Concept of nonintegrable phase factors and global formulation of gauge fields

Tai Tsun Wu*
Gordon McKay Laboratory, Harvard University, Cambridge, Massachusetts 02138

Chen Ning Yang†
Institute for Theoretical Physics, State University of New York, Stony Brook, New York 11794
(Received 8 September 1975)

Through an examination of the Bohm-Aharonov experiment an intrinsic and complete description of electromagnetism in a space-time region is formulated in terms of a nonintegrable phase factor. This concept, in its global ramifications, is studied through an examination of Dirac's magnetic monopole field. Generalizations to non-Abelian groups are carried out, and result in identification with the mathematical concept of connections on principal fiber bundles.

I. MOTIVATION AND INTRODUCTION

The concept of the electromagnetic field was conceived by Faraday and Maxwell to describe electromagnetic effects in a space-time region. According to this concept, the field strenght $f_{\mu\nu}$ describes electromagnetism. It was later realized,[1] however, that $f_{\mu\nu}$ by itself does not, in quantum theory, completely describe all electromagnetic effects on the wave function of the electron. The famous Bohm-Aharonov experiment, first beautifully performed by Chambers,[2] showed that in a multiply connected region where $f_{\mu\nu}=0$ everywhere there are physical experiments for which the outcome depends on the loop integral

$$\frac{e}{\hbar c}\oint A_\mu dx^\mu \tag{1}$$

around an unshrinkable loop. This raises the question of what constitutes *an intrinsic and complete description* of electromagnetism. In the present paper we wish to discuss this question and also its generalization to non-Abelian gauge fields.

An examination of the Bohm-Aharonov experiment indicates that in fact only *the phase factor*

$$\exp\left(\frac{ie}{\hbar c}\oint A_\mu dx^\mu\right), \tag{2}$$

and *not the phase* (1), is physically meaningful. In other words, the phase (1) contains more information than the phase factor (2). But the additional information is not measurable. This simple point, probably implicitly recognized by many authors, is discussed in Sec. II. It leads to the concept of nonintegrable (i.e., path-dependent) phase factor as the basis of a description of electromagnetism.

This concept has been taken[3] as the basis of the definition of a gauge field. The discussions in Ref. 3, however, centered only on the local properties of gauge fields. To extend the concept to global problems we analyze in Sec. III the field produced by a magnetic monopole. We demonstrate how the quantization of the pole strength, a striking result due to Dirac,[4] is understood in this concept of electromagnetism. The demonstration is closely related to that in the original Dirac paper. Dirac discussed the phase factor of the wave function of an electron (which, among other things, depends on the electron energy). Our emphasis is on the nonintegrable electromagnetic phase factor (which does not depend on such quantities as the energy of the electron).

The monopole discussion leads to the recognition that in general the phase factor (and indeed the vector potential A_μ) can only be properly defined in each of many overlapping regions of space-time. In the overlap of any two regions there exists a gauge transformation relating the phase factors defined for the two regions. This discussion is made more precise in Sec. IV. It leads to the definition of global gauges and global gauge transformations.

In Sec. V generalizations to non-Abelian gauge groups are made. The special cases of SU_2 and SO_3 gauge fields are discussed in Secs. VI and VII. A surprising result is that the monopole types are quite different for SU_2 and SO_3 gauge fields and for electromagnetism.

The mathematics of these results is in fact well known to the mathematicians in *fiber bundle theory*. An identification table of terminologies is given in Sec. V. We should emphasize that our interest in this paper does not lie in the beautiful, deep, and general mathematical development in fiber bundle theory. Rather we are concerned with the necessary *concepts to describe the physics of gauge theories*. It is remarkable that these concepts have already been intensively studied as mathematical constructs.

Section VII discusses a "*gedanken*" generalized Bohm-Aharonov experiment for SU_2 gauge fields.

Unfortunately, the experiment is not feasible unless the mass of the gauge particle vanishes. In the last section we make several remarks.

II. DESCRIPTION OF ELECTROMAGNETISM

The Bohm-Aharonov experiment explores the electromagnetic effect on an electron beam (Fig. 1) in a doubly connected region where the electromagnetic field is zero. As predicted[1] by Aharonov and Bohm, the fringe shift is dependent on the phase factor (2), which is equal to

$$\exp\left(\frac{-ie}{\hbar c}\Omega\right),$$

where Ω is the magnetic flux in the cylinder. Thus two cases a and b for which

$$\Omega_a - \Omega_b = \text{integer} \times (hc/e) \tag{3}$$

give the same interference fringes in the experiment. This we shall state and prove as follows.

Theorem 1: If (3) is satisfied, no experiment outside of the cylinder can differentiate between cases a and b.

Consider first an electron outside of the cylinder. We look for a gauge transformation on the electron wave function ψ_a and the vector potential $(A_\mu)_a$ for case a, which changes them into the corresponding quantities for case b, i.e. we try to find $S = e^{-i\alpha}$ such that

$$S = S_{ab} = (S_{ba})^{-1},$$

$$\psi_b = S^{-1}\psi_a, \quad \text{or } \psi_b = e^{i\alpha}\psi_a, \tag{4}$$

$$(A_\mu)_b = (A_\mu)_a - \frac{i\hbar c}{e} S\frac{\partial S^{-1}}{\partial x^\mu}, \quad \text{or } (A_\mu)_b = (A_\mu)_a + \frac{\hbar c}{e}\frac{\partial \alpha}{\partial x^\mu}. \tag{5}$$

For this gauge transformation to be definable, S must be *single-valued*, but α itself need not be. Now $(A_\mu)_b - (A_\mu)_a$ is curlless; hence (5) can always be solved for α. But it is multiple-valued with an increment of

$$\Delta\alpha = \frac{e}{\hbar c}\oint [(A_\mu)_b - (A_\mu)_a]\, dx^\mu = \frac{e}{\hbar c}(\Omega_b - \Omega_a) \tag{6}$$

every time one goes around the cylinder. If (3) is satisfied, $\Delta\alpha = 2\pi \times$ integer and S is single-valued. Case a and case b outside of the cylinder are then gauge-transformable into each other, and no physically observable effects would differentiate them. The same argument obviously holds if one studies the wave function of an interacting system of particles provided the charges of the particles are all integral multiples of e. Thus we have shown the validity of Theorem 1.

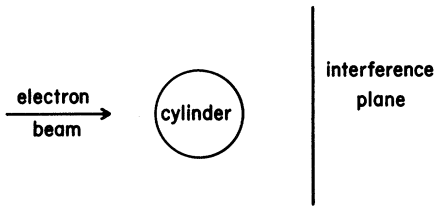

FIG. 1. Bohm-Aharonov experiment (Refs. 1, 2). A magnetic flux is in the cylinder. Outside of the cylinder the field strength $f_{\mu\nu} = 0$.

We conclude: (a) The field strength $f_{\mu\nu}$ underdescribes electromagnetism, i.e., different physical situations in a region may have the same $f_{\mu\nu}$. (b) The phase (1) overdescribes electromagnetism, i.e., different phases in a region may describe the same physical situation. What provides a complete description that is neither too much nor too little is the phase factor (2).

