IJMMS 2004:34, 1777–1836
PII. S0161171204304400
http://ijmms.hindawi.com

GEOMETRICAL AND TOPOLOGICAL FOUNDATIONS OF THEORETICAL PHYSICS: FROM GAUGE THEORIES TO STRING PROGRAM

LUCIANO BOI

Received 20 April 2003 and in revised form 27 October 2003

We study the role of geometrical and topological concepts in the recent developments of theoretical physics, notably in non-Abelian gauge theories and superstring theory, and further we show the great significance of these concepts for a deeper understanding of the dynamical laws of physics. This work aims to demonstrate that the global topological properties of the manifold's model of spacetime play a major role in quantum field theory and that, therefore, several physical quantum effects arise from the nonlocal metrical and topological structure of this manifold. We mathematically argue the need for building new structures of space with different topology. This means, in particular, that the "hidden" symmetries of fundamental physics can be related to the phenomenon of topological change of certain classes of (presumably) nonsmooth manifolds.

2000 Mathematics Subject Classification: 14-xx, 55-xx, 81-xx, 83-xx.

1. Introduction. We analyze the role of geometrical and topological concepts in the developments of theoretical physics, especially in gauge theory and string theory, and we show the great significance of these concepts for a better understanding of the dynamics of physics. We claim that physical phenomena very likely emerge from the geometrical and topological structure of spacetime. The attempts to solve one of the central problems in twentieth century theoretical physics, that is, how to combine gravity and the other forces into a unitary theoretical explanation of the physical world, essentially depend on the possibility of building a new geometrical framework conceptually richer than Riemannian geometry. In fact, this geometrical framework still plays a fundamental role in non-Abelian gauge theories and in superstring theory, thanks to which a great variety of new mathematical structures has emerged. A very interesting hypothesis is that the global topological properties of the manifold's model of spacetime play a major role in quantum field theory and that, consequently, several physical quantum effects arise from the nonlocal metrical and topological structure of these manifold. Thus the unification of general relativity and quantum theory requires some fundamental breakthrough in our understanding of the relationship between spacetime and quantum process. In particular the superstring theories lead to the guess that the usual structure of spacetime at the quantum scale must be dropped out from physical thought. Non-Abelian gauge theories satisfy the basic physical requirements pertaining to the symmetries of particle physics because they are geometric in character. They profoundly elucidate the fundamental role played by bundles, connections, and curvature in explaining the essential laws of nature. Kaluza-Klein theories and, more remarkably,

superstring theory showed that spacetime symmetries and internal (quantum) symmetries might be unified through the introduction of new structures of space with a different topology. This essentially means, in our view, that "hidden" symmetries of fundamental physics can be related to the phenomenon of topological change of a certain class of (presumably) nonsmooth manifolds.

2. The geometrization of theoretical physics: from Cartan's theory of gravitation to geometric quantum theories. This expository article, which summarizes the main subject of a book in progress on the same topic, is aimed at analyzing some of the most important mathematical developments and the conceptual significance of the *geometrization of theoretical physics*, from the work of Cartan and Weyl to the recent non-Abelian gauge theories. The starting point of our reflections is the question of how to characterize the properties of space (topological and algebraic invariants, group structures, symmetries and symmetry breaking) at the quantum level physics. More generally, we will try to highlight some striking aspects of the mathematical developments inspired by the attempts to solve one of the central problems in twentieth century theoretical physics: *how to combine general relativity and quantum field theory into a unitary theoretical description of the physical world*. Another point, which is in all likelihood intimately connected to the above, is the question of how to determine the topological (global) structure of the universe, as well as the physical conditions for its early formation. Finally, we seek to outline some theoretical remarks which raised the recent developments in theoretical physics concerned by the above questions.

Moreover, these two questions lead to the fundamental issue of the nature of space and spacetime: is it a purely formal structure, or does it include a generative principle for physical phenomena? What relation is there among the physical properties of microscopic and macroscopic matters, the kind of extended (or pointless) objects they yield, and features of space into which they are embedded? Generally, an answer to these fundamental questions and an explanation of the *basic aporias* such as continuous/discrete, local/global, deterministic/nondeterministic, linear/nonlinear, depend on a satisfactory geometric theory whose concepts are somewhat different from the ones underlying the progress of physics at the beginning of this century (general relativity and quantum mechanics). In particular, it seems necessary to build a geometry conceptually richer than Riemannian geometry. This has been partly achieved in the last two decades, and we can now see the possibility of unifying theory of gravitation with quantum mechanics. The enriched geometry plays a basic role in non-Abelian gauge theories and in superstring theory, for which a great variety of new mathematical structures has emerged.

This more general post-Riemannian geometry is based upon two very interesting ideas I would like now to stress:

(1) space has ten or eleven dimensions—according to which we deal with superstring theory or supergravity—rather than four, an assumption made more plausible by internal mathematical reasons as well as experimental physical evidence;
(2) the structure of spacetime at the quantum level is not that of a differentiable manifold C^∞, but apparently the equivalent of an arbitrary topological space

constructed from a complex (infinite-dimensional) Riemann surface and on which some fundamental mathematical objects are defined.

The latter hypothesis implies particularly that the global topological properties of the (Lorentzian or Riemannian) manifold M play a major role in quantum field theory and that, consequently, several (physical) quantum effects arise from the nonlocal metrical and topological structure of M. It seems reasonable to think that general relativity and quantum theory are intrinsically incompatible and that, rather than merely developing technique, what is required is some fundamental breakthrough in our understanding of the relationship between spacetime structure and quantum process.

Concerning the first idea, one may add that, in fact, eleven-dimensional spacetime recommends itself as the habitat of the maximal supergravity theory. Remarkably, there is also a phenomenological argument for eleven as the minimal number of spacetime dimensions, which was pointed out by Witten. If the familiar $SU(3)_{\text{color}} \times (SU(2) \times U(1))_{\text{electroweak}}$ gauge symmetry in four dimensions is to originate in isometries of a compact manifold in N "hidden" dimensions, then these extra dimensions must be at least seven in number. This follows from the observation that no manifold of dimension three or smaller can have more than six isometries, and thus the eight-parameter group $SU(3)$ can most economically appear as the isometry group of the four-dimensional manifold $CP(2)$. Similarly an $SU(2) \times U(1)$ gauge symmetry in four dimensions is most economically obtained from the isometries of the three-dimensional manifold $S^2 \times S^1$. Thus the gauge group of low energy physics is obtainable from the isometries of the seven-dimensional manifold $CP(2) \times S^2 \times S^1$. This is not an Einstein manifold, however, and as such, it is not relevant as the compactification of eleven-dimensional supergravity. Fortunately coset spaces of type $SU(3) \times SU(2) \times (U(1)/SU(2)) \times U(1) \times U(1)$ other than $CP(2) \times S^2 \times S^1$ exist (S^7 for instance) and some of these are Einstein's manifolds. Alas, because there is no supersymmetric Yang-Mills theory in eleven dimensions, the fermion spectrum is necessarily nonchiral.

One of the most remarkable discoveries in the last decades is that bosons and fermions can be placed in the same multiplet of a "supergroup" whose infinitesimal parameters contain anticommuting elements [59]. Such a theory predicts a boson-fermion mass degeneracy that is not observed in nature and thus the supersymmetry must be broken. The Goldstone fermions associated with spontaneous breaking have the wrong property to be neutrinos and hence the symmetry needs to be implemented as a local gauge invariance with the Higgs-Kibble mechanism in action. On the other hand, the construction of a successful quantum theory of gravity seems to depend largely on our capacity to give an answer to the following questions: is there some profound breakdown of spacetime continuum and quantum concepts at the Planck length (10^{-33} cm)? Some of the current geometrical and physical works have this idea. In particular, the superstring theories lead to the guess that the usual structure of spacetime (to which we have been used since the general relativity) at the Planck length must be dropped out from physical thought.

The main issue can be put in the following terms. How much of the mathematical and conceptual structure of classical general relativity do we expect to retain? In particular, one thinks of the underlying smooth C^∞ manifold, the metric tensor, the local field

equations $G_{\mu\nu} - \Delta g_{\mu\nu} = T_{\mu\nu}$, the global topological properties of the manifold and global metric features such as lightcone structure and the existence of event horizons. It is very difficult to judge how many of these classical concepts should be present, in some form or another, in a quantized theory. On the other hand, how much of the technical and conceptual structure of conventional quantum field theory do we expect to retain? For example, does the usual idea of a local quantum field $\phi(x)$ make any sense at all or should we decide from the outset that, at a Planck scale, spacetime structure is not that of a smooth manifold and therefore the local properties of fields become very unconventional? Similarly, what remains of the local commutativity of quantum fields in a theory where the lightcone is determined by the metric tensor, which is itself a dynamical variable? To sum up all that precedes, two questions then naturally arise.

(i) Which kind of mathematical framework could underly the quantum field theory? If it is a topological space, then the latter is not a smooth manifold.

(ii) If "something peculiar" happens to spacetime topology at Planck distances, the possibility arises that spaces should be considered not to be in any sense differentiable manifolds.

Attempts to solve the above problems have given rise to a fertile program of geometrization which in turn has increasingly influenced theoretical physics, especially quantum field theory, gauge theory, and string theory, as well as branches of mathematics, notably differential and algebraic geometry and topology. This idea of geometrization, which is already present in Riemann's work on abstract manifolds (endowed with metrics) and "Riemannian surfaces," and in Poincaré's work on the topology of differentiable manifolds and the geometrical theory of differential equations, has in our time taken on new directions and greater importance. However, in some fundamental work appeared in 1918–1930, E. Cartan and H. Weyl introduced some concepts and theories which subsequently gave rise to many recent and important contributions to the comprehension of the relation between geometry and theoretical physics. These ideas included projective, affine, and conformal connections and fibre spaces. The global theory of connections over a differentiable fibre space was created by Ch. Ehresmann around 1940–1950. (On this subject, see [8]).

In this perspective, one may recall that in 1923 Cartan proposed to modify the Einstein theory of gravitation by allowing spacetime to have torsion and relating it to the density of intrinsic angular momentum of a continuous medium (see [26, 53]). The idea of connecting torsion to spin has known new developments around 1960, mainly thanks to the work of D. W. Sciama and T. W. B. Kibble. There was considerable interest in this problem from 1966 to 1976. All available evidence from experiments in macrophysics attests to the validity of Einstein's general theory of relativity as a description of this interaction. The need to propose alternative or more general gravitational theories stems from a dichotomy in theoretical physics. Strong, electromagnetic, and weak interactions find their successful description within the framework of relativistic quantum field theory in flat Minkowski spacetime. These quantum fields reside in spacetime but are separate from it. Gravitation, according to Einstein, deforms Minkowski space and inheres in the dynamic Riemannian geometry of spacetime. One branch of fundamental physics is highly successful in a flat and rigid spacetime, but gravitation

requires a nonflat and dynamic spacetime. This state of affairs seems, at least from a theoretical point of view, to be unsatisfactory. Stated differently, there is no "logical" or experimental compelling need to modify Einstein's theory, but one can advance good heuristic arguments in favor of the Cartan idea.

(i) The geometrical independence of the metric g and linear connection Γ leads to the idea of treating these quantities as independent variables in the sense of a principle of least action. If g and Γ are assumed to be compatible, then the freedom in the choice of Γ reduces to that of the torsion tensor Q.

(ii) According to relativistic quantum theory, the Poincaré group—or the inhomogeneous Lorentz group—is physically more significant than the Lorentz group itself. The Poincaré group has two fundamental invariants: mass and spin. The first of them is related to translations and to energy momentum. In Einstein's theory, the density of energy momentum is source of curvature whereas spin has no such direct dynamical significance. In a sense, Einstein-Cartan theory restores—to some extent—the symmetry between mass and spin. It introduces also an unexpected "duality": via Noether's theorem, energy momentum is generated by translations whereas Einstein's equation relates it to curvature, which is responsible for rotations of vectors undergoing parallel transport. Conversely, spin is generated by rotations, but torsion induces translations in the tangent space to a manifold ("Cartan displacement"). This duality can be traced to the fact that the Einstein-Cartan Lagrangian is linear in curvature.

(iii) There is an interesting analogy between the description of magnetic moments in electrodynamics and spin in the theory of gravitation. In a phenomenological description of electromagnetism, the external magnetic field produced by a ferromagnet may be obtained in at least three ways: by considering a surface current equivalent to the actual distribution of microscopic currents and magnetic moments, by replacing the latter by a volume distribution of "Ampère currents," or, finally, by introducing a smooth field of the magnetization vector. In the Einstein theory, there are analogues for the first two descriptions, whereas the Einstein-Cartan theory provides the third.

The Einstein-Cartan theory assumes, as a model of spacetime, a four-dimensional manifold with a linear connection Γ compatible with a metric tensor g. The gravitational part of the Lagrangian, $\sqrt{-g}R$, is formed from the curvature tensor of Γ. The left-hand sides of the field equations are obtained by varying this Lagrangian with respect to g and Q. Variation with respect to g may be replaced by that relative to the field of frames. The sources of gravitational field are described by expressions resulting either from phenomenology or by varying an action integral obtained by applying the principle of minimal gravitational coupling to a special-relativistic Lagrangian. The Einstein-Cartan equations are

$$R_{\mu\nu} - \frac{1}{2} g_{\mu\nu} R = \frac{8\pi G}{c^4 T_{\mu\nu}}, \tag{2.1}$$

$$Q^{\theta}_{\mu\nu} - \delta^{\theta}_{\mu} Q^{\sigma}_{\sigma\nu} - \delta^{\theta}_{\nu} Q^{\sigma}_{\mu\sigma} = \frac{8\pi G}{c^3 s^{\theta}_{\mu\nu}}. \tag{2.2}$$

The Cartan equation (2.1) is trivial in the sense that if the spin density vanishes, $s^{\theta}_{\mu\nu} = 0$, then so does torsion, $Q^{\theta}_{\mu\nu} = 0$. Quite independently of this, torsion is topologically

trivial: any linear connection can be deformed into a connection without torsion. The Einstein-Cartan theory may be physically relevant only when the density of energy is of the same order of magnitude as the spin density squared. For matter consisting of particles of mass m and spin $h/2$, this will occur at densities of order m^2c^4/Gh^2.

3. Introduction to Kaluza-Klein theories. In another direction, there have been in the 1920s very interesting attempts by Theodor Kaluza then by Oskar Klein to unify the relativistic theory of gravitation with Maxwell's theory by introducing a new geometrical framework within which electromagnetism could be coupled with gravity (at least theoretically). The Kaluza-Klein theories are purely geometrical in character and have been worked out in order to encompass two apparently inconsistent physical theories into a unitary theoretical explanation. Actually, even before Einstein's general relativity, the physicist Gunnar Nordström in 1914 proceeded to unify his theory of gravitation (in which gravity was described by a scalar field coupled with the trace of the energy momentum tensor) with Maxwell's theory in a most imaginative way. Inspired by Minkowski's four-dimensional spacetime continuum, Nordström added yet another space dimension, thus obtaining a flat five-dimensional world. There he introduced an Abelian five-vector gauge field for which he wrote down the Maxwell equations including a conserved five-current. He then identified the fifth component of the five-vector potential with scalar gravity, whereas he identified the first four components of the five-vector potential with the Maxwell four-potential. With these interpretations he then noticed that in the cylindrical case (when all dynamical variables become independent of the fifth coordinate) the equations of his five-dimensional Maxwell theory reduced to those of the four-dimensional Maxwell-Nordström electromagnetic gravitational theory. It is then fair to say that higher-dimensional unification starts with Nordström, who assumed scalar gravity in our four-dimensional world to be a remnant of an Abelian gauge theory in a five-dimensional flat spacetime.

The next step was taken by the mathematician Theodor Kaluza in 1919 in the wake of Einstein's general relativity. Kaluza proposed that one pass to an Einstein-type theory of gravity in five dimensions, from which ordinary four-dimensional Einstein gravity and Maxwell electromagnetism are to be obtained upon imposing a cylindrical constraint. More precisely, what this amounts to is starting with a five-dimensional manifold M which is the product of $M^4 \times S^1$ of a four-dimensional spacetime M^4 with a circle S^1. The metric $\gamma_{mn}(x,y)$ on the five-manifold M $(m,n = 0,1,2,3,5)$ is a function of both the coordinates x^μ $(\mu = 0,1,2,3)$ on M^4 and $y \equiv x^5$, the coordinate of the circle S^1. It is convenient to replace the fifteen field variables $\gamma_{mn}(=\gamma_{nm})$ by fifteen new field variables $g_{\mu\nu} = g_{\nu\mu}, A_\mu, \phi$ according to the field redefinitions

$$\begin{aligned} \gamma_{\mu\nu} &= g_{\mu\nu} + e^2\kappa^2\phi A_\mu A_\nu, \\ \gamma_{\mu 5} &= \gamma_{5\mu} = e\kappa\phi A_\mu, \\ \gamma_{55} &= \phi. \end{aligned} \tag{3.1}$$

All field quantities, old and new, are periodic functions of the coordinate y on the circle. If $y = \rho\theta$, where θ is the usual angular coordinate and ρ the radius of the circle,

then the period is $2\pi\rho$. Thus any field quantity $F(x,y)$ (F being any of the $g_{\mu\nu}$'s, A_μ's, ϕ's, or γ_{mn}'s) admits a Fourier expansion

$$F(x,y) = \sum_{n=-\infty}^{+\infty} F^{(n)}(x) e^{iny/\rho}. \tag{3.2}$$

Kaluza assumed the five-dimensional dynamics to be governed by a gravitational Einstein-Hilbert action

$$I_5 = -\frac{1}{16\pi G_5} \int \sqrt{|\gamma_5|}\, R_5 d^5x, \tag{3.3}$$

with $\gamma_5 = \det(\gamma_{mn})$, R_5 the five-dimensional curvature scalar, and G_5 a five-dimensional counterpart of the gravitational constant. Using the Fourier expansions, the y-dependence becomes explicit so that the y-integration can be carried out. A four-dimensional action involving an infinity of fields—the Fourier components $A_\mu^{(n)}$, $g_{\mu\nu}^{(n)}(x)$, $\phi^{(n)}$—then emerges. At this point Kaluza imposed a "cylindricity" condition: he truncated the action by dropping all harmonics with $n \neq 0$, retaining only the zero modes:

$$g_{\mu\nu}(x,y) = g_{\mu\nu}^{(0)}(x), \qquad A_\mu(x,y) = A_\mu^{(0)}(x), \qquad \phi(x,y) = \phi^{(0)}(x). \tag{3.4}$$

The five-dimensional line element then takes the form

$$ds_5^2 \equiv \gamma_{mn} dx^m dx^n = ds_4^2 + \phi^{(0)}(x)\left(dx^5 + e\kappa A_\mu^{(0)}(x) dx^\mu\right)^2, \tag{3.5}$$

where

$$ds_4^2 \equiv g_{\mu\nu}^{(0)}(x) dx^\mu dx^\nu \tag{3.6}$$

is the four-dimensional line element corresponding to the metric $g_{\mu\nu}^{(0)}(x)$. The line element (3.6) is invariant under the transformations

$$\begin{aligned} x_5^\mu &\longrightarrow x_5^\mu, \\ x &\longrightarrow x + e\kappa\alpha(x^\rho), \\ A_\mu^{(0)} &\longrightarrow A_\mu^{(0)} - \partial_\mu \alpha(x^\rho), \\ \phi^{(0)} &\longrightarrow \phi^{(0)}, \\ g_{\mu\nu}^{(0)} &\longrightarrow g_{\mu\nu}^{(0)}, \end{aligned} \tag{3.7}$$

which we recognize as Abelian gauge transformations à la Weyl (see Section 5). Here these transformations assume a geometrical meaning as shifts in the fifth coordinate by an amount $\alpha(x^\rho)$, which depends on ordinary four-spacetime. The Abelian gauge symmetry in four dimensions originates in the isometries of the small circle in the fifth dimension.

When the y integration is carried out with the cylindric truncation enforced, the action (3.3), invariant under five-dimensional general coordinate transformation, reduces to a four-dimensional action invariant under both four-dimensional general coordinate

transformations and Abelian gauge transformations. This four-dimensional action is, up to a surface term,

$$I_4 = \int \sqrt{|g_4^{(0)}|}\sqrt{|\phi^{(0)}|}\left[\left(-\frac{1}{16\pi G}\right)R_4^{(0)} + \left(\frac{e^2\kappa^2}{16\pi G}\right)\phi^{(0)}g^{(0)\mu\rho}g^{(0)\nu\sigma}F_{\mu\nu}^{(0)}F_{\rho\sigma}^{(0)}\right] \tag{3.8}$$

with

$$G = \frac{G_5}{2\pi\rho}, \qquad g_4^{(0)} = \det\left(g_{\mu\nu}^{(0)}\right), \qquad F_{\mu\nu}^{(0)} = \partial_\mu A_\nu^{(0)} - \partial_\nu A_\mu^{(0)}, \tag{3.9}$$

and $R_4^{(0)}$ = scalar curvature calculated from the four-metric $g_{\mu\nu}^{(0)}$; (our metric convention calls for a minus (plus) sign for a time (spacelike) dimension).

