\mu (A^\mu + g_\lambda A_D^\mu) \ . \end{aligned} \tag{5.43}$$

This expression requires some explanation. Firstly, as to notation: All fields are unrenormalized, α is the usual gauge-fixing parameter, $A_D^\mu = (0, \vec{A}_D)$ is the standard Dirac monopole potential, and J^μ is an external c-number conserved current,

$$\partial_\mu J^\mu = 0 \ , \tag{5.44}$$

which we will use to construct test charges and current loops. Secondly, as to physics: This is supposed to contain only the degrees of freedom which are relevant at distances larger than λ. Thus, although it is not indicated explicitly, the theory is supposed to be cut off in some gauge-invariant way (say, with regulator fields) at λ. It is for this reason that the places of the bare coupling constants are taken by e_λ and g_λ, and also why the monopole is replaced by an external Dirac potential.

I emphasize that Eq. (5.43) is not intended to give an exact description of the theory at distances greater than λ. It is certainly possible to construct an effective Lagrangian that would give such a description, but it would be much more complicated, full of nonlocal and nonpolynomial interactions. Equation (5.43) is intended only to give a simple model of the relevant physics.

Arguments which by now should be excessively familiar show that the string singularity in the monopole potential is undetectable only if $e_\lambda g_\lambda$ is a half-integer. I will therefore assume that this is the case. It is not at all clear at this stage that this implies that eg is a half-integer, but we have a well-defined Lagrangian, so let's just go ahead and calculate.

To warm up, let us make sure that our test apparatus is properly calibrated by computing the effects of the external current far from the monopole, or, equivalently, with g_λ set equal to zero. In this case, the vacuum-to-vacuum transition matrix element is

$$\langle 0|S|0\rangle = 1 - \frac{e_\lambda^2}{2}\int d^4x\, d^4y\, J^\mu(x) J^\nu(y) T\langle 0|A_\mu(x) A_\nu(y)|0\rangle + O(J^4) \ . \tag{5.45}$$

I emphasize that this is not an (illegitimate) expansion in powers of e_λ; it is an expansion in powers of J^μ. There is nothing wrong with this; the interactions with J^μ are gauge-invariant and thus incapable of detecting the Dirac string whatever the magnitude of J^μ, so long as it is conserved. (This is just another way of saying that there is no quantization condition for classical charged particles.)

The relevant object in Eq. (5.45) is the propagator for the un-

renormalized field,

$$T\langle 0|A_\mu(x)A_\nu(y)|0\rangle = Z_3^\lambda D^0_{\mu\nu}(x-y) + \dots , \tag{5.46}$$

where $D^0_{\mu\nu}$ is the free propagator, the triple dots indicate terms that fall off at large distances more rapidly than the term displayed, and the superscript is on Z_3 to remind you of its cutoff-dependence. Thus, two distantly separated current elements interact as they would in free electromagnetism, with interaction strength given by

$$e_\lambda^2 Z_3^\lambda = e^2 . \tag{5.47}$$

This is famously the right result.

Let us now return to nonzero g_λ. There is now a linear term in the transition matrix element,

$$\langle 0|S|0\rangle = 1 - i e_\lambda \int d^4x\, J^\mu(x)[g_\lambda A^D_\mu(x) + \langle 0|A_\mu(x)|0\rangle] + O(J^2) . \tag{5.48}$$

I emphasize that $|0\rangle$ is now the ground state of the theory computed to zeroth order in J^μ but to all orders in the external monopole field. We wish to calculate eg in the same way we calculated e^2 in the preceding paragraph. This can be done if the total integral in Eq. (5.48) has the same form as its first term when the support of the current is far from the monopole; we can then identify the coefficient of the total expression with eg. In equations,

$$\langle 0|S|0\rangle - 1 \xrightarrow[\text{large distance}]{?} -ieg\int d^4x\, J^\mu(x)A^D_\mu(x) + O(J^2) . \tag{5.49}$$

I've put a question mark here because we don't yet know that things will have the right form in this limit. (Let's hope they do; if they don't, we're in real trouble.)

