ANNALES DE L'I. H. P., SECTION A

J. ANANDAN

Non local aspects of quantum phases

Annales de l'I. H. P., section A, tome 49, nº 3 (1988), p. 271-286
<http://www.numdam.org/item?id=AIHPA_1988__49_3_271_0>

Ann. Inst. Henri Poincaré,
Vol. 49, n° 3, 1988, p. 271-286. *Physique théorique*

Non Local Aspects of Quantum Phases

by

J. ANANDAN (*)
Departments of Physics, Chemistry and Mathematics,
University of California, Berkeley, CA 94720
and Department of Physics, New York University, New York, NY 10003

ABSTRACT. — Non local aspects of the geometric, gauge field and gravitational quantum phases are discussed. The geometric phase is generalized to one obtained from a connection with an arbitrary Chern class over the projective Hilbert space. A new geometric interpretation is given to it as an area in the projective Hilbert space, regarded as a phase space, and its relationship to the classical phase space area is pointed out. It is argued that gauge fields provide a fundamental reason for introducing Planck's constant. Remarks are also made on the gravitational quantum phase and its implication to Weyl's theory of gravity.

RÉSUMÉ. — Les aspects non-locaux des phases quantiques géométriques, de champs de jauge et de gravitation sont discutés. La phase géométrique est généralisée à une classe arbitraire du second groupe de cohomologie de De Rham. Une nouvelle interprétation géométrique lui est donnée en terme de surface dans l'espace de Hilbert projectif, vu comme espace de phase, et sa relation à l'espace de phase classique est signalée. On soutient l'idée que les champs de jauge fournissent une raison fondamentale d'introduire la constante de Planck. Des remarques sont aussi faites sur la phase quantique gravitationnelle et ses implications au sujet de la théorie de Weyl.

(*) On leave from University of South Carolina, Columbia, SC 29208.

Annales de l'Institut Henri Poincaré - Physique théorique - 0246-0211
Vol. 49/88/03/271/16/$ 3,60/© Gauthier-Villars

1. INTRODUCTION : NON LOCAL ASPECTS OF QUANTUM THEORY

The nature of human experience makes our thinking essentially local. For example, it was believed for a long time that the earth is flat, which turned out to be a rather unreasonable extrapolation of our local experiences. Similarly, it was believed that the flat Euclidean geometry is the geometry of physical space (regarded by Immanuel Kant as being necessarily true as an « *a priori* synthetic » proposition) until Einstein's great discovery that space-time, though locally flat, is in fact curved.

Quantum phenomena have revealed, since then, new types of « curvature » which can be most simply understood by their effects on quantum phases. In 1959, Aharonov and Bohm [*1*] predicted that there would be a phase shift in the interference of two coherent wave functions of a charged particle due to an enclosed electromagnetic field, even if the wave functions are in a region in which the field strength is zero. This caused tremendous surprise and disbelief in the physics community. But it was pointed out that this is analogous to the rotation of a vector when it is parallel transported around a curve on a cone enclosing the apex, even though the cone is intrinsically flat everywhere except at the apex. This suggests that the enclosed electromagnetic field strength may be similar to the curvature of the apex of the cone which may be regarded as the cause of the rotation of the vector. Indeed, the most important lesson to be learned from the Aharonov-Bohm effect is, perhaps, that the electromagnetic field is a connection, called a gauge field, whose curvature is the field strength.

Another curvature which was overlooked for about six decades since quantum mechanics was created is due to the nature of the Hilbert space $\mathcal{H}$ itself. Even though $\mathcal{H}$ is a linear space which is flat, a physical state, as pointed out long ago by Dirac [*2*], is represented by a ray, i. e. a one dimensional subspace of $\mathcal{H}$. And the set $\mathcal{P}$ of all physical states of $\mathcal{H}$, called the projective Hilbert space, has a curvature arising from the inner product in $\mathcal{H}$. This causes the geometric phase in a cyclic evolution of a quantum system which was discovered by Berry [*3*] for adiabatic evolutions and generalized to all cyclic evolutions by Aharonov and Anandan [*4*], who also pointed out the role of the curved projective Hilbert space in producing this phase.

There are also two other non local aspects of quantum mechanics, which are not related to any « curvature », at least at present. One is the quantum correlation between two particles due to interference between products of wave functions of the particles. This was discussed qualitatively by Einstein, Podolsky and Rosen [*5*] and quantitatively by Bell [*6*], who derived a set of inequalities that a local, realistic theory must satisfy. These are violated by quantum mechanics. The second non local aspect seems to occur when a measurement is made on a quantum system. According to the usual Copenhagen interpretation, the wave function « collapses » instantaneously which, for a position measurement, must be a non local phenomenon. Also, the outcome can only be predicted statistically, so that a pure state undergoes a transition to a mixed state during the measurement [*7*].

The last mentioned phenomenon has been subject to heated controversy during the six decades since quantum mechanics was discovered. There are basically two possible ways in which we may try to avoid this strange transition from a pure state to a mixed state. We may let a pure state evolve to a pure state or a mixed state evolve to a mixed state during a measurement.