This action involves a graviton ($g_{\mu\nu}^{(0)}$), an Abelian gauge boson ($A_\mu^{(0)}$), and a scalar field $\phi^{(0)}$. Kaluza arbitrarily set $\phi^{(0)}$ = constant, in which case I_4 turns into the four-dimensional Einstein-Maxwell action. To be sure, one has to have $\phi^{(0)} > 0$ in order to have the proper relative sign of the Einstein and Maxwell terms, so that energy is positive. This in turn means that the fifth dimension must be spacelike; in fact, the extra dimensions must all be spacelike. In addition to the invariances under general coordinate transformations and gauge transformations, the action (3.8) also exhibits an invariance under global scale transformations:

$$\begin{aligned} g_{\mu\nu}^{(0)} &\longrightarrow \lambda^{-1}g_{\mu\nu}^{(0)}, \\ A_\mu^{(0)} &\longrightarrow \lambda^{-3/2}A_\mu^{(0)}, \\ \phi^{(0)} &\longrightarrow \lambda^2\phi^{(0)}. \end{aligned} \tag{3.10}$$

The field equations of the original five-dimensional theory have a solution in which the five-dimensional spacetime is the direct product of a circle with flat four-dimensional Minkowski spacetime. Then

$$g_{\mu\nu} = \eta_{\mu\nu}, \qquad A_\mu = 0, \qquad \phi = 1 \tag{3.11}$$

($\eta_{\mu\nu}$ is the four-dimensional Minkowski metric $(-1,+1,+1,+1)$). This solution serves as a natural vacuum, and it spontaneously breaks the scale invariance (3.9). The massless $\phi^{(0)}$-field is the Nambu-Goldstone boson associated with this spontaneous symmetry breaking. So the zero-mode spectrum includes spin 2 and spin 1 gauge fields and a spin 0 Nambu-Goldstone boson. In the full quantum theory the spin 0 boson is expected to acquire a mass. Of course, the full classical theory contains not only the zero modes, but also the $n \neq 0$ harmonics (equation (3.2)). The action (3.3) determines their spins, masses, and couplings. They all have spin less than or equal to 2, and they are all massive. The nth harmonics have mass

$$m_n = \frac{|n|}{\rho}, \tag{3.12}$$

where ρ, as before, is the radius of the small circle in the fifth dimension. The couplings of these harmonics with the gauge field $A_\mu^{(0)}$ are determined from the action (3.3), and

these harmonics do carry electric charge

$$g_n = \frac{n(4\sqrt{\pi G})}{\rho}. \tag{3.13}$$

Remarkably, electric charge is quantized because the fifth dimension is compact. We see that the elementary charge is

$$e : 4\frac{\sqrt{\pi G}}{\rho} \tag{3.14}$$

and the corresponding fine-structure constant is

$$\alpha = \frac{4G}{\rho^2}. \tag{3.15}$$

If α is to correspond to the $U(1)$ subgroup of grand-unification group, then $\alpha \sim 1/100$ so that the circumference of the small circle $l \equiv 2\pi\rho \sim 100\sqrt{G} \sim 10^{-17} GeV^{-1}$. The circle must be very small indeed; a size about 100 Planck lengths could hardly have been detected as yet. Nevertheless, this is large enough to call into question grand-unification in four-dimension: the scales at which the grand-unification group is to reveal itself unbroken are close to the scales at which the extra dimensions would become manifest. To make all this applicable in a world with strong and electroweak interactions, one of course has to introduce more than one extra dimension.

Kaluza's work has been unknown until when Oskar Klein, in 1926, rediscovered Kaluza's theory. (Einstein delayed the publication of Kaluza's paper for two years.) Klein noted the quantization of the electric charge and hoped Kaluza theory would underlie quantum mechanics (see Section 9). The relativistic generalization of Schrödinger's equation was carried out independently by many authors: Schrödinger, Klein, Gordon, Fock, and others. This equation, now commonly known as the Klein-Gordon equation, was arrived at by both Klein and Fock starting from Kaluza's theory: a zero-mass wave equation in five dimensions yields four-dimensional Klein-Gordon equations for the individual harmonics. It must be noted that this early work is viewed as a mathematical trick devoid of any physical significance. Nevertheless, this mathematical idea will prove very fruitful for the further developments of the theory, especially in supergravity and string theories. Oskar Klein comes closest to the modern point of view: he discusses the higher harmonics and the size of the small circle. Later Einstein and Bergman also adopted such a point of view. A purely mathematical approach (a projective interpretation of the fifth coordinate) was developed by Veblen, Pauli, Jordan, and others. Jordan appears to have been the first to realize the importance of including the scalar field $\phi^{(0)}$ into the new five-dimensional theory.

Remarkably, the most recent work on superstrings incorporates both the ideas of Nordström and the subsequent ideas of Kaluza and Klein (see Section 11). However, there was no real reason to extend the Kaluza-Klein idea beyond the five dimensions until the emergence of non-Abelian gauge field theories invented by Yang and Mills in 1954 (see Section 5). In 1963, DeWitt suggested that a unification of Yang-Mills theories and gravitation could be achieved in a higher-dimensional Kaluza-Klein framework. Trautman was independently aware of this possibility as were others. A detailed

discussion of the Kaluza-Klein unification of gravity and Yang-Mills theories, including the correct form of the $(4+N)$-dimensional metric, first appeared in the work of Kerner. The first complete derivation of the four-dimensional gravitational plus Yang-Mills plus scalar theory from a $(4+N)$-dimensional Einstein-Hilbert action was finally given by Cho and Freund in 1975. The weakness of this higher-dimensional work was the absence of any good reason as to why any dimension would compactify, let alone the right number, so as to leave the ordinary four-dimensional "large" world. While the five-dimensional theory at least admitted the compactified fifth dimension along with Minkowski space as a solution to the five-dimensional equations of motion, even this was not true of the higher-dimensional theories. The essential reason for this is that the higher-dimensional manifolds that give rise to Yang-Mills theories have curvature. If a $(4+N)$-dimensional Einstein theory is to compactify into the direct product of four-dimensional spacetime M^4 and a compact internal space with isometries, the metric $\gamma_{mn}(x,y)$ can be written as follows in the zero-mode approximation:

$$\gamma_{mn}(x,y)=\begin{bmatrix} g_{\mu\nu}(x)+\gamma_{mn}(y)\zeta^m_\alpha(y)\zeta^n_\beta(y)A^\alpha_\mu(x)A^\beta_\nu(x) & \gamma_{mn}(y)\zeta^m_\alpha(y)A^\alpha_\mu(x) \\ \gamma_{mn}(y)\zeta^n_\beta(y)A^\beta_\nu(x) & \gamma_{mn}(y) \end{bmatrix}. \tag{3.16}$$

The metric $\gamma_{mn}(y)$ is that of the corresponding N-dimensional symmetric space and the Killing vectors $\zeta^n_\alpha(y)$ have upper indices running over the dimension of the symmetry group. If four-space is to be flat (and, actually, it cannot be flat!), the Ricci tensor $R_{mn}=0$ for the spacetime indices, and therefore $R+\Lambda=0$. But then R_{mn} must vanish for the internal indices as well, and this cannot be the case if the internal space is curved.

Cremer and Scherk began to address this problem by pointing out that inclusion of additional Yang-Mills and scalar matter fields in the higher-dimensional theory would allow classical solutions in which spacetime is the direct product of Minkowski space and a compact internal space of constant curvature. This "spontaneous compactification" was achieved, however, by going beyond the pure Kaluza-Klein framework and including extra fields in just such a way as to induce the desired compactification. The program of seeking solutions to the combined Einstein-Yang-Mills equations in $4+D$ dimensions was generalized to a larger class of internal spaces by Luciani, Salam, Duff, and others. All this work on classical, higher-dimensional Kaluza-Klein theories provided a springboard for the study of both Kaluza-Klein supergravity and the quantum dynamics of Kaluza-Klein theories.

Roughly, supergravity is an attempt to unify matter and force as different components of the same agency. This is a kind of supersymmetric theory in which, because of the fact that the numbers of Bose and Fermi degrees of freedom have to be equal in supersymmetric theory, Bose fields beyond gravity appear in eleven dimensions. In fact supersymmetry dictates that the missing Bose degrees of freedom be supplied in the form of a massless antisymmetric tensor field with three indices A_{mnp} which indeed have $(11-2/3)=84=128-44$ degrees of freedom. Moreover, in eleven dimensions, there exist no matter and no Yang-Mills supermultiplets, so that besides gravity one only has its supersymmetric partner A_{mnp} and gravitino fields as "matter." The source

of gravity is thus fixed by supersymmetry. Furthermore, it is supersymmetry that determines the dimension of spacetime in eleven-dimensional supergravity. Force and matter uniquely determine each other; they are but different components of the same supermultiplet. In ten dimensions a similar argument can be made, but there we encounter Yang-Mills supermultiplets whose gauge group is fixed, though not uniquely, by the requirement of anomaly cancellation. For superstring theories similar considerations apply. To find the possible vacuum of the eleven-dimensional theory, we look for a solution of the classical equations in which the eleven-dimensional world manifold M_{11} is of the form $M_{11} = M_d \times M_{11-d}$, where M_d is the spacetime and M_{11-d} the small compact manifold. In the vacuum we require the spacetime M_d to be maximally symmetric. This then fixes the metric of $M_d(M_{11-d})$. The antisymmetric tensor potential A_{mnp} has its own gauge invariance under

$$A_{mnp}(x,y) \longrightarrow A_{mnp}(x,y) + \partial_m \alpha_{np}(x,y) + \partial_n \alpha_{pm}(x,y) + \partial_p \alpha_{mn}(x,y) \qquad (3.17)$$

with $\alpha_{mn} = -\alpha_{nm}$. The corresponding gauge-invariant quantities are the field strengths F_{mnpr} given by the curl of A_{mnp}:

$$F_{mnpr} = \partial_m A_{npr} + \partial_n A_{prm} + \partial_p A_{rmn} + \partial_r A_{mnp}. \qquad (3.18)$$

If F or its dual F^* is to have a nonvanishing vacuum expectation value on d-dimensional spacetime without destroying the maximal symmetry, then either d or $11-d$ must equal the number of indices of F, that is, $d = 4$ or $d = 7$. In the case $d = 4$, once one has fixed the maximally symmetric form of F, the simplest solutions are obtained by setting $h(y) = 1$ and $F^{mnpq} = 0$. Then the field equations and Bianchi identities of eleven-dimensional supergravity require $F(x,y) = \mathbf{F}$ = constant, and the energy momentum tensor of the A-field is equivalent to two cosmological terms, one on M^4 and one on M^7, with cosmological constants of opposite signs. Provided M^4 contains the time dimension, the cosmological constant on M^7 will have the sign appropriate to a compact manifold so that spontaneous compactification really does occur. M^4 is the maximally symmetric noncompact anti-de Sitter space and M^7 a compact Einstein space. The scale is set by the expectation value $\mathbf{F}$ of the field strengths. The rest of four-dimensional physics is determined by the shape of the small seven-dimensional manifold M^7. If, for instance, M^7 is the seven-sphere S^7, then the gauge group in four dimensions is $SO(8)$—the isometry group of S^7—and one finds eight supersymmetries. Just as gauge symmetries in four dimensions are related to the Killing vectors of the small manifold, so the supersymmetries are related to the Killing spinors. All the solutions with one or more surviving supersymmetries are stable with respect to classical perturbations. Some of the solutions without any surviving supersymmetry are stable, and others are unstable. Several problems stand in the way of producing a realistic theory. First of all, the four-dimensional anti-de Sitter space has a too large cosmological constant that has to be eliminated somehow. Of course Higgs mechanisms in four dimensions further affect the cosmological constant and it is the endproduct that has to be very small or zero. Another serious problem is the lack of chiral fermions at least as long as bound states and solitons are ignored. It seems that these problems can be solved in the context of higher-dimensional superstring theory, which recently demonstrated

the possibility of a finite quantum theory of gravity, whereas the eleven-dimensional supergravity theory, while trivially finite at one loop, is questionable in this regard.

4. The role of topological concepts in physics. We return to the geometrization of mathematics and theoretical physics. It must be stressed that the geometrization movement appears today to be more influenced by this century's concepts and methods, than by those of ordinary geometry and even the new non-Euclidean geometries developed in the 19th century. A new family of geometrical and topological invariants (Betti numbers, Euler-Poincaré characteristic, Whitney-Stiefel characteristic classes, Pontrjagin and Chern characteristic forms) is at the heart of twentieth-century mathematical progress, as well as at the foundation of recent physical theories, especially non-Abelian gauge theories. The introduction of these concepts and the development of a host of new notions and techniques in geometry and algebraic and differential topology in the 1940s—homology, cohomology, and homotopy (Whitney, Lefschetz, Hopf), fibre space, and characteristic classes (Ehresmann, Pontrjagin, Steenrod, Thom, Chern, Milnor), categories (Eilenberg, MacLane)—mark the passage from the local to the global study of mathematical objects.

One of the great mathematical advances of this century was the introduction of *characteristic classes* by Whitney and Stiefel in 1935, and *characteristic forms* by Pontrjagin (over a real fibre space) and Simons and Chern (complex). Intuitively, these objects are geometrical and topological invariants that can be classified in different families even though they are all mutually related. Thus we say that we can topologically transform a surface (or a manifold) into another if they have the same characteristic invariants (and if the dimension of the space is compatible with the type of transformation). Two manifolds satisfying these conditions are said to be topologically equivalent. These invariants and their corresponding algebraic structures can be technically very complicated. The two invariants mentioned above globally characterize the two mathematical objects of fibre spaces and connections, which in turn imply several basic algebraic and geometric notions such as homology (homology group, intrinsic homology, singular homology, algebraic homology, functors, etc.), cohomology classes, homotopy, and so forth. Two examples of homology and cohomology that are very important in contemporary physics are *bordism* and *cobordism* as developed by R. Thom, and ordinary homology $H(X)$. In fact, several notions of classical field theory can be expressed by cohomology. The more recent quantum field theories, reinterpreted in the mathematical framework of gauge theory, show a remarkable presence of cohomological ideas, seen in some cases as a generalization of characteristic classes such as those of Euler-Poincaré. A very interesting example in our time is that of a non-Abelian cohomology space of Riemannian surfaces with boundary.

We first give some basic notions and definitions on principal bundles, connection, curvature, and characteristic classes (we follow closely Steenrod [48] and Husemoller [27]). Among other things, a principal bundle has a structure group G, which is a Lie group, and a base B, which is a topological space. Notice that on the product $B \times G$ there is a natural right action of G by right multiplication on the second factor. This is a free action, and the quotient is defined with B.

Definition 4.1. A (right) principal G-bundle consists of a triple (P,B,π), where $\pi : P \to B$ is a map and a continuous, free right action $P \times G \to P$ with respect to which π is invariant and so that π induces a homeomorphism between the quotient space of this action and B. Furthermore, there is an open covering $\{U_\alpha\}$ of B over which all the above data are isomorphic to the product data. That is to say, for each α, there exists a commutative diagram

$$\begin{array}{ccc} \pi^{-1}(U_\alpha) & \xrightarrow{\varphi_\alpha} & U_\alpha \times G \\ \Big\downarrow{\scriptstyle \pi} & & \Big\downarrow{\scriptstyle p_1} \\ U_\alpha & \xrightarrow{\approx} & U_\alpha, \end{array} \tag{4.1}$$

where φ_α is a homeomorphism which is equivariant with respect to the right G-action and p_1 is the projection onto the first factor.

The space P is called the *total space* of the principal bundle, π is called the *projection*, and B is called the *base*. The maps φ_α are called *local trivializations*. Lastly, G is called the *structure group of the bundle*. If P and B are smooth manifolds, if the action of G on P is a smooth action, and if π is a smooth submersion, then the principal bundle is said to be a *smooth principal bundle*. In this case it follows automatically that one can choose the local trivializations so that the φ_α are diffeomorphisms. An *isomorphism* of G-bundles with the same base is a homeomorphism between their total spaces, which is G-equivariant and which commutes with the projections to the base. A map between G-bundles over possibly different bases is a G-equivariant map between the total spaces. Such a map must be an isomorphism on each fibre and it induces a map between the base spaces. In order to consider a general example of principal bundle, let M be a smooth manifold. Let E be the frame space for the tangent bundle of M. A point of E consists of a point $p \in M$ and a basis $\{v_1,\dots,v_n\}$ for the tangent space TM_p to M at p. The topology, and indeed the smooth structure, of E is induced in the obvious projection of E to M and an obvious action of $GL(n,\mathbb{R})$ on E. The action of $A = (a_{ij}) \in GL(n,\mathbb{R})$ on the point $(x,\{v_1,\dots,v_n\})$ gives the point $(x,\{w_1,\dots,w_n\})$, where

$$w_j = \sum_{i=1}^{n} a_{ij} v_i. \tag{4.2}$$

That is to say, the matrix A acts on the basis to produce a new basis for the same space; the expression for the new basis in terms of the old basis is given by the columns of the matrix A. This defines a right action of $GL(n,\mathbb{R})$ on E.

Let $\pi : P \to B$ be a smooth principal G-bundle over an n-dimensional manifold. A connection for this bundle is an infinitesimal version of an equivariant family of cross sections. It is an n-dimensional distribution $\mathcal{H}$ (i.e., smooth family of n-dimensional linear subspaces of the tangent bundle TP of P) which is *horizontal* in the sense that the restriction of $D\pi$ to each plane in the distribution is an isomorphism onto the corresponding tangent plane to B and which is invariant under the G-action. This distribution induces an isomorphism $TP_p \equiv TP_p^v \bigoplus TB_\pi(p)$. Suppose that Γ is a connection for $P \to B$. Let $\gamma : [0,1] \to B$ be a smooth path and $e \in \pi^{-1}(\gamma(0))$. Then there is a unique

path $\gamma : [0,1] \to P$ such that $\gamma(0) = e$, $\pi \circ \gamma = \gamma$, and $\tilde{\gamma}'(t)$ is contained in the horizontal space $\mathcal{H}_{\tilde{\gamma}(t)}$. Given a smooth curve in the base $\gamma : [0,1] \to B$ from b_0 to b_1, a connection determines an isomorphism between the fibers $\pi^{-1}(b_0) \to \pi^1(b_1)$, which is equivariant with respect to the G-actions on these fibers. Therefore a connection gives a manner to connect distinct fibers, albeit one needs a path in the base between the image points in the base. Let $\mathcal{G}$ be the Lie algebra for G. Then there is a unique one-form $\omega_{MC} \in \Omega^1(G;\mathcal{G})$ which is invariant under left multiplication by G and whose value at the identity element of G is the identity linear map from $TG_e \to \mathcal{G}$. This form is called the *Maurer-Cartan form*. It is often denoted by $g^{-1}dg$. Its value on a tangent vector $\tau \in TG_g$ is equal to $g^{-1} \cdot \tau \in TG_e = \mathcal{G}$.

Lemma 4.2. *A connection on a smooth principal bundle $\pi : P \to B$ is equivalent to a differential one-form $\omega \in \Omega^1(P;\mathcal{G})$ with the following properties.*

(i) *Under right multiplication by G, the form ω transforms via the adjoint representation of G on $\mathcal{G}$; that is,*

$$\omega_{pg}(\tau \cdot g) = g^{-1}\omega_p(\tau) \cdot g \tag{4.3}$$

for any $p \in P$, any $\tau \in TP_p$, and any $g \in G$.

(ii) *For any $p \in P$, consider the embedding $R_p : G \to P$ given by $R_p(g) = p \cdot g$. Then the pullback $R_p^*(\omega) = \omega_{MC}$.*

Suppose that A is a connection on a principal bundle $\pi : P \to B$, and suppose that $W \to B$ is a vector bundle associating to this principal bundle and a linear action of G on a vector space V. We can use the connection to differentiate sections of W, producing one-forms with values in W. This covariant differentiation is a linear operator

$$\nabla_A : \Omega^0(B;W) \longrightarrow \Omega^1(B;W). \tag{4.4}$$

The curvature arises as the obstruction to integrating the horizontal distribution of a connection over two-dimensional submanifolds of the base. Let $P \to B$ be a smooth principal G-bundle and let $\mathrm{ad}P$ be the vector bundle associated to P and the adjoint action of G on its Lie algebra $\mathcal{G}$. Suppose that A is a connection on P, and $\mathcal{H} \subset TP$. We can integrate $\mathcal{H}$ along paths in B to give a lifting of paths from B to P. If we try to perform the same construction over higher-dimensional subspaces of B, then it is not always possible to lift—there is an obstruction which is the curvature of the connection. We fix a point $b \in B$ and two linearly independent tangent vectors τ_1, τ_2 at b. Consider a local coordinate system $(x_1,\dots,x_k)$ centered at a point $b \in B$ with the property that $(\partial/\partial x_i)|_0 = \tau_i$ for $i = 1,2$. We consider a rectangle $[0,\varepsilon]\times[0,\varepsilon]$ in the (x_1,x_2)-subspace. We lift the four sides of this rectangle in counterclockwise fashion beginning with the side on the x_1-axis. We do this so that the initial point lifts to a point $p \in P$ and so that each side begins where the previous side ends. There is no guarantee that the end of the last side will be equal to p, but it will be of the form $p \cdot g$ for some unique $g = g(\varepsilon) \in G$. If ε is sufficiently close to zero, then $g(\varepsilon)$ will be close to the identity in G, and hence

we can form $\log(g(\varepsilon)) \in \mathcal{G}$. We consider the element

$$K_A(\varepsilon) = -\frac{\log(g(\varepsilon))}{\varepsilon^2}. \tag{4.5}$$

Lemma 4.3. *The element in g given by*

$$K_A(p,\tau_1,\tau_2) = \lim_{\varepsilon\to 0} K_A(\varepsilon) \tag{4.6}$$

depends only on p, τ_1, τ_2. Furthermore, the point

$$[p, K_A(e,\tau_1,\tau_2)] \in \operatorname{ad} P \tag{4.7}$$

depends only on τ_1, τ_2, and is bilinear and skew-symmetric in these variables. It is given by evaluating a two-form on B with values in $\operatorname{ad} P$*, denoted by F_A, on (τ_1,τ_2). This two-form F_A is called the curvature of A.*

We can use the curvature to define cohomology classes in B which measures the nontriviality of the bundle. These are called *characteristic classes*. The first result we need in order to define characteristic classes from the curvature is the so-called *Bianchi identity*.

Lemma 4.4 (Bianchi identity)**.** $\nabla_A F_A = 0$.

Suppose that

$$\varphi : \underbrace{\mathcal{G}\otimes\cdots\otimes\mathcal{G}}_{k \text{ times}} \longrightarrow \mathbb{R} \tag{4.8}$$

is a linear map which is symmetric and invariant under the simultaneous adjoint action of G on $\mathcal{G}$; that is,

$$\varphi(F_1,\dots,F_k) = \varphi(g^{-1}F_1 g,\dots,g^{-1}F_k g). \tag{4.9}$$

Then we can form

$$\varphi(F_A,\dots,F_A) \in \Omega^{2k}(B;\mathbb{R}). \tag{4.10}$$

Lemma 4.5. *The form $\varphi(F_A,\dots,F_A)$ is closed. If another connection A' for P is chosen, then the difference*

$$\varphi(F_A,\dots,F_{A'}) - \varphi(F_A,\dots,F_A) \tag{4.11}$$

is exact.

For the special orthogonal group $SO(n)$, a basis for the invariant polynomials on the Lie algebra is given by the even coefficients of the characteristic polynomial together with the Pfaffian if n is even. Thus, we get one characteristic class in each degree $4i$, and if $n = 2k$, we also get one characteristic class in degree $2k$. If we normalize properly, then these classes are, respectively, the ith Pontrjagin class and the Euler class. There is a similar result for complex-valued symmetric, multilinear functions on the Lie algebra.

Applying this to the unitary group, we see that a basis for the complex-valued invariant polynomials is given by the coefficients of the characteristic polynomials. Thus, in this case, we have one characteristic class in each degree $2i$. Correctly normalized, these are the Chern classes.

