Gauge theory: Historical origins and some modern developments

Lochlainn O'Raifeartaigh

Dublin Institute for Advanced Studies, Dublin 4, Ireland

Norbert Straumann

Institut für Theoretische Physik der Universität Zürich-Irchel, Zürich, Switzerland

One of the major developments of twentieth-century physics has been the gradual recognition that a common feature of the known fundamental interactions is their gauge structure. In this article the authors review the early history of gauge theory, from Einstein's theory of gravitation to the appearance of non-Abelian gauge theories in the fifties. The authors also review the early history of dimensional reduction, which played an important role in the development of gauge theory. A description is given of how, in recent times, the ideas of gauge theory and dimensional reduction have emerged naturally in the context of string theory and noncommutative geometry.

CONTENTS

I. INTRODUCTION

It took decades until physicists understood that all known fundamental interactions can be described in terms of gauge theories. Our historical account begins with Einstein's general theory of relativity, which is a non-Abelian gauge theory of a special type (see Secs. III and VII). That other gauge theories emerged, in a slow and complicated process, gradually from general relativity and their common geometrical structure—best expressed in terms of connections of fiber bundles—is now widely recognized. Thus H. Weyl was right when he wrote in the preface to the first edition of *Space, Time, Matter (Raum.· Zeit.· Materie)* early in 1918: "Wider expanses and greater depths are now exposed to the searching eye of knowledge, regions of which we had not even a presentiment. It has brought us much nearer to grasping the plan that underlies all physical happening" (Weyl, 1922).

It was Weyl himself who in 1918 made the first attempt to extend general relativity in order to describe gravitation and electromagnetism within a unifying geometrical framework (Weyl, 1918). This brilliant proposal contains the germs of all mathematical aspects of a non-Abelian gauge theory, as we shall make clear in Sec. II. The words gauge (*Eich-*) transformation and gauge invariance appeared for the first time in this paper, but in the everyday meaning of change of length or change of calibration.[1]

Einstein admired Weyl's theory as "a coup of genius of the first rate . . . ," but immediately realized that it was physically untenable: "Although your idea is so beautiful, I have to declare frankly that, in my opinion, it is impossible that the theory corresponds to Nature." This led to an intense exchange of letters between Einstein (in Berlin) and Weyl [at the Eidgenössische Technische Hochschule (ETH) in Zürich], part of which has now been published in Vol. 8 of *The Collected Papers of Albert Einstein* (1987). [The article of Straumann (1987) gives an account of this correspondence, which is preserved in the Archives of the ETH.] No agreement was reached, but Einstein's intuition proved to be right.

Although Weyl's attempt was a failure as a physical theory, it paved the way for the correct understanding of gauge invariance. Weyl himself reinterpreted his original theory after the advent of quantum theory in a seminal paper (Weyl, 1929), which we shall discuss at length in Sec. III. Parallel developments by other workers and interconnections are indicated in Fig. 1.

[1] The German word *eichen* probably comes from the Latin *aequare*, i.e., equalizing the length to a standard one.

0034-6861/2000/72(1)/1(23)/$19.60 ©2000 The American Physical Society

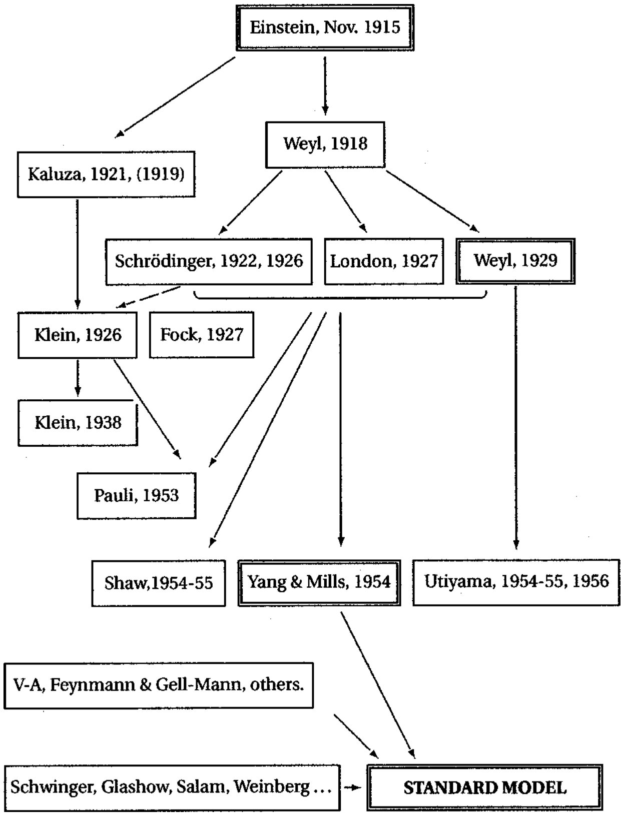


FIG. 1. Key papers in the development of gauge theories.

At the time Weyl's contributions to theoretical physics were not appreciated very much, since they did not really add new physics. The attitude of the leading theoreticians was expressed with familiar bluntness in a letter by Pauli to Weyl of July 1, 1929, after he had seen a preliminary account of Weyl's work:

> Before me lies the April edition of the *Proc. Nat. Acad. (US)*. Not only does it contain an article from you under "Physics" but shows that you are now in a "Physical Laboratory": from what I hear you have even been given a chair in "Physics" in America. I admire your courage; since the conclusion is inevitable that you wish to be judged, not for success in pure mathematics, but for your true but unhappy love for physics. (Translated from Pauli, 1979.)