The Everett interpretation [*8*] of quantum mechanics belongs to the first category, but it pays the tremendous cost of introducing infinite number of unobservable « worlds », which is unacceptable to most physicists. Also, it is not clear how the probabilities, observed in a laboratory in our own « world », can be obtained in this purely deterministic picture. This cannot be obtained as in classical statistical mechanics which becomes statistical on « coarse graining » a more fundamental deterministic theory. For example, suppose that every electron that passes through a Stern-Gerlach apparatus is in a quantum state so that the probability of it being found to have spin up is 1/3. According to the Everett interpretation, as each electron passes through the apparatus the world splits into *two* worlds corresponding to the electron having spin up or down. In an arbitrarily chosen world, therefore, the probability of an electron having spin up is the same as the probability of it having spin down, if we define probability as a relative frequency as we normally do, which is in conflict with the fact that in the laboratory 1/3 rd of the electrons have spin up (*).

An interpretation of quantum mechanics that belongs to the second category is the « causal interpretation » of quantum mechanics [*9*]. In this view, a quantum system, such as an electron, has a definite trajectory which is guided by a non local quantum potential. But it is not possible to know which of the many possible trajectories corresponding to a given wave function that the particle actually takes without changing the wave function and therefore the quantum potential. Hence this description really represents an ensemble of particles and therefore describes a mixed state throughout because it is not possible to *predict* the trajectory of a given particle.

Perhaps, the solution to the measurement problem requires a modification of quantum theory that would be non local and non linear. This non locality would be fundamentally different from the non locality of quantum theory that will be discussed in this paper, which is really non

(*) I first heard this argument from Yakir Aharonov.

local aspects of a fundamentally local theory. An interesting question that will not be answered in this paper is whether there are any connections between the different non local aspects mentioned above.

2. THE GEOMETRIC PHASE

2.1. The Geometric Phase as a Consequence of Curvature.

Suppose a quantum system undergoes cyclic evolution in the interval $[0, \tau]$, by which we mean that its state vector $|\psi(t)\rangle$ evolves so that $|\psi(\tau)\rangle = e^{i\phi}|\psi(0)\rangle$. Then the curve γ in $\mathscr{H}$ representing the evolution projects to a closed curve C in $\mathscr{P}$. If the evolution is according to the Schrödinger equation with Hamiltonian H, then it can be shown that [*4*] $\phi = \beta + \delta$, where

$$\delta = -\hbar^{-1}\int_0^\tau \langle \psi | \mathrm{H} | \psi \rangle \, dt \tag{2.1}$$

and

$$\beta = \oint_C \mathrm{B} \tag{2.2}$$

where B is a 1-form field on $\mathscr{P}$ defined as follows: if C is a piece-wise smooth curve then we can choose an open set U of $\mathscr{P}$ containing C such that $|\tilde{\psi}\rangle$ is a differentiable function on U with values in $\mathscr{H}$ satisfying $\langle \tilde{\psi} | \tilde{\psi} \rangle = 1$. Then

$$\mathrm{B} = i\langle \tilde{\psi} | d | \tilde{\psi} \rangle \tag{2.3}$$

where d is the exterior differential operator on $\mathscr{P}$. Now, δ depends on H and is therefore called the dynamic phase. But β depends only on C and is therefore called the geometric phase. β and δ have been generalized to evolution according to any first order differential equation which may be non linear [*10*], [*11*].

More generally, we can choose an open covering $\{\mathrm{U}_\alpha\}$ of $\mathscr{P}$ such that on each U_α a differentiable function $|\tilde{\psi}_\alpha\rangle$, called a section or « gauge », satisfying $\langle \tilde{\psi}_\alpha | \tilde{\psi}_\alpha \rangle = 1$, can be defined. It is a consequence of the geometry of $\mathscr{H}$ and $\mathscr{P}$ that one section cannot cover all of $\mathscr{P}$. Define also $\mathrm{B}_\alpha = i\langle \tilde{\psi}_\alpha | d | \tilde{\psi}_\alpha \rangle$ on each U_α. On $\mathrm{U}_\alpha \cap \mathrm{U}_\beta$, if it is non empty, $|\tilde{\psi}_\alpha\rangle = \exp(if_{\alpha\beta})|\tilde{\psi}_\beta\rangle$, where $\exp(if_{\alpha\beta})$ is called a transition function. Now C is a union of segments C_α contained in U_α. Then it is easily seen that $\exp(i\beta)$ is a product of the factors $\exp\left(i\int_{\mathrm{C}_\alpha} \mathrm{B}_\alpha\right)$ and the values of the transition functions $\exp(if_{\beta\alpha})$ at each point where two neighboring segments C_α and C_β meet. Also, on using Stoke's theorem, we can write

$$\beta = \int_S G, \tag{2.4}$$

where S is a surface in $\mathscr{P}$ spanned by C and

$$G = dB = i \langle d\tilde{\psi} | \Lambda | d\tilde{\psi} \rangle . \tag{2.5}$$

In an overlapping region $U_\alpha \cap U_\beta$, G is independent of which section $| \tilde{\psi}_\alpha \rangle$ is used to evaluate G, which is why we have omitted the subscript α in the definition of G. This makes it clear that $\exp(i\beta)$ is independent of the choice of the open covering and the sections. But to keep our formalism simple, we shall work with just one section that covers C, as described in the previous paragraph.

A geometric interpretation that has already been given to the phase factor $\exp(i\beta)$ is the following: B may be regarded as the pullback of a connection with respect to the section $| \tilde{\psi} \rangle$ defined on any one of three different bundles over $\mathscr{P}$ that are described elsewhere [*4*], [*12*]. Then $\exp(i\beta)$ is the holonomy transformation or the operator that parallel transports a state vector around C with respect to this connection. Clearly, $\exp(i\beta)$ is independent of the choice of the chosen section only if C is a closed curve and is different from 1, in general, because of the curvature G of this connection. In this sense, $\exp(i\beta)$ is a non local consequence of this curvature in the same way that the Aharonov-Bohm effect is a non local consequence of the curvature of the electromagnetic gauge field.