We further recall some fundamental geometric-differential facts regarding the notions of bordism and cobordism. For each topological space X, the commutative group $\Omega(X)$ can be defined as follows (Thom [51]). Continuous mappings $f : Y \to X$ from oriented and compact manifolds with boundary Y in X are called *chains*. The sum and difference of chains are defined by disjoint union and change of orientation, respectively. The boundary of f is its restriction to the boundary of Y. It is well known (thanks to a fundamental theorem of algebraic geometry) that the boundary of a boundary is empty: $\partial \circ \partial = 0$. A cycle is a chain whose source has no boundary. The equivalence classes of cycles form the group $\Omega(X)$. The *Thom ring* is the bordism of a point. With the product $Y \times X$, one can see that $\Omega(X)$ is a module over Ω, so that Ω acts on $\Omega(X)$. To every continuous mapping $X_1 \to X_2$ is associated a linear transformation of $\Omega(X_1)$ into $\Omega(X_2)$; we then have a *functor*. The cobordism cohomology $\Omega^*(X)$ is obtained by taking the homotopy classes of continuous mappings of X into a given space, in fact, a nested sequence of topological spaces, the *Thom spectrum*. The cohomology (the theory Ω^*) is richer than the homology (Ω), dealing with rings having all sorts of operations. Cobordism is an equivalence relation on the set of submanifolds, say N and N' of M, which means that the cobordism Z transforms N into N'. M designates a compact oriented manifold with $H_1(M) = 0$. We say then that two submanifolds N, N' are cobordant in M if there is a compact $Z = M \times [0,1]$ so that $\partial Z = N \times \{0\} \subseteq N' \times \{1\}$ (see [35]).

Even these short considerations suffice to highlight some basic characteristics of cohomology that make it a good basis for building a richer and more sophisticated theory of the spatial continuum and of spacetime, with enormous theoretical implications for physics. Some of these characteristics are listed below (Bennequin [6]).

(1) Homology is constructed by quotienting a part of the data (cut and gluing). It stabilizes forms.

(2) It shows the close relationship that can exist between figures and numbers, especially coefficients. We can reconstruct a new ring Ω from combinations of chains with rational ($\mathbb{Q}$) or complex coefficients ($\mathbb{C}$). This lets us "localize" and "complete," respectively.

(3) The most remarkable property is probably the universality. There are many cohomologies that all give the same results. More exactly, different definitions lead to isomorphic (or related, at the very least) theories. This means that axiomatic constructions are permitted (Atiyah (1968)).

(4) Cohomology realizes forms; in a certain sense, it defines forms. In any case, it ensures certain stability and genericity. Several notions from classical field theory can be expressed cohomologically. Furthermore, the more recent quantum field theories, reinterpreted in the common mathematical framework of gauge theory, highlight the basic role played by cohomology and characteristic classes (see the work of Atiyah and Bott [4], Manin [34], Uhlenbeck [54], and Taubes [50]). These concepts are also used in the attempts to give a consistent mathematical formulation and an intelligible physical

interpretation of other quantum gauge theories such as the quantum electrodynamics of Dirac, Feynman, and Schwinger.

5. The birth and development of gauge theory. A review of the origin and development of gauge theory is in order (for more details, see [37, 67]). Two major geometrical advances of Weyl must be mentioned. In 1918–1919 he outlined what he called a "*purely infinitesimal geometry*" (for the history of this theory, see [44, 49, 57]), which should know a transfer principle for length measurements between infinitely close points only, and which should admit a conformal structure. The allusion is of course to Levi-Civita parallel displacement principle in a Riemannian manifold embedded in a sufficiently high-dimensional Euclidean space, locally given by

$$\xi'^i = \xi^i - \Gamma^i_{jk}\xi^j dx^k \tag{5.1}$$

with the dx^i to be interpreted as the coordinate representation of a displacement vector between two infinitesimally close points so that the direction vector ξ^i is transferred to ξ'^i. According to Weyl, one has to separate logically the concept of parallel displacement from metrics and to introduce what he called an *affine connection* Γ on a (differentiable) manifold as a linear torsion-free connection. Thus, Weyl proposes a generalization of Riemannian geometry which seemed to be the most natural mathematical framework for the construction of a unified theory of gravitational and electromagnetic forces. This generalized Riemannian metric; a *Weylian metric* on a differentiable manifold M is given by

(i) a *conformal structure* on M, that is, a class of (semi-) Riemannian metrics $[g]$ in local coordinates given by $g_{ij}(x)$ or $g_{ij}(x) = \lambda(x)g_{ij}(x)$, with multiplication by $\lambda(x) > 0$ (real-valued) representing what Weyl considered to be *gauge transformation* of the representative of $[g]$,

(ii) a *length connection* on M, that is, a class of differential forms φ in local coordinates represented by $\varphi_i dx_i$, $\varphi_i dx_i - d\log\lambda$ (representing the *gauge transformation* of the representative of j).

This new infinitesimal geometry enfolds in fact the first formulation of a gauge theory. The idea of *gauge* was introduced by Weyl in a very influential paper of 1918 [60] (See also the interesting paper on the same subject published thereafter by Pauli [38].) The background of this thinking at that time can be retraced through the preface of the various editions of his landmark book *Raum, Zeit, Materie* (first edition, 1918) (Hermann Weyl has evidently been inspired by the work of Einstein on gravity (1915–1916), but also by the work of Felix Klein, who introduced the general mathematical concept of group of transformations in his famous Erlangen Program in 1872, by D. Hilbert, and mostly by Levi-Civita and Elie Cartan, who introduced, respectively, the concepts of parallel-transport and of connection, which turned out to play a more and more important role in the mutual relations of mathematics and physics. He was also influenced by the German physicist Gustav Mie who tried—in a series of articles published in 1912–1913—to explain the basic phenomena of matter on a purely electromagnetic basis, in particular the existence, mass, and stability of electrons. Besides, Mie attempted to formulate a theory of the electron that does not involve divergent field quantities

inside of the electron). Weyl showed that while Einstein's gravity theory depended on a quadratic differential form

$$ds^2 = \sum_{ik} g_{ik} dx_i dx_k, \tag{5.2}$$

electromagnetism depended on a linear differential form

$$\phi = \sum_i \phi_i dx_i, \quad 1 \le i,k \le 4, \tag{5.3}$$

(which in today's notation is $\sum A_\mu dx^\mu$) defined up to the gauge transformation

$$ds^2 \longrightarrow \lambda ds^2, \qquad \phi \longrightarrow \phi + d\log\lambda. \tag{5.4}$$

Thus the idea of a nonintegrable scalar factor

$$e^{\int_P^Q d\phi} \tag{5.5}$$

was born. Weyl argued that the addition of a gradient $d(\log\lambda)$ to $d\phi = \sum \phi_\mu dx_\mu$ does not change the physical content of the theory, thus concluding that

$$F_{\mu\nu} = \frac{\partial \phi_\mu}{\partial x_\nu} - \frac{\partial \phi_\nu}{\partial x_\mu} \tag{5.6}$$

has "invariant significance." He naturally then identified $F_{\mu\nu}$ with the electromagnetic field and put

$$\phi_\mu = (\text{constant}) A_\mu, \tag{5.7}$$

where A_μ is the electromagnetic potential. Thus electromagnetism is conceptually incorporated into Weyl's theory and, in particular, into the geometric idea of a nonintegrable scalar factor (see [9]).

Here, an important aspect of the relationship between Einstein's general relativity and Weyl's gauge theory is worth noting. In general relativity, one uses a kind of mathematical relationships, known as a "*connection*," which specifies that if the spacetime orientation of a frame at x is given, then the relative orientation of a frame at $x + dx$ can be calculated. Since the frames are in a gravitational field, the connection itself is determined by the strength of the field. In fact, the connection can replace the gravitational field entirely so that all motion can be described in terms of the connection alone. This replacement of the field by a mathematical connection leads to the well-known geometrical picture of general relativity. The familiar "curvature" of spacetime can be calculated directly from the connection. Now Weyl went a step beyond general relativity and asked the following question: if the effects of a gravitational field can be described by a connection which gives the relative orientation between local frames in spacetime, can other forces of nature such as electromagnetism also be associated with similar connections? Generalizing the concept that all physical magnitudes are relative, Weyl proposed that the absolute magnitude or norm of a physical vector also should

not be an absolute quantity but should depend on its location in spacetime. A new connection would then be necessary in order to relate the lengths of vectors at different positions. This connection is associated with the idea of scale or "gauge" invariance. It is important to note that the true significance of Weyl's proposal lies in the "local" property of gauge symmetry and not in the particular choice of the norm or "gauge" as a physical variable. Actually, the assumption of locality is an enormously powerful condition that determines not only the general structure but many of the specific features of gauge theory.

Thus, after Einstein has developed his theory of general relativity, in which a dynamical role was given to geometry, Weyl conjectured that perhaps the scale length, indeed the scale of all dimensional quantities, would vary from point to point in space and in time. His motivation was to unify gravity and electromagnetism to find a geometrical origin for electrodynamics (see the good presentations of Moriyasu [36] and Gross [25]). He assumed that a translation in spacetime dx^μ would be accompanied by a change of scale or gauge, $1 \to 1 + S_\mu(x)dx^\mu$. The gauge function $S_\mu(x)$ would determine the relative scale of lengths so that a certain function would transform as $f(x) \to f(x) + [\partial_\mu + S_\mu(x)]f(x)dx_\mu$. The hope was to identify the connection S_μ with the vector potential of electrodynamics, thus unifying this theory with gravity. This did not work but only temporally! In fact, in 1927, after the development of quantum mechanics, Fock and London noticed that the $p_\mu - eA_\mu$, when p_μ is replaced with ∂_μ by $\partial_\mu - (ie/hc)A_\mu$, looked very much like Weyl's change of scale, but with a complex coefficient for the connection. Two years later Weyl completed the discussion, showing how electrodynamics was invariant under the gauge transformation of the gauge field and of the wave function Ψ of a charged particle,

$$A_\mu \longrightarrow A_\mu + \partial_\mu \alpha, \qquad \Psi \longrightarrow e^{ie\alpha/hc}\Psi. \tag{5.8}$$

The concept of gauge invariance, and therefore the principle of local gauge symmetry, was born. Accompanying the translation of charged particle, there is a phase change. The fact that the physics, at least at Planck scale, remain unchanged with respect to a gauge transformation lies at the heart of different forms of matter.

The most remarkable thing mathematically is that all the objections to the Weyl's theory disappear if we interpret it, as will be done later, as based on the geometry of a circle bundle over a Lorentzian manifold. Then the form ϕ above (see (5.3)), subject to the gauge transformation, can be interpreted as defining a connection in the circle bundle and thus the metric remains unaltered. More generally, the characteristic features of gauge theories can be described in terms of the topological and geometrical differential concept of fibre bundles and the connections in them. The connection is an intrinsic local structure that can be imposed on the bundle; it gives an elementary but fundamental example of a gauge field. Since gauge fields, including in particular the electromagnetic field, are fibre bundles, all gauge fields are thus based on topology and geometry. Starting in the 1970s, 20 years after the discovery by Yang and Mills of a non-Abelian gauge theory for strong force (nuclear interactions) in which the local gauge group was the $SU(3)$ isotopic-spin group, the physicists were able to express the concept of a gauge field in such a way that it could be recognized as an instance of more

abstract structures known to mathematicians as connections in fibre bundles. The discovery of this equivalence has made it possible to understand why and how powerful mathematical concepts and structures are necessary and suitable for the description and explanation of physical reality.

In a very important paper, Wu and Yang introduced the fundamental concept of nonintegrable—that is, path-dependent—phase factor as the basis of a description of electromagnetism [66]. Further this concept is made to correspond to the definition of a gauge field; to extend it to global problems, they analyzed, in relation with the original Dirac's result, the field produced by a magnetic monopole. 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 spacetime. In the overlap of any two regions, there exists a gauge transformation relating the phase factors defined for the two regions. The concept of monopole leads to the definition of global gauges and global gauge transformations. 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 the fiber bundle theory. Furthermore gauge fields, including in particular the electromagnetic field, are fiber bundles, and all gauge fields are thus based on geometry. So maybe all of the fundamental interactions of the physical worlds could be based on these geometrical and topological objects.

The exact formulation of the concept of a nonintegrable phase factor depends on the definition of global gauge transformations, that is, on the choice of the overlapping regions of R (where R is a region of spacetime, precisely, all spacetime minus the origin $r = 0$) and of the potential A_μ in this region. Through a certain kind of operations, called *distorsion*, one arrives 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. Notice that the gauge transformation factor in the overlap between R_a and R_b does not refer to any specific A_μ. (The gauge transformation in the overlap of the two regions is $S = S_{ab} = \exp(-i\alpha) = \exp(2ige/hc)\phi$.) Thus two different gauges may share the same characterizations (a) and (b). In the case of the monopole field, one can attach to the gauge any $(A_\mu)_a$ and $(A_\mu)_b$ provided they are gauge-transformed into each other in the region of overlap. Thus a gauge is a concept not tied to any specific vector potential. Wu and Yang called the process of distorsion leading from one gauge to another a *global gauge transformation*. It is also a concept not tied to any specific vector potential. The collection of gauges that can be globally gauge-transformed into each other will be said to belong to the same gauge type. The phase factor $\exp(ie/hc \int A_\mu dx^\mu)$ (which is nonintegrable, i.e., path-dependent) around a loop starts and ends at the same point in the same region. Thus it does not change under any global transformation, so that we have the following theorem for Abelian gauge fields.

Theorem 5.1. *The phase factor around any loop is invariant under a global gauge transformation.*

The next two theorems follow trivially from this by taking an infinitesimal loop.

THEOREM 5.2. *The field strength $f_{\mu\nu}$ is invariant under a global gauge transformation.*

THEOREM 5.3. *Between two gauge fields defined on the same gauge there exists a continuous interpolating gauge field defined on the same gauge.*

THEOREM 5.4. *Consider gauge $\mathcal{G}_D$ and define any gauge field on it. The total magnetic flux through a sphere around the origin $r = 0$ is independent of the gauge field and depends on the gauge only:*

$$\iint f_{\mu\nu} dx^\mu dx^\nu = -\frac{ihc}{e} \int \frac{\partial}{\partial x^\mu} (lnS_{ab}) dx^\mu, \tag{5.9}$$

where S is the gauge transformation defined by (5.8) for the gauge $\mathcal{G}_D$ in question, and the integral is taken around any loop around the origin $r = 0$ in the overlap between R_a and R_b, such as the equation on a sphere $r = 1$.

As in the case of electromagnetism, in the non-Abelian gauge fields both the concept of a gauge and the concept of a global gauge transformation are not tied to any specific gauge potentials. The nonintegrable phase factor for a given path is now an element of the gauge group (see [41]). Since these phase factors do not in general commute with each other, Theorems 5.1 and 5.2 for the Abelian case need to be modified as follows.

THEOREM 5.5. *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 5.6. *The field strength $f^k_{\mu\nu}$ is covariant under a global gauge transformation.*

Theorem 5.5 defines the *class of a loop*. This concept is a generalization of the phase factor for electromagnetism around a loop with the magnetic flux as the exponent. It is a gauge-invariant concept.

Rigorously the mathematical structure of gauge theory is that of a vector bundle E with structure group G over a compact Riemannian manifold M. We assume that $G \in O(m)$ and E carries an inner product compatible with G. Let E be the space of G-connections on E, and let $\mathcal{G}$ be the space of G-automorphisms of E. Then $\mathcal{G}$ acts on E as before, and we have a quotient space $B \equiv E/\mathcal{G}$. To each connection $\nabla \in E$ there is associated a curvature 2-form R^∇, and at each point x, we can take its norm

$$||R^\nabla||^2_x \equiv \sum_{i<j} ||R^\nabla_{e_i,e_j}||^2_x, \tag{5.10}$$

where $\{e_1, \ldots, e_n\}$ is an orthonormal basis of T_xM and the norm of $R^\nabla_{e_i,e_j}$ is the usual one on $\mathrm{Hom}(E,E)$—namely, $\langle A,B\rangle \equiv \mathrm{trace}(A^t \circ B)$. Given any $g \in \mathcal{G}$, we recall that $R^{g(\nabla)} = g \circ R^\nabla \circ g^{-1}$, so

$$||R^{g(\nabla)}|| \equiv ||R^\nabla|| \quad \text{on } M. \tag{5.11}$$

This says that the pointwise norm of the curvature is gauge-invariant.

Definition 5.7. The *Yang-Mills functional* is the mapping $YM : E \to \mathbb{R}^+$ given by

$$YM(\nabla) \equiv \frac{1}{2}\int_M ||R^\nabla||^2. \tag{5.12}$$

(Note that, by gauge invariance of the density (5.11), this functional descends to a functional $YM : B \to \mathbb{R}^+$.)

An important verified fact is that if M is four-dimensional, then YM is *conformally invariant*; that is, if we replace the metric ds^2 on M by a new metric $ds^2 = f^2 ds^2$, for some positive function f on M, then the Yang-Mills functional is unchanged. We think of YM as an "action integral" and seek its stationary points.

Definition 5.8. A connection $\nabla \in E$ is called a Yang-Mills connection, and its curvature R^∇ is called a Yang-Mills field if $\operatorname{grad}_\nabla(YM) = 0$.

Lemma 5.9. *The following are equivalent:*
(1) ∇ *is Yang-Mills,*
(2) $\delta^\nabla R^\nabla = 0$,
(3) $\Delta(R^\nabla) = 0$, *where* $\Delta \equiv d^\nabla\delta^\nabla + \delta^\nabla d^\nabla$.

The equations $\delta^\nabla R^\nabla = 0$ are called the *Yang-Mills equations*. The equivalent condition (3) states that the curvature R^∇ is *harmonic* (with respect to its own Laplacian). It is naturally appealing to a geometer to study connections with harmonic curvature. The subject has even more allure when one learns that the "classical fields" corresponding to basic forces of nature (the electromagnetic, weak, and strong interactions) can, and should, be formulated in these terms. A basic case is electromagnetic, where $G = U_1$ and E is a complex line bundle over a four-dimensional Lorentzian manifold. In relativistic terms, the six components of the curvature 2-form Ω of a connection on E represent the six components of the electromagnetic tensor. The equations

$$d\Omega = 0, \qquad \delta\Omega = 0 \tag{5.13}$$

are exactly Maxwell's field equations.

6. The geometric nature of gauge invariance: gauge theory and quantum mechanics. General relativity was discovered around 1916 by Einstein. Its complete mathematical formulation was due to his friend the mathematician Marcel Grassmann. Soon after, this theory was recognized a major scientific event. Hermann Weyl of course agreed with that. However he was still looking for a more complete formulation of Einstein's theory. Thus, under the influence of general relativity, he aimed to a search for a unification of gravitation and electromagnetism. In other words, he asked himself whether one could not find an extended geometry which would likewise allow to accommodate the only other force field known at the time, the electromagnetic field $F_{ij} = \partial_i A_j - \partial_j A_i$ with its potential A_i, into the geometrical structure of spacetime. His main idea was that of a local gauge invariance, which I will try to explain.

Let me start with a historical note. In a letter to Einstein, 1 March 1916, Weyl wrote: "These days, I believe to have succeeded in deriving electricity and gravitation from a

common source. One obtains a wholly determined principle of action which within the electricity-free field leads to your expression for gravitation. On the other hand, within a gravitation-free field, one gets an expression which, at first sight, is in agreement with Maxwell's theory." The response of Einstein came soon: "I received your paper. It is a stroke of genius (*Genie-Streich*) of the highest rank. Besides, I was not able so far to settle my objection concerning the measure standard .... (We translated both quotations from German).

The importance of the gauge invariance (*Eichinvarianz*) can be measured by what the theoretical physicist Abdus Salam wrote in Nobel conference of 1975: "One of the most revolutionary events in the history of science of the last century is the idea of gauge unification of the electromagnetic force with the weak nuclear forces" (Salam [42]); or by what the outstanding theoretical physicist T. W. B. Kibble wrote in 1982: "Revolutions are hard to recognize till they are past. This is surely true of the changes that have occurred in elementary particle physics over the last two decades. The development of gauge theories may well come to be seen as constituting one of the most fundamental revolutions of this century, rivaling the development of quantum mechanics itself. Yet so far its significance is not widely understood outside the ranks of specialists" (Kibble [28]).

Now we have to return back to Weyl. We said that the Weyl's work was aimed at extending the physical significance of general relativity and consequently to propose a generalization of Riemannian geometry. According to Weyl, these generalizations may be possible by introducing the main idea that length of vectors, and not only direction, must depend on the path. In other words, length ceases to be an action-at-distance concept. Mathematically, the idea of local gauge invariance amounts to introducing a nonintegrable scale factor or a function, which should supply the fact that in Riemannian geometry the invariance of the length of each two vectors gets lost. So Weyl proposes a procedure for recalibrating the displacement of a vector at each point of spacetime, in order to leave the length as well as the direction of this vector locally unchanged. Furthermore, Weyl had the ingenious idea of associating the metric tensor with the strength of the electromagnetic field, and the scale vector with the electromagnetic potential.

The idea of Weyl runs as follows. The parallel transport of the two vectors $\mathbf{V}'$ and $\mathbf{W}'$ from x' to $x'+dx'$ and, consequently, around a closed contour is generalized. The angle between the two vectors is still kept fixed under parallel transport, but the assumption of the invariance of the length of both vectors is dropped. The length of a vector—in contrast to the angle between two of them—ceases to be an action-at-distance concept. How should one change the expression

$$\delta\|\mathbf{V}^k = -\left\{ \begin{matrix} k \\ ij \end{matrix} \right\} \mathbf{V}^j dx^i \tag{6.1}$$

(which expresses the change of the components of a vector $\mathbf{V}^j(x)$, if displaced or transported parallely on a Riemannian manifold M^r of r dimensions from the point with coordinates x^i to the one with coordinates x^i+dx^i)? One would like to uphold the

bilinearity of $\delta\|\mathbf{V}^k$ in $\mathbf{V}^j$ and dx^i, thereby arriving at

$$\delta\|\mathbf{V}^k = -\Gamma^k_{ij}\mathbf{V}^j dx^i \tag{6.2}$$

with the so far unknown connection coefficients Γ^k_{ij}. On the basis of the last expression, we can define once more a covariant differentiation, which we still denote by ∇_i. Then the change of a vector around the contour turns out to be

$$\nabla\mathbf{V}^l = -\left(\frac{1}{2}\nabla A^{ij}F^l_{ijk}\right)\mathbf{V}^k, \tag{6.3}$$

this time involving the curvature tensor, and we get

$$F^l_{ijk} = 2\partial_i\Gamma^l_{jk} + 2\Gamma^l_{im}\Gamma^m_{jk} = -F^l_{jik}. \tag{6.4}$$

We thus see that in [61] Weyl enlarged the Riemannian spacetime of general relativity by an independent vector field of geometric origin—in modern terms, a one-form. This additional geometric object is intimately linked with the geometrical structure of spacetime. In addition, the Weyl vector is the compensating potential for allowing invariance with respect to local recalibration of lengths, that is, with respect to conformal changes of the metric. One can furthermore generalize the Weyl geometry to the metric-affine geometry, which is based on a (symmetric) metric and an independent (nonsymmetric) linear connection. In Weyl geometry, one geometrical object, the metric tensor, stands for the gravitational potential, as in general relativity, whereas the other one, the linear connection, was surmised to represent the electromagnetic potential known from Maxwell's theory. Together with a suitable (gravitational and electromagnetic) field Lagrangian, which turns out to be quadratic in the curvature of the underlying Weyl spacetime, this builds up Weyl's unified theory of 1918. The idea of gauge invariance, or the so-called principle of recalibration, which applies first to length of vectors in spacetime, transmuted to the concept of local gauge invariance of the phase of a wave function in 1929, and represents, in the last form, one of the underlying principles of all modern gauge theories, such as the Weinberg-Salam theory of electroweak interactions.