Expression (2) is less easy to use (especially when one makes generalizations to non-Abelian groups) as a fundamental concept than the concept of a phase factor for any path from P to Q

$$\Phi_{QP} = \exp\left(\frac{ie}{\hbar c}\int_P^Q A_\mu\, dx^\mu\right) \tag{7}$$

provided that an arbitrary gauge transformation

$$\exp\left(\frac{ie}{\hbar c}\int_P^Q A_\mu\, dx^\mu\right) \rightarrow \exp\left(\frac{ie}{\hbar c}a(Q)\right)\exp\left(\frac{ie}{\hbar c}\int_P^Q A_\mu\, dx^\mu\right)\exp\left(\frac{-ie}{\hbar c}a(P)\right) \tag{8}$$

does not change the prediction of the outcome of any physical measurements. Following Ref. 3, we shall call the phase factor (7) a nonintegrable (i.e., path-dependent) phase factor.

Electromagnetism is thus the gauge-invariant manifestation of a nonintegrable phase factor. We shall develop this theme further in the next section.

III. FIELD DUE TO A MAGNETIC MONOPOLE

The definition of a nonintegrable phase factor (7) in a general case may present problems. To illustrate the problem, let us study the magnetic monopole field of Dirac.[4] Consider a static magnetic monopole of strength $g \neq 0$ at the origin $\vec{r} = 0$ and take the region R of space-time under consideration to be all space-time minus the origin $\vec{r} = 0$. We shall now show the following:

Theorem 2: There does not exist a singularity-free A_μ over all R.

If a singularity-free A_μ does exist throughout R, consider the loop integral $\oint A_\mu dx^\mu$ for time $t=0$ around a circle at fixed spherical coordinates r and θ with azimuthal angle $\phi=0\to 2\pi$. This integral, denoted by $\Omega(r,\theta)$ for $r>0$, is equal to the magnetic flux through a cap bounded by the loop, or more explicitly $\Omega(r,\theta)=2\pi g(1-\cos\theta)$. At $\theta=0$, $\Omega(r,0)=0$. Increasing θ leads to a continuous increase in Ω till one approaches $\theta=\pi$, at which

$$\Omega(r,\pi)=4\pi g\,. \tag{9}$$

But at $\theta=\pi$ the loop shrinks to a point. Therefore $\Omega(r,\pi)=0$ since A_μ has no singularity. We have thus reached a contradiction and Theorem 2 is proved.

With an A_μ which has singularities, the nonintegrable phase factor becomes undefined if the path goes through a singularity. This difficulty *must* be resolved in order to use a nonintegrable phase factor as a fundamental concept to describe electromagnetism. It can be resolved in the following way. Let us seek to divide R into two overlapping regions R_a and R_b and to define $(A_\mu)_a$ and $(A_\mu)_b$, each singularity-free in their respective regions, so that (i) their curls are equal to the magnetic field and (ii) in the overlapping region $(A_\mu)_a$ and $(A_\mu)_b$ are related by a gauge transformation. One possible choice is to take the regions to be

$$\begin{aligned} R_a&:\ 0\le\theta<\pi/2+\delta\ \ 0<r\,,\ \ 0\le\phi<2\pi\,,\ \text{all } t\\ R_b&:\ \pi/2-\delta<\theta\le\pi\ \ 0<r\,,\ \ 0\le\phi<2\pi\,,\ \text{all } t \end{aligned} \tag{10}$$

with an overlap extending throughout $\pi/2-\delta<\theta<\pi/2+\delta$. (We assume $0<\delta\le\pi/2$.) Take

$$(A_t)_a=(A_r)_a=(A_\theta)_a=0\,,\quad (A_\phi)_a=\frac{g}{r\sin\theta}(1-\cos\theta), \tag{11}$$

$$(A_t)_b=(A_r)_b=(A_\theta)_b=0\,,\quad (A_\phi)_b=\frac{-g}{r\sin\theta}(1+\cos\theta)\,.$$

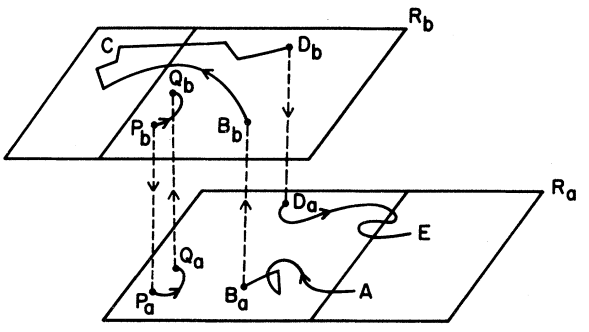

FIG. 2. Schematic diagram illustrating the relationship between R_a and R_b.

The gauge transformation in the overlap of the two regions is

$$S=S_{ab}=\exp(-i\alpha)=\exp\left(\frac{2ige}{\hbar c}\phi\right). \tag{12}$$

This is an allowed gauge transformation if and only if S is single-valued, i.e.,

$$\frac{2ge}{\hbar c}=\text{integer}=D\,, \tag{13}$$

which is Dirac's quantization. With (13) we have

$$S_{ab}=\exp(iD\phi)\,. \tag{12'}$$

To define the phase factor for a path we refer to Fig. 2, where a point in the overlapping region, such as point P, is regarded as two points P_a and P_b. If a path is entirely within region a or b, we define Φ along the path by (7) with $(A_\mu)_a$ or $(A_\mu)_b$ in the integrand in the exponent. If the path $Q\leftarrow P$ is entirely within the overlapping region we have then two possible phase factors $\Phi_{Q_aP_a}$ and $\Phi_{Q_bP_b}$. It is easy to prove that

$$\Phi_{Q_bP_b}=S^{-1}(Q)\Phi_{Q_aP_a}S(P)\,, \tag{14}$$

i.e.,

$$\Phi_{Q_aP_a}S(P)=S(Q)\Phi_{Q_bP_b}\,, \tag{14'}$$

which merely states that $(A_\mu)_a$ and $(A_\mu)_b$ are related by a gauge transformation with the transformation factor (12).

For a path that crisscrosses in and out of the overlapping region, such as $A\to B\to C\to D\to E$ in Fig. 2, the definition of Φ is

$$\Phi_{EDCBA}=\Phi_{ED_a}S_{ab}(D)\Phi_{D_bCB_b}S_{ba}(B)\Phi_{B_aA}\,. \tag{15}$$

Notice that fixing the path but sliding the points B and D along it does not change Φ_{EDCBA} [because of formulas like (14')] so long as B and D remain in the overlapping region.

The phase factor so defined satisfies the group property, e.g.,

$$\begin{aligned}\Phi_{EDCBA}&=\Phi_{ED_a}\Phi_{D_aCBA}\\ &=\Phi_{ED_b}\Phi_{D_bCBA}\\ &=\Phi_{EDC}\Phi_{CBA}\,,\ \text{etc.}\end{aligned} \tag{16}$$

The relationship between the electromagnetic field and the phase factor around a loop is the same as usual. One only has to be careful that if the starting and terminating point A is in the overlapping region, the phase factor is taken to be $\Phi_{A_aBA_a}=\Phi_{A_bBA_b}$, and not $\Phi_{A_aBA_b}$ or $\Phi_{A_bBA_a}$. The phase factor around the loop is then equal to

$$\exp\left(\frac{ie}{\hbar c}\right)\Omega\,,$$

where Ω is the magnetic flux through a cap bor-

dered by the loop. Notice that because of Dirac's quantization condition, the phase factor is the same whichever way one chooses the cap provided it does not pass through the point $\vec{r}=0$ (any t).

We have satisfactorily resolved the difficulty mentioned at the beginning of this section, provided Dirac's quantization condition (13) is satisfied. We shall now prove the following.

Theorem 3: If (13) is not satisfied (the above method of resolving the difficulty would not work since) there exists no division of R into overlapping regions $R_a, R_b, R_c, \ldots$ so that condition (i) and (ii) stated above, properly generalized to the case of more than two regions, would hold.