Weyl's reinterpretation of his earlier speculative proposal had actually been suggested before by London and Fock, but it was Weyl who emphasized the role of gauge invariance as a *symmetry principle* from which electromagnetism can be derived. It took several decades until the importance of this symmetry principle—in its generalized form to non-Abelian gauge groups developed by Yang, Mills, and others—also became fruitful for a description of the weak and strong interactions. The mathematics of the non-Abelian generalization of Weyl's 1929 paper would have been an easy task for a mathematician of his rank, but at the time there was no motivation for this from the physics side. The known properties of the weak and strong nuclear interactions, in particular their short-range behavior, did not point to a gauge-theoretical description. We all know that the gauge symmetries of the standard model are very hidden, and it is therefore not astonishing that progress was very slow indeed.

In this paper we present only the history up to the invention of Yang-Mills theory in 1954. The independent discovery of this theory by other authors has already been described (O'Raifeartaigh, 1997). Later history covering the application of the Yang-Mills theory to the electroweak and strong interactions is beyond our scope. The main features of these applications are well known and are covered in contemporary textbooks. One modern development that we do wish to mention, however, is the emergence of both gauge theory and dimensional reduction in two fields other than traditional quantum field theory, namely, string theory and noncommutative geometry, as their emergence in these fields is a natural extension of the early history. Indeed in string theory both gauge invariance and dimensional reduction occur in such a natural way that it is probably not an exaggeration to say that, had they not been found earlier, they would have been discovered in this context. The case of noncommutative geometry is a little different, as the gauge principle is used as an input, but the change from a continuum to a discrete structure produces qualitatively new features. Amongst these is an interpretation of the Higgs field as a gauge potential and the emergence of a dimensional reduction that avoids the usual embarrassment concerning the fate of the extra dimensions.

A fuller account of the early history of gauge theory is given by O'Raifeartaigh (1997). There one can also find English translations of the most important papers of the early period, as well as Pauli's letters to Pais on non-Abelian Kaluza-Klein reductions. These works underlie the diagram in Fig. 1.

II. WEYL'S ATTEMPT TO UNIFY GRAVITATION AND ELECTROMAGNETISM

On the 1st of March 1918 Weyl writes in a letter to Einstein:

> "These days I succeeded, as I believe, to derive electricity and gravitation from a common source . . . ."

Einstein's prompt reaction by postcard indicates already a physical objection, which he explained in detail shortly afterwards. Before we come to this we have to describe Weyl's theory of 1918.

A. Weyl's generalization of Riemannian geometry

Weyl's starting point was purely mathematical. He felt a certain uneasiness about Riemannian geometry, as is clearly expressed by the following sentences early in his paper:

> But in Riemannian geometry described above there is contained a last element of geometry "at a distance" (*ferngeometrisches Element*)—with no good reason,

> as far as I can see; it is due only to the accidental development of Riemannian geometry from Euclidean geometry. The metric allows the two magnitudes of two vectors to be compared, not only at the same point, but at any arbitrarily separated points. *A true infinitesimal geometry should, however, recognize only a principle for transferring the magnitude of a vector to an infinitesimally close point* and then, on transfer to an arbitrary distant point, the integrability of the magnitude of a vector is no more to be expected than the integrability of its direction.

After these remarks Weyl turns to physical speculation and continues as follows:

> On the removal of this inconsistency there appears a geometry that, surprisingly, when applied to the world, *explains not only the gravitational phenomena but also the electrical.* According to the resultant theory both spring from the same source, indeed *in general one cannot separate gravitation and electromagnetism in a unique manner.* In this theory *all physical quantities have a world geometrical meaning; the action appears from the beginning as a pure number. It leads to an essentially unique universal law; it even allows us to understand in a certain sense why the world is four dimensional.*

In brief, Weyl's geometry can be described as follows (see also Audretsch, Gähler, and Straumann, 1984). First, the space-time manifold M is equipped with a conformal structure, i.e., with a class $[g]$ of conformally equivalent Lorentz metrics g (and not a definite metric as in general relativity). This corresponds to the requirement that it should only be possible to compare lengths at one and the same world point. Second, it is assumed, as in Riemannian geometry, that there is an affine (linear) torsion-free connection which defines a covariant derivative ∇ and respects the conformal structure. Differentially this means that for any $g \in [g]$ the covariant derivative ∇g should be proportional to g:

$$\nabla g = -2A \otimes g \quad (\nabla_\lambda g_{\mu\nu} = -2A_\lambda g_{\mu\nu}), \tag{1}$$

where $A = A_\mu \, dx^\mu$ is a differential 1-form.

Consider now a curve $\gamma:[0,1]\rightarrow M$ and a parallel-transported vector field X along γ. If l is the length of X, measured with a representative $g \in [g]$, we obtain from Eq. (1) the following relation between $l(p)$ for the initial point $p = \gamma(0)$ and $l(q)$ for the end point $q = \gamma(1)$:

$$l(q) = \exp\left(-\int_\gamma A\right) l(p). \tag{2}$$

Thus the ratio of lengths in q and p (measured with $g \in [g]$) depends in general on the connecting path γ (see Fig. 2). The length is only independent of γ if the curl of A,

$$F = dA \quad (F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu), \tag{3}$$

vanishes.