The other fundamental contribution of Weyl is related to his gauge theory but concerns quantum mechanics. In an article in 1927 [62], then in his book *Gruppentheorie und Quantenmechanik* (1928), Weyl proposes developing the mathematical foundations of this newly discovered physical theory by showing its close relationship to group representation theory. (For a very illuminating overview of Weyl's contribution to the theory of Lie groups, see [10].) In Weyl's new mathematical approach, the basic question at that time was to explain the properties of particles (protons and electrons) by the properties of the quantum laws: do these laws satisfy the basic symmetries known at that time (right/left, past/future, positive/negative electric charge)? Mathematically,

that was equivalent to knowing the structure of certain classes of (continuous) groups and their algebras. These three kinds of symmetry were introduced (under other names) into quantum physics in the 1930s by Weyl himself and by E. Wigner, but no one thought then of unifying the three kinds. In 1930, Dirac had detected the existence of a particle (positron) with a charge opposite to that of an electron, and Weyl then generalized to a universal essential equivalence between positive and negative electricity. This idea was reformulated in 1937 as the conjugate invariance of electrical charge. However, in 1957, Lee and Yang found that left-right symmetry (or conservation of parity P), which physicists had always found useful to accept, was not entirely satisfied by the laws of nature, particularly in weak interactions, which are responsible for radioactive (*beta*) disintegration. Since it could be verified theoretically and experimentally that this radioactivity gave a correct description of the neutrino, the conclusion was that the existence of the Weyl-Pauli theory (of the neutrino) violated left-right symmetry. This asymmetry seemed to be a consequence of duplication: massless particles (neutrinos) emitted in a *beta* disintegration existed in only one form (left), while the corresponding antiparticles (antineutrinos) could then only exist in the opposite form. Mathematically, this duplication could appear as the existence of two valid solutions for an equation. Some theoretical physicists interpret this phenomenon to speculate that the world did not have to be symmetrical with respect to every operation which left the laws of nature invariant: the loss of symmetry could be ascribed to the asymmetry of the whole universe. Such an explanation raises several questions. It is just as reasonable to believe that the loss of symmetry, as a characteristic of a transitory phase in which the laws of nature apply, could be explained by a richer, more general mathematical symmetry. Recent research in this field seems to be oriented toward this second outlook.

7. Quantum electrodynamics, gauge theory, and the concept of symmetry. It is now important to emphasize some facts about the quantum electrodynamics, that is, the theory that results from combining electron matter fields with electromagnetic fields—formulation begun in the 1930s by P. Dirac and was essentially completed in about 1949 by S. Tomonaga, J. Schwinger, R. P. Feynman, and F. J. Dyson. (The original papers have been republished in [46]). We recall firstly that it is based on a local gauge symmetry. Another theory, Einstein's general theory of relativity, is based on a local gauge symmetry, which pertains not to a field distributed through space and time but to the structure of spacetime itself. Indeed every point in spacetime can be labeled by four numbers, which give its position in the three spatial dimensions and its sequence in the one time dimension. These numbers are the coordinates of the event, and the procedure for assigning such numbers to each point in spacetime is a coordinate system. The choice of such a coordinate system is clearly a matter of convention. The freedom to move the origin of a coordinate system constitutes a symmetry of nature. Actually there are three relate symmetries: all laws of nature remain invariant when the coordinate system is transformed by translation, by rotation, or by mirror reflection. It is important to note, however, that the symmetries are only global ones. Each symmetry transformation can be defined as a formula for finding the new coordinates of a point

from the old coordinate. Those formulas must be applied simultaneously in the same way to all the points.

In quantum electrodynamics, the symmetry operation consists of a local phase change in the electron field, each such phase shift being accompanied by an interaction with the electromagnetic field. Imagine an electron undergoing two consecutive phase shifts: the emission of a photon and then the absorption of one. If the sequence of the phase shifts was reversed, the final result would be the same. It follows that an unlimited series of phase shifts can be made, and the final result will simply be their algebraic sum, no matter what their sequence is. On the contrary, in the Yang-Mills theory, where the symmetry operation is a local rotation of the isotopic-spin arrow, the result of multiple transformations may be rather different. Suppose a hadron undergoing a gauge transformation A followed by a second transformation B may have an isospin arrow in the orientation of a proton. The same hadron undergoes B; at the end of this sequence the isotopic-spin arrow is found in the orientation that corresponds to a proton. Now suppose the same transformation was applied to the same hadron but in the reverse sequence: B followed by A. In general the final state will not be the same; the particle may be a neutron instead of a proton. Therefore, the net effect of the two transformations depends explicitly on the sequence in which they are applied. Because of this distinction, quantum electrodynamics is called an Abelian theory and the Yang-Mills theory is called a non-Abelian one. Abelian groups are made up of transformations that, when applied one after another, have the commutative property; non-Abelian groups are not commutative (see [16]). (The terms are borrowed from the mathematical theory of groups created by the Norwegian mathematician N. H. Abel.) Like the Yang-Mills theory, the general theory of relativity is non-Abelian. Even the electromagnetic interaction has been incorporated into a larger theory that is non-Abelian. For now, at least, it seems all the forces of nature are governed by non-Abelian gauge theories.

This important and surprising result (i.e., the asymmetry of certain fundamental laws of physics) spurred a vast new investigation, still active today, into spontaneous symmetry breaking. The central question now seems to be the connection between the symmetry breaking occurring in the behavior of certain elementary particles at a certain level of size and temperature, and the geometrical structure of space at that same level. More precisely, it has been hypothesized that a symmetry breaking occurs when there is a change (or degeneration) in the space structure, or, mathematically speaking, a jump from a group to a "poorer" group of the field or the interaction concerned. However, nothing prevents us from believing that if there is a richer group containing the two others as subgroups, the difficulty may be removed (see below for further considerations on this point).

Mathematically the phenomenon of symmetry breaking can be formulated as follows. Let V be a vector bundle with structure group G; it might happen that under some conditions the structure group of V can be reduced to a subgroup G_0. This phenomenon of gauge symmetry breaking plays a central role in particle physics—more precisely, in the Weinberg-Salam-Glashow model of weak interactions. Suppose that at some low mass scale m, the gauge group G is effectively reduced to a subgroup G_0. Even

if the representations R and $\bar{R}$ are inequivalent as representations of G, they may be equivalent as representations of G_0. In this case, the fermions that were kept massless by the inequivalence of R and $\bar{R}$ will be able to gain masses of order m. This is precisely what seems to happen in nature. At a mass scale of order $10^{-17}M_{Pl}$, the gauge group $SU(3)\times SU(2)\times U(1)$ is reduced to $SU(3)\times U(1)$. At this point, some of the gauge fields become massive. At the same time, the representations R and $\bar{R}$ are isomorphic as representations of $SU(3)\times U(1)$, so the light fermions can and do gain mass. Many facts of this symmetry-breaking process are not yet understood, for example, why the mass scale associated with symmetry breaking is so tiny compared to the natural mass scale M_{pl}. It is, however, pretty clear that the idealization in which the masses of the particles are all zero is the situation in which the gauge group $SU(3)\times SU(2)\times U(1)$ is not broken to a subgroup.

Consider further the basic decomposition $S = S_+ \otimes S_-$ of the spinor representation S into spinors $S_\pm$ of positive and negative chirality; the distinction between S_+ and S_- is a matter of convention. Under a change of the orientation of spacetime, called a *parity transformation* by physicists, S_+ and S_- are exchanged. The representations R and $\bar{R}$ are therefore exchanged by parity. If we assume that the laws of nature are invariant under parity, then R and $\bar{R}$ must be isomorphic. The explanation of the lightness of the fermions therefore rests on parity violation. However, in the 1950s it was discovered that the weak interactions violate parity. On the other hand, parity is conserved by strong and electromagnetic interactions; this is the statement that R and $\bar{R}$ are isomorphic as representations of $SU(3)\times U(1)$. In order to overcome this contradiction, one has lately contemplated the possibility of extending the observed gauge group G to a larger group $\bar{G}$, as $SU(5)$ which contains $SU(3)\times SU(2)\times U(1)$, $SO(10)$, or the exceptional group E_6.

In physical terms, however, the problem can be put in a quite different manner. It is well known that one of the serious difficulties of the Yang-Mills theory is that when isotopic-spin symmetry becomes exact, the result is that protons and neutrons are indistinguishable; this situation is obviously contradictory. Even more troubling is the prediction of electrically charged photons. The photon is necessarily massless because it must have an infinity range. Imposing a mass on the quanta of the charged fields does not make the fields disappear, but it does confine them to a finite range. If the mass is large enough, the range can be made as small as wished. As the long-range effects are removed, the existence of the fields can be reconciled with experimental observations. The modified Yang-Mills theory was easier to understand, but the theory still had to be given a quantum-mechanical interpretation.

The problem of infinities turned out to be severer than it had been in quantum electrodynamics, and the standard recipe for renormalization would not solve it. In this respect, the fundamental idea of the Higgs mechanism was to include in the modified Yang-Mills gauge theory an extra field, one having the peculiar property that it does not vanish in the vacuum. One usually thinks of a vacuum as a space with nothing in it, but in physics that vacuum is defined more precisely as the state in which all fields have their lowest possible energy. For most fields the energy is minimized when the value of the field is zero everywhere, or in other words when the field is "turned off."

An electron field, for example, has its minimum energy when there are no electrons. The effect of the Higgs field is to provide a frame of reference in which the orientation of the isotopic-spin arrow can be determined. The Higgs field can be represented as an arrow superposed on the other isotopic-spin indicators in the imaginary internal space of a hadron. What distinguishes the arrow of the Higgs field is that it has a fixed length, established by the vacuum value of the field. The orientation of the other isotopic-spin arrows can then be measured with respect to the axis defined by the Higgs field. In this way a proton can be distinguished from a neutron. It might seem that the introduction of the Higgs field would spoil the gauge symmetry of the theory and thereby lead again to insoluble infinities. In fact, however, the gauge symmetry is not destroyed but merely cancelled. The symmetry specifies that all the laws of physics must remain invariant when the isotopic-spin arrow is rotated in an arbitrary way from place to place. This implies that the absolute orientation of the arrow cannot be determined since any experiment for measuring the orientation would have to detect some variation in a physical quantity when the arrow was rotated. With the inclusion of the Higgs field, the absolute orientation of the arrow still cannot be determined because the arrow representing the Higgs field also rotates during a gauge transformation. All what can be measured is the angle between the arrow of the Higgs field and the other isotopic-spin arrows, or in other words their relative orientation. The Higgs mechanism is an example of the process called spontaneous symmetry breaking, which was already well established in other areas of physics.

The concept was first put forward by W. Heisenberg in his description of ferromagnetic materials (in 1971). Heisenberg pointed out that the theory describing a ferromagnet has perfect geometric symmetry in that it gives nonspecial distinction to any one direction in space. When the material becomes magnetized, however, there is one axis—the direction of magnetization—that can be distinguished from all other axes. The theory is symmetrical but the object it describes is not. Similarly, the Yang-Mills theory retains its gauge symmetry with respect to rotations of the isotopic-spin arrow, but the objects described—protons and neutrons—do not express the symmetry. Philosophically, this fact leads to making the distinction between the “ontological” or “objectal” level and the “operational” or “theoretical” level of physical entities; moreover, the first level cannot be reduced to the latter one.

Despite all these difficulties, the Yang-Mills theory had begun as a model of the strong interactions, but by the time it had been renormalized, interest in it centered on applications to weak interactions. In 1967, S. Weibeng, A. Salam, and C. Ward proposed a model of the weak interactions based on a version of the Yang-Mills theory in which the gauge quanta take on mass through the Higgs mechanism. The Weinberg-Salam-Ward model actually embraces both the weak force and electromagnetism (Salam [43]). The conjecture on which the model is ultimately founded is a postulate of local invariance with respect to isotopic spin; in order to preserve that invariance, four photonlike fields are introduced, rather than the three of the original Yang-Mills theory. The fourth photon could be identified with some primordial form of electromagnetism. It corresponds to a separate force, which had to be added to the theory without explanation. For this reason the model should not be called a unified field theory.

If one were to search for a nonlinear generalization of Maxwell's equation to explain elementary particles, there are various symmetry properties one would require (see [15]):

(i) *external symmetries* under the Lorentz and Poincaré groups and under the conformal group if one is taking the rest-mass to be zero,

(ii) *internal symmetries* under groups like $SU(2)$ or $SU(3)$ to account for the known features of elementary particles,

(iii) *covariance* or the ability to be coupled with gravitation by working on a curved spacetime.

Gauge theories satisfy these basic requirements because they are geometric in character. In fact, on the mathematical side, gauge theory is a well-established branch of differential geometry, known as the theory of fibre bundles with connection. (On this topic, see [14, 18, 48].) It has much in common with Riemannian geometry which provided Einstein with the basis for his theory of general relativity. If the current expectations of Yang-Mills theory are eventually fulfilled, it will in some measure justify Einstein's point of view that the basic laws of physics should all be in geometrical form (Atiyah [3] and Wheeler [63]).

8. Topological aspects of gauge theory, and invariants of four-manifold topology and quantum field theory. We need once more to emphasize this fundamental fact. The mathematical basis of gauge field theory lies in vector bundles and the connections in them. In fact, one of the most striking developments in mathematical physics over the past quarter century has been the discovery of the fundamental role played by bundles, connections, and curvature in expressing and eventually explaining the basic laws of nature. (See [11, 12, 32].) The so-called Yang-Mills theory does reflect in the deeper way the intimate relationship between geometrical concepts and physical ideas. The key feature of Yang-Mills theory is the invariance of the physical properties of particles under a group, but in this case an infinite-dimensional group (whereas the Maxwell's equations in vacuum for the electromagnetic field are invariant under the finite-dimensional Lorentz group of linear isometries of $R^{3,1}$). Consider the classical electromagnetic field in terms of an exterior 2-form ω on $R^{3,1}$. Notice that since $d\omega = 0$, we may express ω as

$$\omega = d\alpha, \tag{8.1}$$

where α is a 1-form on R^4 called electromagnetic potential. The form α is defined only up to an exact form, that is, we may replace α by $\alpha + df$, where f is any smooth function on R^4. Such a replacement is called a *change of gauge* or a *gauge transformation*. So, the invariance of physical laws of particles interactions with respect to the group of gauge transformations lies at the heart of matter. It is called the "principle of local invariance." In an attempt to describe strong interactions at the classical level, C. N. Yang and R. Mills proposed in 1954 that the Lagrangian of the interaction should involve a potential with values in the Lie algebra of the non-Abelian group $SU(2)$, which describes the degrees of freedom of isotopic spin, the first quantum number to be understood in relation to strong interactions. Moreover, this Lagrangian should be invariant under the group

of local internal symmetries, again called gauge transformations. One of the striking features of the Yang-Mills proposal was that the potential A (an $SU(2)$-valued 1-form) was required to transform like a connection. Specifically, a gauge transformation is here defined as a map $\rho : R^{3,1} \to SU(2)$, and the transformed potential is given by

$$A\rho = \rho A \rho^{-1} + \rho d\rho^{-1}. \tag{8.2}$$

Furthermore, the proposed (Lagrangian) density was just $\|\Omega\|^2$, where $\Omega = dA + 1/2[A,A]$ was the curvature of the connection A. This connection lives on the trivialized principal $SU(2)$-bundle over $R^{3,1}$. The gauge transformation ρ simply amounts to a principal bundle automorphism.

There is at present no doubt that some mathematical concepts of fibre bundle theory have become an established part of mathematical physics because fibre bundles provide a natural and very deep framework for discussing the concepts of relativity and invariance, describing gravitation and other gauge fields, and giving a geometrical interpretation to quantization and the canonical formalism of particles and fields. Fibre bundles provide the language which is needed for dealing with local problems of differential geometry and field theory. They are necessary to understand and solve global, topological problems, such as those arising in connection with magnetic monopoles and instantons. For example, in an attempt to understand the properties of Donaldson invariants of four manifolds, E. Witten presented a new approach to using physics to illuminate Donaldson theory (Witten [64] and Donaldson [19]). (For a very illuminating survey of the Seiberg-Witten equations and their relation to topological invariants of four-manifolds, see [19]. Donaldson has stressed the importance of these equations by the following words: "Since 1982 the use of gauge theory, in the shape of the Yang-Mills instanton equations, has permeated research in 4-manifold topology. (…) A body of techniques has built up through the efforts of many mathematicians, producing results which have uncovered some of the mysteries of 4-manifold theory, and leading to substantial internal conundrums within the field itself. In the last three months of 1994 a remarkable thing happened: this research was turned on its head by the introduction of a new kind of differential-geometric equations by Seiberg and Witten: in the space of a few weeks, long-standing problems were solved, new and unexpected results were found, along with simpler new proofs of existing ones, and new vistas for research opened up" [19, page 45].) He suggested that, instead of computing the Donaldson invariants by counting $SU(2)$ instanton solutions, one can obtain the same invariants by cutting the solutions of the dual equations, which involve $U(1)$ gauge fields and monopoles. From a physical point of view, the dual description via monopoles and Abelian gauge fields should be simpler than the microscopic $SU(2)$ description since in the renormalization group sense it arises by "integrating out the irrelevant degrees of freedom."

The new monopole equations and the topological invariants of four-manifolds introduced by Witten involve two entities, a $U(1)$ connection and a "spinor" field. Thus a main prerequisite for their study is a knowledge of spinors on four-manifolds. More precisely, the most relevant notion is that of Spin^c structure. Recall that if X is an oriented, closed Riemannian four-manifold, a spin structure on X is a lift of the structure

group of the tangent bundle from $SO(4)$ to its double cover Spin(4). The exceptional isomorphism $\mathrm{Spin}(4) = SU(2)\times SU(2)$ means that this can be given a more concrete description in terms of vector bundles. Giving a spin structure is the same as giving a pair of complex 2-plane bundles S^+, $S^- \to X$, each with structure group $SU(2)$ and related to the tangent bundle by a structure map $c : TX \to \mathrm{Hom}(S^+, S^-)$. Now the map $e\wedge f \to c(e)^*c(f) - c(f)^*c(e)$ induces a map ρ from the self-dual 2-forms Λ^+ to $\mathrm{Hom}(S^+, S^-)$, which corresponds to the standard isomorphism between the Lie algebras of $SU(2)$ and $SO(3)$.

The map c is the symbol of the Dirac operator $D : \Gamma(S^+) \to \Gamma(S^-)$, and one of the most fruitful calculations in differential geometry leads to the Lichnerowicz-Weitzenbock formula for the Dirac operator:

$$D^*D\psi = \nabla^*\nabla\psi + \frac{1}{4}R\psi. \tag{8.3}$$

Here, ∇ is the covariant derivative on spinors, induced by Levi-Civita connection, and R is the scalar curvature, which acts in (12.1) by scalar multiplication at each point. If we have an additional auxiliary bundle $E \to X$, with a Hermitian metric and connection, we may consider spinors with values in E—sections of $S^\pm \otimes E$. The Dirac operator on these coupled spinors satisfies

$$D^*D\psi = \nabla^*\nabla\psi + \frac{1}{4}R\psi - F_E^+(\psi), \tag{8.4}$$

where F_E^+ is the self-dual part $1/2(F_E + *F_E)$ of the curvature E. Here, the self-dual forms act on spinors in the way described above. Now a spin structure may not exist globally—the Stiefel-Whitney class $w_2(X) \in H^2(X;\mathbb{Z}/2)$ is the obstruction—but a variant, a Spin^c structure, always does. A Spin^c structure is given by a pair of vector bundles $W^\pm$ over X with an isomorphism, say $\Lambda^2W^+ = \Lambda^2W^- = L$, such that locally $W^\pm = S^\pm \otimes L^{1/2}$, where $L^{1/2}$ is a local square root of $L : L^{1/2}\otimes L^{1/2} = L$. An old result of Hirzebruch and Hopf assures the existence of Spin^c structures on any oriented, closed four-manifold; up to an action of the finite group $H^1(X;\mathbb{Z}/2)$, they are classified by the lifts of $w^2(X)$ to $H^2(X;\mathbb{Z})$, the first Chern class of the line bundle L. A connection on L gives a Dirac operator $D : \Gamma(W^+) \to \Gamma(W^-)$, which is locally just the same as the Dirac operator on $L^{1/2}$-valued spinors. In particular we get the Lichnerowicz formula

$$D^*D\psi = \nabla^*\nabla\psi + \frac{1}{4}R\psi - \frac{1}{2}F_L^+(\psi), \tag{8.5}$$

where the factor of $1/2$ comes from the square root of L. Note that $\mathrm{Hom}(W^+, W^+) \sim \mathrm{Hom}(S^+, S^+)$.

Now, the Seiberg-Witten equations for a four-manifold X with Spin^c structure $W^\pm$ are equations for a pair (A, ψ), where

(1) A is a unitary connection on $L = \Lambda^2W^\pm$,

(2) ψ is a section of W^+.

If ξ and η are in W^+, we write $\xi\eta^*$ for the endomorphism $\theta \to \langle\theta,\eta\rangle\xi$ of W^+. The trace-free part of this endomorphism lies in the image of the map ρ, and we write $\tau(\xi,\eta)$ for the corresponding element of $\Lambda^+\otimes\mathbb{C}$. So, τ is a sesquilinear map $\tau : W^+\times W^+ \to \otimes\mathbb{C}$.

The Seiberg-Witten equations are

$$D_A\psi = 0, \qquad F_A^+ = -\tau(\psi,\psi). \tag{8.6}$$

The sign of the quadratic form $\tau(\psi,\psi)$ is crucial and underpins the whole theory.