The equation of motion for A_μ implies that

$$\langle 0|A_\mu(x)|0\rangle = -ie_\lambda\int d^4y\, D^0_{\mu\nu}(x-y)\langle 0|J^\nu(y)|0\rangle . \tag{5.50}$$

The object on the right is the matrix element of a gauge-invariant operator; thus, despite the Dirac string, it is rotationally invariant and CP invariant. (Note that this would not be true if we had foolishly attempted to expand things in powers of the monopole field.)

By rotational invariance,

$$\langle 0|\bar\psi\gamma^0\psi|0\rangle = f(r^2) , \tag{5.51}$$

and

$$\langle 0|\bar\psi\gamma^i\psi|0\rangle = r^i g(r^2) , \tag{5.52}$$

for some functions f and g. CP invariance implies f vanishes. The conservation equation

$$\partial_\mu\langle 0|\bar\psi\gamma^\mu\psi|0\rangle = 0 , \tag{5.53}$$

implies g vanishes. Thus Eq. (5.49) is satisfied trivially, and

$$eg = e_\lambda g_\lambda . \tag{5.54}$$

This is in striking contrast to Eq. (5.42), but, as I have explained, the contradiction is only apparent; despite the appearance of similar symbols, the two equations are answers to different questions in different theories.

Equation (5.54) is very satisfying. Monopole theory says that eg must be a half integer, but a modern field theorist would ask, "eg defined at what distance scale?" The answer is, "At any distance scale, from infinity down to the monopole core. It doesn't matter."

5.5 The Effects of Confinement on Non-Abelian Magnetic Charge

We have analyzed the renormalization of Abelian magnetic charge, computed the electromagnetic monopole field at large distances in terms of the electromagnetic monopole field just outside the monopole core. To complete the analysis, we should now extend the argument to non-Abelian magnetic charge, and compute, for example, the chromomagnetic monopole field at large distances from the grand unified monopoles of Sec. 5.3. But such a computation would be folly. We know very well that there is no chromomagnetic monopole field at large distances; there is no chromodynamic field of any kind proportional to any inverse power of distance. Because of confinement, all chromodynamic forces fall off exponentially with distance, with a coefficient given by the lightest hadron mass. (If we ignore fermions, as we have been doing, this would be the glueball mass.)

This leads to an apparent paradox. For grand unified monopoles,

the simultaneous existence of chromomagnetic and electromagnetic monopoles was necessary to keep the Dirac string undetectable; when we diffracted a fractionally charged particle around the string, the chromomagnetic phase factor was needed to cancel the electromagnetic one. (I remind you that the fact that fractionally charged particles are themselves confined is irrelevant to this point, as we showed in Sec. 5.3.) But at large distances the chromomagnetic monopole field disappears, while the electromagnetic one does not. Does this mean that at large distances the Dirac string is observable?

To ask this question is to be carried away by the momentum of my own rhetoric. Of course, the string can not become observable; confinement does not spoil gauge invariance. Nevertheless, we would like to understand how effects that fall off exponentially with distance are nevertheless able to produce a distance-independent phase factor. The remainder of this section is devoted to this problem.

Although the relevant group for grand unified monopoles is $SU(3) \otimes U(1)/Z_3$, to keep things simple I will work here with SO(3), the smallest non-Abelian group. The extension is trivial. I shall first give general arguments and then back them up with explicit computations in a highly simplified but exactly soluble two-dimensional model.

The Lubkin construction associates a path in group space, $g(\tau)$, with every sphere in ordinary space; the homotopy class of this path tells us whether the sphere contains a monopole. Although the homotopy class is gauge-invariant, the path itself is not. When investigating tricky questions in gauge field theory, it's wise to work as much as possible with gauge-invariant entities. Thus I will work with the Wilson loop factor,

$$W(\tau) = \tfrac{1}{2} \operatorname{Tr} g(\tau) \ . \qquad (5.55)$$

Although the loop factor is independent of the gauge, it does depend on the particular matrix representation we use to realize the group. For the problem at hand, it will turn out to be advantageous to choose the spin-$\frac{1}{2}$ representation of SO(3), that is to say,

to choose g to be an element of SU(2). This is a double-valued representation, so there is a binary ambiguity in computing the loop factor. We will eliminate this ambiguity by adopting the convention that $g(0) = 1$, so $W(0) = 1$. Of course, we could have chosen the opposite convention, $g(0) = -1$, in which case $W(\tau)$ would everywhere have been replaced by $-W(\tau)$.