2.2. Generalization of the Geometric Phase to an Arbitrary Chern Class in the Second De Rham Cohomology Group.

There is a beautiful global topological aspect of the relation between $\mathscr{H}$ and $\mathscr{P}$ that can already be seen in the simplest non trivial case of $\mathscr{H}$ being a two dimensional Hilbert space, so that $\mathscr{P}$ is the one dimensional complex projective space $P_1(C)$, which when regarded as a real manifold is a two dimensional sphere. Then, $\mathscr{P}$ may be given polar coordinates (θ, ϕ) and we may choose $| \tilde{\psi} \rangle = \left(\cos \frac{\theta}{2}, e^{-i\phi} \sin \frac{\theta}{2} \right)$. Then

$$B \equiv i \langle \tilde{\psi} | d\tilde{\psi} \rangle = (1/2)(1 - \cos\theta) d\phi . \tag{2.6}$$

Clearly $| \tilde{\psi} \rangle$ and B are well defined on $\mathscr{P}$ except at $\theta = \pi$. The corresponding point on the sphere can be covered by a different section or gauge $|\tilde{\psi}'\rangle = \left(e^{i\phi}\cos\frac{\theta}{2}, \sin\frac{\theta}{2}\right)$. Then

$$|\tilde{\psi}'\rangle \quad \text{and} \quad B' \equiv i\langle\tilde{\psi}'|d\tilde{\psi}'\rangle = (1/2)(-1-\cos\theta)d\phi$$

are defined everywhere on $\mathscr{P}$ except along $\theta = 0$.

Also, $G = dB$ or dB' is given everywhere on $\mathscr{P}$ by

$$G \equiv \frac{\sin\theta}{2}\, d\theta\Lambda d\phi\,. \tag{2.7}$$

Now, (2.6) and (2.7) are the vector potential and field strength of a magnetic monopole [13] of unit charge multiplied by e/hc on a sphere surrounding the monopole. This can be verified by computing the flux on this sphere:

$$\int_{\mathscr{P}} G = \int_{\mathscr{P}} \frac{\sin\theta}{2}\, d\theta d\phi = 2\pi\,. \tag{2.8}$$

More generally, if $\mathscr{H}$ has arbitrary dimension then $\int_{\mathscr{S}} G$ over any closed 2-surface $\mathscr{S}$ that is obtained by smoothly deforming an embedding of $P_1(C)$ in $\mathscr{P}$ is 0 or 2π. This means that G, like the electromagnetic field strength of a magnetic monopole, belongs to a Chern class that is an element of the second de Rham cohomology group. These classes are classified by integers, which in the electromagnetic case corresponds to the strength of the magnetic monopole. The geometric phase, as treated so far, is obtained from a curvature for which this integer is -1. This raises the question whether there is a geometric phase corresponding to a Chern class of arbitrary integer. We answer this question now in the affirmative.

Suppose there are n identical bosons in the state $|\psi\rangle$. Then $|\Psi\rangle = |\psi\rangle|\psi\rangle\ldots|\psi\rangle$, i.e. the tensor product of $|\psi\rangle$ with itself n times, represents this n-particle state. Let $|\tilde{\Psi}\rangle = |\tilde{\psi}\rangle|\tilde{\psi}\rangle\ldots|\tilde{\psi}\rangle$. The corresponding connection coefficient that replaces (2.3) is

$$B = i\langle\tilde{\Psi}|d|\tilde{\Psi}\rangle = ni\langle\tilde{\psi}|d|\tilde{\psi}\rangle, \tag{2.9}$$

on using Leibniz's rule and $\langle\tilde{\psi}|\tilde{\psi}\rangle = 1$. Similarly, the curvature

$$G = dB = i\langle d\tilde{\Psi}|\Lambda|d\tilde{\Psi}\rangle = ni\langle d\tilde{\psi}|\Lambda|d\tilde{\psi}\rangle\,. \tag{2.10}$$

Hence the geometric phase β for $|\Psi\rangle$ is n times the geometric phase for the single particle state. To obtain a geometric phase corresponding to a negative integer, take the dual of $|\Psi\rangle$, i.e. $\langle\Psi|$, which may be regarded as representing the time reversed state. Then

$$B = id(\langle\tilde{\Psi}|)|\tilde{\Psi}\rangle = -ni\langle\tilde{\psi}|d|\tilde{\psi}\rangle \tag{2.11}$$

on using Leibniz's rule and $\langle\tilde{\psi}|\tilde{\psi}\rangle = 1$.

Hence, in mathematical terms, there is a geometric phase arising from a curvature belonging to a Chern cohomology class of any non zero integer. This curvature G may be regarded as due to the canonical connection on the tensor product of the natural line bundle [*4*] over $\mathscr{P}$ or the dual of this line bundle with itself n times. Indeed, these tensor product bundles are, up to isomorphism, all possible line bundles that we can have over $\mathscr{P}$. Also, G satisfies

$$\int_{\mathscr{S}} \mathrm{G} = 2\pi n \quad \text{or} \quad 0, \tag{2.12}$$

where $\mathscr{S}$ is a closed 2-surface in $\mathscr{P}$, defined above, and n is an integer, positive or negative, that is determined entirely by the tensor product bundle chosen or, in physical terms, on the number of particles or time reversed particles in the quantum state of interest.