Witten showed that (Witten [64], Seiberg and Witten (1999)), in general, the number of solutions of a system of equations weighted by the sign of the determinant of the operator analogous to T (an elliptic operator $T : \Lambda^1 \otimes (S^+ \otimes L) \to \Lambda^0 \otimes \Lambda^{2,+} \otimes (S^- \otimes L)$ is defined by $T = s^* \otimes t$) is always a topological invariant if a suitable compactness holds. If one has a gauge-invariant system of equations, and one wishes to count gauge orbits of solutions up to gauge transformations, then one requires (i) compactness, (ii) free action of the gauge group on the space of solutions. By contrast with Donaldson theory, according to which for $SU(2)$ instantons, compactness fails precisely because an instanton can shrink to zero size, the monopole equations are scale-invariant but they have no nonconstant L^2 solutions on flat $\mathbb{R}^4$.

The general problem behind the above result is that of finding topological invariant defined by solutions of partial differential equations. In differential topology one is familiar with many contexts in which the solutions of an equation $f(x) = y$ are, at the level of homology, unchanged by continuous variations of parameters. For example, f might be a map $f : P \to Q$ between oriented manifolds, then the homology class in $H_*(P)$ of $f^{-1}(y)$, for generic y in Q, is a homotopy invariant of f—just the Poincaré dual of the pullback of the fundamental cohomology class of Q. Or f might be a section of an oriented vector bundle $V \to P$, and $y = 0$, so the solutions are the zero set of the section which, assuming transversality, gives a submanifold representing the Poincaré dual of the Euler class of V. Now if we have a family of partial differential equations, depending on continuous parameters, we may hope to find similar invariants from the homology class of the solution space. This can be developed abstractly in the framework of differential topology in certain manifolds. The key points one needs to establish in order to find invariants analogous to the finite-dimensional case are the following.

(1) The maps involved should be *Fredholm* maps, which in practice means that the linearization of the equations about a solution should be represented by linear *elliptic* differential equations, say over a compact manifold. The index of the linearized equation gives the "expected dimension" of the solutions space.

(2) One needs to establish the *compactness* of the space of solutions, or some weaker analog of this.

(3) One needs to establish orientability, analogous to the finite-dimensional case; otherwise one only gets invariants modulo 2. This can be set up in terms of the index theory of families of operators. In the cases arising from gauge theory, the equations are invariant under the action of the gauge group of bundle automorphisms, and one studies spaces of solutions modulo this action.

(4) One must not encounter reducible solutions in generic one-parameter families of equations.

Now one can show that the essential features of Seiberg-Witten equations listed above define differential-topological invariants of the underlying four-manifold. Indeed, the theory is significantly simpler than for the Donaldson instanton equations (Donaldson

and Kronheimer [20]). To check the Fredholm property we can ignore the quadratic term $\tau(\psi,\psi)$ since this does not affect the symbol (leading term) of the linearization. At the level of the symbol, the linearization is given by the sum of the linearization of the $U(1)$ instanton equation, which modulo gauge is represented by the operator d^*+d^+ acting on ordinary forms, and the Dirac operator D_A. Regarding compactness, unlike the instanton case, the Seiberg-Witten moduli spaces are compact, without qualification. This follows from a priori estimates on the solutions. These can be obtained from energy estimates using integration by parts as in the previous section, or, more directly, by the maximum principle applied to second-order equations. The remaining issues are reducibles and orientations. If a nontrivial gauge transformation $g \in \mathrm{Aut}(L)$ fixes a pair (A,ψ), then ψ must be zero and $g \in U(1)$ a constant scalar. Thus, the only reducible Seiberg-Witten solutions are the self-dual $U(1)$ connections, and these do not occur in generic r-dimensional families of metrics on X, so long as $b^+(X) > r$. Thus if $b^+ > 1$, reducibles do not interfere with the definition of invariants. Considering orientations, an orientation of the moduli space is furnished by an orientation of the "determinant line" of the relevant index bundle over the space C^* of all irreducible pairs (A,ψ) modulo gauge transformation.

The most straightforward application of the Seiberg-Witten invariants is to distinguish differentiable four-manifolds within the same homeomorphism type. Myriads of examples could be given, the simplest being to show that connected sums $X_{p,q}$, say, of p copies of $\mathbb{CP}^2$ and q copies of $\overline{\mathbb{CP}}^2$, $q > 1$, for which the Seiberg-Witten invariants vanish, are not diffeomorphic to Kähler surfaces (or any other manifolds with nonzero Seiberg-Witten invariants). The Seiberg-Witten equations have led to astounding advances in four-manifold theory (see, e.g., [33]). To some extent they may well have brought the study of the gauge theory invariants to a fairly complete form, resolving many of the main problems that drove research in this area in the last ten years. Perhaps the most exciting challenge is to come to grips with the quantum field theory ideas which led to these new advances—in parallel with other important developments such as mirror symmetry, three-manifold invariants, conformal field theory—and to understand in a rigorous way the intricate structures discovered by Seiberg and Witten. At the same time there are notable questions which are left open at present. One is the question of whether all simply connected manifolds are of simple type. A more wide-ranging problem is to understand the structure of the invariants of families of four-manifolds, and the relation between the instanton and Seiberg-Witten theories, for manifolds with $b^+ = 1$. By considering an r-dimensional family of equations of either kind, one should get invariants which are, roughly speaking, cohomology classes in $H^r(B\,\mathrm{Diff}(X))$, where $B\,\mathrm{Diff}(X)$ is the classifying space of the diffeomorphism group of a four-manifold X. Then the same issues which complicate the story for ordinary invariants when $b^+ = 1$ should arise, for any X, once $r \geq b^+ - 1$. In another direction one may consider four-manifolds which are not smooth. The instanton theory can be extended to the class of *quasiconformal* four-manifolds (where the coordinate change maps are only quasiconformal, not necessarily smooth).

In order to see the relation of these results to quantum field theory, one must recall the analysis of $N = 2$ supersymmetric Yang-Mills theory. To begin, we work on flat

$\mathbb{R}^4$. It has long been known that this theory has a family of quantum vacuum states parametrized by a complex variable u, which corresponds to the four-dimensional class in Donaldson theory. For $u \to \infty$, the gauge group is spontaneously broken down to the maximal torus, the effective coupling is small, and everything can be computed using asymptotic freedom. For small u, the effective coupling is strong. Classically, at $u = 0$, the full $SU(2)$ gauge symmetry is restored. But the classical approximation is not valid near $u = 0$. Quantum mechanically, the u plane turns out to parametrize a family of elliptic curves, in fact, the modular curve of the group $\Gamma(2)$. The family can be described by the equation

$$y^2 = (x^2 - \Lambda^4)(x - u), \tag{8.7}$$

where Λ is the analog of a parameter that often goes by the same name in the theory of strong interactions. (The fact that $\Lambda \neq 0$ means that the quantum theory does not have the conformal invariance of the classical theory.) The curve (12.5) is smooth for generic u, but degenerates to a rational curve for $u = \Lambda^2, -\Lambda^2$, or ∞. Near each degeneration, the theory becomes weakly coupled, and everything is calculable, if the right variables are used. At $u = \infty$, the weak coupling is (by asymptotic freedom) in terms of the original field variables. Near $u = \pm\Lambda^2$, a magnetic monopole becomes massless; the light degrees of freedom are the monopole, dyon and a dual photon, or $U(1)$ gauge boson. In terms of the dyon and dual photon, the theory is weakly coupled and controllable near $u = \pm\Lambda^2$.

LeBrun [33] obtained some very important results concerning Einstein metrics on a generalized hyperbolic 4-space $H^4 = SO(4,1)/SO(4)$ or complex-hyperbolic 2-space $\mathbf{CH}_2 = SU(2,1)/U(2)$. He showed the following.

Theorem 8.1. *Let M^4 be a smooth compact quotient of complex hyperbolic 2-space $\mathbf{CH}_2 = SU(2,1)/U(2)$, and let g_0 be its standard complex-hyperbolic metric. Then every Einstein metric g on M is of the form $g = \lambda\varphi^* g_0$, where $\varphi : M \to M$ is a diffeomorphism and $\lambda > 0$ is a constant.*

This theorem is proved by estimating the scalar curvature of Riemannian metrics by means of the Seiberg-Witten invariants of smooth four-manifolds.

Theorem 8.2. *Infinitely many compact smooth simply connected four-manifolds with $2\chi > 3|\tau|$ do not admit Einstein metrics.*

In fact, it is possible to describe a sequence of smooth manifolds homeomorphic to $k\mathbb{CP}_2 \neq l\mathbb{CP}_2$, where $l : k$ is roughly $4 : 1$, which do not admit Einstein metrics.

Regarding the Seiberg-Witten techniques, one needs first to recall the following facts. If (X^4, J) is a compact complex surface—that is, a complex manifold of real dimension four—then there is a process called *blowing up* which produces a new complex surface by replacing some given point $x \in X$ with a complex projective line $\mathbb{CP}_1$. The resulting surface is diffeomorphic to a connected sum $X\#\mathbb{CP}_2$, where $\mathbb{CP}_2$ is the complex projective plane with the nonstandard orientation. This process can then be iterated, and in particular one may blow up any given collection of k distinct points of X so as to produce new complex surfaces diffeomorphic to $X\#k\mathbb{CP}_2$ for any positive integer k. Conversely, any compact complex surface (M, J) can be expressed as $X\#k\mathbb{CP}_2$ with

$k \geq 0$, an iterated blowup of some complex surface X which is not itself the blowup of anything else. One says that X is a *minimal model* for M. A compact complex surface (M, J) is said to be of *general type* if its minimal model X satisfies

$$(2\chi + 3\tau)(X) > 0 \tag{8.8}$$

and X is neither $\mathbb{CP}_2$-nor a $\mathbb{CP}_1$-bundle over a complex curve. For example, the degree-m hypersurface

$$\{[u : v : w : z] \in \mathbb{CP}_3 \mid u^m + v^m + w^m + z^m = 0\} \tag{8.9}$$

in complex projective three-space is of general type if $m > 4$; these examples are all simply connected and are their own minimal models. Now, starting from these facts, we have the following result.

Theorem 8.3. *Let (M, J) be a compact complex surface of general type, and let X be its minimal model. Then any Riemannian metric g on M satisfies*

$$\int_M s_g^2 d\mu_g \geq (2\chi + 3\tau)(X) \tag{8.10}$$

with equality if and only if $M = X$ and g is Kähler-Einstein with respect to some complex structure on M.

Proof. The complex structure J is a priori completely unrelated to the metric g under discussion, but its deformation class is enough to allow one to define twisted spinor bundles $V_\pm = \mathbf{S}_\pm \otimes L^{1/2}$, where L is a Hermitian line bundle with $c_1(L) = c_1(M, J)$. Now assume for simplicity that $b^+(M) > 1$. For any g, it then turns out that the Seiberg-Witten equations

$$D^\theta \Phi = 0, \qquad F^{\theta +} = i\sigma(\Phi) \tag{8.11}$$

must be satisfied by some smooth connection θ on L and some smooth section Φ of V_+. Here, D^θ is the Dirac operator coupled with θ, the purely imaginary 2-form $F^{\theta +}$ is the self-dual part of the curvature of θ, and the real-quadratic map $\sigma : V_+ \to \Lambda^2_+$ induced by the isomorphism $\Lambda^+ \otimes \mathbb{C} = \Lambda^2 \mathbf{S}_+$ satisfies $|\sigma(\Phi)|^2 = |\Phi|^4/8$. This can be made more explicit by choosing some Hermitian local trivialization of L, so that the connection θ is represented by a purely imaginary 1-form ϑ; in Penrose's spinorial abstract-index notation, the Seiberg-Witten equations then become

$$\left(\nabla_{AA'} + \frac{1}{2}\vartheta_{AA'}\right)\Phi^A = 0, \qquad \nabla^{A'}_{(A}{}^{\vartheta}{}_{B)A'} = \frac{1}{2}\Phi(A^{\Phi}B) \tag{8.12}$$

with the convention that $|\Phi|^2 = \Phi^A \Phi^A$. The number of solutions, modulo gauge equivalence and counted with appropriate multiplicities, can be shown to be independent of g; and because the equations can be solved explicitly when the metric happens to be Kähler, it is not difficult to show that this invariant is 1. It follows that there must be at least one solution for every metric g on M. □

One sees thus that Seiberg-Witten theory gives us differential-topological invariants which allow one to estimate the scalar curvature of a metric in relation to its volume. The entropy method instead allows one to deduce Ricci-curvature estimates from homotopy-theoretic assumptions.

9. The structure of fibre bundles and the topological significance of physical theories. We now return to the concept of fibre bundles or fibre spaces. That notion, being global in character, arose in topology. At first it was an attempt to find new examples of manifolds. Fiber spaces are locally, but not globally, product spaces. The presence of such a distinction is a sophisticated mathematical fact. The development of fibre spaces has to wait until invariants are found to distinguish the fiberings or even to show that globally there are nontrivial ones. The first such invariants are the characteristic classes introduced by H. Whitney and by E. Stiefel in 1935. Topology, however, forgets the algebraic structure, and in applications vector bundles, with the linear structure intact, are more useful.

A vector bundle $\pi : E \to M$ over a manifold M is, roughly speaking, a family of vector spaces parametrized by M such that it is locally a product. The vector space $E_x = \pi^{-1}(x)$ corresponding to $x \in M$ is called the fiber at x. Examples are the tangent bundle M and all tensor bundles associated to it. A more trivial bundle is the product bundle $M \times V$, where V is a fixed vector space and (x, V), $x \in M$, is the fiber at x. A vector bundle is called *real* or *complex* according to whether the fiber is a real or complex vector space. Its dimension is the dimension of the fibers. It is important that the linear structure on the fibers has a meaning so that the general linear group $GL(n, \mathbb{R})$ plays a fundamental role in matching the fibers; it is called the *structure group*. A real (resp., complex) vector bundle is called Riemannian (resp., Hermitian) if the fibers are provided with inner products. In this case the structure group is reduced to $O(n)$ (resp., $U(n)$), with n being the dimension of the fibers; the bundle is then called an $O(n)$-bundle (resp., $U(n)$-bundle). Similarly, we have the notion of an $SU(n)$-bundle. A section of the bundle E is an attachment, in a continuous and smooth manner, to every point $x \in M$, a point of the fiber E_x. In other words, it is a continuous mapping $\sigma : M \to E$ such that the composition $\pi \circ \sigma$ is the identity. This notion is a natural generalization of a vector-valued function and of a tangent vector field. In order to differentiate σ, we need a so-called connection in E. The latter allows the definition of the covariant derivative $D_X \sigma$ (X being a vector field in M), which is a new section of E. Covariant differentiation is generally not commutative; that is, $D_X \circ D_Y \neq D_Y \circ D_X$ for two vector fields X, Y in M. The measure of the noncommutativity gives the curvature of the connection; this is an analytic version of the geometric concept of nonholonomy introduced by Elie Cartan. According to him, it is important to regard the curvature as a matrix-valued exterior quadratic differential form. Its trace is a closed 2-form. More generally, the sum of all its principal minors of order k is a closed $2k$-form. It is called a characteristic class. By the de Rham theory the characteristic form of degree $2k$ determines a cohomology class of dimension $2k$, to be called a *characteristic class*. Whereas the characteristic forms depend on the connection, the characteristic class depends only on the bundle. They are the simplest invariants of the bundle. It must be an act of nature that the nontriviality of

a vector bundle is recognized through the need for a covariant differentiation and that its noncommutativity accounts for the first global invariants. This introduction of the characteristic classes gives emphasis on its local character, and the characteristic forms contain more information than the classes. When M is a compact oriented manifold, a characteristic class of the top dimension (i.e., of dimension equal to that of M) gives by integration a characteristic number. When it is an integer, it is called a topological quantum number.

These differential-geometric notions have been found to be the likely mathematical basis of a unified field theory. Weyl's gauge theory deals with a circle bundle or a $U(1)$-bundle, that is, a complex Hermitian bundle of dimension one. In studying the isotopic spin, Yang and Mills used what is essentially a connection in an $SU(2)$-bundle. It is the first instance of a non-Abelian gauge theory. From the connection the "action" can be defined. A connection in an $SU(2)$-bundle at which the action takes the minimum is called an *instanton*. (On this new theory, see [20, 24].) Its curvature has a simple expression and is called self-dual. An instanton is thus a self-dual solution of the Yang-Mills equation. When the space R^4 is compactified into the four-dimensional sphere S^4, the $SU(2)$-bundles are determined up to an isomorphism by a topological quantum number k, which is an integer. It has been proved that over S^4 the moduli (or parameter) space for the set of connections with self-dual curvature on the $SU(2)$-bundle with given $k > 0$ is a smooth manifold of dimension $8k-3$ (Atiyah et al. [5]). In physical terms this is the dimension of the space of instantons with topological quantum number $k > 0$. Instantons can claim a relation to Einstein through the following result. The group $SO(4)$ is locally isomorphic to $SU(2)\times SU(2)$, so that a Riemannian metric on a four-dimensional manifold M gives rise through projection to connections in the $SU(2)$-bundles. M is an Einstein manifold if and only if these connections are self-dual or antidual.

The notion of fibre bundle generalizes that of a Cartesian product on a manifold. Two examples from physics and geometry will clarify the need for such a generalization (for a more detailed presentation, see [21, 39].

(i) In *Aristotelian physics* both space and time are absolute, every event being defined by an instant of time and a location in space. This is equivalent to saying that spacetime E is a Cartesian product $T\times S$, where T is the time axis and S is the three-dimensional space.

(ii) In *Galilean physics* time remains absolute, but space is relative. This can be described by saying that there is a *projection* $\pi : E \to T$, that is, a surjective (onto) map π that associates to any event $p \in E$ the corresponding instant of time $t = \pi(p) \in T$. The set (line) T is called the *base space* and the set $\pi^{-1}(t)$ of all events simultaneous with p is called the *fiber* over t. Each fiber is isomorphic to the Euclidean three-dimensional space $\mathbb{R}^3$, which is therefore called the *typical fiber*. The total space E of this bundle may be trivialized, that is, represented as the Cartesian product $T\times\mathbb{R}^3$. Any such trivialization (map) $h : E \to T\times\mathbb{R}^3$ is of the form $h(p) = (\pi(p),\mathbf{r}(p))$, where $\mathbf{r}(p) = (x(p),y(p),z(p))$ are the space coordinates of the event p relative to an inertial observer. One can say that Galilean spacetime E is the total space of a fibre bundle

TABLE 9.1

Electromagnetism	Gravitation
$A'_\mu = A_\mu + \partial_\mu \chi$	$\Gamma'^{\mu}_{\nu\rho} = \Gamma^{\alpha}_{\beta\gamma}(\partial x'^{\mu}/\partial x^{\alpha})(\partial x^{\beta}/\partial x'^{\nu})(\partial x^{\gamma}/\partial x'^{\rho}) + (\partial x'^{\mu}/\partial x^{\alpha})(\partial^2 x^{\alpha}/\partial x'^{\nu}\partial x'^{\rho})$
$\partial_\mu - iA_\mu$	∇_μ
$F_{\mu\nu}$	$R^{\alpha}_{\beta\mu\nu}$
$F_{(\mu\nu,\rho)} = 0$	$R^{\alpha}_{\beta(\mu\nu,\rho)} = 0$

which is *trivial*, that is, isomorphic to the product bundle $T \times \mathbb{R}^3$, without a natural isomorphism between these bundles.

(iii) Consider now the two-dimensional sphere S^2 with a preferred orientation. Define a "dyad" as a pair of unit orthogonal vectors tangent to S^2 at a point. Let P be the set of all dyads whose orientation agrees with that of S^2. One can make P into the total space of a bundle in such a way that $\pi : P \to S^2$ is the map sending a dyad into the point at which its vectors are attached to S^2. If $e = (e_1, e_2)$ is a dyad at $x \in S^2$, then so is the pair (e'_1, e'_2), where

$$e'_1 = e_1 \cos\varphi + e_2 \sin\varphi, \qquad e'_2 = -e_1 \sin\varphi + e_2 \cos\varphi, \tag{9.1}$$

and all dyads at x may be obtained in this manner from (e_1, e_2). Therefore, $SO(2)$ is the typical fiber of the bundle $\pi : P \to S^2$. Equation (9.1) defines an *action* of the (*structure*) *group* $SO(2)$ on P. The bundle $\pi : P \to S^2$ is a simple example of a *principal bundle*. Moreover, this bundle is nontrivial in the following sense: there is no diffeomorphism $k : S^2 \times SO(2) \to P$ such that $\pi \circ k(x, a) = x$. Indeed, if such a k existed, then $s : S^2 \to P$, defined by $s(x) = k(x, a_0)$, would determine a smooth field of unit vectors on S^2. By the "no combing of S^{2n}" theorem of Brouwer, such a field σ does not exist. In general, if $\pi : E \to M$ is a bundle and N is an open subset of M, then a smooth map $\sigma : N \to P$, such that $\pi \circ \sigma = \mathrm{id}_N$, is called a (local) *section* of π. If $N = M$, then σ is a *global section*. For a principal bundle, the existence of a global section is equivalent to its triviality. Incidentally, the bundle of dyads occurs in the description of a magnetic pole of unit strength. The nontrivial nature of the bundle $\pi : P \to S^2$ shows up in the occurrence of a "string singularity" in the expression for the vector potential of the magnetic pole.

The last remark leads to what is probably the most important domain of applications of fibre bundles in theoretical physics: infinitesimal connections on principal bundles provide good geometrical models of classical gauge fields. This has been known among mathematicians and physicists for some time but, for the sake of completeness, we recall some of the arguments in favor of this view. In a notation that is standard in physics, one can consider the analogies between electromagnetism and gravitation (see Table 9.1).