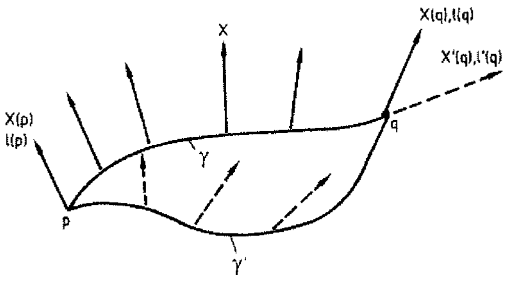


FIG. 2. Path dependence of parallel displacement and transport of length in Weyl space.

The compatibility requirement (1) leads to the following expression for the Christoffel symbols in Weyl's geometry:

$$\Gamma^\mu_{\nu\lambda} = \frac{1}{2} g^{\mu\sigma}(g_{\lambda\sigma,\nu} + g_{\sigma\nu,\lambda} - g_{\nu\lambda,\sigma}) + g^{\mu\sigma}(g_{\lambda\sigma}A_\nu + g_{\sigma\nu}A_\lambda - g_{\nu\lambda}A_\sigma). \tag{4}$$

The second A-dependent term is a characteristic new piece in Weyl's geometry, which has to be added to the Christoffel symbols of Riemannian geometry.

Until now we have chosen a fixed, but arbitrary, metric in the conformal class $[g]$. This corresponds to a choice of calibration (or gauge). Passing to another calibration with metric $\bar{g}$, related to g by

$$\bar{g} = e^{2\lambda} g, \tag{5}$$

we find that the potential A in Eq. (1) will also change to $\bar{A}$, say. Since the covariant derivative has an absolute meaning, $\bar{A}$ can easily be worked out: On the one hand we have, by definition,

$$\nabla \bar{g} = -2\bar{A} \otimes \bar{g}, \tag{6}$$

and on the other hand we find for the left side with Eq. (1)

$$\nabla \bar{g} = \nabla(e^{2\lambda} g) = 2\, d\lambda \otimes \bar{g} + e^{2\lambda} \nabla g = 2\, d\lambda \otimes \bar{g} - 2A \otimes \bar{g}. \tag{7}$$

Thus

$$\bar{A} = A - d\lambda \quad (\bar{A}_\mu = A_\mu - \partial_\mu \lambda). \tag{8}$$

This shows that a change of calibration of the metric induces a *gauge transformation* for A:

$$g \rightarrow e^{2\lambda} g, \quad A \rightarrow A - d\lambda. \tag{9}$$

Only gauge classes have an absolute meaning. [The Weyl connection is, however, gauge invariant. This is conceptually clear, but can also be verified by direct calculation from Eq. (4).]

B. Electromagnetism and gravitation

Turning to physics, Weyl assumes that his "purely infinitesimal geometry" describes the structure of space-time and consequently he requires that physical laws satisfy a double invariance: (1) They must be invariant with respect to arbitrary smooth coordinate transformations;

(2) They must be *gauge invariant*, i.e., invariant with respect to the substitutions of Eq. (9) for an arbitrary smooth function λ.

Nothing is more natural to Weyl than identifying A_μ with the vector potential and $F_{\mu\nu}$ in Eq. (3) with the field strength of electromagnetism. In the absence of electromagnetic fields $(F_{\mu\nu}=0)$ the scale factor, $\exp(-\int_\gamma A)$ in Eq. (2), for length transport becomes path independent (integrable) and one can find a gauge such that A_μ vanishes for simply connected spacetime regions. In this special case, it is the same situation as in general relativity.

Weyl proceeds to find an action that is generally invariant as well as gauge invariant and that would give the coupled field equations for g and A. We do not want to enter into this, except for the following remark. In his first paper Weyl (1918) proposes what we now call the Yang-Mills action:

$$S(g,A)=-\frac{1}{4}\int \mathrm{Tr}(\Omega\wedge *\Omega). \tag{10}$$

Here Ω denotes the curvature from and $*\Omega$ its Hodge dual.[2] Note that the latter is gauge invariant, i.e., independent of the choice of $g\in[g]$. In Weyl's geometry the curvature form splits as $\Omega=\hat{\Omega}+F$, where $\hat{\Omega}$ is the metric piece (Audretsch, Gähler, and Straumann, 1984). Correspondingly, the action also splits,

$$\mathrm{Tr}(\Omega\wedge *\Omega)=\mathrm{Tr}(\hat{\Omega}\wedge *\hat{\Omega})+F\wedge *F. \tag{11}$$

The second term is just the Maxwell action. Weyl's theory thus contains formally all aspects of a non-Abelian gauge theory.

Weyl emphasizes, of course, that the Einstein-Hilbert action is not gauge invariant. Later work by Pauli (1919) and by Weyl himself (1918, 1922) soon led to the conclusion that the action of Eq. (10) could not be the correct one, and other possibilities were investigated (see the later editions of *Space, Time, Matter*).

Independent of the precise form of the action, Weyl shows that in his theory gauge invariance implies the conservation of electric charge in much the same way as general coordinate invariance leads to the conservation of energy and momentum.[3] This beautiful connection pleased him particularly: " . . . [it] seems to me to be the strongest general argument in favour of the present theory—insofar as it is permissible to talk of justification in the context of pure speculation." The invariance principles imply five "Bianchi-type" identities. Correspondingly, the five conservation laws follow in two independent ways from the coupled field equations and may be "termed the eliminants" of the latter. These structural connections hold also in modern gauge theories.