The behavior of W tells us whether our sphere contains a monopole. If $g(\tau)$ is in the trivial homotopy class, $W(1) = W(0) = 1$. Figure 6a shows this behavior for a typical field of this class, the zero field. If $g(\tau)$ is in the nontrivial homotopy class, $W(1) = -W(0) = -1$. Figure 6b shows this behavior for a typical field of this class, the GNO monopole field. Phrased in terms of W, our problem is to understand how effects that fall off exponentially with distance are nevertheless able to produce a distance-independent change in W.

But this formulation is not quite right. Confinement is a quantum effect, and in quantum theory, W is an operator. Except in the leading semi-classical approximation (which does not display confinement) the one-monopole state is not an eigenstate of W. The object we must study is $\langle W \rangle$, the expectation value of W.

Of course, for a small sphere, one whose radius is much less than the confinement length, the semi-classical approximation is valid, and Fig. 6 gives accurate plots of $\langle W \rangle$ for the vacuum and the one-monopole state. But for a large sphere, one whose radius is much greater than the confinement length, even the graph for the vacuum, Fig. 6a, is wrong. As our loop sweeps around a large sphere, it necessarily becomes, at the half-way point, a loop of large area, and there-

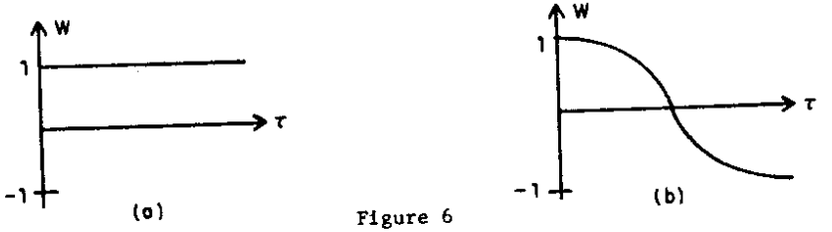

Figure 6

fore <W> becomes very small, because of Wilson's area law. This is sketched in Fig. 7a. The sketch is very much not to scale. <W(½)> is $O(\exp-\pi K r^2)$, where K is the string tension. This is so small that if the graph were drawn to scale you couldn't see that <W(½)> wasn't zero.

Our problem is solved. The monopole does not need to produce a distance-independent change in <W>. The area law reduces <W> from 1 to a small positive value. All the monopole field has to do is make a small additional change to bring <W> to a small negative value. The area law then brings <W> back to -1. This is sketched in Fig. 7b. I emphasize that the right side of this graph represents the same physics as the left side; the evolution of negative values of <W> can be deduced from that of positive values by a change of convention.

It's amusing to note that if we had defined W using an integer-spin representation of SO(3), we would not have had the area law, but neither would we have had a test for monopoles.

This concludes the general arguments. Now, as promised, I'll do some explicit computations in a two-dimensional Euclidean gauge theory.

Two-dimensional gauge theories have two big advantages. One is that they are easy to work with. This is basically because they have very few degrees of freedom; in Minkowski space, there is no radiation field, just a Coulomb field. The other is that they trivially display confinement. In one spatial dimension, even ordinary Abelian electrodynamics yields a linear potential. For our purposes, they have one big disadvantage; they don't possess monopoles. I will take

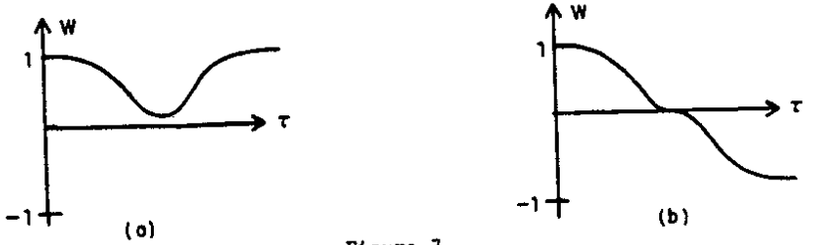

Figure 7

care of this by studying SO(3) gauge field theory not on flat two-space but on a two-dimensional sphere. Ordinary two-dimensional Euclidean field theory can be thought of as obtained from four-dimensional theory by holding z and t fixed, leaving x and y to vary. Likewise, the spherical two-dimensional theory can be obtained by holding t and r fixed, leaving θ and ϕ to vary. The monopole is in the part of space we have discarded, near r = 0, but its effects are still felt on the sphere we have retained.