The issue raised in the discussion on the significance of the electromagnetic potentials becomes clear when electromagnetism is interpreted as an (infinitesimal) connection in the space of phases. Namely, the experiments proposed by Aharonov and Bohm [1] have a very simple analog in elementary differential geometry: the surface of a cone

is locally flat, but a vector undergoing parallel transport along a loop enclosing the vertex does not return to its original position. Similarly, the phase of a wave function of a charged particle undergoes parallel transport determined by the potential. The region with the magnetic field is analogous to the vertex of the cone. Electromagnetism potentials should not be slighted, but considered for what they are: *the coefficients of a connection.*

A heuristic approach to the notion of a connection on a principal bundle shows how this concept is related to the physicist's view of gauge potentials (see [52]). Let $\pi : P \to M$ be a principal bundle with structure group G. The result of action of $a \in G$ on $p \in P$ is another point $pa \in P$, lying in the same fiber as $p, \pi(pa) = \pi(p)$. A local section $s : N \to P$ defines a diffeomorphism $k : N \times G \to \pi^{-1}(N)$ by $k(x,a) = s(x)a = p$. With the section s fixed for the moment, we may identify $s(x)$ with (x,ε) and $s(x)a$ with $(x,a) = (x,\varepsilon)a$, where ε is the unit element of G. An infinitesimal connection on P defines parallel displacement of elements of P. If $dx = (dx^\mu)$ is a small displacement at $x = \pi(p) \in N$, then the parallel transport of (x,ε) along dx results in $(x+dx, \varepsilon - A)$, where $A = A_\mu dx^\mu$ is a 1-form on N, with values in the Lie algebra $\mathcal{G}$ of G. Parallel transport should commute with the action of G such that (x,a) displaced along dx becomes $(x+dx, a - Aa)$. If $s' : N' \to P$ is another section, then there is a map $U : N \cap N' \to G$ such that

$$s'(x) = s(x)U(x) \tag{9.2}$$

for $x \in N \cap N'$. The section s' leads to the diffeomorphism $k' : N' \times G \to \pi^{-1}(N')$, $k'(x,a) = s'(x)a = s(x)U(x)a$, and

$$\begin{aligned} k'(x,a) &= k(x,Ua), \\ k'(x+dx,a) &= k(x+dx,(U+dU)a). \end{aligned} \tag{9.3}$$

Relative to k', parallel transport is described by a 1-form $A' = A'_\mu dx^\mu$. By parallel transport, the point $k'(x,\varepsilon)$ becomes $k'(x+dx, \varepsilon - A')$, which is the same as $k(x+dx,(U+dU)(\varepsilon - A'))$. On the other hand, $k'(x,\varepsilon) = k(x,U)$ is parallel to $k(x+dx, U - AU)$. Since parallel displacement in P should not depend on the choice of section (gauge), $(U+dU)(\varepsilon - A') = U - AU$. This leads to the transformation law

$$A' = U^{-1}(dU + AU) \tag{9.4}$$

of the potential under gauge transformations of the second kind. It follows from (9.4) that the G-valued 1-form

$$\omega = a^{-1}(da + Aa) \tag{9.5}$$

is independent of the section. The form ω has a simple geometric interpretation: $\varepsilon + \omega$ is the element of G that moves the point (x,a) into the point $(x,a)(\varepsilon + \omega) = (x, a + da + Aa)$ parallel to $(x+dx, a+da)$. The section-independent 1-form ω on P is called the connection form; it is the gauge-independent counterpart of the potential A. Relation (9.4) contains, as special cases, the transformation laws of the coefficients of

a linear connection (Christoffel symbols, Ricci rotation coefficients) of the electromagnetic potentials and of non-Abelian gauge potentials of Yang-Mills type. The advantage of the connection form ω, defined on P, over the potential A, defined on $N \in M$, results from the following considerations: the connection form ω is defined independently of any section, whereas A refers to a (local) section of the bundle. As a consequence, for a nontrivial bundle, the potentials are defined only locally, whereas the connection form ω is defined globally, all over P.

An interesting application of the bundle approach to gauge fields is the construction of Riemannian geometries of Kaluza-Klein type. If there is a connection form ω on P, $g = g_{\mu\nu}dx^{\mu}dx^{\nu}$ is a metric tensor on M, and h is a bi-invariant metric on G, then one can define a metric tensor γ on P by the formula

$$\gamma(u,v) = g\big(T\pi(u), T\pi(v)\big) + \text{const}\, h\big(\omega(u), \omega(v)\big), \tag{9.6}$$

where u and v are vectors tangent to P, and $T\pi : TP \to TM$ is the projection of such vectors on M, induced by π. The metric γ is invariant under the action of G on P. For $G = SO(2)$, it coincides with the metric considered in five-dimensional, "unified" theories of gravitation and electromagnetism.

Relativistic theories of gravitation—such as Einstein's theory of general relativity—may also be considered as gauge theories. The bundle P consists in this case of orthonormal linear frames (tetrads, *vierbein*) of the spacetime manifold M and G is the Lorentz group. Alternatively, one can take P to be the bundle of orthonormal affine frames, in which case G is the inhomogeneous Lorentz group. There are, however, important differences between Einstein's theory and gauge theories such as electrodynamics or the Yang-Mills theory. First of all, the bundle of frames is soldered to the base M, whereas in other gauge theories the bundle is rather loosely connected to M. The soldering results in the appearance, in theories of gravitation, of *torsion*, in addition to *curvature*, which occurs in gauge theory. (Torsion is zero in Riemannian geometry, but being zero is different from not existing at all.) Moreover, the form of Einstein's equations of gravitation is different from the "generic" form of the field equations assumed in gauge theories. The latter are derived from Lagrangians quadratic in curvature, whereas the former are based on a linear Lagrangian. The possibility of constructing such a linear Lagrangian is also related to the existence of the soldering form on P.

In the past, there were much research and discussion on whether and in what sense gravitation is a gauge theory. Recently, this problem has been considered in connection with the program of constructing a "supersymmetric" theory of gravitation. In classical relativity, the following questions have been raised and given diverse answers by different authors.

(1) What is the gauge group of gravitation?
(2) What are the corresponding gauge potentials; what is the status of the metric tensor?
(3) Can the form of the field equations be derived from arguments of gauge invariance?

Utiyama [55] was the first to say that gravitation may be looked upon as a gauge theory; he identified its potentials with the coefficients of the Riemannian connection on spacetime. Using gauge arguments, Sciama argued in favor of an asymmetric connection as the basis of gravitation and showed that spin may be the source of torsion. Independently, on the ground of heuristic considerations invoking a gauge group with translations (in addition to Lorentz transformations), Kibble derived the full set of field equations of gravitation with spin and torsion; the Sciama-Kibble theory was later recognized as being essentially equivalent to Cartan's theory of 1923 (see [13]). Chen Ning Yang pointed out that Einstein's theory is different from other gauge theories in being based on a Lagrangian that is linear, rather than quadratic, in curvature. He proposed considering a theory of gravitation based on Riemannian geometry and a Lagrangian of the form

$$ {}^*\Omega^{\mu}_{\nu} \wedge \Omega^{\nu}_{\mu} \tag{9.7} $$

(the dual ${}^*\Omega^{\mu}_{\nu}$ of the curvature form Ω^{μ}_{ν} (where $\Omega^{\mu}_{\nu} = d\omega^{\mu}_{\nu} + \omega^{\mu}_{\rho} \wedge \omega^{\rho}_{\nu}$) and the conformally invariant Lagrangian density). The source-free equations of this theory, $\nabla_{\mu} R_{\nu\rho} = \nabla_{\nu} R_{\mu\rho}$, appear to be too weak; for example, they admit as a solution the de Sitter universe with an arbitrary radius of curvature. There is a modification of Yang's theory based on a metric connection with torsion and two sets of field equations, as in the Einstein-Cartan theory. It is clear, from the diversity of results and views, that there is no unique "gauge theory of gravitation." This is due to the fact that gravitation is a "rich" theory from the geometrical point of view: it contains several invariants which may be used to build the kinetic part of the gravitational Lagrangian. The correspondence principle of relativistic gravity to the Newtonian theory suggests—but probably does not require—a Lagrangian linear in curvature, whereas the analogy with electrodynamics leads to the idea of a quadratic Lagrangian.

According to Regge [40], there is no difficulty in writing the modern (gauge) form of electromagnetism (with the compact group $SO(1)$ or $U(1)$) on a Riemannian manifold and it is possible to write, à la Cartan, general relativity as an $SO(3,1)$ gauge theory. Besides, it may be useful to recall that Cartan was largely responsible for the introduction of the concept of *torsion* in Physics. Torsion remains a very interesting idea. We need to use it, even by just declaring it to vanish, if we want to write general relativity as a gauge theory in which all fields, and not only the spin connection, appear as gauge potentials. The interesting feature of general relativity is that the associate curvature of the *vierbein*, that is, torsion, vanishes as a consequence of the variational principle of Hilbert, Einstein, and Cartan. And in fact the Lagrangian density is not invariant under all gauge transformations of the Poincaré group but only under those of the Lorentz subgroup. Although nature has prepared the gauge potentials for the full group, it ends up by requiring invariance under a subgroup only. A world with torsion would appear inescapable if we have around enough density of high-spin particles which act as sources, but this density seems at the moment well below the limit of observability. Regarding the kind of space in which torsion is supposed to appear, one can remark that it would not be any more a Riemannian manifold or, rather, none of the Riemannian

structures existing on the manifold would be directly related to Physics and the theory would not be a geometrical theory in the sense envisaged by Einstein. One could yet consider general relativity as $GL(4,\mathbb{R})$ theory with the Christoffel connection playing the role of a Yang-Mills potential. If the torsion vanishes, it follows that the Christoffel symbol is symmetrical into the two lower indices whose role is however quite different. The first index is a $GL(4,\mathbb{R})$ gauge index; the second labels instead the differentials on spacetime. We may relate them because of the accidental and marvelous fact that the Jacobian group of derivatives on a differentiable manifold is isomorphic to $GL(4,\mathbb{R})$ and that we use the same indexing for differentials and vectors in $GL(4,\mathbb{R})$. Once the symmetry is established, the theory becomes almost by definition geometrical. If there is no symmetry but we can control torsion by introducing suitable norms and bounds, then we may still speak of an almost-geometrical theory whose exact mathematical definition is still lacking. (About the work of Christoffel, see [22].)

A gauge theory is any physical theory of a dynamical variable which, at the classical level, may be identified with a connection on a principal bundle. The structure group G of the bundle P is the group of gauge transformations of the first kind; the group $\mathcal{G}$ of gauge transformations of the second kind may be identified with a subgroup of the group $\mathrm{Aut}\,P$ of all automorphisms of P. In this sense, gravitation is a gauge theory: the basic gauge field is a linear connection ω. In addition to ω, there is a metric tensor g which plays the role of a Higgs field. The most important difference between gravitation and other gauge theories is due to the soldering of the bundle of the frames LM to the base manifold M. The bundle LM is constructed in a natural and unique way from M, whereas a noncontractible M may be the base of inequivalent bundles with the same structure group. For example, LS^2 reduced to $SO(2)$ is isomorphic to $SO(3)$, but there is a denumerable set of inequivalent $SO(2)$ bundles over S^2, corresponding to the different elements of $\pi_1(SO(2)) = \mathbb{Z}$. The soldering form θ leads to torsion which has no analog in nongravitational theories. Moreover, it affects the group $\mathcal{G}$, which now consists of the automorphisms of LM preserving θ. This group contains no vertical automorphism other than the identity; it is isomorphic to the group $\mathrm{Diff}\,M$ of all diffeomorphisms of M. In a gauge theory of Yang-Mills type over Minkowski spacetime, the group $\mathcal{G}$ is isomorphic to the semidirect product of the Poincaré group by the group $\mathcal{G}_0$ of vertical automorphisms of P. In other words, in the theory of gravitation, the group $\mathcal{G}_0$ of "pure gauge" transformations reduces to the identity; all elements of $\mathcal{G}$ correspond to diffeomorphisms of M. What is the structure group G of the gravitational principal bundle? Since spacetime M is four-dimensional, if $P = LM$, then $G = GL(4,\mathbb{R})$. But one can equally well take for P the bundle AM of affine frames; in this case, G is the affine group. There is a simple correspondence between affine and linear connections, which makes it really immaterial whether one works with LM or AM. If one assumes—as one usually does—that ω and g are compatible, then the structure group of LM or AM can be restricted to the Lorentz or the Poincaré group, respectively. It is also possible to take, as the underlying bundle for a theory of gravitation, another bundle attached in a natural manner to spacetime, such as the bundle of projective frames or the first extension of LM. The corresponding structure groups are natural extensions of $GL(4,\mathbb{R})$, $O(1,3)$, or the Poincaré group.

TABLE 10.1

Gauge field terminology	Bundle terminology
Gauge (or global gauge)	Principal fibre bundle
Gauge type	Principal fibre bundle
Gauge potential b_μ^k	Connection on a principal fibre bundle
S_{ba}	Transition function
Phase factor Φ_{QP}	Parallel displacement
Field strength $f_{\mu v}^k$	Curvature
Source J_μ^k	—
Electromagnetism	Connection in a $U_1(1)$ bundle
Isotopic spin gauge field	Connection in an 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 to a nontrivial $U_1(1)$ bundle

The importance of gauge theories in modern theoretical physics is well known. Yang and Mill's new gauge theory should especially serve as a model for the study of strong interactions, including the quantum effects on them. The main feature of this gauge theory is the use of a non-Abelian Lie group, the simplest of the noncommutative continuous groups, as its invariance group. This mathematical property of the symmetry group gives a very rich structure to the theory, whose field equations are more general than Maxwell's. This already illustrates the fundamental role of both geometrical and internal symmetries in physical problems which can be handled by gauge theories. In Weyl's theory, in addition to the position variables of spacetime, there is already an internal space parameter on which the phase group acts. The field identified with the particle's wave function can therefore be seen as associating to each point of spacetime a point of the internal space, or an angle (of rotation) in the case of electromagnetism. A gauge requires that the coordinates of spacetime be combined with the parameters of the internal space. Weyl's theory satisfies the "principle of local invariance": that is, the field equations are invariant under a gauge shift.

10. Some open mathematical problems in gauge theory. In the last thirty years, elementary particle physics turned to modern mathematics. To emphasize the developments of the past decades, we reproduce the Wu and Yang dictionary [66] (see Table 10.1).

So, theoretical physics is more and more concerned with the following topics: Riemannian surfaces and their moduli spaces, the topology of compact Lie groups, Calabi-Yau spaces (Ricci flat Kähler manifolds), representation theory of affine algebra, knot theory, and so forth. If one looks carefully to some of the basic problems in theoretical physics, which heavily involve mathematics, one is reinforced in the idea that the quantization of gauge theories and the string theory require analysis and geometry of

special infinite-dimensional manifolds. Many problems can be formulated as the missing infinite-dimensional analogues of finite-dimensional results.

Some examples of infinite-dimensional geometries. (i) For gauge theories, the geometric object is $a/\mathcal{G}$. Here, a is the set of connections of a principal G-bundle P over a compact Riemannian three-manifold M. $\mathcal{G}$ is the group of gauge transformations, the automorphism of the G-bundle; it acts on a. G is the compact Lie group. $a/\mathcal{G}$ is the orbit space. Since the tangent space $T(a,\mathcal{G})$ of a at A is the space of equivariant 1-forms on P with values in the Lie algebra of G, there is a natural inner product on $T(a,A)$ invariant under $\mathcal{G}$. Therefore, $a/\mathcal{G}$ has a Riemannian structure.

(ii) For the so-called σ-model, the natural geometric object is $L(M)$, the set of free loops on M, that is, the smooth maps of S^1 into M, M a Riemannian manifold, usually compact. However, M might be $\mathbb{R}^d$ or Minkowski space $\mathbb{R}^{d-1,1}$. The tangent space of $L(M)$ at γ, $T(L(M),\gamma)$, is the set of smooth vector fields along γ (sections of $\gamma^*(T(M))$). This tangent space has an inner product

$$\langle V_1, V_2\rangle = \int \langle V_1(\gamma(t)), V_2(\gamma(t))\rangle dt \tag{10.1}$$

for $V_1, V_2 \in T(L(M),\gamma)$. Note that the inner product is *not* invariant under the action of Diff S^1, the diffeomorphisms of S^1, on $L(M)$. Here, Diff $S^1 \times L(M) \to L(M)$ with $(\phi,\gamma) \to \phi\cdot\gamma$, where $(\phi\cdot\gamma)(t) = \gamma(\phi^{-1}(t))$.

(iii) In quantum mechanics, one studies the Schrödinger operator $\Delta/2+V$ on $L_2(M)$, where Δ is the Laplacian and V is multiplication by a potential function. In quantum field theory, the operators should act on L_2 of certain function spaces or mapping spaces: **a** in (i) and $L(M)$ in (ii). One can emphasize that an alternate to the canonical formalism, studying $\Delta/2+V$ directly, is to use the Feynman-Kac formula, which expresses the heat kernel $K_T(x,y)$ of $e^{-T(\Delta/2+V)}$ as a path integral over paths from x to y:

$$K_T(x,y) = \int_{\substack{\text{paths } \gamma \\ \gamma(0)=x \\ \gamma(T)=y}} e^{-\int_0^T V(\gamma(t))dt} e^{-\gamma^2/2} D\gamma. \tag{10.2}$$

Here, $e^{-\gamma^2/2}D\gamma$ means the Wiener measure of this path space. The path integral approach for operators on $L_2(L(M))$ requires paths in $L(M)$, that is, maps $X : S^1\times[0,T] \to M$. So the measure space analogous to the space of paths is

$$\chi = \left[X : S^1\times[0,T] \longrightarrow M;\ X(\theta,0) = \gamma_0(\theta),\ X(\theta,T) = \gamma_1(\theta)\right]. \tag{10.3}$$

(iv) For gauge theories, the situation is a little more complicated. Note that a path $t \to f_t(x)$ of functions on M is a function $f(t,x)$ on $[0,T]\times M$. A connection $A = (A_\mu)$ on $[0,T]\times M$ can be transformed by a gauge transformation on $[0,T]\times M$ so that $A_0 = A(d/dt)$ is 0 (the temporal gauge; integrate the differential equation $dA_0/dt = U(t,x)A_0(t,x)$). Connections on $[0,T]\times M$ become paths of connections on M. Although there are some technical complications, one is led very quickly for path integral purposes to $\mathbf{a}/\mathcal{G}$ based on a four-dimensional manifold, usually $M\times\mathbb{R}$ (interpreted

as paths on $\mathbf{a}/\mathcal{G}$ based on M). The last geometric objects we consider are homogeneous spaces of $\mathrm{Diff}_0 S^1$, the orientation-preserving diffeomorphisms of S^1. $\mathrm{Diff}\, S^1$ enters string theory because the theory, involving as it does maps of S^1, should be invariant under reparameterizations of S^1. It is supposed to play a role similar to gauge transformations in gauge theories and $\mathrm{Diff}(M)$ for metrics on M, gravity.

(v) The space $\mathrm{Diff}_0 S^1/S^1$ can be made into a Kähler manifold: the Lie algebra of $\mathrm{Diff}_0 S^1$ is $\mathrm{Vect}(S^1)$. The tangent space of $\mathrm{Diff}_0 S^1/S^1$ at the identity coset is the set of vector fields whose 0th Fourier coefficient is 0. Thus

$$J = \frac{(d/d\theta)}{|d/d\theta|} \tag{10.4}$$

makes $\mathrm{Diff}_0 S^1/S^1$ into an invariant almost complex structure. It is easy to see that J is integrable and one assumes the Nirenberg-Newlander theorem will hold. There is a family of Kähler metrics given by the cocycles (of the Lie algebra of vector fields on S^1 after complexification) with either $a = 0$, $b \neq 0$ or $a \neq 0$, $-b/a \neq n^2$. Other interesting homogeneous spaces are $\mathrm{Diff}_0 S^1/K_n$, where K_n is the subgroup with Lie algebra generated by L_0, L_n, and L_{-n}. The case $n = 0$ is (v) above and the case $n = 1$ gives $K_n = Sl(2,\mathbb{R}) \subseteq \mathrm{Diff}_0(S^1)$. (For good introductions to the theory of Kählerian manifolds, see [30, 58].)

Mathematical note on almost complex structures and Kähler manifolds

Definition 10.1. Let M be a Hausdorff space. Let $\{U_\alpha\}_{\alpha\in A}$ be an open cover of M and suppose that for each U_α there is a homeomorphism ψ_α from U_α onto an open set D_α of $\mathbb{C}^n$ satisfying the following property: if $U_\alpha \cap U_\beta \neq \emptyset$, then the map $f_{\beta\alpha} = \psi_\beta \circ \psi_\alpha^{-1}$ from the open set $\psi_\alpha(U_\alpha \cap U_\beta)$ of $\mathbb{C}^n$ onto the open set $\psi_\beta(U_\alpha \cap U_\beta)$ of $\mathbb{C}^n$ and the map $f_{\alpha\beta} = \psi_\alpha \circ \psi_\beta^{-1}$ from $\psi_\beta(U_\alpha \cap U_\beta)$ onto $\psi_\alpha(U_\alpha \cap U_\beta)$ are both holomorphic. If M has an open cover $\{U_\alpha\}_{\alpha\in A}$ and a set of maps $\{\psi_\alpha\}_{\alpha\in A}$ with this property, then M is called a complex manifold of complex dimension n, and $\{(U_\alpha, \psi_\alpha)\}_{\alpha\in A}$ is called a holomorphic coordinate neighborhood system of M.

If we identify $\mathbb{C}^n$ with $\mathbb{R}^{2n}$, then a holomorphic map of an open set of $\mathbb{C}^n$ to an open set of $\mathbb{C}^n$, considered as a map between open sets in $\mathbb{R}^{2n}$, is analytic (because the real part and imaginary part of holomorphic function are analytic). Hence, of course, a complex manifold of complex dimension n is a $2n$-dimensional (real) analytic manifold. Let M be a complex manifold of complex dimension n and let $\{(U_\alpha, \psi_\alpha)\}_{\alpha\in A}$ be a holomorphic coordinate neighborhood system. Let U be an open set of M, ψ a homeomorphism from U onto an open set D of $\mathbb{C}^n$, and suppose they satisfy the following property: if $U \cap U_\alpha \neq \emptyset$ $(\alpha \in A)$, then the maps $\psi_\alpha \circ \psi_\alpha^{-1}$ from $\psi_\alpha(U \cap U_\alpha)$ to $\psi(U \cap U_\alpha)$ are both holomorphic. If this is the case, (U, ψ) is called a *holomorphic coordinate neighborhood* of M. For $q \in U$, set $\psi(q) = (z^1(q), \ldots, z^n(q))$. Then z^k $(k = 1, \ldots, n)$ is a complex-valued function defined on U, and we call $(z^1, \ldots, z^n)$ the *complex local coordinate system* on (U, ψ).

Let f be a complex-valued function defined on an open set E of a complex manifold M. For each point p of E, we can choose a holomorphic coordinate neighborhood

(U,ψ) such that $p \in E$. If the function $f \circ \psi^{-1}$ defined on the open set $\psi(U)$ of $\mathbb{C}^n$ is holomorphic, then f is said to be *holomorphic* in a neighborhood of p. This definition does not depend on the choice of the holomorphic coordinate neighborhood (U,ψ). Let f be holomorphic at all points of E, and let a complex local coordinate system in a neighborhood of p be $(z^1,\dots,z^n)$. Then we can write $f(q) = f(z^1(q),\dots,z^n(q))$, and the right-hand member is a holomorphic function of n variables.