C. Einstein's objection and reactions of other physicists

After this sketch of Weyl's theory we come to Einstein's striking counterargument, which he first communicated to Weyl by postcard (see Fig. 3). The problem is that if the idea of a nonintegrable length connection (scale factor) is correct, then the behavior of clocks would depend on their history. Consider two identical atomic clocks in adjacent world points and bring them along different world trajectories which meet again in adjacent world points. According to Eq. (2) their frequencies would then generally differ. This is in clear contradiction with empirical evidence, in particular with the existence of stable atomic spectra. Einstein therefore concludes (see Straumann, 1987):

> . . . (if) one drops the connection of the ds to the measurement of distance and time, then relativity loses all its empirical basis.

Nernst shared Einstein's objection and demanded on behalf of the Berlin Academy that it be printed in a short amendment to Weyl's article. Weyl had to accept this. One of us has described elsewhere (Straumann, 1987; see also Vol. 8 of Einstein, 1987) the intense and instructive subsequent correspondence between Weyl and Einstein. As an example, let us quote from one of the last letters of Weyl to Einstein:

> This [insistence] irritates me of course, because experience has proven that one can rely on your intuition; so unconvincing as your counterarguments seem to me, as I have to admit . . .
>
> By the way, you should not believe that I was driven to introduce the linear differential form in addition to the quadratic one by physical reasons. I wanted, just to the contrary, to get rid of this "methodological inconsistency (*Inkonsequenz*)" which has been a bone of contention to me already much earlier. And then, to my surprise, I realized that it looked as if it might explain electricity. You clap your hands above your head and shout: But physics is not made this way! (Weyl to Einstein 10 December 1918).

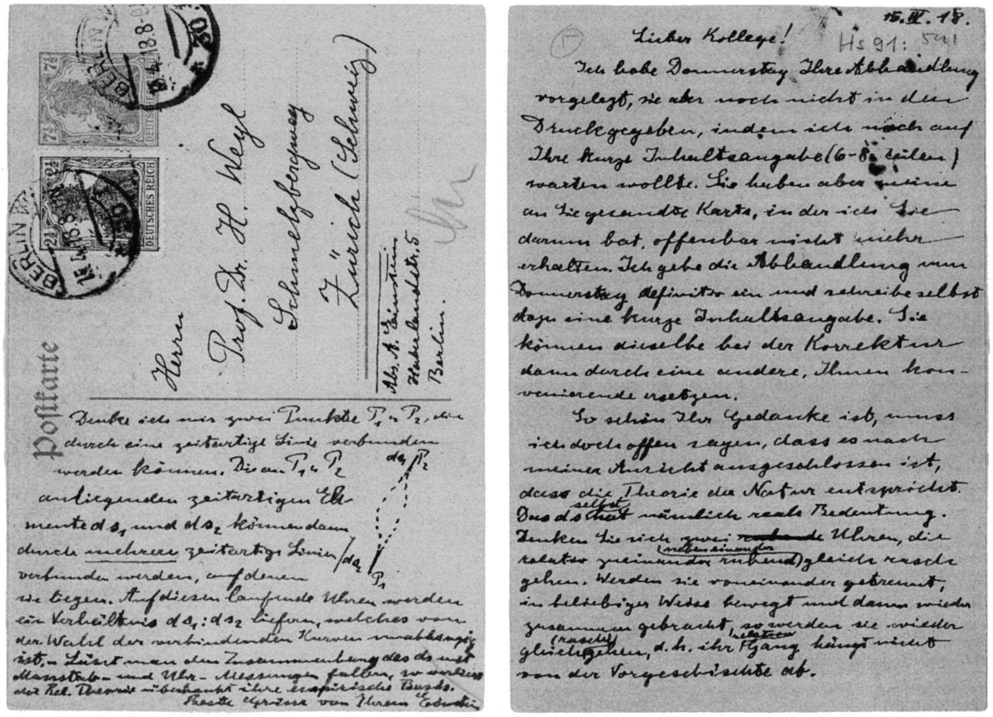
Weyl's reply to Einstein's criticism was, generally speaking, this: The real behavior of measuring rods and clocks (atoms and atomic systems) in arbitrary electromagnetic and gravitational fields can be deduced only from a dynamical theory of matter.

Not all leading physicists reacted negatively. Einstein transmitted a very positive first reaction by Planck, and Sommerfeld wrote enthusiastically to Weyl that there was " . . . hardly doubt, that you are on the correct path and not on the wrong one."

In his encyclopedia article on relativity Pauli (1921) gave a lucid and precise presentation of Weyl's theory, but commented on Weyl's point of view very critically. At the end he states:

[2] The integrand in Eq. (10) is indeed just the expression $R_{\alpha\beta\gamma\delta}R^{\alpha\beta\gamma\delta}\sqrt{-g}\,dx^0\wedge\cdots\wedge dx^3$ in local coordinates which is used by Weyl ($R_{\alpha\beta\gamma\delta}$=the curvature tensor of the Weyl connection).

[3] We adopt here the somewhat naive interpretation of energy-momentum conservation for generally invariant theories of the older literature.



Postkarte

DEUTSCHES REICH

BERLIN 15.4.18.8-9

Herrn
Prof. Dr. H. Weyl
Schmelzbergweg
Zürich (Schweiz)

Abs. A. Einstein
Haberlandstr. 5
Berlin.