Now let's make this more precise. We are dealing with an SO(3) gauge field theory on a sphere. Just as in Sec. 3, we will go to string gauge, so A_θ vanishes everywhere. In general, this gauge introduces a Dirac string singularity at the south pole, a nonzero $A_\phi(\pi,\phi)$. We define $g(\phi)$, an element of SU(2), by

$$\frac{dg}{d\phi} = - A_\phi(\pi,\phi) g(\phi) \ , \tag{5.56}$$

and

$$g(0) = 1 \ . \tag{5.57}$$

Then, for the string singularity to be just a gauge artifact,

$$g(2\pi) = \pm 1 \ . \tag{5.58}$$

The plus sign corresponds to no monopole (inside the sphere), the minus sign to a monopole. If we choose our Lubkin loops as in Fig. 4,

$$W(\tau) = \tfrac{1}{2} \operatorname{Tr} g(2\pi\tau) \ . \tag{5.59}$$

The Euclidean action, S_E, is the conventional one, restricted to the sphere,

$$\begin{aligned} S_E &= - \frac{1}{f^2} \int \operatorname{Tr} F_{\theta\phi} F^{\theta\phi} \sqrt{g}\, d\theta\, d\phi \\ &= - \frac{1}{r^2 f^2} \int \operatorname{Tr} (\partial_\theta A_\phi)^2 \frac{d\theta\, d\phi}{\sin\theta} \ , \end{aligned} \tag{5.60}$$

by Eqs. (3.43) and (3.45).

As always, the expectation value of a gauge-invariant observable, $\mathcal{O}$, is given by the ratio of two functional integrals,

$$\langle \mathcal{O} \rangle = \frac{\int (d A_\phi) \mathcal{O}\, e^{-S_E}}{\int (d A_\phi)\, e^{-S_E}} \ . \tag{5.61}$$

We will compute $\langle \mathcal{O} \rangle$ in the absence (presence) of a monopole by integrating over all gauge field configurations with $g(2\pi) = +1$ (-1). This guarantees from the very beginning that $\langle W \rangle$ will have the appropriate behavior, $\langle W(1) \rangle = \pm 1$.

We will begin with the denominator in Eq. (5.61); the extension to the numerator will be straightforward. We write A_ϕ as a sum,

$$A_\phi = \hat{A}_\phi + \tfrac{1}{2}(1-\cos\theta)A_\phi(\pi,\phi) \ . \tag{5.62}$$

Thus,

$$\hat{A}_\phi(0,\theta) = \hat{A}_\phi(\pi,\theta) = 0 \ . \tag{5.63}$$

This separation diagonalizes S_E:

$$S_E = \hat{S} + S_1 \ , \tag{5.64}$$

where

$$\hat{S} = \frac{\mathrm{Tr}}{r^2 f^2}\int d\theta\, d\phi\, \frac{(\partial_\theta \hat{A}_\phi)^2}{\sin\theta} \ , \tag{5.65}$$

and

$$\begin{aligned} S_1 &= -\frac{\mathrm{Tr}}{2r^2 f^2}\int_0^{2\pi} d\phi [A_\phi(\pi,\phi)]^2 \\ &= -\frac{\mathrm{Tr}}{2r^2 f^2}\int_0^{2\pi} d\phi [g^{-1}\, dg/d\phi]^2 \ . \end{aligned} \tag{5.66}$$

Our theory has been revealed as the sum of two independent systems. One, described by $\hat{S}$, is a two-dimensional field theory, but it is a trivial one; the dynamical variables obey linear boundary conditions, and the action is a quadratic functional of these variables. The other, described by S_1, is a nontrivial system, but it is a one-dimensional one. That is to say, it is not a Euclidean field theory at all, but simply the imaginary-time version of an ordinary mechanical system, with ϕ the imaginary time.