If the closed curve C is on $\mathscr{S}$ which it divides into $\mathscr{S}_1$ and $\mathscr{S}_2$ then (2.12) may be rewritten as

$$\beta_1 - \beta_2 = 2\pi n \tag{2.12'}$$

where β_1 and β_2 are the geometric phases evaluated on $\mathscr{S}_1$ and $\mathscr{S}_2$ using Stoke's theorem on the integral around C of B, given by (2.9) or (2.11), by means of the sections that are defined on $\mathscr{S}_1$ and $\mathscr{S}_2$. Then $\exp(i\beta_1) = \exp(i\beta_2)$ and since it is $\exp(i\beta)$ that is experimentally observed it does not matter which surface spanned by C is used to evaluate β, provided (2.12) is satisfied. Conversely, this physical requirement may be used to derive (2.12) which is a restriction that any connection on $\mathscr{P}$ must satisfy. This is analogous to the derivation [*14*] of Dirac's quantization condition for magnetic monopoles, in which case $\mathscr{P}$ is replaced by space-time and β is the Aharonov-Bohm phase.

2.3. **The Uniqueness of the Geometric Phase.**

It may be asked why we cannot split the phase change during a cyclic evolution as, say, $\phi = (2/3)\beta + \delta'$ and call $\beta' = (2/3)\beta$, which is as gauge invariant as β, the geometric phase. We shall give two answers, one mathematical and the other physical. Mathematically, the requirement (2.12) places a restriction on the connection from which the geometric phase is obtained so that the latter is an integer multiple of (2.2). This can be understood as due to the canonical connection on the tensor product bundle described above. But the « connection » represented by $\mathrm{B}' = (2/3)\mathrm{B}$ that would give β', for all cyclic evolutions, is not mathematically admissible in any bundle over $\mathscr{P}$, because it would not satisfy (2.12). On the other hand, we can add to B any 1-form field A on $\mathscr{P}$ that is differentiable everywhere and thereby obtain a connection which satisfies (2.12). But A is then arbitrary so that the corresponding connection is not naturally given by the geometry of $\mathscr{H}$ or its dual $\mathscr{H}^*$. The requirement that the geometric phase should be determined naturally by the geometry of the relevant natural tensor bundle over $\mathscr{P}$ makes it unique.

The above argument shows the advantage of the projective Hilbert space over the parameter space, which was originally used as the arena for the Berry-Simon connection [3], [15], even in the adiabatic limit. The use of (2.12) in this argument to isolate the geometric phase uniquely from the total phase is similar to Dirac's quantization [13] of the magnetic monopole that placed a restriction on the possible magnetic monopoles that could occur in nature. So far, magnetic monopoles have not been observed. But the beautiful mathematical structure that made Dirac propose its possibility is in fact realized in nature through the geometry of the Hilbert space. Nature often takes advantage of beautiful mathematics, although not necessarily in the way that physicists envisage it would.

Our second answer to the above question is that the Hamiltonian defines a natural dynamical frequency ω according to

$$\hbar\omega = \langle \psi(t) \mid \mathrm{H}(t) \mid \psi(t) \rangle . \tag{2.13}$$

This generalizes the Planck-Einstein-De Broglie law $\mathrm{E} = \hbar\omega$, that was originally formulated for a free particle in a stationary state, to arbitrary states. It is then remarkable that if the dynamical phase $-\int_0^\tau \omega dt$ is subtracted from the total phase ϕ acquired during a cyclic evolution then the remainder β is geometrical [14].

Alternatively, we may postulate the above connection on $\mathscr{P}$ so that a vector field $| \psi^p(t) \rangle$ along a curve in $\mathscr{P}$ is parallel transported with respect to this connection if

$$\left\langle \psi^p(t) \left| \frac{d}{dt} \right| \psi^p(t) \right\rangle = 0 \tag{2.14}$$

and measure the phase θ of the dynamically evolving $| \psi(t) \rangle$ with respect to $| \psi^p(t) \rangle$, i. e. $| \psi(t) \rangle = \exp(i\theta) | \psi^p(t) \rangle$. Then, Schrödinger's equation implies that $\omega = d\theta/dt$ satisfies (2.13). This is like a quantum principle of equivalence in that $| \psi(t) \rangle$, has now been given a globally well defined frequency as if it is a free particle state with energy $\langle \psi(t) \mid \mathrm{H}(t) \mid \psi(t) \rangle$. The only previous physical interpretation of $\langle \psi(t) \mid \mathrm{H}(t) \mid \psi(t) \rangle$ is that it is an ensemble average of the energies of the eigenstates of H in the mixed state obtained when a measurement of energy is performed, whereas we now have an interpretation of it for the pure state $| \psi \rangle$.

2.4. The Geometric Phase as an Area in a Phase Space.

The n complex dimensional space $\mathscr{H}$ may be regarded as a $2n$ real dimensional space. On this space, define the complex coordinates $\mathrm{Q}_j = \psi_j$ and

$P_j = i\psi_j^*$, where ψ_j are the components of $|\psi\rangle$ in some orthonormal basis of $\mathscr{H}$. Let $B \equiv i\langle\psi|d|\psi\rangle = P_j dQ_j$ and $\omega \equiv dB = dP_j \Lambda dQ_j$. (We are using Einstein's summation convention.) Then $\omega(\psi, \phi)$ is twice the imaginary part of the Hilbert space inner product of the tangent vectors ψ and ϕ. Thus, $\mathscr{H}$ may be regarded as a phase space $\{(Q, P)\}$ with a natural symplectic structure defined by ω, which is derived naturally from the inner product in $\mathscr{H}$, Now, the section $|\tilde{\psi}_\alpha\rangle$ may be regarded as a map from $U_\alpha \subset \mathscr{P}$ into $\mathscr{H}$ and the skew symmetric 2-form G on U_α is the pull back of ω with respect to $|\tilde{\psi}_\alpha\rangle$. But since G is independent of which section $|\tilde{\psi}_\alpha\rangle$ is chosen, G is non degenerate and $dG = 0$, G is a natural symplectic 2-form on $\mathscr{P}$. Hence $\mathscr{P}$ is a phase space with this natural symplectic structure.