To prove this statement, observe that if such a division is possible, one could generalize (15) and arrive at a satisfactory definition of the phase factor. The phase factor around a loop is then a continuous function of the loop. Take the loop to be a parallel on the sphere r fixed, $t=0$, θ fixed, $\phi=0\rightarrow 2\pi$. The phase factor defined by the generalization of (15) is equal to

$$\exp\left[\frac{ie}{\hbar c}\Omega(r,\theta)\right]=\exp\left[\frac{ie}{\hbar c}2\pi g(1-\cos\theta)\right]. \quad (17)$$

This is not equal to unity when $\theta=\pi$, since (13) is assumed to be invalid. Thus we have a contradiction.

Theorem 3 shows that if Dirac's quantization condition (13) is not satisfied, then the field of a magnetic monopole of strength g cannot be taken as a realizable physical situation in R. (Of course, if one excludes the half-line $x=y=0$, $z<0$, or any half-line starting from $\vec{r}=0$ leading to infinity, then it is possible to have any value for g.) This conclusion is the same as Dirac's, but viewed from a somewhat different point of emphasis.

IV. GENERAL DEFINITION OF GAUGE AND GLOBAL GAUGE TRANSFORMATION

Assuming that (13) holds, to round out our concept of a nonintegrable phase factor the question of the flexibility in the choice of the overlapping regions and the flexibility in the choice of A_μ in the regions must be faced. Both of these questions are related to gauge transformations.

Consider a gauge transformation ξ in R_b (ξ will be assumed to be many times differentiable, but not necessarily analytic), resulting in a new potential $(A_\mu)'_b$. We shall illustrate schematically the transformation by "elevating" the region b in Figure 3(a).

One could extend the region b. One could also contract it, provided the whole R remain covered.

One could create a new region by considering a subregion of b as an additional region R_c [Figure 3(b)], and define the gauge transformation connecting them as the identity transformation so that $(A_\mu)_c=(A_\mu)_b$. One can then "elevate" R_c and contract R_b, which results in Fig. 3(c).

Through operations of the kind mentioned in the last three paragraphs, which we shall call *distortions*, we arrive at a large number of possibilities, each with a particular choice of overlapping regions and with a particular choice of gauge transformation from the original $(A_\mu)_a$ or $(A_\mu)_b$ to the new A_μ in each region. Each of such possibilities will be called a *gauge* (or *global gauge*). This definition is a natural generalization of the usual concept, extended to deal with the intricacies of the field of a magnetic monopole.

For each choice of gauge there is a definition of a nonintegrable phase factor for every path. The group condition $\Phi_{C_cBA_a}=\Phi_{C_cB_b}\Phi_{B_bA_a}$ is always satisfied.

Notice that the original gauge we started with was characterized by (a) specifying [in (10)] the regions [R_a and R_b] and (b) specifying the gauge transformation factor (12′) in the overlap (between R_a and R_b). *It does not refer to any specific* A_μ. [A distortion may of course lead to no changes in characterizations (a) and (b). Thus two different gauges may share the same characterizations (a) and (b).] In the case of the monopole field, we had chosen the vector potential to be given by (11). But, in fact, we can attach to this gauge any $(A_\mu)_a$ and $(A_\mu)_b$ provided they are gauge-transformed into each other by (12′) in the region of overlap. (The resultant $f_{\mu\nu}$ is, of course, not a monopole field in general.) *Thus a gauge is a concept not tied to any specific vector potential.* We shall call the process of distortion leading from one gauge to another a *global gauge transformation*. It is also a concept not tied to any specific vector potential. It is a natural generalization of the usual gauge transformation.

The collection of gauges that can be globally gauge-transformed into each other will be said to

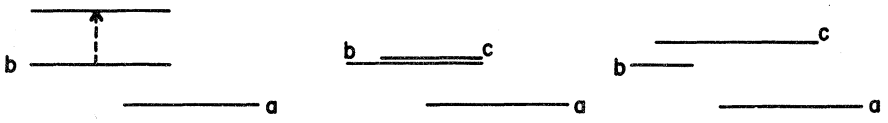

FIG. 3. Distortions allowed in gauge transformation.

belong to the same *gauge type*.

The phase factor around a loop starts and ends at the same point in the same region. Thus it does not change under any global gauge transformation, i.e. we have, for Abelian gauge fields, the following.

Theorem 4a: The phase factor around any loop is invariant under a global gauge transformation.

It follows trivially from this, by taking an infinitesimal loop, that

Theorem 5a: The field strength $f_{\mu\nu}$ is invariant under a global gauge transformation.

For a given value of D, the gauge defined by (10) and (12) will be denoted by $\mathcal{G}_D$. For $D \neq D'$, the relationship, or rather the lack of relationship, between $\mathcal{G}_D$ and $\mathcal{G}_{D'}$ is shown by Theorem 6.

Theorem 6: For $D \neq D'$, $\mathcal{G}_D$ and $\mathcal{G}_{D'}$ are not related by a global gauge transformation, i.e., they are not of the same gauge type.

To prove this theorem we use Theorem 7.

Theorem 7: Between two gauge fields defined on the same gauge there exists a continuous interpolating gauge field defined on the same gauge.

To prove Theorem 7, we simply make a linear interpolation between the two original gauge fields which we shall denote by $(A_\mu)^{(\alpha)}$ and $(A_\mu)^{(\beta)}$:

$$A^{(\gamma)} = t(A_\mu)^{(\alpha)} + (1-t)(A_\mu)^{(\beta)}, \quad 0 \leq t \leq 1. \tag{18}$$

In an overlap between regions a and b this interpolating vector potential assumes values $(A_\mu)_a^{(\gamma)}$ and $(A_\mu)_b^{(\gamma)}$ which are related by the proper gauge transformation belonging to this overlap. Thus we have proved Theorem 7.

Now go back to Theorem 6 and assume it to be invalid. Then we can gauge-transform the vector potential belonging to the monopole of strength $D'\hbar c/2e$ to the gauge $\mathcal{G}_D$. For this gauge we have then two monopole fields of different pole strengths. Using Theorem 7 we interpolate between them and obtain unquantized magnetic monopoles, which contradict Theorem 3.

Notice that although in this proof of Theorem 6 we have used two specific gauge fields, the theorem itself does not refer to any specific gauge fields at all.

By the same argument as used in the proof of Theorem 7, any gauge field defined on $\mathcal{G}_D$ must have a magnetic monopole of strength $D\hbar c/2e$ at the excluded point $\vec{r}=0$, in addition to possible fields produced by electric charges and currents. Thus the total magnetic flux around the origin $\vec{r}=0$ is equal to $(2\pi\hbar c/e)D$ for any gauge field defined on $\mathcal{G}_D$. We shall state this as a theorem and give another proof of it.

Theorem 8: Consider gauge $\mathcal{G}_D$ and define any gauge field on it. The total magnetic flux through a sphere around the origin $\vec{r}=0$ is *independent of the gauge field* and only depends on the gauge:

$$\oint\!\!\oint f_{\mu\nu}\,dx^\mu\,dx^\nu = \frac{-i\hbar c}{e}\oint \frac{\partial}{\partial x^\mu}(\ln S_{ab})\,dx^\mu, \tag{19}$$

where S is the gauge transformation defined by (12) for the gauge $\mathcal{G}_D$ in question, and the integral is taken around any loop around the origin $\vec{r}=0$ in the overlap between R_a and R_b, such as the equator on a sphere $r=1$.