Denke ich mir zwei Punkte P_1 u. P_2, die durch eine zeitartige Linie verbunden werden können. Die an P_1 u. P_2 anliegenden zeitartigen Elemente ds_1 und ds_2 können dann durch mehrere zeitartige Linien verbunden werden, auf denen sie liegen. Auf diesen laufende Uhren werden ein Verhältnis $ds_1 : ds_2$ liefern, welches von der Wahl der verbindenden Kurven unabhängig ist. – Lässt man den Zusammenhang des ds mit Masstab- und Uhr-Messungen fallen, so verliert die Rel. Theorie überhaupt ihre empirische Basis.

Beste Grüsse von Ihrem Einstein

15.IV.18.

Hs 91: 541

Lieber Kollege!

Ich habe Donnerstag Ihre Abhandlung vorgelegt, sie aber noch nicht in den Druck gegeben, indem ich noch auf Ihre kurze Inhaltsangabe (6-8 Zeilen) warten wollte. Sie haben aber meine an Sie gesandte Karte, in der ich Sie darum bat, offenbar nicht mehr erhalten. Ich gebe die Abhandlung nun Donnerstag definitiv ein und schreibe selbst dazu eine kurze Inhaltsangabe. Sie können dieselbe bei der Korrektur dann durch eine andere, Ihnen konvenierende ersetzen.

So schön Ihr Gedanke ist, muss ich doch offen sagen, dass es nach meiner Ansicht ausgeschlossen ist, dass die Theorie der Natur entspricht. Das ds selbst hat nämlich reale Bedeutung. Denken Sie sich zwei ~~ruhende~~ Uhren, die relativ zueinander ruhend (nebeneinander) gleich rasch gehen. Werden sie voneinander getrennt, in beliebiger Weise bewegt und dann wieder zusammen gebracht, so werden sie wieder (rasch) gleich gehen, d. h. ihr relativer Gang hängt nicht von der Vorgeschichte ab.

FIG. 3. Postcard from Einstein to Weyl 15 April 1918. From Archives of Eidgenössische Technische Hochschule, Zürich.

> . . . In summary one may say that Weyl's theory has not yet contributed to getting closer to the solution of the problem of matter.

Eddington's reaction was at first very positive but he soon changed his mind and denied the physical relevance of Weyl's geometry.

The situation was later appropriately summarized by London (1927) as follows:

> In the face of such elementary experimental evidence, it must have been an unusually strong metaphysical conviction that prevented Weyl from abandoning the idea that Nature would have to make use of the beautiful geometrical possibility that was offered. He stuck to his conviction and evaded discussion of the above-mentioned contradictions through a rather unclear re-interpretation of the concept of "real state," which, however, robbed his theory of its immediate physical meaning and attraction.

In this remarkable paper, London suggested a reinterpretation of Weyl's principle of gauge invariance within the new quantum mechanics: The role of the metric is taken over by the wave function, and the rescaling of the metric has to be replaced by a phase change of the wave function.

In this context an astonishing early paper by Schrödinger (1922) has to be mentioned, which also used Weyl's "world geometry" and is related to Schrödinger's later invention of wave mechanics. This precursor relation was discovered by Raman and Forman (1969). [See also the discussion by C. N. Yang in Schrödinger (1987).]

Simultaneously with London, Fock (1927) arrived along a completely different line at the principle of gauge invariance in the framework of wave mechanics. His approach was similar to that of Klein, which will be discussed in detail (in Sec. IV).

The contributions of Schrödinger (1922), London (1927), and Fock (1927) are discussed in the book of O'Raifeartaigh (1997), where English translations of the original papers can also be found. Here, we concentrate on Weyl's seminal paper "Electron and Gravitation."

III. WEYL'S 1929 CLASSIC: "ELECTRON AND GRAVITATION"

Shortly before his death late in 1955, Weyl wrote for his *Selecta* (Weyl, 1956) a postscript to his early attempt in 1918 to construct a unified field theory. There he expressed his deep attachment to the gauge idea and adds (p. 192):

> Later the quantum-theory introduced the Schrödinger-Dirac potential ψ of the electron-positron field; it carried with it an experimentally-based principle of gauge-invariance which guaranteed the conservation of charge, and connected the ψ with the electromagnetic potentials ϕ_i in the same way that my speculative theory had connected the gravitational potentials g_{ik} with the ϕ_i, and measured the ϕ_i in known atomic, rather than unknown cosmological units. I have no doubt but that the correct context for the principle of gauge-invariance is here and not, as I believed in 1918, in the intertwining of electromagnetism and gravity.

This reinterpretation was developed by Weyl in one of the great papers of this century (Weyl, 1929). Weyl's classic not only gives a very clear formulation of the gauge principle, but contains, in addition, several other important concepts and results—in particular his two-component spinor theory. The richness and scope of the paper is clearly visible from the following table of contents:

> Introduction. Relationship of General Relativity to the quantum-theoretical field equations of the spinning electron: mass, gauge-invariance, distant-parallelism. Expected modifications of the Dirac theory. $-I$. Two-component theory: the wave function ψ has only two components. −§1. Connection between the transformation of the ψ and the transformation of a normal tetrad in four-dimensional space. Asymmetry of past and future, of left and right. −§2. In General Relativity the metric at a given point is determined by a normal tetrad. Components of vectors relative to the tetrad and coordinates. Covariant differentiation of ψ. −§3. Generally invariant form of the Dirac action, characteristic for the wave-field of matter. −§4. The differential conservation law of energy and momentum and the symmetry of the energy-momentum tensor as a consequence of the double-invariance (1) with respect to coordinate transformations (2) with respect to rotation of the tetrad. Momentum and moment of momentum for matter. −§5. Einstein's classical theory of gravitation in the new analytic formulation. Gravitational energy. −§6. The electromagnetic field. From the arbitrariness of the gauge-factor in ψ appears the necessity of introducing the electromagnetic potential. Gauge invariance and charge conservation. The space-integral of charge. The introduction of mass. Discussion and rejection of another possibility in which electromagnetism appears, not as an accompanying phenomenon of matter, but of gravitation.