An n-dimensional complex manifold M is a $2n$-dimensional manifold, so that at each point p of M, the tangent space $T_p(M)$ and its dual $T_p^*(M)$ are defined. Let $(z^1,\dots,z^n)$ be a complex local coordinate system, and let x^k and y^k be the real and imaginary parts of z^k, respectively. Then $\{(\partial/\partial x^1)_p,(\partial/\partial y^1)_p,\dots,(\partial/\partial x^n)_p,(\partial/\partial y^n)_p\}$ is a basis of $T_p(M)$ and $\{(dx^1)_p,(dy^1)_p,\dots,(dx^n)_p,(dy^n)_p\}$ is a basis of $T_p^*(M)$ dual to the former. Let M and M' be complex manifolds of complex dimensions n and m, respectively, and let ϕ be a continuous map from M to M'. If for each point p of M and each holomorphic function f on M' defined on a neighborhood of $\phi(p)$, $\phi^* f$ is also holomorphic in a neighborhood of p, then ϕ is called a *holomorphic map* from M to M'. Holomorphic maps are naturally differentiable. If ϕ is a one-to-one holomorphic map from M to M' and if the inverse map ϕ^{-1} is also a holomorphic map from M' to M, then ϕ is called a *holomorphic isomorphism* (or *holomorphism*) from M to M'.

Let $(z^1,\dots,z^n)$ be a complex local coordinate system on a neighborhood U of a point p of M. Define a linear transformation J_p of $T_p(M)$ by

$$J_p\left(\frac{\partial}{\partial x^k}\right)_p = \left(\frac{\partial}{\partial y^k}\right)_p, \quad J_p\left(\frac{\partial}{\partial y^k}\right)_p = -\left(\frac{\partial}{\partial x^k}\right)_p \quad (k = 1,\dots,n). \tag{10.5}$$

We prove that the definition of J_p does not depend on the choice of the complex local coordinate system $(z^1,\dots,z^n)$. To see this, extend J_p to a linear transformation of the complex vector space $T_p^C(M)$ set $J_p(u+iv) = J_p u + iJ_p v$ $(u,v \in T_p(M))$. Then, by (10.5) we have

$$J_p\left(\frac{\partial}{\partial z^k}\right)_p = i\left(\frac{\partial}{\partial z^k}\right)_p, \quad J_p\left(\frac{\partial}{\partial z^k}\right)_p = -i\left(\frac{\partial}{\partial z^k}\right)_p \quad (k = 1,\dots,n). \tag{10.6}$$

Hence, if an element a of $T_p^C(M)$ is a linear combination of $(\partial/\partial z^k)_p$ $(k = 1,\dots,n)$ only, then we have $J_p a = ia$, and if a is a linear combination of $(\partial/\partial z^k)_p$ $(k = 1,\dots,n)$ only, then we have $J_p a = -ia$. Now, if $(w^1,\dots,w^n)$ is also a complex local coordinate system on the neighborhood U of p and if $w^k = u^k + iv^k$, then we can define a new linear transformation I_p of $T_p(M)$ in the same manner as above. Hence J_p and I_p coincide, and this shows that the definition of J_p does not depend on the choice of the complex local coordinate system in the neighborhood of p. From (10.5), it is clear that J_p satisfies

$$J_p^2 = -1, \tag{10.7}$$

where 1 denotes the identity transformation of $T_p(M)$. The correspondence J, which assigns to each point p of M the linear transformation J_p of $T_p(M)$, is called the *almost complex structure* attached to M, which is defined more abstractly as follows.

Definition 10.2. An almost complex structure on a real differentiable manifold M is a tensor field J which is, at every point p of M, an endomorphism of the tangent space $T_p(M)$ such that $J^2 = -1$.

Now, let M and M' be almost complex manifolds with almost complex structures J and J', respectively. A mapping $f : M \to M'$ is said to be *almost complex* if $J' \circ f_* = f_* \circ J$. In this case, f is differentiable and holomorphic.

Definition 10.3. A Hermitian metric on an almost complex manifold M is a Riemannian metric g invariant by the almost complex structure J, that is, $g(JX, JY) = g(X, Y)$ for any vector fields X and Y.

A Hermitian metric thus defines a Hermitian inner product on each tangent space $T_p(M)$ with respect to the complex structure defined by J. An almost complex manifold (resp., a complex manifold) with a Hermitian metric is called an *almost Hermitian manifold* (resp., a *Hermitian manifold*).

Proposition 10.4. *Let M be an almost Hermitian manifold with almost complex structure J and metric g. Let Φ be the fundamental 2-form, N the torsion of J, and ∇ the covariant differentiation of the Riemannian connection defined by g. Then, for any vector fields X, Y, and Z on M,*

$$4g\big((\nabla_X J)Y, Z\big) = 6d\Phi(X, JY, JY) - 6d\Phi(X, Y, Z) + g\big(N(Y, Z), JX\big). \tag{10.8}$$

We now state an important theorem.

Theorem 10.5. *For an almost Hermitian manifold M with almost complex structure J and metric g, the following conditions are equivalent:*

(i) *the Riemannian connection defined by g is almost complex;*

(ii) *the almost complex structure has no torsion and the fundamental 2-form Φ is closed.*

A Hermitian metric on an almost complex manifold is called a *Kähler metric* if the fundamental 2-form is closed. An almost complex manifold (resp., a complex manifold) with a Kähler metric is called an *almost Kähler manifold* (resp., a *Kähler manifold*). An almost Hermitian manifold with $d\Phi = 0$ and $N = 0$ used to be called a *pseudo-Kähler manifold*. Since an almost complex manifold with $N = 0$ is a complex manifold, a pseudo-Kähler manifold is necessarily a Kähler manifold.

Proposition 10.6. *The curvature R and the Ricci tensor S of a Kähler manifold possess the following properties:*

(i) *$R(X, Y) \circ J = J \circ R(X, Y)$ and $R(JX, JY) = R(X, Y)$ for all vector fields X and Y;*

(ii) *$S(JX, JY) = S(X, Y)$ and $S(X, Y) = 1/2\{\text{trace of } J \circ R(X, JY)\}$ for all vector fields X and Y.*

Theorem 10.7. *For a Kähler manifold M of complex dimension n, the restricted linear holonomy group is contained in $SU(n)$ if and only if the Ricci tensor vanishes identically.*

Lemma 10.8. *For an almost complex linear connection Γ with curvature tensor R on a two-dimensional almost complex manifold M, the restricted linear holonomy group is contained in (the real representation of) $SL(n;\mathbb{C})$ if and only if*

$$\operatorname{trace} R(X,Y) = 0, \qquad \operatorname{trace} J \circ R(X,Y) = 0 \tag{10.9}$$

for all vector fields X and Y, where J denotes the almost complex structure.

Theorem 10.9. *An almost Hermitian manifold M is a Kähler manifold if and only if the bundle $U(M)$ of unitary frames admits a torsion-free connection (which is necessarily unique).*

On each almost complex manifold M, one can construct the bundle $C(M)$ of complex linear frames and study connections in $C(M)$ and their torsion. Let M be an almost complex manifold of dimension $2n$ with almost complex structure J and let J_0 be the canonical complex structure over the vector space $\mathbb{R}^{2n}$. Then a *complex linear frame* at a point x of M is a nonsingular linear mapping $u : \mathbb{R}^{2n} \to T_x(M)$ such that $u \circ J_0 = J \circ u$. One easily shows that J defines the structure of a complex vector space in $T_x(M)$, and $u : \mathbb{R}^{2n} \to T_x(M)$ is a complex linear frame at x if and only if it is a nonsingular complex linear mapping of $\mathbb{C}^n = \mathbb{R}^{2n}$ onto $T_x(M)$. The set of complex linear frames forms a principal fibre bundle over M with group $GL(n;\mathbb{C})$; it is called the *bundle of complex linear frames* and is denoted by $C(M)$. Since a bundle $C(M)$ is a subbundle of the bundle $L(M)$ of linear frames, each almost complex structure gives rise to a reduction of the structure group $GL(2n,\mathbb{R})$ of $L(M)$ to $GL(n;\mathbb{C})$. Then one gets the following results.

Proposition 10.10. *Given a $2n$-dimensional manifold M, there is a natural one-to-one correspondence between the almost complex structures and the reductions of the structure group of $L(M)$ to $GL(n;\mathbb{C})$.*

Proposition 10.11. *Given a $2n$-dimensional manifold M, there is a natural one-to-one correspondence between the almost complex structures of M and the cross-sections of the associated bundle $L(M)/GL(n;\mathbb{C})$ over M.*

We know that, given a Riemannian manifold M with metric tensor g, a linear connection Γ of M is a metric connection, that is, Γ comes from a connection in the bundle $O(M)$ of orthonormal frames if and only if g is parallel with respect to G.

Proposition 10.12. *For a linear connection Γ on an almost complex manifold M, the following conditions are equivalent:*

(i) *Γ is a connection in the bundle $C(M)$ of complex linear frames;*

(ii) *the almost complex structure J is parallel with respect to Γ.*

Theorem 10.13. *Every almost complex manifold M admits an almost complex affine connection such that its torsion T is given by $N = 8T$, where N is the torsion of the almost complex structure J of M.*

Corollary 10.14. *An almost complex manifold M admits a torsion-free almost complex affine connection if and only if the almost complex structure has no torsion.*

11. A new era in the relationship between geometry and physics: topology as a guiding principle. Mathematical and conceptual issues. Beginning in the 1970s, it was recognized that, mathematically, gauge theory is essentially one branch of differential geometry that uses the new concept of "fibre spaces" with "connections." This notion is absolutely central in the understanding of the relation between mathematical structures and physical theories, and *it directly links geometry and physics to the point that it can be said that the two are coextensive*.

Consider the mathematical concept of a space with a connection and its curvature. Let $f : M \to N$ be a map between spaces M, N, where M, say, represents a model of spacetime, and at each point p of M, there is localized a physical system with the space of internal states $f^{-1}(p)$. A connection on a geometrical object is a rule permitting the transport of the system along the curves in M. In other words, if we know part of the world lines and the initial internal state of a system in M, then, thanks to the corresponding displacement determined by the connection, we can know the future states of the system. According to recent physical theories, a gravitational field is a connection in the space of internal degrees of freedom of a gyroscope; the connection allows us to follow the evolution of the gyroscope in spacetime. An electromagnetic field is also a connection in the space of internal degrees of freedom of a quantum electron; the connection allows us to follow the evolution of the electron in spacetime. A Yang-Mills field is yet a connection in the space of internal degrees of freedom of a quark.

This geometrical image seems now to be the most universal mathematical model of an ideal universe with a small number of basic interactions. The state of matter in spacetime, at each point and each moment, is described by a section of an appropriate fibre space $N \to M$. A field is described by a connection on this fibre space. Matter acts on the connection by imposing restrictions on its curvature, and the connection acts on matter by forcing it to propagate by "parallel displacement" along world lines. The famous equations of Einstein, Maxwell and Dirac, and Yang and Mills are exactly the embodiment of this idea. The geometrical concept of connection has thus become an essential element of physics.

One can see that to each physical entity corresponds a geometrical or global differential concept. For example, field strength is identified with the curvature of the connection; the action integral is but a global measure of curvature. Certain topological and algebraic invariants in the theory of characteristic classes have been seen to be most appropriate to describe the charge of the particle in the sense of Yang and Mills. More generally, we can establish a direct *correspondence* from the concepts of gauge field theory to those of the differential geometry—and topology of fibre spaces. But how can we understand precisely the nature of such a correspondence? Inspired by an idea already proposed by Weyl in another manner [61], we support the thesis that, essentially, *physics is but geometry in act*. This implies not only that geometry yields mathematical abstract concepts like manifolds, groups, curvature, connections, and bundles, but also that it is, in a way, ontologically (or, if you wish, physically) rooted in reality, because it is an integral part of the properties of physical entities and the features of phenomena.

One could go so far as to postulate that there must be a geometrical structure, continuous or discrete according to the theory and the class of phenomena considered, underlying any given physical family of phenomena, or maybe a topological structure which would encompass at the same time the continuous and discrete characters of space and of nature into a more general mathematical scheme. To convince oneself of this, it suffices to remember that some principles of geometrical symmetry (or, equivalently, some groups) can be transformed into dynamical principles that are in turn responsible for changes in the phenomena. Should we then affirm "*in the beginning was the symmetry or the group* …?" However, this concept is not just abstract, and mathematical properties related to it have simultaneously an explanatory power and a capacity to generate a world of forces, interactions, and energy …, so that the mathematical understanding of this world cannot be separate from the understanding of reality itself. Indeed, at a deeper level, one is increasingly led to believe that symmetry may, in a hidden sense, determine almost everything. Moreover, in view of all this, it is not unreasonable to look on topology, like symmetry, as some kind of underlying or unifying principle which helps us to understand natural phenomena at the microscopic as well as the macroscopic levels.

In this regard, we note here that a connection, which is a well-defined geometrical object, is more primitive than the curvature. Therefore, we should consider the gauge potential to be more primitive than the gauge field. In fact, in electromagnetism we can show experimentally that the field can be identically zero but physical effects can still be detected; this is because the parallel transport need not be trivial if the region of space is not simply connected. The vanishing of curvature only gives information about the parallel transport around very small closed paths. Physically, the parallel transport is generally described in terms of a nonintegrable phase factor. The property of nonintegrability refers locally to the existence of a nonvanishing field, whereas large-scale nonintegrability is of a topological nature and may arise even if the field is zero. Classically, the concept of potential was introduced as a mathematical device to simplify the field equations, and the arbitrary nature of the gauge characteristic in the choice of potential indicated that the potential did not really have a physical meaning. But, geometrically, one can in fact show that such an interpretation is not satisfactory. The connection is a geometrical object and so the potential should be considered as having a physical nature. It is the choice of gauge describing the potential which has no physical meaning, and this corresponds to the fact that the geometrical fibre space where the connection sits has no (natural) horizontal sections.

A more general problem concerns the relation between purely mathematical geometry and physical geometry. According to an idea going back to Riemann and Clifford and next developed by T. Levi-Civita, E. Cartan, H. Weyl, and A. Einstein, physical concepts cannot be dissociated from geometrical ones, and inversely. Some remarks about the general relativity theory can help to understand what we mean by that. In this theory, the gravitational field is seen as the effect of a geometric distortion, a curvature or warping of spacetime. In this theory, as is well known, freely falling bodies are not treated as subject to gravitational forces, but are instead regarded as following the straightest possible path (a geodesic) in an underlying curved spacetime. In Newton's

theory of gravitation, the earth's orbit curves around the sun because the sun's gravity forces it to depart from its natural straight line motion. In Einstein's theory, there are nongravitational forces as such. The sun produces a warping of spacetime in its vicinity and the earth travels freely along a geodesic in this curved spacetime. Gravity is treated as a geometrical effect precisely because it is universal; it affects all test objects in the same way. Thus, even light will follow a curved path in a gravitational field. On a large scale, the distribution of galaxies throughout the universe will depend on the geometry of space. The fact that there might be a systematic curvature of space on a cosmological scale raises the interesting question of the *topology* of the universe. So long as space is considered to be flat, it must be either infinite in extent or else possess some sort of boundary. But if space is curved, there are other possibilities. Think of the situation with a two-dimensional sheet. A curved sheet could be closed into a sphere, for example, or a torus. It is possible to envisage a three-dimensional version of a closed spherical surface, called a hypersphere. If the universe had the topology of a hypersphere, it would posses a finite volume, but there would be no boundary or edge to space. It is not known what topology space actually possesses, but the issue is crucial to the superstring theory. (On this very interesting subject, see [23, 31].)

One of the basic assumptions in modern cosmology, the *cosmological principle*, is that on large-scale average, our universe is spatially homogeneous and isotropic. The apparent isotropy on large scales is normally explained as a consequence of spatial homogeneity, which in turn is understood as a natural result of an "inflationary" period of the early universe. An alternative approach to explaining the apparent homogeneity is to assume an expanding universe with small and finite space sections with a nontrivial topology, the "small universe" model. From the theoretical point of view, it is possible to have quantum creation of the universe with a multiply connected topology. From the observational side, this model has been used to explain the "observed" periodicity in the distribution of quasars and galaxies.

It is also worthwhile noting that to the generation of new space dimensions and structures corresponds changes in the physical state of phenomena. For example, we know that the qualitative properties of a certain physical (dynamical) system are sensitive to the dimension of the space, and that the geometrical and topological structure of the space puts constraints on the evolution of the system (see [7, 47]). We mention only one outstanding example. In 1984, the British physicist Michael Berry showed that the adiabatic evolution of energy eigenfunctions, with respect to a time-dependent quantum Hamiltonian $H(t)$, contains a phase of deeply geometrical origin in addition to the familiar dynamical phase

$$\exp -\frac{1}{h}\int E(t)dt. \tag{11.1}$$

The additional phase approaches a finite, nonzero limit as the Hamiltonian is taken more and more slowly around a closed path in its parameter space. The *geometric phase* $\gamma(C)$ (where C is a closed circuit on a sphere) measures the *anholonomy* of a physical (classical or quantum) system. Anholonomy is a geometrical phenomenon in which nonintegrability causes some variables to fail to return to their original values when others, which drive them, are altered around a cycle. The simplest anholonomy

is in the parallel transport of vectors, two examples being the change in the direction of swing of a Foucault pendulum after one rotation of the earth, and the change in the direction of linear polarization of light along a twisting ray or coiled optical fibre whose direction is altered in a cycle. *Adiabaticity* is slow change and therefore denotes phenomena at the border between dynamics and statics. Adiabatic change provides the simplest way to make quantum parallel transport happen. The variables which are cycled are parameters in the Hamiltonian of a system. If the cycling is slow, the adiabatic theorem guarantees that the system returns to its original state. But it usually acquires a nontrivial phase, a manifestation of anholonomy.

Moreover, some mathematical ideas can provide a deep and powerful connection between, on the one hand, the geometrical symmetries of space, and on the other, the dynamical behavior of material bodies. In fact, forbidding the absence of spontaneous changes in motion amounts to a statement of the laws of conservation of momentum and regular momentum. The translation symmetry of space leads directly to momentum conservation for particles, whereas the rotational symmetry implies angular momentum conservation. In addition to this, the conservation of energy can be shown to follow from the translation symmetry of time. Thus, the most fundamental and comprehensive laws of physics are seen to follow from the basic fact that empty space and time are featureless. It illustrates well the power of symmetry in ordering the natural world. An interesting question now arises. Do all the forces of nature necessarily respect the geometrical symmetries of space and time? Certainly, Maxwell's electromagnetic theory, as well as Einstein's general relativity theory, incorporates all the symmetries we have just mentioned. What about the discrete (quantic) geometrical symmetries? How can the laws of physics be tested for them?

A last remark about the possibility of discovering a deeper, yet unknown level of theory and experience is where the discrete and the continuous characters of the laws of physics are but special cases according to each other in the framework of a new unitary mathematical theory. The theory of supergravity, developed mathematically in the 1970s, which generalizes a theory of gravitation conceived by Weyl in 1923 and another by Kaluza and Klein about the same time, as well as the more recent superstring theory, gives some hope (only in theory, actually) of unifying the laws of physics (see [56]). In fact, at the base of this last theory, there is a new symmetry called *supersymmetry* that acts even on a global level. It links the two large classes of elementary particles, the fermions (such as the electron, the proton, and the neutron) and the bosons (such as the photon), which, as is well-known, have very different properties. Since supersymmetry extends from the global to the local level, it leads to a theory which includes gravity and which suggests the possibility of unifying it with the other forces. In this new perspective, it would be very interesting to study particularly the relation between the topological structure of certain (local and global) groups acting on a certain family of nonsmooth (quasiconformal or symplectic) manifolds and the corresponding kinds of physical symmetries and symmetries breaking. In fact, the study of the gauge theory invariants seems intimately related to the problem of constructing diffeomorphisms between four-manifolds, or finding embedded surfaces of a given genus, which would complement the obstructions and invariants which have been found.

12. Further remarks on the Kaluza-Klein program. Probably the best geometrical and physical—but hardly unified—theory resting on some global, topological ideas is the one due to Kaluza and Klein. Its underlying geometry is that of a five-dimensional Riemannian space with a one-parameter group of isometries. It turns out that the Kaluza-Klein space is the total space of a circle bundle and that the electromagnetic potentials play a double role: they define a connection form over the bundle and, together with the metric of spacetime, determine the five-dimensional Riemannian geometry. Gauge theories such as those based on $SU(n)$ group have a similar geometry. Since the recent views of the role of gauge field in strong and weak interactions are more and more confirmed, one is reinforced in the guess that the theory of fibre bundles with connection should provide the framework for a geometrical understanding of all fundamental physical forces. This unification seems to be considerably different from Einstein's own attempt but may be close in spirit to his program of geometrizing physics.

More specifically, in the 1920s, Kaluza and Klein proposed to further unify the concepts of internal and spacetime symmetries by reducing the former to the latter through the introduction of some extra dimension of space. The main point can be reviewed as follows. Assume that spacetime contains a fifth (spacelike) dimension, which has the topology of a circle, that is, we write

$$x^A = (x^\mu, x^5) \tag{12.1}$$

and make the identification

$$x^5 \equiv x^5 + 2\pi R. \tag{12.2}$$

Any sensible wave function will have to be periodic in x^5 and thus of the form

$$\sum_{p_5 = n/R} e^{ip_5 x^5} \psi_{p_5}(x^\mu). \tag{12.3}$$

Consider now the particular coordinate transformation

$$x^5 \longrightarrow x^5 + l_P \alpha(x^\mu), \tag{12.4}$$

where, for dimensional reasons, we have introduced a length l_P. Using (12.3), this will imply

$$\psi_{p_5}(x^\mu) \longrightarrow e^{il_P p_5 \alpha(x)} \psi_{p_5}(x^\mu) \tag{12.5}$$

which looks like the gauge transformation

$$\psi(x) \longrightarrow e^{iq\alpha(x)} \psi(x), \qquad A_\mu \longrightarrow A_\mu - \partial_\mu \alpha \tag{12.6}$$

for a field carrying charge

$$q = l_P p_5 = \frac{n l_P}{R}. \tag{12.7}$$

Furthermore, Kaluza and Klein showed that the μ^5 components of the five-dimensional metric transform like the gauge field in (12.6) and that the five-dimensional gravitational action generates the four-dimensional gravity-plus-gauge action

$$S_{\text{gravity}} = \frac{1}{16\pi G_N}\int d^4x\sqrt{-g}R(g) + \cdots,$$
$$S_{\text{gauge}} = -\frac{1}{4}\int d^4x F_{\mu\nu}^2 + \cdots, \tag{12.8}$$

provided l_P is identified with the so-called Planck length, $\sqrt{G_N\hbar} \sim 10^{-33}$ cm. Besides its conceptual beauty, Kaluza-Klein theory has two interesting consequences:

(i) electric charge is automatically quantized, thanks to quantization of momentum on a circle,

(ii) electromagnetic and gravitational interactions get unified at energies $M_c = 1/R$ since, using (12.7) for $n = 1$, $G_N M_c^2 = l_P^2/R^2 = q^2$.