To prove this theorem we observe that the flux through the upper half of the sphere $r=1$ is equal to the following integral around the equator:

$$\oint (A_\mu)_a\,dx^\mu. \tag{20a}$$

The flux through the lower half is equal to a similar integral around the equator:

$$-\oint (A_\mu)_b\,dx^\mu. \tag{20b}$$

Hence

$$\text{total flux} = \oint [(A_\mu)_a - (A_\mu)_b]\,dx^\mu$$

$$= \frac{-i\hbar c}{e}\oint \frac{\partial}{\partial x^\mu}(\ln S_{ab})\,dx^\mu, \tag{21}$$

which completes the proof. Using (13) and (12), the right-hand side of (21) is equal to $4\pi g$, as expected.

If one starts with any gauge which is of the same gauge type as $\mathcal{G}_D$, and makes a global gauge transformation on it, the total flux is not changed by Theorem 5a. Thus (19), which depends only on the gauge, is in fact the same for all gauges of the same type. Notice that if there are more regions in a gauge than two, (19) should be replaced by a sum of line integrals along paths that are in the various overlaps between the regions. For a case of three regions there are three paths, which are illustrated in Fig. 4. Along each path the integral is of the form (19) with S denoting the gauge transformation factor, such as (12), between the two regions containing the path. To prove Theorem 8 in this case one need only add three loop integrals to-

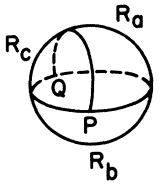

FIG. 4. Case of three regions for Theorem 8. The three paths from P to Q are in the three overlapping regions between (R_a, R_b), (R_b, R_c), and (R_c, R_a).

gether, each of the form of (20a) and (20b), and notice that along each path the integrand is always the difference of the vector potential A_μ between two regions, very much as in (21).

The first proof we gave above of Theorem 8 is easy and is "obvious" to a physicist. The second proof is more involved but is more intrinsic. The theorem is a special case of the Chern-Well theorem which evolved from the famous Gauss-Bonnet-Allendoerfer-Weil-Chern theorem, a seminal development in contemporary mathematics.[5] We want to emphasize two consequences of the theorem. (i) The right-hand side of (19) is independent of the gauge field, and only depends on the gauge type. (ii) The right-hand side of (19) has as integrand the gradient of $\ln S$. Since S is single-valued, the integral must be equal to an *integral multiple* of a constant (in this case $2\pi i$). A remarkable fact is that these consequences remain valid in the general mathematical theorem, which is very deep.

V. GENERALIZATION TO NON-ABELIAN GAUGE FIELD

So far we have only considered electromagnetism and described it in terms of an Abelian gauge field that corresponds to the group U_1, or equivalently SO_2. On the basis of the discussions in the preceding section, the generalization to the non-Abelian case can be carried out without much difficulty. For a local region this has been done in Ref. 3. Extension to global considerations is our present focus of interest.

A gauge is defined by (a) a particular choice of overlapping regions and (b) a particular choice of *single-valued* gauge transformations S_{ab} in the overlapping regions. The choice of gauge transformations clearly must satisfy the following two conditions.

(1) In the overlapping region $R_a \cap R_b$, the gauge transformations S_{ba} from a to b and S_{ab} from b to a are related by

$$S_{ab}S_{ba}=1\,,$$

where 1 is the identity element of the gauge group.

(2) If three regions R_a, R_b, and R_c overlap, then there are gauge transformations $S_{ab}, S_{ba}, S_{ac}, S_{ca}$, S_{bc}, S_{cb} so that

$$S_{ab}S_{bc}S_{ca}=1\,,\ \text{etc.}$$

in $R_a \cap R_b \cap R_c$.

As in the case of electromagnetism, both the concept of a gauge and the concept of a global gauge transformation are not tied to any specific gauge potentials, denoted in general by b^k_μ.

The *nonintegrable phase factor* for a given path is now an element of the gauge group. We shall still call it a phase factor. Since these phase factors do not in general commute with each other, Theorems 4a and 5a for the Abelian case need to be modified as follows.

Theorem 4: Under a global gauge transformation, the phase factor around any loop remains in the same class. The class does not depend on which point is taken as the starting point around the loop.

Theorem 5: The field strength $f^k_{\mu\nu}$ is covariant under a global gauge transformation.

Only theorem 4 is not immediately transparent. For a loop $ABCA$, under a gauge transformation[3]

$$\Phi_{ABCA} \to \Phi'_{ABCA} = \xi(A)\Phi_{ABCA}\xi^{-1}(A)\,.$$

Thus Φ'_{ABCA} and Φ_{ABCA} are in the same class. Also around the same loop if we change the starting point from A to C,

$$\Phi_{CABC} = \Phi_{CA}\Phi_{ABCA}\Phi_{AC}\,.$$

Hence changing the starting point does not change the class.

Theorem 4 defines *the class of a loop*. This concept is the generalization of the phase factor for electromagnetism around a loop with the magnetic flux as the exponent. It is a gauge-invariant concept.

These concepts have been extensively studied by the mathematicians in the framework of more general[6] mathematical constructs. A translation of terminology is given in Table I.

VI. CASE OF SU_2 GAUGE FIELD

For the SU_2 case we take the infinitesimal generators X_k to satisfy

$$X_1X_2 - X_2X_1 = X_3\,,\ \text{etc.} \tag{22}$$

and define the phase factor, as a generalization of (7), by[7]

$$\Phi_{QP} = \left[\exp\left(\int_P^Q \frac{-e}{\hbar c} b^k_\mu X_k dx^\mu\right)\right]_{\text{ordered}}\,, \tag{23}$$

i.e., we make the replacement

$$ieA_\mu \to -eb^k_\mu X_k\,, \tag{24}$$

or

$$A_\mu \to ib^k_\mu X_k\,. \tag{25}$$

[The subscript "ordered" means that, in the definition of the exponential in terms of a power series, the factors $b^k_\mu X_k$ are ordered along the path from P to Q with the factor $b^k_\mu(P)X_k$ at the right end of the product.] The algebraic operators X_k can be thought of as the collection of all irreducible representations of (22). The eigenvalue of iX_k with the

TABLE I. Translation of terminology.

Gauge field terminology	Bundle terminology
gauge (or global gauge)	principal coordinate bundle
gauge type	principal fiber bundle
gauge potential b_μ^k	connection on a principal fiber bundle
S_{ba} (see Sec. V)	transition function
phase factor Φ_{QP}	parallel displacement
field strength $f_{\mu\nu}^k$	curvature
source [a] J_μ^K	?
electromagnetism	connection on a $U_1(1)$ bundle
isotopic spin gauge field	connection on a SU_2 bundle
Dirac's monopole quantization	classification of $U_1(1)$ bundle according to first Chern class
electromagnetism without monopole	connection on a trivial $U_1(1)$ bundle
electromagnetism with monopole	connection on a nontrivial $U_1(1)$ bundle

[a] I.e., electric source. This is the generalization (see Ref. 3) of the concept of electric charges and currents.

minimum absolute value is $\pm\frac{1}{2}$. Therefore the minimum "charge" of all physical states can be read off from (24) by taking the 2×2 irreducible representation of X_k:

$$X_k = -\frac{i\sigma_k}{2}\,, \tag{26}$$

where σ_k are the Pauli matrices. Thus

$$\text{minimum "charge"} = \frac{e}{2}\,. \tag{27}$$

The particle of the gauge field belongs to the adjoint representation. Its "charges" are e, 0, and $-e$. Thus

$$\frac{\text{"charge" of gauge particle}}{\text{minimum "charge"}} = 2 \quad \text{for } SU_2\,. \tag{28}$$

We shall now try to define a Dirac monopole field as a special SU_2 field along only one isospin direction $k=3$, i.e., we define

$$b_\mu^1 = b_\mu^2 = 0\,, \quad b_\mu^3 = A_\mu\,, \tag{29}$$

where A_μ is given in the two regions (10) by (11). In the overlapping region, transformation factor S of (12) and (14) now becomes

$$S_{ab} = \exp\left(-\frac{2ge}{\hbar c}\phi X_3\right) \tag{30}$$

by replacement (25). This is single-valued if and only if the quantization condition

$$\frac{eg}{\hbar c} = \text{integer} = D \tag{31}$$

is satisfied because for SU_2

$$\exp(4\pi X_3) = 1\,, \quad \exp(2\pi X_3) \neq 1\,,$$

which follows from the existence of half-integral representations such as (26).