Also, on writing $\psi_j = \psi_j^R + i\psi_j^I = \alpha_j \exp(i\phi_j)$, where ψ_j^R, ψ_j^I, α_j and ϕ_j are real $\omega = 2d\psi_j^I \Lambda d\psi_j^R = -d(\alpha_j^2)\Lambda d\phi_j$. Hence, two other convenient choices for the generalized coordinates and momenta for $\mathscr{H}$, regarded as a phase space, are *i*) $Q_j = \sqrt{2}\psi_j^R$, $P_j = \sqrt{2}\psi_j^I$ and *ii*) $Q_j = \alpha_j^2$, $P_k = -\phi_j$. These are real coordinates, unlike the earlier choice. For any of the above three pairs of (Q_j, P_j) evaluated on $|\tilde{\psi}\rangle$, defined on U containing the closed curve C, it can be easily shown that

$$\beta = \oint_C P_j dQ_j = \int_S dP_j \Lambda dQ_j. \tag{2.15}$$

Since, as mentioned, β is independent of the chosen $|\tilde{\psi}\rangle$, (2.15) is the area of S determined by the symplectic structure in $\mathscr{P}$, defined above. Thus, it is not necessary to leave the projective Hilbert space in order to give a geometric interpretation to β. Also, (2.15) is invariant under *canonical transformations* provided, of course, that C is a closed curve. This shows, again, that β is geometrical only when it is associated with a *closed* curve in the projective Hilbert space, and is therefore a non local phase.

At first sight it would seem that the symplectic structure in $\mathscr{P}$ has nothing to do with the symplectic structure in the classical phase space $\mathscr{C}$. For a particle in one dimension, for example, $\mathscr{C}$ is two dimensional whereas the phase spaces of $\mathscr{H}$ and $\mathscr{P}$ described above are infinite dimensional. However, the submanifold $\mathscr{G}$ of $\mathscr{P}$ corresponding to Gaussian wave packets peaked around all possible (q, p), with a given width $\Delta q = \xi$ and $\Delta p = \hbar^{-1}\xi$, can be identified with the classical phase space $\mathscr{C} = \{(q, p)\}$ in the classical limit in which the particle has large enough mass so that a wave packet undergoing Schrödinger evolution has negligible spread during the time interval of interest. Then, for a cyclic evolution of such a Gaussian wave packet [*22*], [*23*], to a very good approximation the geometric phase is

$$\beta = \hbar^{-1} \oint_C p_j dq_j. \tag{2.16}$$

Therefore, the usual symplectic structure on $\mathscr{C}$ may be regarded as the symplectic structure induced on $\mathscr{G}$ due to the above mentioned symplectic structure on $\mathscr{P}$. The observation of Berry and Hannay [*18*] that Hannay's angles [*19*], [*20*] and their non adiabatic generalization [*18*], [*21*] can be obtained from an area in the classical phase space averaged over a torus is now not surprising.

Eq. (2.16) suggests that

$$\eta = \oint_C p_j dq_j \tag{2.17}$$

is the classical analog [*21*] of the quantum geometric phase β. Even if we were ignorant of quantum theory, we may conclude that η is geometrical because it is a canonical invariant. However, for η to be in the exponent of a holonomy transformation of a connection, it is necessary to make it dimensionless by dividing it by a constant having the dimension of action. This raises the question of whether the geometric nature of η requires the introduction of Planck's constant. We shall return to this question in the next section.

3. GAUGE FIELDS AND THE AHARONOV-BOHM EFFECT

3.1. Structure of a Gauge Theory.

Symmetry groups were very important in physics, even prior to the introduction of gauge theory, partly due to the fact that they gave conserved quantities via Noether's theorem. This connection is even stronger in quantum field theory where the conserved quantities generate the symmetry group that acts on state vectors. It was also known that these conserved quantities or « charges » sometimes acted as sources of fields according to certain field equations. These fields in turn influence the motion of particles or fields which carry these charges. For example, the invariance of the Lagrangian of a matter field under the U(1) group of electromagnetism implied the conservation of electric charge, which is the source of the electromagnetic field. The charged fields are in turn influenced by the electro-

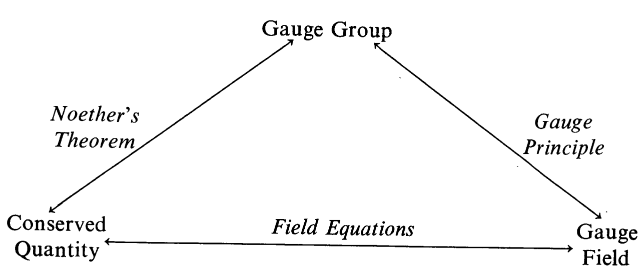

FIG. 1. — The Gauge Triangle.

magnetic field, as in the Aharonov-Bohm effect [*1*]. The profound importance of the « gauge principle », due to Weyl [*16*] and Yang and Mills [*17*], lies in the fact that it completes the third side of this triangle by establishing the field as a connection corresponding to the group (fig. 1), which are therefore called gauge field and gauge group respectively.