The modern version of the gauge principle is already spelled out in the introduction:

> The Dirac field-equations for ψ together with the Maxwell equations for the four potentials f_p of the electromagnetic field have an invariance property which is formally similar to the one which I called gauge-invariance in my 1918 theory of gravitation and electromagnetism; the equations remain invariant when one makes the simultaneous replacements
>
> $$\psi \quad \text{by} \quad e^{i\lambda}\psi \quad \text{and} \quad f_p \quad \text{by} \quad f_p - \frac{\partial\lambda}{\partial x^p},$$
>
> where λ is understood to be an arbitrary function of position in four-space. Here the factor e/ch, where $-e$ is the charge of the electron, c is the speed of light, and $h/2\pi$ is the quantum of action, has been absorbed in f_p. The connection of this "gauge invariance" to the conservation of electric charge remains untouched. But a fundamental difference, which is important to obtain agreement with observation, is that the exponent of the factor multiplying ψ is not real but purely imaginary. ψ now plays the role that Einstein's ds played before. It seems to me that this new principle of gauge-invariance, which follows not from speculation but from experiment, tells us that the electromagnetic field is a necessary accompanying phenomenon, not of gravitation, but of the material wave-field represented by ψ. Since gauge-invariance involves an arbitrary function λ it has the character of "general" relativity and can naturally only be understood in that context.

We shall soon enter into Weyl's justification, which is, not surprisingly, strongly associated with general relativity. Before this we have to describe his incorporation of the Dirac theory into general relativity, which he achieved with the help of the tetrad formalism.

One of the reasons for adapting the Dirac theory of the spinning electron to gravitation had to do with Einstein's recent unified theory, which invoked a distant parallelism with torsion. Wigner (1929) and others had noticed a connection between this theory and the spin theory of the electron. Weyl did not like this and wanted to dispense with teleparallelism. In the introduction he says:

> I prefer not to believe in distant parallelism for a number of reasons. First my mathematical intuition objects to accepting such an artificial geometry; I find it difficult to understand the force that would keep the local tetrads at different points and in rotated positions in a rigid relationship. There are, I believe, two important physical reasons as well. The loosening of the rigid relationship between the tetrads at different points converts the gauge-factor $e^{i\lambda}$, which remains arbitrary with respect to ψ, from a constant to an arbitrary function of space-time. In other words, only through the loosening of the rigidity does the established gauge-invariance become understandable.

This thought is carried out in detail after Weyl has set up his two-component theory in special relativity, including a discussion of P and T invariance. He emphasizes thereby that the two-component theory excludes a linear implementation of parity and remarks: "It is only the fact that the left-right symmetry actually appears in Nature that forces us to introduce a second pair of ψ components." To Weyl the mass problem is thus not relevant for this.[4] Indeed he says: "Mass, however, is a gravitational effect; thus there is hope of finding a substitute in the theory of gravitation that would produce the required corrections."

[4] At the time it was thought by Weyl, and indeed by all physicists, that the two-component theory required a zero mass. In 1957, after the discovery of parity nonconservation, it was found that the two-component theory could be consistent with a finite mass. See Case (1957).

A. Tetrad formalism

In order to incorporate his two-component spinors into general relativity, Weyl was forced to make use of local tetrads (*Vierbeine*). In Sec. 2 of his paper he develops the tetrad formalism in a systematic manner. This was presumably independent work, since he does not give any reference to other authors. It was, however, mainly E. Cartan (1928) who demonstrated the usefulness of locally defined orthonormal bases—also called moving frames—for the study of Riemannian geometry.

In the tetrad formalism the metric is described by an arbitrary basis of orthonormal vector fields $\{e_\alpha(x);\ \alpha=0,1,2,3\}$. If $\{e^\alpha(x)\}$ denotes the dual basis of 1-forms, the metric is given by

$$g=\eta_{\mu\nu}e^\mu(x)\otimes e^\nu(x), \quad (\eta_{\mu\nu})=\mathrm{diag}(1,-1,-1,-1). \tag{12}$$

Weyl emphasizes, of course, that only a class of such local tetrads is determined by the metric: the metric is not changed if the tetrad fields are subject to space-time-dependent Lorentz transformations:

$$e^\alpha(x)\rightarrow\Lambda^\alpha_\beta(x)e^\beta(x). \tag{13}$$

With respect to a tetrad, the connection forms $\omega=(\omega^\alpha_\beta)$ have values in the Lie algebra of the homogeneous Lorentz group:

$$\omega_{\alpha\beta}+\omega_{\beta\alpha}=0. \tag{14}$$

(Indices are raised and lowered with $\eta^{\alpha\beta}$ and $\eta_{\alpha\beta}$, respectively.) They are determined (in terms of the tetrad) by the first structure equation of Cartan:

$$de^\alpha+\omega^\alpha_\beta\wedge e^\beta=0. \tag{15}$$

(For a textbook derivation see Straumann, 1984.) Under local Lorentz transformations [Eq. (13)] the connection forms transform in the same way as the gauge potential of a non-Abelian gauge theory:

$$\omega(x)\rightarrow\Lambda(x)\omega(x)\Lambda^{-1}(x)-d\Lambda(x)\Lambda^{-1}(x). \tag{16}$$

The curvature forms $\Omega=(\Omega^\mu_\nu)$ are obtained from ω in exactly the same way as the Yang-Mills field strength from the gauge potential:

$$\Omega=d\omega+\omega\wedge\omega \tag{17}$$

(second structure equation).