The phase factor (30) describes a great circle, wound D times, on the manifold of SU_2 when ϕ varies from $0 \to 2\pi$. Such a circle can be continuously shrunk to the identity element, in contrast with the situation for electromagnetism. Thus, by a global gauge transformation S may be changed to $S'=1$, and the two regions a and b after the global gauge transformation can be *fused into one single region*. The gauge potential b_μ^k is then defined *everywhere* in R as a single region. Thus we have the following theorem.

Theorem 9: For the SU_2 gauge group, the gauges $\mathcal{G}_D$ for different D can be transformed into each other by global gauge transformations. The different monopole fields are therefore of the same type.

We shall only exhibit the global transformation for the case $\mathcal{G}_{-1}$ for which

$$S_{ba} = \exp(-2\phi X_3)\,, \tag{32}$$

$$\frac{e}{\hbar c} = \frac{-1}{g}\,. \tag{33}$$

The gauge transformations we shall seek are illustrated in Fig. 5. We shall choose

$$\xi = \exp[\theta(X_1\sin\phi - X_2\cos\phi)]\,, \tag{34}$$

$$\eta = \exp[(\pi-\theta)(X_1\sin\phi - X_2\cos\phi)]\exp(\pi X_2)\,. \tag{35}$$

It is easy to see that ξ is analytic in the coordinates x^μ at all points in R_a. (One only has to verify this statement at $\theta=0$, which is easily done.) Similarly η is analytic in R_b. ξ and η are therefore allowed gauge transformations in, respectively, R_a and R_b.

Now one can prove after some algebra that[8]

$$S'_{ba}=\eta S_{ba}\xi=1\,.$$

Thus after the gauge transformations ξ and η, which together form a global gauge transformation, regions R_a and R_b are related by the identity gauge transformation in their overlap, i.e., the two regions can be fused into one. To calculate the gauge potentials $b^{k'}_{\mu}$ after the global gauge transformation we use

$$\xi(Q)\left[1+\frac{1}{g}(b^k_\mu)'_a X_k dx^\mu\right]\xi^{-1}(P)=1+\frac{1}{g}(b^k_\mu)_a X_k dx^\mu$$

$$=1+\frac{1}{g}(A_\mu)_a X_3 dx^\mu\,, \tag{36}$$

where A_μ is given by (11) and $Q=P+dx$. By choosing dx^μ to be along the t and r directions, one obtains $b^{k'}_t=b^{k'}_r=0$. By choosing dx^μ to be along the θ direction, one obtains

$$b^{1'}_\theta=\frac{-g}{r}\sin\phi\,,\quad b^{2'}_\theta=\frac{g}{r}\cos\phi\,,\quad b^{3'}_\theta=0\,. \tag{37}$$

Now take dx^μ to be along the ϕ direction. We obtain, to order $d\phi$,

$$1+\frac{1}{g}(b^k_\phi)'X_k r\sin\theta\, d\phi$$

$$=\xi^{-1}(Q)\xi(P)+\frac{1}{g}(A_\phi)_a r\sin\theta\, d\phi\,\xi^{-1}(P)X_3\xi(P)\,. \tag{38}$$

The first term on the right-hand side can be[8] computed in a straightforward manner:

$$\xi^{-1}(Q)\xi(P)=1-\sin\theta\, d\phi[(X_1\cos\phi+X_2\sin\phi)-X_3\tan\tfrac{1}{2}\theta]\,.$$

The second term also can be easily computed since

$$\xi^{-1}(P)X_3\xi(P)=\sin\theta(X_1\cos\phi+X_2\sin\phi)+X_3\cos\theta$$

and $(A_\phi)_a$ was given by (11). Finally one arrives at

$$b^{1'}_\phi=\frac{-g}{r}\cos\theta\cos\phi\,,$$

$$b^{2'}_\phi=\frac{-g}{r}\cos\theta\sin\phi\,, \tag{39}$$

$$b^{3'}_\phi=\frac{g}{r}\sin\theta\,.$$

Combining these results and remembering (33), we obtain

$$\frac{e}{\hbar c}b^{k'}_\mu X_k dx^\mu=\frac{-1}{r^2}\epsilon_{ikj}x^j dx^i X_k\,,$$

i.e.,

$$b^{k'}_4=0\,,\quad \frac{e}{\hbar c}b^{k'}_i=-\frac{1}{r^2}\epsilon_{ikj}x^j\,. \tag{40}$$

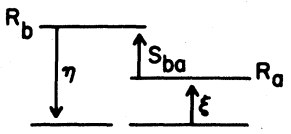

FIG. 5. A global transformation after which R_a and R_b can be fused.

Thus the new potential $b^{k'}_\mu$ is analytic in R_a. Because $\eta S_{ba}\xi=1$ the new potential (in the overlapping region) for R_b must be the same as (40). By analyticity (40) is seen to be valid throughout R. Notice that (40) is the same potential as one of the solutions [solution (12a)], for a sourceless gauge field, in Ref. 9.

The global gauge transformation that transforms $\mathcal{G}_D$ into $\mathcal{G}_0$ for $D\neq-1$ can be obtained by slightly modifying (34) and (35).

We shall discuss Theorem 9 further in the next section.

VII. CASE OF SO_3 GAUGE FIELD

We turn to SO_3, which is locally the same as SU_2, but for which

$$e^{2\pi X_k}=1\,. \tag{41}$$

Equations (22) to (25) remain unaltered. The minimum "charge" of all physical states is now

$$\text{minimum "charge"}=e\,, \tag{42}$$

giving

$$\frac{\text{"charge" of gauge particle}}{\text{minimum "charge"}}=1\ \text{ for } SO_3\,. \tag{43}$$

This last formula differentiates physically the SO_3 case from the SU_2 case.

We emphasize here a point already made in the literature[10] for electromagnetism: The local character of the gauge group is of course determined by the interactions (which determine the conservation laws). We want to ask what determines the global character. The global character (compact or noncompact in the case of electromagnetism, SU_2 or SO_3 in the isospin case) *is determined by* the representations for all states which *physically* exist. For example, in electromagnetism, if all charges are integral multiples of a single unit, the gauge group is compact,[10] because the group is *physically defined* as the simultaneous local phase factor change of *all* charge fields. There is then no physically definable meaning to the noncompact group. In the case of SU_2 or SO_3, if (43) is satisfied, then all representations of X_k physically realizable are integral representations.

Thus the simultaneous local changes of isospin phase factor of all physical systems *cannot differentiate* the group element $e^{2\pi X_k}$ from the identity. *Therefore*, the physical definition of $e^{2\pi X_k}$ is unity and the group must be SO_3.

Turning now to the monopole field for SO_3 we find that (30) is still correct. Equation (41) then leads to the quantization condition

$$\frac{2eg}{\hbar c} = \text{integer} = D \tag{44}$$

in order that S (as an element of SO_3) be a single-valued function of the coordinates x^μ in R.