The gauge principle was stated by Weyl and Yang-Mills as the enlargement of the global gauge group G to a local gauge group, consisting of transformations of the form $\psi(x) \to U(x)\psi(x)$, $U(x) \in G$. This leads to the introduction of gauge potentials as compensating fields for the Lagrangian to be invariant. But perhaps a more physical way of stating this principle [*24*] is that, due to the locality of the laws of physics, we cannot compare the directions of $\psi(x)$ with $\psi(x')$ when x and x' are two different space-time points, unless we introduce a connection so that $\psi(x')$ can be parallel transported along a curve joining x and x' and compared with $\psi(x)$.

The necessity to compare them, formally, arises from the need to take derivatives of $\psi(x)$ in forming the Lagrangian, whose (global) gauge invariance leads to the conserved quantities and which also yields the field equations. When the parallel transport between two points is path dependent, a non trivial gauge field is present. More generally, by gauge principle we may simply mean the introduction of a connection. On the other hand, $\psi^+(x)\psi(x)$ represents the probability density (which should not depend on the path along which $\psi(x)$ is parallel transported because number is a global concept unlike the direction of a vector which is a local concept.

The triangle in fig. 1 is essential for a gauge theory, and may be used as its definition. We shall call it the gauge triangle. If we encounter one side of this triangle then it is reasonable to look for the other two sides in order to see if we have a gauge theory. But there are four important examples in which we know of only one or two sides of this triangle. One is the isospin conservation in strong interactions which corresponds to the first side of this triangle with SU(2) as the symmetry group. It was this invariance that originally led Yang-Mills [*17*] to introduce non abelian gauge fields. But so far there is no clear evidence for the other two sides of this triangle and strong interactions are now explained by a gauge theory based on the color SU(3) gauge group.

A second example is the magnetic monopole [*13*] which would be an additional conserved quantity that generates a field, but it is not necessary to introduce a new U(1) group and the new electromagnetic field may be regarded as a connection of the old U(1) gauge group. Indeed, the presence of a clear gauge triangle in electromagnetism only if no magnetic monopoles exist may be used as an argument against introducing magnetic charges, which have not been observed so far.

A third example is the geometric connection described in section 2 which, as mentioned, contains the beautiful mathematics of magnetic monopoles, But the other two sides of the gauge triangle seem to be missing. This is also different from gauge fields in that it is a connection on the projective Hilbert space and not on space-time.

Finally, for gravitation, which we shall deal with in the next section, the first two sides of the triangle are present. In general relativity, the energy-momentum which is the conserved quantity of translation invariance acts as the source of space-time curvature. In the modified Sciama-Kibble theory [*25*] and its generalizations, energy-momentum and spin, which are the conserved quantities of the Poincare group, are the sources of curvature and torsion. But it would be fair to say that the third side of the gauge triangle has not been clearly formulated [*26*].

3.2. Aharonov-Bohm Effect, Charge Quantization and Planck's Constant.

Electromagnetism is the simplest and the most dramatic example of a gauge theory. The U(1) symmetry group not only gives a conserved charge but also requires that this charge must be quantized, which is due to the compactness [*27*] of the gauge group. Also, the gauge principle then implies the existence of the electromagnetic field which must interact with the charge to complete the gauge triangle. Thus the *existence, conservation, quantization and interaction* of the electric charge all follow from the chosen gauge group. Also, the path dependence of parallel transport in the electromagnetic connection results in the Aharonov-Bohm effect which is a non local manifestation of the hypothesis that electromagnetism is a U(1) gauge field [*28*].

Wu and Yang [*29*] have argued, using the Aharonov-Bohm effect, that the complete description of the electromagnetic field is provided by the phase factor

$$\exp\left(i\frac{e}{\hbar c}\oint \mathrm{A}_\mu dx^\mu\right) \tag{3.1}$$

which parallel transports $\psi(x)$ around a closed curve (holonomy transformation) and physically determines the fringe shift in the Aharonov-Bohm experiment. This can be generalized to an arbitrary gauge field by the generalization [*30*] of the Aharonov-Bohm effect for an arbitrary gauge field and the theorem [*31*] that the gauge potential can be reconstructed from the holonomy transformations and it is then unique up to gauge transformations.

Now the Aharonov-Bohm effect is a quantum phenomenon. But even in classical electromagnetism, we can argue that

$$\mathrm{I} = \frac{e}{c}\oint \mathrm{A}_\mu dx^\mu \tag{3.2}$$

has more information than the field strength $F_{\mu\nu}$ in a non simply connected region and it is invariant under the gauge transformation $A_\mu \rightarrow A_\mu - \partial_\mu \Lambda$. However, from this we cannot conclude that the electromagnetic field is a gauge field (connection) because the exponent of the holonomy transformation of a gauge field, which is an element of a Lie algebra, must be dimensionless. So, in order to describe the electromagnetic field as a gauge field, we need to introduce a new constant, $\hbar$, having the dimension of the action so that $I/\hbar$ is dimensionless.