In fact, it is a very well-studied mechanical system. The states of the system are labeled by elements of the rotation group, that is to say, by sets of three Euler angles. The system is a rigid body with one point fixed, or more properly, since g is an element of SU(2) and not SO(3), it is the double covering of a rigid body; half-odd-integral angular momenta are allowed as well as integral ones.

If we consider motions near $g = 1$,

$$g = 1 - i\varepsilon^a\sigma^a/2 + O(\varepsilon^2) \ , \tag{5.67}$$

then

$$S_1 = \frac{1}{4r^2f^2}\int \frac{d\varepsilon^a}{d\phi}\,\frac{d\varepsilon^a}{d\phi}\,d\phi + O(\varepsilon^3) \ , \tag{5.68}$$

from which we see that the rigid body is isotropic; it has equal principal moments of inertia,

$$I_1 = I_2 = I_3 \equiv I = 1/(2r^2f^2) \ . \tag{5.69}$$

This enables us to analyze the system totally for all values of r. However, the answer is especially simple for very small r and very large r, so I will restrict myself to these two cases here.

For very small r, $f^2r^2 \ll 1$, the functional integral is dominated by the stationary points of the action, the solutions of the classical equations of motion. For a rigid body, these are steady rotations about a fixed axis, for example, the positive 3-axis,

$$g = \exp(-\tfrac{i}{2}\omega\sigma_3\,\phi) \ , \tag{5.70}$$

with ω an arbitrary non-negative constant, the angular velocity. (It's a good thing the moments of inertia turned out to be all equal; otherwise we'd be worrying about force-free precession.) Of these motions, the dominant one is the one of minimum action, that is to say, minimum ω, consistent with the boundary conditions. In the absence of a monopole, the boundary conditions are $g(0) = g(2\pi) = 1$. The dominant motion is $\omega = 0$, and the corresponding vector potential is $A_\phi = 0$. In the presence of a monopole, the boundary conditions are $g(0) = 1$, $g(2\pi) = -1$. The dominant motion is $\omega = 1$, and the corresponding vector potential is

$$A_\phi = -\tfrac{1}{4}\sigma_3(1-\cos\theta) \ , \tag{5.71}$$

the GNO monopole field. At small distances, quantum field theory looks like classical physics, just as it should.

To study large r, we need the energy eigenstates and eigenvalues for the (doubly covered) isotropic rigid body. These are discussed

in standard texts,[21] so I will merely give the results here. The eigenstates are of the form $|j,m,m'\rangle$, where $j = 0, \frac{1}{2}, 1 \ldots$, and m and m' individually range from $-(2j+1)$ to $2j+1$ by unit steps. The eigenvalues are given by

$$H|j,m,m'\rangle = E_j|j,m,m'\rangle \ , \tag{5.72}$$

where

$$E_j = j(j+1)/2I \ ,$$

and I is the moment of inertia. In any one of these states, the amplitude for finding the system in the configuration labeled by the group element g is

$$\langle g|j,m,m'\rangle = (2j+1)^{\frac{1}{2}} D^{(j)}_{mm'}(g) \ , \tag{5.73}$$

where D is the usual representation matrix.

We are now ready to go. The contribution of the rigid body to the functional integral is

$$\int (dg) e^{-S_1} = \langle g = \pm 1|e^{-2\pi H}|g = 1\rangle \ , \tag{5.74}$$

by Feynman's path-integral formula. Inserting a complete set of energy eigenstates,

$$\begin{aligned}\langle g = \pm 1|e^{-2\pi H}|g = 1\rangle &= \sum_{mm'j} (2j+1) D^{(j)}_{mm'}(\pm 1) D^{(j)*}_{mm'}(1) e^{-2\pi E_j} \\ &= \sum_j (2j+1)^2 (\pm 1)^{2j} e^{-\pi j(j+1)/I} \ . \end{aligned} \tag{5.75}$$

This is exact. For large r, $f^2r^2 \gg 1$, this expression is dominated by the first term in the series, $j = 0$, which is totally insensitive to the presence or absence of the monopole. To see the monopole at all, we have to retain the second term, $j = \frac{1}{2}$. We find

$$\int (dg) e^{-S_1} = 1 \pm 4e^{-3\pi f^2 r^2/2} \ . \tag{5.76}$$

The effect is miniscule.