For a vector field V, with components V^α relative to $\{e_\alpha\}$, the covariant derivative DV is given by

$$DV^\alpha=dV^\alpha+\omega^\alpha_\beta V^\beta. \tag{18}$$

Weyl generalizes this in a unique manner to spinor fields ψ:

$$D\psi=d\psi+\frac{1}{4}\omega_{\alpha\beta}\sigma^{\alpha\beta}\psi. \tag{19}$$

Here, the $\sigma^{\alpha\beta}$ describe infinitesimal Lorentz transformations (in the representation of ψ). For a Dirac field these are the familiar matrices

$$\sigma^{\alpha\beta}=\frac{1}{2}[\gamma^\alpha,\gamma^\beta]. \tag{20}$$

(For two-component Weyl fields, one has similar expressions in terms of the Pauli matrices.)

With these tools the action principle for the coupled Einstein-Dirac system can be set up. In the massless case the Lagrangian is

$$\mathcal{L}=\frac{1}{16\pi G}R-i\bar{\psi}\gamma^\mu D_\mu\psi, \tag{21}$$

where the first term is just the Einstein-Hilbert Lagrangian (which is linear in Ω). Weyl discusses, of course, immediately the consequences of the following two symmetries:

(i) local Lorentz invariance,

(ii) general coordinate invariance.

B. The new form of the gauge principle

All this is a kind of a preparation for the final section of Weyl's paper, which has the title "electric field." Weyl says:

> We come now to the critical part of the theory. In my opinion the origin and necessity for the electromagnetic field is in the following. The components ψ_1,ψ_2 are, in fact, not uniquely determined by the tetrad but only to the extent that they can still be multiplied by an arbitrary "gauge-factor" $e^{i\lambda}$. The transformation of the ψ induced by a rotation of the tetrad is determined only up to such a factor. In special relativity one must regard this gauge-factor as a constant because here we have only a single point-independent tetrad. Not so in general relativity; every point has its own tetrad and hence its own arbitrary gauge-factor; because by the removal of the rigid connection between tetrads at different points the gauge-factor necessarily becomes an arbitrary function of position.

In this manner Weyl arrives at the gauge principle in its modern form and emphasizes "From the arbitrariness of the gauge factor in ψ appears the necessity of introducing the electromagnetic potential." The first term $d\psi$ in Eq. (19) now has to be replaced by the covariant gauge derivative $(d-ieA)\psi$, and the nonintegrable scale factor (2) of the old theory is now replaced by a phase factor:

$$\exp\left(-\int_\gamma A\right)\rightarrow\exp\left(-i\int_\gamma A\right),$$

which corresponds to the replacement of the original gauge group **R** by the compact group $U(1)$. Accordingly, the original Gedankenexperiment of Einstein translates now to the Aharonov-Bohm effect, as was first pointed out by Yang (1980). The close connection between gauge invariance and conservation of charge is again revealed. The current conservation follows, as in the original theory, in two independent ways: On the one hand, it is a consequence of the field equations for matter plus gauge invariance. On the other hand, however, it is also a consequence of the field equations for the electromagnetic field plus gauge invariance. This corresponds to an identity in the coupled system of field equations that has to exist as a result of gauge invariance. All this is now familiar to students of physics and does not need to be explained in more detail.

Much of Weyl's paper appeared also in his classic book *The Theory of Groups and Quantum Mechanics* (Weyl, 1981). There he mentions the transformation of his early gauge-theoretic ideas: ''This principle of gauge invariance is quite analogous to that previously set up by the author, on speculative grounds, in order to arrive at a unified theory of gravitation and electricity. But I now believe that this gauge invariance does not tie together electricity and gravitation, but rather electricity and matter.''

When Pauli saw the full version of Weyl's paper he became more friendly and wrote (Pauli, 1979, p. 518):

> In contrast to the nasty things I said, the essential part of my last letter has since been overtaken, particularly by your paper in *Z. f. Physik*. For this reason I have afterward even regretted that I wrote to you. After studying your paper I believe that I have really understood what you wanted to do (this was not the case in respect of the little note in the *Proc. Nat. Acad.*). First let me emphasize that side of the matter concerning which I am in full agreement with you: your incorporation of spinor theory into gravitational theory. I am as dissatisfied as you are with distant parallelism and your proposal to let the tetrads rotate independently at different space-points is a true solution.

In brackets Pauli adds:

> Here I must admit your ability in Physics. Your earlier theory with $g'_{ik}=\lambda g_{ik}$ was pure mathematics and unphysical. Einstein was justified in criticizing and scolding. Now the hour of your revenge has arrived.

Then he remarks, in connection with the mass problem,

> Your method is valid even for the massive [Dirac] case. I thereby come to the other side of the matter, namely, the unsolved difficulties of the Dirac theory (two signs of m_0) and the question of the 2-component theory. In my opinion these problems will not be solved by gravitation . . . the gravitational effects will always be much too small.