The above argument, which seems to suggest a fundamental reason for introducing the Planck's constant h, does not tell us if the gauge group of electromagnetism is U(1) or T(1), the translation group in one dimension which is non compact. The holonomy transformation

$$\exp\left(\frac{e}{\hbar c}\oint A_\mu dx^\mu\right) \tag{3.3}$$

would correspond to the T(1) group, unlike (3.1) which corresponds to the U(1) group. But we note now the empirical fact that charge is quantized. This implies that the homomorphism defined on the loop group in the reconstruction theorem [*31*], mentioned above, gives a representation [*24*] of a fundamental U(1) group if (3.1) is used instead of (3.3) as the holonomy transformation. With this choice, the reconstruction theorem implies that electromagnetism is a U(1) gauge field.

Now, (3.2) is similar to (2.17) if we regard $(e/c)A_\mu$ as a « potential energy-momentum » due to the electromagnetic field. So, we may conclude in the spirit of the above arguments that in order for the area in the classical phase space, which is a canonical invariant, to be proportional to the exponent of the holonomy transformation of a geometric connection, we must introduce Planck's constant.

4. GRAVITATIONAL QUANTUM PHASE

It is sometimes said that the difference of phases at two different space-time points is not well defined because of local gauge invariance. This statement needs to be qualified. Consider for simplicity a scalar field for which the wave function is $\psi = \alpha \exp(i\phi)$ where $\alpha(x)$ and $\phi(x)$ are real scalar functions of space-time. If a particle has charge q, then $\Delta\phi = \phi(x') - \phi(x)$ is not invariant under the local gauge transformation $\phi(x) \rightarrow \phi(x) + q\Lambda(x)$. However, $\Delta\phi$ can be split into the electromagnetic phase $-\frac{q}{\hbar c}\int A_\mu dx^\mu$ and the remainder

$$\gamma = \int_{C_{xx'}} \left(\partial_\mu \phi + \frac{q}{\hbar c} A_\mu\right) dx^\mu \tag{4.1}$$

where $C_{xx'}$ is a path joining x and x'. Then γ is a path dependent but gauge invariant phase difference. We shall call it the gravitational phase because in the WKB approximation $p_\mu = -\hbar\left(\partial_\mu \phi + \frac{q}{\hbar c} A_\mu\right)$ is the energy-momentum which determines the inertial mass and couples to the gravitational field because of the equivalence of inertial and active gravitational masses.

The existence of this gauge invariant phase associated with open curves C distinguishes gravity from gauge fields. This γ is a local quantity unlike the Aharonov-Bohm phase or the geometric phase β that depends on information around an entire closed curve and is therefore non local. To understand the origin of this difference, consider a wave function in the WKB limit and choose γ to be an integral curve of p^μ, which is a possible classical trajectory. Then using the eikonal equation $p_\mu p^\mu = m^2c^2$,

$$\gamma = -\int_C \frac{mc}{\hbar}\, ds \tag{4.2}$$

i. e.-γ is the length along the path in units of Compton wave length of the particle. (4.2) has been experimentally observed for a closed curve C in the non relativistic limit in neutron interference [*32-35*] and in the relativistic case, for a charged particle, using superconductors [*36*], [*37*]. (4.2) suggests that the reason for γ to be an invariant for open curves is because distance is meaningful for open curves in general relativity.

Weyl [*38*], on the other hand, proposed a generalization of general relativity in which distance is subject to the same type of gauge invariance that the electromagnetic phase is subject to. We noted in section 3 that the direction of a vector is a local concept whereas the magnitude of a vector is a global concept for gauge fields. This is true also for gravitation, as described in general relativity, except that the vectors involved are tangent vectors. This results in the invariant length along any path. But Weyl removed the restriction that the magnitudes of two vectors at x and x' can be compared in a path independent way. If Weyl's theory is correct, then we cannot define a gauge invariant observable γ associated with open curves as we did above. However γ associated with an open curve is measured by the Josephson effect [*24*] and, in the WKB approximation, in kaon decay [*35*]. Hence, these phenomena are experimental evidences against Weyl's theory.

It may be noted that in the determination of the geometry of space-time by the classical motions of freely falling particles by Ehlers, Pirani and Schild [*39*], the Weyl Structure is obtained naturally, but to obtain the Riemannian structure it was necessary to introduce the additional assumptions of the absence of « the second clock effect » and vanishing torsion. In quantum mechanics, the existence of the invariant, observable gravitational phase γ provides another, perhaps a more elegant, reason for eliminating the gauge freedom of the metric which Weyl postulated. But as far as we can see, quantum mechanics provides no natural reason for setting torsion to be zero.

There is, however, a non local phase due to the gravitational field. This is the phase shift due to the coupling of spin to space-time curvature [34], [30] of a connection, that may contain torsion, which can be understood as due to parallel transport of a spinor or a vector around a closed curve through the interfering beams. This is a gravitational analog of the Aharonov-Bohm effect.

This raises the question of whether the curvature that is measured by this effect is a non local concept. To answer this question, it is useful to distinguish between intrinsic and extrinsic curvature. The cone that was used in section 1 to illustrate the Aharonov-Bohm effect has zero intrinsic curvature, except at the apex; but its extrinsic curvature due to its embedding in a three dimensional space is non zero everywhere. The extrinsic curvature of a manifold can be recognized by a local experiment such as the deviation between a geodesic in the manifold and a geodesic in the embedding space which initially have the same tangent vector. But the measurement of intrinsic curvature, e. g. curvature of space-time, needs a non local experiment such as parallel transporting a vector around a closed curve.

The curvature of the earth, mentioned at the beginning of this article, was discovered by observations such as the gradual disappearance of a ship as it sails away from the shore or the ancient Greek measurement based on the simultaneous shadows of vertical objects at two different locations on the earth's surface. These were really measurements of extrinsic curvature. This may be why it was easier to convince people of the fact that the earth is curved than of the reality of the Aharonov-Bohm effect.