Later on, the Kaluza-Klein idea was widely generalized, for example, to generate larger (non-Abelian) gauge groups from even higher-dimensional spaces endowed with suitable isometries. Kaluza-Klein theory leads to a unified *classical* theory but is based, in an essential way, on quantum mechanics: the *quantization* of momentum gives the quantization of electric charge. This means that there is no way to ignore quantum mechanics within the Kaluza-Klein theory. But are the two consistent with each other? Unfortunately, when we go from the semiclassical approximation to full-fledged quantum field theory, the problem of ultraviolet infinities immediately shows up. How do we handle that? In $D = 4$, gauge theories can be dealt with through the process of renormalization; however, no such recipe is known for gravity. As we move to $D > 4$, both gauge and gravity become nonrenormalizable. In Kaluza-Klein theory, in particular, both diverge in a similar way in the ultraviolet, another expected consequence of Kaluza-Klein unification. We thus face a kind of paradoxical situation. On the one hand, quantum mechanics is essential to the success of the Kaluza-Klein idea. At the same time, quantum field theory gives meaningless infinities and spoils the nice semiclassical results. If the beautiful Kaluza-Klein idea is to be saved, we need a better quantum theory than quantum field theory. Now such theory already exists; it is called superstring theory.

13. Superstring theory, physics, and spacetime. It seems more and more justified to believe that superstring achieves remarkable progress in the search for a theory of all fundamental interactions in nature, going all the way from gravity, which is responsible for keeping the planets in orbit around the Sun, through electromagnetism which keeps electrons in orbit around nuclei, through the strong interactions of the nuclear forces which are responsible for many forms of radioactive decay. (See [17, 45] and especially [65] which we follow here closely.)

One of the most important features of string theories is the unification of gauge couplings. There are in particular two reasons why this is a particularly compelling feature to study. On the one hand, the unification of gauge coupling—like the appearance of gravity or of gauge symmetry in the first place—is a feature intrinsic to string theory. On the other hand, viewing the situation from an experimental perspective, the unification of gauge couplings is arguably the highest-energy phenomenon that any extrapolation

from low-energy data can uncover; in this sense, it sits at what is believed to be the frontier between our low-energy $SU(3)\times SU(2)\times U(1)$ world and whatever may lie beyond. Thus, the unification of gauge couplings provides a fertile meeting ground where string theory can be tested against the results of low-energy experimentation.

Superstring theory relies crucially on the two ideas of *supersymmetry* and a *spacetime structure of* eleven *dimensions*. Supersymmetry requires that for each known particle having integer spin—0, 1, 2, and so on, measured in quantum units—there is a particle with the same mass but half-integer spin $(1/2, 3/2, 5/2$, and so on), and *vice versa*. Supersymmetry transforms the coordinate of space and time such that the laws of physics are the same for all observers. Einstein's general theory of relativity derives from this condition, and so supersymmetry implies gravity. In fact, supersymmetry predicts "supergravity," in which a particle with a spin of 2—the graviton—transmits gravitational interactions and has as a partner a graviton, with a spin of $3/2$.

Superstring theory is based on the very fundamental notion of *T-duality*, which relates two kinds of particles that arise when a string loops around a compact dimension. One kind (call them "vibrating particles") is analogous to those predicated by Kaluza and Klein and comes from vibrations of the loop of the string (see [2, 29]). Such particles are more energetic if the circle is small. In addition, the string can wind many times around the circle, like a rubber band on a wrist; its energy becomes higher the more times it wraps around and the larger the circle. Moreover, each energy level represents a new particle (call them "winding particles"). *T*-duality states that the "winding particles" for a circle of radius R are the same as the "vibration particles" for a circle of radius $1/R$, and *vice versa*. So, to a physicist, the two sets of particles are indistinguishable: a fat, compact dimension may yield apparently the same particles as a thin one.

This duality has a profound implication. For decades, physicists have been struggling to understand nature at the extremely small scales near Planck length of 10^{-33} centimeters. We have always supposed that laws of nature, as we know them, break down at smaller distances. What *T*-duality suggests, however, is that at these scales, the universe looks just the same as it does at large scales. One may even imagine that if the universe were to shrink to less than the Planck length, it would transform into a dual universe that grows bigger as the original one collapses.

Supersymmetry is a conjectured symmetry between fermions and bosons. It is an inherently quantum mechanical symmetry since the very concept of fermions is quantum mechanical. Bosonic quantities can be described by ordinary (commuting) numbers or by operators obeying commutation relations. Fermionic quantities involve anticommuting numbers or operators. Supersymmetry is an updating of special relativity to include fermionic as well as bosonic symmetries of spacetime. In developing relativity, Einstein assumed that the spacetime coordinates were bosonic; fermions had not yet been discovered. In supersymmetry the structure of spacetime is enriched by the presence of fermionic as well as bosonic coordinates. If this is true, supersymmetry explains why fermions exist in nature. Supersymmetry demands their existence. From experiments, we have some hints that nature may be supersymmetric. In string theory, elementary particles are understood as vibrating strings, and the structure of spacetime is coded in

the laws by which the strings propagate. A vibrating string is described by an auxiliary two-dimensional field theory, whose Lagrangian is roughly

$$I = \frac{1}{2}\int d\tau\, d\sigma \left(\left(\frac{\partial X}{\partial \tau} \right)^2 + \left(\frac{\partial X}{\partial \sigma} \right)^2 \right). \tag{13.1}$$

Here, $X(\tau,\sigma)$ is the position of the string at proper time τ, at a coordinate σ along the string. In string theory the auxiliary two-dimensional field theory plays a more fundamental role than spacetime, and spacetime exists only to the extent that it can be reconstructed from the two-dimensional field theory. String theory also leads in a strikingly elegant way to models of particle physics with the qualitative properties of the real world (such as the existence of quarks with electric charge and the structure of weak interactions). String theory, if correct, entails a radical change in our concepts of spacetime. That is what one would expect of a theory that reconciles general relativity with quantum mechanics.

The answer involved duality again. Duality supersymmetries of the two-dimensional field theory put a basic restriction on the validity of classical notions of spacetime. The basic duality is

$$\frac{\partial X}{\partial \tau} \Longleftrightarrow \frac{\partial X}{\partial \sigma} \tag{13.2}$$

and is just analogous to the more familiar electromagnetic duality $E \Leftrightarrow B$. In each case the duality exchanges a regime where familiar ideas in physics are adequate with one where they are not. In the case of electric-magnetic duality, the "easy" region is weak-coupling and the "hard" region is strong-coupling. In the case of the two-dimensional string theory dualities, the "easy" situation is that of large distances and the "hard" region is that in which some distances become very small.

There are at least five consistent relativistic string theories. These theories involve ten spacetime dimensions, some of which can be "compactified" or rolled up into unobservably small manifolds. Each theory consequently has various classical solutions and quantum states, and thus might be manifested in nature in different ways. This can be related notably with the fact that the strong-coupling behavior of supersymmetric string theories and field theories is governed by a web of dualities relating different theories. When one description breaks down because a coupling parameter becomes large, another description takes over. For instance, in uncompactified ten-dimensional Minkowski space, the strong-coupling limit of the type I superstring is the weakly coupled heterotic $SO(32)$ superstring; the strong-coupling limit of the type IIA superstring is related to eleven-dimensional supergravity; the strong-coupling limit of the type IIB superstring theory is equivalent to the same theory at weak coupling; and the strong coupling limit of the $E_8 \times E_8$ heterotic string involves eleven-dimensional supergravity again. Thus, after we compactify some dimensions, we learn that the different theories are all one. That is, they are different manifestations of one underlying and still mysterious theory.

The duality symmetry mentioned above also has a number of nonlinear analogs, such as "mirror symmetry," which is a relationship between two spacetimes that would be quite distinct in ordinary physics but turn out to be equivalent in string theory. The

equivalence is possible because in string theory one does not really have a classical spacetime, but only the corresponding two-dimensional field theory. Two apparently different spacetimes X and Y might correspond to equivalent two-dimensional field theories. The mirror symmetry can be related to the phenomenon of topology change. Here, one considers how space changes as a parameter—which might be the time—is varied. One starts with a spatial manifold X so large that string theory effects are unimportant. As time goes on, X shrinks and strings effects become large; the classical idea of spacetime breaks down. At still later times, the distances are large again and classical ideas are again valid, but one is on an entirely different spatial manifold Y.

Acknowledgments. We would like to warmly thank Jean-Pierre Bourguignon (IHES, Bures-sur-Yvette, and Ecole Polytecnique, Palaiseau), Francis Bailly (CNRS, Laboratoire de Physique des Solides de Bellevue), Marc Lachièze-Rey (CEA Saclay, Astrophysics Program), and Joseph Kouneiher (University of Paris-VII, Physics Department) for their helpful comments and criticisms on an early version of the paper. The author was a Fellow for the year 1997–1998 of the Institute for Advanced Study (Princeton), to whom he is indebted for partial support and for charming hospitality. During the last years, the author was also supported by the John Simon Guggenheim Memorial Foundation, the Social Science and Humanities Research Council of Canada, and the Singer-Polignac Foundation, to whom he would like to express his deep gratitude. Finally, the author warmly acknowledges the suggestions, comments, and criticism of Professors Piet Hut, Chiara Nappi, and Hugo Garcia Compean. In addition, he learned a great deal from attending seminars, especially of Edward Witten and Daniel Fried.

References

[1] Y. Aharonov and D. Bohm, *Significance of electromagnetic potentials in the quantum theory*, Phys. Rev. (2) **115** (1959), 485–491.

[2] Th. Appelquist, A. Chodos, and P. G. O. Freund (eds.), *Modern Kaluza-Klein Theories*, Frontiers in Physics, vol. 65, Addison-Wesley, California, 1987.

[3] M. F. Atiyah, *Geometry on Yang-Mills Fields*, Scuola Normale Superiore di Pisa, Pisa, 1979.

[4] M. F. Atiyah and R. Bott, *The Yang-Mills equations over Riemann surfaces*, R. Soc. Lond. Philos. Trans. Ser. A Math. Phys. Eng. Sci. **308** (1983), no. 1505, 523–615.

[5] M. F. Atiyah, N. J. Hitchin, and I. M. Singer, *Self-duality in four-dimensional Riemannian geometry*, R. Soc. Lond. Proc. Ser. A Math. Phys. Eng. Sci. **362** (1978), no. 1711, 425–461.

[6] D. Bennequin, *Questions de physique galoisienne*, Passion des Formes. Dynamique Qualitative, Sémiophysique et Intelligibilité. À René Thom (M. Porte, ed.), Presses de l'ENS Fontenay aux Roses, Fontenay, 1994, pp. 311–409.

[7] M. V. Berry, *The quantum phase, five years after*, Geometric Phases in Physics (A. Shapere and F. Wilczek, eds.), Adv. Ser. Math. Phys., vol. 5, World Scientific Publishing, New Jersey, 1989, pp. 7–28.

[8] L. Boi, *Le Problème Mathématique de l'Espace* [*The Mathematical Problem of Space*], Springer-Verlag, Berlin, 1995.

[9] ———, *Theories of space-time in modern physics*, Synthese **139** (2004), no. 3, 429–489.

[10] A. Borel, *Hermann Weyl and Lie groups*, Hermann Weyl, 1885–1985, Eidgenössische Tech. Hochschule, Zürich, 1986, pp. 53–82.

[11] J.-P. Bourguignon, *Transport parallèle et connexions en géométrie et en physique* [*Parallel transport and connections in geometry and physics*], 1830–1930: A Century of

Geometry (Paris, 1989) (L. Boi, D. Flament, and J.-M. Salanskis, eds.), Lecture Notes in Phys., vol. 402, Springer, Berlin, 1992, pp. 150–164.

[12] J.-P. Bourguignon and H. B. Lawson Jr., *Yang-Mills theory: its physical origins and differential geometric aspects*, Seminar on Differential Geometry (S.-T. Yau, ed.), Ann. of Math. Stud., vol. 102, Princeton University Press, New Jersey, 1982, pp. 395–421.

[13] E. Cartan, *Sur les variétés à connexion affine et la théorie de la relativité généralisée (première partie)*, Ann. Sci. École Norm. Sup. (3) **40** (1923), 325–412 (French).

[14] S. S. Chern, *Differentiable Manifolds*, Textos de Matemática, no. 4, Instituto de Física e Matemática, Universidade do Recife, Recife, 1959.

[15] S. Coleman, *Aspects of Symmetry: Selected Erice Lectures*, Cambridge University Press, Cambridge, 1988.

[16] A. Connes, *Essay on physics and noncommutative geometry*, The Interface of Mathematics and Particle Physics (Oxford, 1988) (D. G. Quillen, G. B. Segal, and S. T. Tsou, eds.), Inst. Math. Appl. Conf. Ser. New Ser., vol. 24, Oxford University Press, New York, 1990, pp. 9–48.

[17] K. R. Dienes, *String theory and the path to unification: a review of recent developments*, Phys. Rep. **287** (1997), no. 6, 447–525.

[18] S. K. Donaldson, *An application of gauge theory to four-dimensional topology*, J. Differential Geom. **18** (1983), no. 2, 279–315.

[19] ———, *The Seiberg-Witten equations and* 4*-manifold topology*, Bull. Amer. Math. Soc. (N.S.) **33** (1996), no. 1, 45–70.

[20] S. K. Donaldson and P. B. Kronheimer, *The Geometry of Four-Manifolds*, Oxford Mathematical Monographs, The Clarendon Press, Oxford University Press, New York, 1990.

[21] J. Ehlers, *The nature and structure of space-time*, The Physicist's Conception of Nature (J. Mehra ed.), Reidel, Dordrecht, 1973, pp. 71–91.

[22] ———, *Christoffel's work on the equivalence problem for Riemannian spaces and its importance for modern field theories of physics*, E. B. Christoffel (Aachen/Monschau, 1979), Birkhäuser, Basel, 1981, pp. 526–542.

[23] G. F. R. Ellis and D. W. Sciama, *Global and non-global problems in cosmology*, General Relativity (Papers in Honour of J. L. Synge) (L. O'Raifeartaigh, ed.), Clarendon Press, Oxford, 1972, pp. 35–59.

[24] D. S. Freed and K. K. Uhlenbeck, *Instantons and Four-Manifolds*, Mathematical Sciences Research Institute Publications, vol. 1, Springer-Verlag, New York, 1984.

[25] D. J. Gross, *Gauge theory—past, present and future*, Chen Ning Yang: a Great Physicist of the Twentieth Century (C. S. Liu and S.-T. Yau, eds.), International Press of Boston, Massachusetts, 1995, pp. 147–162.

[26] F. W. Hehl, P. von der Heyde, G. D. Kerlick, and J. M. Nester, *General relativity with spin and torsion: foundations and prospect*, Rev. Modern Phys. **48** (1976), no. 3, 393–416.

[27] D. Husemoller, *Fibre Bundles*, Graduate Texts in Mathematics, vol. 20, Springer-Verlag, New York, 1994.

[28] T. W. B. Kibble, *Geometrization of quantum mechanics*, Comm. Math. Phys. **65** (1979), no. 2, 189–201.

[29] O. Klein, *1938 Conference on New Theories in Physics*, Poland, 1938, reprinted in 1988 Conference on New Theories in Physics, Proc. 11th Warsaw Symposium on Elementary Particle Physics, (Z. Aiduk, S. Pokorski, and A. Trautman eds.), World Scientific, Singapore, 1989.

[30] S. Kobayashi and K. Nomizu, *Foundations of Differential Geometry. Vol. II*, Interscience Tracts in Pure and Applied Mathematics, no. 15, vol. II, Interscience Publishers, John Wiley & Sons, New York, 1969.

[31] M. Lachièze-Rey and J.-P. Luminet, *Cosmic topology*, Phys. Rep. **254** (1995), no. 3, 135–214.

[32] H. B. Lawson Jr., *The Theory of Gauge Fields in Four Dimensions*, CBMS Regional Conference Series in Mathematics, vol. 58, American Mathematical Society, Rhode Island, 1985.

[33] C. LeBrun, *Four-manifolds without Einstein metrics*, Math. Res. Lett. **3** (1996), no. 2, 133–147.
[34] Y. I. Manin, *Gauge Field Theory and Complex Geometry*, Grundlehren der Mathematischen Wissenschaften, vol. 289, Springer-Verlag, Berlin, 1988.
[35] J. Milnor, *Lectures on the h-Cobordism Theorem*, Notes by L. Siebenmann and J. Sondow, Princeton University Press, New Jersey, 1965.
[36] K. Moriyasu, *The renaissance of gauge theory*, Contemp. Phys. **23** (1982), 553–581.
[37] L. O'Raifeartaigh (ed.), *The Dawning of Gauge Theory*, Princeton Series in Physics, Princeton University Press, New Jersey, 1997.
[38] W. Pauli, *Zur Theorie der Garvitation und der Elektrizität von Hermann Weyl*, Physikalische Zeitschrift **20** (1919), 457–467.
[39] R. Penrose, *Structure of space-time*, Battelle Rencontres: 1967 Lectures in Mathematics and Physics (C. M. DeWitt and J. A. Wheeler, eds.), Benjamin, New York, 1968, pp. 121–235.
[40] T. Regge, *Physics and differential geometry*, 1830–1930: A Century of Geometry (Paris, 1989) (L. Boi, D. Flament, and J.-M. Salanskis, eds.), Lecture Notes in Phys., vol. 402, Springer, Berlin, 1992, pp. 270–272.
[41] A. Salam, *Invariance properties in elementary particle physics*, Lectures in Theoretical Physics (Boulder, Colo, 1959) (W. E. Brittin and B. W. Downs, eds.), Interscience, New York, 1960, pp. 1–30.
[42] _____, *Gauge unification of fundamental forces*, Rev. Modern Phys. **52** (1980), no. 3, 525–538.
[43] _____, *Unification of Fundamental Forces. The First of the 1988 Dirac Memorial Lectures*, Cambridge University Press, Cambridge, 1990.
[44] E. Scholz, *Hermann Weyl's "purely infinitesimal geometry"*, Proceedings of the International Congress of Mathematicians, Vol. 1, 2 (Zürich, 1994), Birkhäuser, Basel, 1995, pp. 1592–1603.
[45] J. H. Schwarz (ed.), *Superstrings: the First 15 Years of Superstring Theory. Vol. 1, 2*, World Scientific Publishing, New Jersey, 1985.
[46] J. Schwinger (ed.), *Selected Papers on Quantum Electrodynamics*, Dover Publications, New York, 1958.
[47] A. Shapere and F. Wilczek (eds.), *Geometric Phases in Physics*, Advanced Series in Mathematical Physics, vol. 5, World Scientific Publishing, New Jersey, 1989.
[48] N. Steenrod, *The Topology of Fibre Bundles*, Princeton Mathematical Series, vol. 14, Princeton University Press, New Jersey, 1951.
[49] N. Straumann, *Zum Ursprung der Eichtheorien bei Hermann Weyl*, Physik. Blätter **43** (1987), no. 11, 414–421 (German).
[50] C. H. Taubes, *Self-dual Yang-Mills connections on non-self-dual* 4*-manifolds*, J. Differential Geom. **17** (1982), no. 1, 139–170.
[51] R. Thom, *Quelques propriétés globales des variétés différentiables*, Comment. Math. Helv. **28** (1954), 17–86 (French).
[52] A. Trautman, *Foundations and current problems of general relativity*, Lectures on General Relativity (Brandeis Summer Institute in Theoretical Physics), Prentice-Hall, New Jersey, 1965, pp. 1–248.
[53] _____, *On the structure of the Einstein-Cartan equations*, Symposia Mathematica, Vol. XII (Convegno di Relatività, INDAM, Rome, 1972), Academic Press, London, 1973, pp. 139–162.
[54] K. K. Uhlenbeck, *Removable singularities in Yang-Mills fields*, Comm. Math. Phys. **83** (1982), no. 1, 11–29.
[55] R. Utiyama, *Invariant theoretical interpretation of interaction*, Phys. Rev. (2) **101** (1956), 1597–1607.

[56] P. van Nieuwenhuizen, *An introduction to simple supergravity and the Kaluza-Klein program*, Relativity, Groups and Topology, II (Les Houches, 1983) (B. S. DeWitt and R. Stora, eds.), North-Holland, Amsterdam, 1984, pp. 823–932.
[57] V. P. Vizgin, *Unified Field Theories in the First Third of the 20th Century*, Science Networks. Historical Studies, vol. 13, Birkhäuser Verlag, Basel, 1994.
[58] A. Weil, *Introduction à l'Étude des Variétés Kählériennes*, Publications de l'Institut de Mathématique de l'Université de Nancago, VI. Actualités Sci. Ind. no. 1267, Hermann, Paris, 1958.
[59] J. Wess and B. Zumino, *A Lagrangian model invariant under supergauge transformations*, Phys. Lett. **49B** (1974), 52–75.
[60] H. Weyl, *Gravitation und Elektrizität*, Sitzber. Preuss. Akad. Wiss. Berlin **26** (1918), 465–480 (German).
[61] ———, *Reine Infinitesimalgeometrie*, Math. Z. (1918), no. 2, 384–411 (German).
[62] ———, *Quantenmechanik und Gruppentheorie*, Z. für Phys. **46** (1927), 1–46 (German).
[63] J. A. Wheeler, *Einstein Vision*, Springer-Verlag, Berlin, 1968.
[64] E. Witten, *Monopoles and four-manifolds*, Math. Res. Lett. **1** (1994), no. 6, 769–796.
[65] ———, *Duality, spacetime and quantum mechanics*, Phys. Today **50** (1997), no. 5, 28–33.
[66] T. T. Wu and C. N. Yang, *Concept of nonintegrable phase factors and global formulation of gauge fields*, Phys. Rev. D (3) **12** (1975), no. 12, 3845–3857.
[67] C. N. Yang, *Hermann Weyl's contribution to physics*, Hermann Weyl, 1885–1985 (K. Chandrasekharan, ed.), Eidgenössische Tech. Hochschule, Zürich, 1986, pp. 7–21.

Luciano Boi: Ecole des Hautes Etudes en Sciences Sociales, Centre de Mathématiques, 54 boulevard Raspail, 75006 Paris, France
E-mail address: boi@ehess.fr