Of course, this is just the denominator of the equation for the expectation value of an operator, Eq. (5.61), but similar reasoning applies to the total expression. A particularly easy case to work out is the expectation value of the action itself,

$$\begin{aligned}\langle S_E\rangle &= \langle \hat{S}\rangle + \langle S_1\rangle \\ &= \langle \hat{S}\rangle + f^2 \frac{d}{df^2} \ln \int (dg) e^{-S_1} \\ &= \langle \hat{S}\rangle \mp 6\pi f^2 r^2 \ e^{-3\pi f^2 r^2/2} \ , \end{aligned} \tag{5.77}$$

where, as before, I've dropped terms exponentially small compared to the terms retained. Another easy computation is

$$\langle W(\tau)\rangle = e^{-3\pi f^2 r^2 \tau/2} \pm e^{-3\pi f^2 r^2 (1-\tau)/2} \ , \tag{5.78}$$

where again I've retained only the leading terms. (Here I've left the details of the calculation as an exercise.) This expression has a very simple physical interpretation; it is the sum of the area factors for the two areas into which the loop divides the sphere.

All of this is in perfect agreement with our general reasoning. At small distances, everything looks like classical physics; at large distances, all effects of the monopole are miniscule, but nevertheless the monopole makes $\langle W\rangle$ go from +1 to -1.

Gravity theorists say, "A black hole has no hair." What this means is that a black hole has limited hair; the only things that stick out of a black hole are massless gauge fields associated with strictly conserved quantities, the gravitational fields associated with total energy and angular momentum, and the electromagnetic fields associated with total electric and magnetic charge.

Chromodynamic forces are short-range forces, because of confinement. Nevertheless, topological charge is strictly conserved, and can be computed from chromodynamic effects at arbitrarily large distances. These two statements would be in contradiction in classical physics, but are perfectly compatible in quantum mechanics, as we have seen. Problem for the student: Can a black hole have colored hair?

Footnotes and References

1. P. A. M. Dirac, Proc. Roy. Soc. (London) Ser. A, 133, 60 (1931).
2. G. 't Hooft, Nucl. Phys. B79, 276 (1974). A. M. Polyakov, JETP

Lett. 20, 194 (1974).