ACKNOWLEDGEMENTS

It is a pleasure to thank A. Bhattacharjee, J. Block, J. Marsden, A. Pines and especially A. Shimony for stimulating conversations.

REFERENCES

[1] Y. Aharonov and D. Bohm, *Phys. Rev.*, t. **115**, 1959, p. 485.
[2] P. A. M. Dirac, *The Principles of Quantum Mechanics* (Oxford, 1930).
[3] M. V. Berry, *Proc. Soc. London Ser.*, t. **A392**, 1984, p. 45.
[4] Y. Aharonov and J. Anandan, *Phys. Rev. Lett.*, t. **58**, 1987, p. 1593.
[5] A. Einstein, Podolsky and N. Rosen, *Phys. Rev.*, t. **47**, 1935, p. 777.
[6] J. S. Bell, *Physics*, t. **1**, 1964, p. 195.
[7] J. Von Neumann, *Mathematical Foundations of Quantum Mechanics* (Princeton, 1955).
[8] H. Everett, *Rev. Mod. Phys.*, t. **29**, 1957, p. 454.

[9] See, for example, J.-P. VIGIER, C. DEWDNEY, P. R. HOLLAND and A. KYPRIANIDIS in *Quantum Implications*, edited by B. J. Hiley and F. D. Peat (Routledge and Paul, London, New York, 1987).
[10] J. GARRISON and R. Y. CHIAO, *Phys. Rev. Lett.*, t. **60**, 1988, p. 165.
[11] J. ANANDAN, *Phys. Rev. Lett.*, t. **60**, 1988, p. 2555.
[12] D. N. PAGE, *Phys. Rev.*, *t.* **A 36**, 1987, p. 3479.
[13] P. A. M. DIRAC, *Proc. Roy. Soc. London Ser.*, t. **A 133**, 1931, p. 60.
[14] J. ANANDAN, *Int. J. Th. Phys.*, t. **19**, 1980, p. 537.
[15] B. SIMON, *Phys. Rev. Lett.*, t. **51**, 1983, p. 2167.
[16] H. WEYL, *Gruppen Theorie und Quanten Mechanik* (Leipzig, S. Hirzel, 1929).
[17] C. N. YANG and R. L. MILLS, *Phys. Rev.*, t. **96**, 1954, p. 191.
[18] M. V. BERRY and J. H. HANNAY, *J. Phys.*, t. **A 21**, 1988, L325.
[19] J. H. HANNAY, *J. Phys.*, t. **A 18**, 1985, p. 221.
[20] M. V. BERRY, *J. Phys.*, t. **A 18**, 1985, p. 15.
[21] J. ANANDAN, *Phys. Lett.*, t. **A 129**, 1988, p. 201.
[22] Y. AHARONOV and M. VARDI, *Phys. Rev.*, t. **D 21**, 1980, p. 2235.
[23] J. ANANDAN and Y. AHARONOV, *Submitted to phys. Rev.* (1988).
[24] J. ANANDAN, *Phys. Rev.*, t. **D 33**, 1986, p. 2280.
[25] D. W. SCIAMA in *Recent Developments in General Relativity* (Oxford, 1962); T. W. B. KIBBLE, *J. Math. Phys.*, t. **2**, 1961, p. 212.
[26] See, however, J. ANANDAN in *Quantum Theory and Gravitation*, edited by A. R. Marlow (Academic Press, New York, 1980).
[27] C. N. YANG, *Phys. Rev.*, t. *D* **1**, 1970, p. 2360.
[28] See the contributions of Y. AHARONOV and C. N. YANG in *Foundations of Quantum Mechanics in the Light of New Technology*, edited by S. Kamefuchi *et al.* (Physical Society of Japan, 1984).
[29] T. T. WU and C. N. YANG, *Phys. Rev.*, t. **D 12**, 1975, p. 3843.
[30] J. ANANDAN, *Il Nuov. Cim.*, t. **53 A**, 1979, p. 221.
[31] J. ANANDAN in *Conference on Differential Geometric Methods in Theoretical Physics*, edited by G. Denardo and H. D. Doebner (World Scientific, 1983).
[32] R. COLELLA, A. W. OVERHAUSER and S. A. WERNER, *Phys. Rev. Lett.*, t. **34**, 1975, p. 1472.
[33] B. LINET and P. TOURRENC, *Can. J. Phys.*, t. **54**, 1976, p. 1129.
[34] J. ANANDAN, *Phys. Rev.*, t. **D 15**, 1977, p. 1448.
[35] L. STODOLSKY, *Gen. Rel. and Grav.*, t. **11**, 1979, p. 391.
[36] J. ANANDAN, *Phys. Lett.*, t. **A 105**, 1984, p. 280 ; J. ANANDAN in *New Techniques and Ideas in Quantum Measurement Theory*, Annals of New York, Academy of Sciences, t. **480**, 1986, p. 224, edited by D. Greenberger.
[37] A. K. Jain, J. E. Lukens and J.-S. TSAI, *Phys. Rev. Lett.*, t. **58**, 1987, p. 1165.
[38] H. WEYL, Sitzungsberichte d. k. Preuss. *Akad. d. Wiss.*, 1918, p. 465.
[39] J. EHLERS, F. A. E. PIRANI and A. SCHILD, Papers in Honour of J. L. Synge, edited by L. O'Raifeartaigh (Clarendon Press, Oxford, 1972).

(Manuscrit reçu le 10 juillet 1988)