As ϕ increases from 0 to 2π, the phase factor (30) describes a closed circuit in the group space of SO_3, starting from the identity element and returning to it. If one continuously traced the corresponding element of the group SU_2, one would have started from the identity and ended with the element that corresponds to

$$\begin{pmatrix} -1 & \\ & -1 \end{pmatrix}$$

in the 2×2 representation of SU_2 when D is odd. In such a case, no distortion of the closed circuit in SO_3 described by the phase factor (30) can shrink it to the identity element. This means that the gauge type for even D is not the same as that for odd D. By constructing explicit gauge transformations like (34) and (35) one can then complete the proof of the following theorem.

Theorem 10: For SO_3, all gauges $\mathcal{G}_D$ for D = even are of one type, and all gauges $\mathcal{G}_D$ for D = odd are of one type. These two types are different.

Summarizing the situation for U_1, SU_2, and SO_3 we find that in each case the "magnetic" monopole fields have quantized strengths. They belong to, respectively, infinitely many types for U_1 gauge group (electromagnetism), one type for the SU_2 gauge group, and two types for SO_3 gauge groups.

The physical meaning of these statements are as follows. In the SU_2 case, all magnetic monopole fields can be continuously changed into each other by the process of continuous changes[19] of "electric" sources. For example, starting with the "magnetic monopole" field for $\mathcal{G}_{-1}$ of Theorem 9 we can, by a gauge transformation, obtain the potentials b' [on $\mathcal{G}_0$] given in (40). We can then consider the potential (on $\mathcal{G}_0$): $b'' = \alpha b$, where $0 \le \alpha \le 1$. The gauge field for b'' is no longer electrically sourceless outside of the origin, but is magnetically sourceless except at the orgin, where it is not sourceless either magnetically or electrically. As α changes from 1 to 0 we thus have a continuous change of the original magnetic monopole field to empty space through a process during which there are continuous changes of electric charge-current distributions. Such a process is not possible for electromagnetism, by Theorem 6. (In the SO_3 case it is also not possible, although it is possible to change the magnetic monopole strength by two units by a similar process.) Thus the meaning of a magnetic monopole field in the non-Abelian case is quite different from that in electromagnetism.

It is not really surprising that in the case of electromagnetism one cannot change the magnetic monopole strength by changing electric sources: In the region R there are no magnetic monopoles. The continuity of magnetic lines of forces in R is guaranteed by the equation $\nabla\cdot\vec{H}=0$. No continuous movement of magnetic lines of force could therefore increase or decrease the net total flux around the origin. That this state of affairs does not obtain for SU_2 and SO_3 is due to the fact that in general $\nabla\cdot\vec{H}^k \neq 0$ in the non-Abelian case, so that one cannot define the magnetic flux through a loop. However, we had seen before (Theorem 4) that in the case of a non-Abelian gauge field what takes the place of the magnetic flux is the *phase factor of a loop*. One may then ask what takes the place of the total magnetic flux outwards from a sphere around the origin $\vec{r}=0$. To answer this question consider the loop

$$r=1\,,\quad \theta=\text{fixed}\,,\quad \phi=0\to 2\pi\,. \tag{45}$$

As θ changes from 0 to π the phase factor of the loop changes and it describes a continuous circuit (in the space of the group) starting from and ending at the identity element. Clearly any other way of "looping" over the sphere only leads to a distortion of this circuit, without changing the starting and ending point. We shall call this circuit the *total circuit* for the gauge field around the origin $\vec{r}=0$. It is a concept that replaces the total magnetic flux around $\vec{r}=0$ in electromagnetism.

We can now prove the following generalization of Theorem 8.

Theorem 11: Consider region R and the group SU_2 or SO_3. Consider a gauge $\mathcal{G}$ and define any gauge field on it. The total circuit for the gauge field around the origin $\vec{r}=0$ is independent of the gauge field and only depends on the gauge type of $\mathcal{G}$. For the case of $\mathcal{G}_D$,

total circuit of the gauge field

$$\simeq [S_{ba}(\phi) \text{ for } \phi = 2\pi \to 0]\,, \tag{46}$$

where $\simeq$ means "can be continuously distorted into."

This last formula is the generalization of (19).

To prove Theorem 11, consider first the loop (45). The phase factor in R_a and R_b will be denoted

by $\Phi^a(\theta)$ and $\Phi^b(\theta)$. They are related in the overlap by

$$\Phi^b(\theta) = \Phi^a(\theta) \tag{47}$$

since $S(\phi = 0) = I$. [One uses a generalization of (14).] Next consider the loop $L(\theta)$ which lies on the sphere $r = 1$ with its projection onto the x-y plane given in Fig. 6. It consists of a first part $(BA)_1$ around the equator and a second part $(AB)_2$ not on the equator except for points A and B. It is clear that

$$[\text{loop(45) for } \theta = 0 \to \pi/2] \simeq [L(\theta) \text{ for } \theta = 0 \to \pi/2]$$

because both sides "loop over" the upper hemisphere. Thus

$$[\Phi^a(\theta) \text{ for } \theta = 0 \to \pi/2] \simeq [\Phi^a_{L(\theta)} \text{ for } \theta = 0 \to \pi/2]$$
$$= [\Phi^a{}_{(AB)_2} \Phi^a_{(BA)_1} \text{ for } \theta = 0 \to \pi/2].$$

$\Phi^a_{(AB)_2}$ is continuous in θ, and

$$[\Phi^a_{(AB)_2} \text{ for } \theta = 0 \to \pi/2] \simeq \text{identity element}.$$

Thus

$$[\Phi^a(\theta) \text{ for } \theta = 0 \to \pi/2] \simeq [\Phi^a_{(BA)_1} \text{ for } \theta = 0 \to \pi/2]. \tag{48}$$

Similarly

$$[\Phi^b(\theta) \text{ for } \theta = \pi/2 \to \pi] \simeq [\Phi^b_{(BA)_1} \text{ for } \theta = \pi/2 \to \pi]. \tag{49}$$

At $\theta = \pi/2$, the left-hand sides of (48) and (49) match because of (47). Also the right-hand sides match. Thus we can take (48) and (49) in tandem, obtaining

$$\text{total circuit of gauge field} \simeq [\Phi^a_{(BA)_1} \text{ for } \theta = 0 \to \pi/2 \text{ followed by } S_B \Phi^a_{(BA)_1} \text{ for } \theta = \pi/2 \to \pi], \tag{50}$$

where we have used

$$\Phi^b_{(BA)_1} = S_B \Phi^a_{(BA)_1} S^{-1}{}_A = S_B \Phi^a_{(BA)_1}. \tag{51}$$

Now

$$[\Phi^a_{(BA)_1} \text{ for } \theta = 0 \to \pi/2 \text{ followed by } \Phi^a_{(BA)_1} \text{ for } \theta = \pi/2 \to \pi] \tag{52}$$

is a loop *that doubles back* on itself, i.e., (52) can be distorted to the identity element. Applying this fact to (50) one obtains

$$\text{total circuit of gauge field} \simeq [I \text{ for } \theta = 0 \to \pi/2 \text{ followed by } S_B \text{ for } \theta = \pi/2 \to \pi]. \tag{53}$$

Now $S_B = S_{ba}(\phi = 4\pi - 4\theta)$. As $\theta = \pi/2 \to \pi$, $S_B = S_{ba}(\phi)$ for $\phi = 2\pi \to 0$. Substitution into (53) leads to (46).