3. J. Preskill, Phys. Rev. Lett. 43, 1365 (1979).
4. S. Coleman, "Classical Lumps and Their Quantum Descendants", in *New Phenomena in Subnuclear Physics*, edited by A. Zichichi (Plenum, 1977).
5. Ref. 4 has a more extensive bibliography. See also the reviews of P. Goddard and D. Olive, Rep. Prog. Phys. 41, 1357 (1978), and E. Amaldi and N. Cabibbo, in *Aspects of Quantum Theory*, edited by A. Salam and E. Wigner (Cambridge Univ. Press, 1972).
6. Note that there is no 4π in the definition of magnetic charge.
7. Y. Aharonov and D. Bohm, Phys. Rev. 115, 485 (1959).
8. This is a truism for non-Abelian gauge theories, where everyone thinks of gauge transformations as labeled by functions from space-time into the gauge group. For the Abelian theory at hand, the gauge group is U(1), and the function into the group is $\exp(-ieX)$.
9. T. T. Wu and C. N. Yang, Nucl. Phys. B107, 365 (1976).
10. This problem was solved very early on by I. Tamm, Z. Phys. 71, 141 (1931), and my results are the same as his, although my method is somewhat different.
11. These commutators are not so innocuous as they seem. If we attempt to verify the Jacobi identity for three D's, we obtain, instead of zero, a term proportional to $\delta^{(3)}(r)$. This is no real problem, for two reasons. Firstly, we don't believe the monopole field all the way down to the origin; it's just the long-range part of something that is more complicated at short distances. (In Sec. 4 we'll see what that something is.) Secondly, even if we believe the field all the way down to the origin, there is, as we shall see, a centrifugal barrier in *all* partial waves that keeps the particle away from the origin.
12. The analysis given here follows that of A. Goldhaber, Phys. Rev. Lett. 36, 1122 (1976). Goldhaber's work was stimulated by investigations by R. Jackiw and C. Rebbi (*ibid.*, 1116) and by P. Hassenfratz and G. 't Hooft (*ibid.*, 1119).
13. P. Goddard, J. Nuyts, and D. Olive, Nucl. Phys. B125, 1 (1977).
14. I'm being very sloppy here about singularities. Here's a more careful argument: We define a gauge-field configuration to be locally non-singular in some region if the region is the union of a family of open sets, such that in each set the gauge field is non-singular and such that in the intersection of any two sets the gauge fields in the two sets are connected by a non-singular gauge transformation. It is possible to show that a gauge-field configuration that is locally non-singular in all of space except for the origin is gauge-equivalent to one that is non-singular (in the ordinary sense) everywhere except for the south polar axis. (This theorem is proved in Ref. 4.) That is to say, if all singularities, other than ones at the origin, are gauge artifacts, then they can all be shoved onto the Dirac string by an appropriate choice of gauge.
15. E. Lubkin, Ann. Phys. (N.Y.) 23, 233 (1963). See especially Sec. XV.
16. A somewhat longer (though still hopelessly vulgar) course can be found in Ref. 4, together with references to the mathematical literature.
17. This stability analysis was first done by R. Brandt and F. Neri, Nucl. Phys. B161, 253 (1979).
18. W. Nahm and D. Olive (private communication, summer 1979).
19. The work described here is drawn from many sources; for references, see Ref. 4.
20. H. Georgi, and S. L. Glashow, Phys. Rev. Lett. 32, 438 (1974).
21. E. Bogomol'nyi, Sov. J. Nucl. Phys. 24, 449 (1976). S. Coleman, S. Parke, A. Neveu, and C. Sommerfield, Phys. Rev. D15, 544 (1977).
22. M. Prasad and C. Sommerfield, Phys. Rev. Lett. 35, 760 (1975).
23. Much of this section is blatant plagiarism from Ref. 4. (Copyright holder take note!)

24. Actually, one can avoid this computational horror by group-theoretic tricks, just as we did for the Laplace operator in Sec. 2.

25. This note is for experts only. The discussion in the text leaves the impression that the detailed computation of the isorotational spectrum is easier than it is. There are technical complications associated with gauge invariance. These can all be dealt with, but they make the calculation lengthier than it would be if they weren't around. For example, in temporal gauge, there are many invariances of the Hamiltonian that do not leave the monopole solution unchanged, to wit, time-independent gauge transformations. No one in his right mind expects these to lead to isorotational levels. However, to demonstrate this, and to disentangle the spurious excitations from the genuine ones, requires fiddling around with the subsidiary condition that is the bane of temporal-gauge quantization. Subsidiary conditions can be avoided by working in Coulomb gauge, for example, where none are needed, but then the form of the Hamiltonian is more complicated, and this makes things messy. (For a careful Coulomb-gauge treatment of the dyons discussed immediately below, see E. Tomboulis and G. Woo, Nucl. Phys. B107, 221 (1976).

26. These dyons were first discovered by B. Julia and A. Zee, Phys. Rev. D11, 2227 (1975), using quite different methods from these.

27. E. Witten, Phys. Lett. 86B, 283 (1979).

28. C. Dokos and T. Tomaras, Phys. Rev. D21, 2940 (1980).

29. This trick was discovered by G. 't Hooft, Nucl. Phys. B105, 538 (1976) and by E. Corrigan, D. Olive, D. Fairlie, and J. Nuyts, Nucl. Phys. B106, 475 (1976).

30. J. Schwinger, Phys. Rev. 144, 1087; 151, 1048; 151, 1055 (1966).

31. For example, see L. Landau and E. Lifshitz, *Quantum Mechanics* (3rd ed.)(Pergamon, 1977) p. 410.