To complete the proof of Theorem 11 we need the generalization of (46) to gauges that contain more than two regions. This can be done without much difficulty, e.g., for the case that region b is further divided into regions c and d, as schematically illustrated in Fig. 7(a), (46) should be replaced by

$$\text{total circuit of gauge field} \simeq [S_{dc}(B) S_{ca}(x) \text{ for } x = A \to A \text{ along direction of arrow},$$
$$\text{followed by } S_{dc}(y) S_{ca}(A) \text{ for } y = B \to B \text{ along direction of arrow}]. \tag{54}$$

For the case that $\mathcal{G}$ has four regions a, b, c, d as illustrated in Fig. 7(b), (46) should be replaced by

$$\text{total circuit of gauge field} \simeq [S_{da}(x) \text{ for } x = A \to B,$$
$$\text{followed by } S_{db}(y) S_{ba}(x) \text{ for } x = B \to C,\ y = B \to D,$$
$$\text{followed by } S_{dc}(D) S_{cb}(x) S_{ba}(C) \text{ for } x = D \to C,$$
$$\text{followed by } S_{dc}(y) S_{ca}(x) \text{ for } x = C \to E,\ y = D \to E,$$
$$\text{followed by } S_{da}(x) \text{ for } x = E \to A]. \tag{55}$$

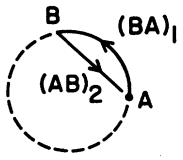

FIG. 6. Projection onto x–y plane of loop $L(\theta)$. The loop lies entirely on sphere $r=1$, and is in the upper (lower) hemisphere for $0\leq\theta\leq\pi/2$ ($\pi/2<\theta\leq\pi$). The portion $(BA)_1$ lies on the equator. Coordinates for $A: r=1$, $\theta=\pi/2$, $\phi=0$. Coordinates for $B: r=1$, $\theta=\pi/2$, $\phi=h(\theta)$, where $h(\theta)=4\theta$ for $0\leq\theta\leq\pi/2$ and $h(\theta)=4\pi-4\theta$ for $\pi/2<\theta\leq\pi$.

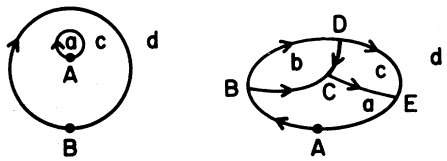

FIG. 7. Schematic diagrams for division lines in overlaps between three or more regions. The drawings are projections from the sphere $r=1$. The projection is from the south pole of the sphere onto the tangent plane at the north pole. The south pole is underneath the plane of the paper.

Notice that the right-hand sides of (54) and (55) are dependent only on the gauge type, and not on the specific gauge field.

VII. GENERALIZED BOHM-AHARONOV EXPERIMENT

The concept of an SU_2 gauge field was first discussed in 1954. In recent years many theorists, perhaps a majority, believe that SU_2 gauge fields do exist. However, so far there is *no experimental proof* of this theoretical idea, since conservation of isotopic spin only suggests, and does not require, the existence of an isotopic spin gauge field. What kind of experiment would be a definitive test of the existence of an isotopic spin gauge field? A generalized Bohm-Aharonov experiment would be.

If the gauge particle for isospin group SU_2 is massless, it is possible to design a *gedanken* generalized Bohm-Aharonov experiment as illustrated in Fig. 1. One constructs the cylinder of material for which the total I_z spin is not zero, e.g., a cylinder made of heavy elements with a neutron excess. One spins the cylinder around its axis, setting up a "magnetic" flux inside the cylinder, along the I_z "direction." If one scatters a proton beam around the cylinder, the fringe shift would be in the opposite direction from the corresponding shift observed with a neutron beam. To be more specific, imagine that one spins the cylinder clockwise. The magnetic flux would be emerging from the diagram towards the reader, since the cylinder has a net negative value for I_z. This means that for a proton (neutron) beam, the flux produces an increment (decrement) of path length counterclockwise around the cylinder. This increment (decrement) produces a net downward (upward) shift of the fringes, i.e., a shift toward the bottom (top) of the diagram.

If one scatters a coherent mixture of neutron and proton in a pure state, in the interference plane one would observe not only fluctuations of nucleon intensity, but also fluctuations of the neutron-proton mixing ratio. A variation of this phenomenon obtains if one imagines rotating a cylinder which has an average $\langle\vec{I}\rangle$ which is not zero, and is not in the I_z direction. A magnetic flux would then be set up which is in a "direction" other than I_z. Scattering a beam of protons would then produce some neutrons as well as protons in the interference plane. This implies, of course, that there is electric charge transfer between the beam and the cylinder together with the gauge field around it.

If the gauge particle has a finite mass $m>0$, then the experiment becomes difficult because the return flux would hug the outside surface of the cylinder, to a distance $\sim\hbar/mc$. Unless the fringe plane lies within this distance of the cylinder, the effect of the flux will be negligible.

IX. REMARKS

(a) From the viewpoint of the present paper, the electric charge and the magnetic charge play completely unsymmetrical roles. This matter deserves further comments. In the non-Abelian case it was in fact already pointed out[11] that the dual of an unquantized sourceless gauge field is not necessarily a gauge field. Thus the asymmetry between electric and magnetic phenomena is not due to the formalism, but is of an intrinsic nature in the non-Abelian case. In contrast, in the Abelian case the asymmetry is only formal since the electric and magnetic charges interact with the electromagnetic field in entirely symmetrical ways. In other words, one can use the phase factor Φ^M associated with the magnetic charges to describe the electromagnetic field, rather than the phase factor Φ discussed in the present paper. The mathematical relationship between these two kinds of phase factors (or between the associated vector potentials A_μ^M and A_μ) remains to be explored. So does the corresponding question in any second-quantized theory[12] of all the fields.

(b) In the proof of Theorem 9 we had shown explicitly how a magnetic monopole field for the SU_2 gauge group can be gauge-transformed into the solution (12a) of Ref. 9. Now a magnetic monopole field is not a gauge field at the origin $\vec{r}=0$ since

it does not satisfy the Bianchi identity[3] at the origin. Thus, although solution (12a) of Ref. 9 is (electrically) sourceless at all points, including the origin, it is not a proper gauge field at the origin, a fact we did not realize before. All three solutions, (12a), (12d), and (12e), are, of course, of the same gauge type.

(c) In Sec. II it was emphasized that $f_{\mu\nu}$ underdescribes electromagnetism because of the Bohm-Aharonov experiment which involves a doubly connected space region. For non-Abelian cases, the field strength $f^k_{\mu\nu}$ underdescribes the gauge field even in a singly connected region. An example of this underdescription was given in Ref. 13.

(d) For the region of space-time outside of the cylinder of Fig. 1 there is only one gauge type. All electromagnetic fields in the region can be continuously distorted into each other by the movement of electric charges and currents inside and outside the cylinder.

(e) The phase factor for the group U_1 is the phase factor of the algebra of complex numbers. It is perhaps not accidental that such a phase factor provides the basis for the description of a physically realized gauge field—electromagnetism. Now the only possible more complicated division algebra is the *algebra of quaternions*. The phase factors of the quaternions form the group SO_3. It is tempting to speculate that such a phase factor provides the basis for the description of a physically realized gauge field—the SU_2 gauge field. Speculation about the possible relationship between quaternions and isospin has been made before.[14] Such speculations were, however, not made with reference to gauge fields. If one believes that gauge fields give the underlying basis for strong and/or weak interactions, then the fact that gauge fields are fundamentally *phase factors* adds weight to the speculation that quaternion algebra is the real basis of isospin invariance.

(f) It is a widely held view among mathematicians that the fiber bundle is a natural geometrical concept.[15] Since gauge fields, including in particular the electromagnetic field, are fiber bundles, *all gauge fields are thus based on geometry*.[16] To us it is remarkable that a geometrical concept formulated without reference to physics should turn out to be exactly the basis of one, and indeed maybe all, of the fundamental interactions of the physical world.

ACKNOWLEDGMENTS

It is a pleasure to thank Professor Shiing-shen Chern for correspondence and discussions. We are especially indebted to Professor J. Simons, whose lectures and patient explanations have revealed to us glimpses of the beauty of the mathematics of fiber bundles.

While we were making corrections on the draft of this paper, a report on the experimental discovery of a magnetic monopole[17] reached us.

Additional references to fiber bundles, monopoles and quaternions are given in footnote 18.

---

*Work supported in part by the U. S. ERDA under Contract No. AT(11-1)-3227.

†Work supported in part by the National Science Foundation under Grant No. MPS74-13208 A01.

[1] Y. Aharonov and D. Bohm, Phys. Rev. 115, 485 (1959). See also W. Ehrenberg and R. E. Siday, Proc. Phys. Soc. London B62, 8 (1949).

[2] R. G. Chambers, Phys. Rev. Lett. 5, 3 (1960).

[3] Chen Ning Yang, Phys. Rev. Lett. 33, 445 (1974). This paper introduced the formulation of gauge fields in terms of the concept of nonintegrable phase factors. The differential formulation of gauge fields for Abelian groups was first discussed by H. Weyl, Z. Phys. 56, 330 (1929); for non-Abelian groups it was first discussed by Chen Ning Yang and Robert L. Mills, Phys. Rev. 96, 191 (1954). See also S. Mandelstam, Ann. Phys. (N.Y.) 19, 1 (1962); 19, 25 (1962); I. Białynicki-Birula, Bull. Acad. Pol. Sci., Ser. Sci. Math. Astron. Phys. 11, 135 (1963); N. Cabibbo and E. Ferrari, Nuovo Cimento 23, 1146 (1962); R. J. Finkelstein, Rev. Mod. Phys. 36, 632 (1964); N. Christ, Phys. Rev. Lett. 34, 355 (1975); and A. Trautman, in *The Physicist's Conception of Nature*, edited by J. Mehra (Reidel, Boston, 1973), p. 179.

[4] P. A. M. Dirac, Proc. R. Soc. London A133, 60 (1931). Since this brilliant work of Dirac, there have been several hundred papers on the magnetic monopole. For a listing of papers until 1970, see the bibliography by D. M. Stevens, Virginia Polytechnic Institute Report No. VPI-EPP-70-6, 1970 (unpublished).

[5] See J. Milnor and J. Stasheff, *Characteristic Classes* (Princeton Univ. Press, Princeton, N.J., 1974); C. B. Allendoerfer and A. Weil, Trans. Am. Math. Soc. 53, 101 (1943); Shiing-shen Chern, Ann. Math. 45, 747 (1944). See also H. Weyl, Amer. J. Math. 61, 461 (1939), and Ref. 6 below.

[6] There are many books on fiber bundles. See e.g., N. Steenrod, *The Topology of Fibre Bundles* (Princeton Univ. Press, Princeton, N. J., 1951). For connection, see, e.g., S. Kobayashi and K. Nomizu, *Foundations of Differential Geometry* (Interscience, New York, Vol. I-1963, Vol. II-1969).

[7] The notation here is the same as that in Ref. 3, except for the normalization factor $e/\hbar c$ which was absorbed into b in Ref. 3. To avoid confusion with the azimuthal angle, we write Φ for the ϕ of Ref. 3. Notice that

$$\Phi_{(A+dx)A} = I - \frac{e}{\hbar c} b^k_\mu(x) X_k dx^\mu .$$

All formulas are the same as in Ref. 3, but the name for Φ_{QP} will now be "the phase factor from P to Q." (In Ref. 3 the same name applied to Φ_{PQ}.) The new name is in accordance with the usual convention of time ordering.

[8]Since the 2×2 representation is faithful, it is sufficient for computational purposes to use the representation (26) for X_k. This makes the algebra quite simple since one can apply the formula $e^{i\theta\sigma_k} = \cos\theta + i\sigma_k \sin\theta$.

[9]Tai Tsun Wu and Chen Ning Yang, in *Properties of Matter under Unusual Conditions*, edited by H. Mark and S. Fernbach (Wiley, New York, 1969), p. 349.

[10]Chen Ning Yang, Phys. Rev. D 1, 2360 (1970).

[11]Gu Chao-hao and Chen Ning Yang, Sci. Sin. 18, 483 (1975).

[12]P. A. M. Dirac, Phys. Rev. 74, 817 (1948); J. Schwinger, *Particles, Sources and Fields* (Addison-Wesley, Reading, Mass., Vol. 1-1970, Vol. 2-1973).

[13]Tai Tsun Wu and Chen Ning Yang, preceding paper, Phys. Rev. D 12, 3843 (1975).

[14]Cheng Ning Yang, comments after J. Tiomno's talk, session 9, *Proceedings of the Seventh Annual Rochester Conference on High-Energy Nuclear Physics, 1957* (Interscience, New York, 1957); Chen Ning Yang, in *The Physicist's Conception of Nature*, edited by J. Mehra (Reidel, Boston, 1973), p. 447.

[15]See, e.g., Shiing-shen Chern, Geometry of Characteristic Classes, Proceedings of the 13th Biennial Seminar, Canadian Mathematics Congress, 1972, p. 1.

[16]This is in sharp contrast with an interaction (if it exists), which is not related to gauge concepts.

[17]P. B. Price, E. K. Shirk, W. Z. Osborne, and L. S. Pinsky, Phys. Rev. Lett. 35, 487 (1975).

[18]There have been many papers on fiber bundles, monopoles, and quaternions in the physics literature. The following is only a partial list: Elihu Lubkin, Ann. Phys. (N.Y.) 23, 233 (1963); J. Math. Phys. 5, 1603 (1964); D. Finkelstein, J. M. Jauch, S. Schiminovich, and D. Speiser, J. Math. Phys. 4, 788 (1963); articles by J. A. Wheeler, B. S. DeWitt, A. Lichnerowicz, and C. W. Misner, in *Relativity, Groups and Topology*, edited by C. DeWitt and B. S. DeWitt (Gordon and Breach, New York, 1963); also C. Misner, K. Thorne, and J. A. Wheeler, *Gravitation* (Freeman, San Francisco, 1973); A. Trautman, Rep. Math. Phys. 1, 29 (1970); G. 't Hooft, Nucl. Phys. B79, 276 (1974); Hendricus G. Loos, Phys. Rev. D 10, 4032 (1974); J. Arafune, P. G. O. Freund, and C. J. Goebel, J. Math. Phys. 16, 433 (1975); B. Julia and A. Zee, Phys. Rev. D 11, 2227 (1975); M. K. Prasad and Charles M. Sommerfield, Phys. Rev. Lett. 35, 760 (1975).

[19]*Footnote added in proof.* Professor A. Lenard has raised an interesting question in this connection: In the continuous changes of "electric" sources, are sources quantized according to (44)? The answer to this question is "no," and requires explanations. The "electric" charges play two separate roles. They act as sources of gauge fields, and they also act as responders to gauge fields. In a physical situation the two roles are of course interrelated. In the discussion here, however, we separate the two roles. We therefore do not require quantization of electric charges as sources, but require quantization of electric charges as responders.