Many years later, Weyl summarized this early tortuous history of gauge theory in an instructive letter (Seelig, 1960) to the Swiss writer and Einstein biographer C. Seelig, which we reproduce in an English translation.

> The first attempt to develop a unified field theory of gravitation and electromagnetism dates to my first attempt in 1918, in which I added the principle of gauge invariance to that of coordinate invariance. I myself have long since abandoned this theory in favour of its correct interpretation: gauge invariance as a principle that connects electromagnetism not with gravitation but with the wave-field of the electron. —Einstein was against it [the original theory] from the beginning, and this led to many discussions. I thought that I could answer his concrete objections. In the end he said ''Well, Weyl, let us leave it at that! In such a speculative manner, without any guiding physical principle, one cannot make Physics.'' Today one could say that in this respect we have exchanged our points of view. Einstein believes that in this field [Gravitation and Electromagnetism] the gap between ideas and experience is so wide that only the path of mathematical speculation, whose consequences must, of course, be developed and confronted with experiment, has a chance of success. Meanwhile my own confidence in pure speculation has diminished, and I see a need for a closer connection with quantum-physics experiments, since in my opinion it is not sufficient to unify Electromagnetism and Gravity. The wave-fields of the electron and whatever other irreducible elementary particles may appear must also be included.

Independently of Weyl, Fock (1929) also incorporated the Dirac equation into general relativity using the same method. On the other hand, Tetrode (1928), Schrödinger (1932), and Bargmann (1932) reached this goal by starting with space-time-dependent γ matrices, satisfying $\{\gamma^\mu,\gamma^\nu\}=2g^{\mu\nu}$. A somewhat later work by Infeld and van der Waerden (1932) is based on spinor analysis.

IV. THE EARLY WORK OF KALUZA AND KLEIN

Early in 1919 Einstein received a paper of Theodor Kaluza, a young mathematician (Privatdozent) and consummate linguist in Königsberg. Inspired by the work of Weyl one year earlier, he proposed another geometrical unification of gravitation and electromagnetism by extending space-time to a five-dimensional pseudo-Riemannian manifold. Einstein reacted very positively. On 21 April 1919 he writes, ''The idea of achieving [a unified theory] by means of a five-dimensional cylinder world never dawned on me . . . . At first glance I like your idea enormously.'' A few weeks later he adds: ''The formal unity of your theory is starting.'' For unknown reasons, Einstein submitted Kaluza's paper to the Prussian Academy after a delay of two years (Kaluza, 1921).

Kaluza was actually not the first who envisaged a five-dimensional unification. It is astonishing to note that G. Nordström had this idea already in 1914 (Nordström, 1914). We recall that Nordström had worked out in several papers (Nordström, 1912, 1913a, 1913b) a scalar theory of gravitation that was regarded by Einstein as the only serious competitor to general relativity.[5] (In collaboration with Fokker, Einstein gave this theory a generally covariant, conformally flat form.) Nordström started in his unification attempt with five-dimensional electrodynamics and imposed the "cylinder condition," that the fields should not depend on the fifth coordinate. Then the five-dimensional gauge potential ${}^{(5)}A$ splits as ${}^{(5)}A = A + \phi\, dx^5$, where A is a four-dimensional gauge potential and ϕ is a space-time scalar field. The Maxwell field splits correspondingly, ${}^{(5)}F = F + d\phi \wedge dx^5$, and hence the free Maxwell Lagrangian becomes

$$-\frac{1}{4}({}^{(5)}F|{}^{(5)}F) = -\frac{1}{4}(F|F) + \frac{1}{2}(d\phi|d\phi). \tag{22}$$

In this manner Nordström arrived at a unification of his theory of gravity and electromagnetism. [The matter source (five-current) is decomposed correspondingly.] It seems that this early attempt left, as far as we know, no traces in the literature.

We now return to Kaluza's attempt. Like Nordström he assumes the cylinder condition. Then the five-dimensional metric tensor splits into the four-dimensional fields $g_{\mu\nu}$, A_μ, and ϕ. Kaluza's identification of the electromagnetic potential is not quite the right one, because he chooses it equal to $g_{\mu 5}$ (up to a constant), instead of taking the quotient $g_{\mu 5/}g_{55}$. This does not matter in his further analysis, because he considers only the linearized approximation of the field equations. Furthermore, the matter part is only studied in a nonrelativistic approximation. In particular, the five-dimensional geodesic equation is only written in this limit. Then the scalar contribution to the four-force becomes negligible and an automatic split into the usual gravitational and electromagnetic parts is obtained.

Kaluza was aware of the limitations of his analysis, but he was confident of being on the right track, as becomes evident from the final paragraph of his paper:

> In spite of all the physical and theoretical difficulties which are encountered in the above proposal it is hard to believe that the derived relationships, which could hardly be surpassed at the formal level, represent nothing more than a malicious coincidence. Should it sometime be established that the scheme is more than an empty formalism this would signify a new triumph for Einstein's General Theory of Relativity, whose suitable extension to five dimensions is our present concern.

For good reasons the role of the scalar field was unclear to him, except in the limiting situation of his analysis, where ϕ becomes the negative of the gravitational potential. Kaluza was, however, well aware that the scalar field could play an important role, and he makes some speculative remarks in this direction.

In the classical part of his first paper, Klein (1926a) improves on Kaluza's treatment. He assumes, however, beside the condition of cylindricity, that g_{55} is a constant. Following Kaluza, we keep here the scalar field ϕ and write the Kaluza-Klein ansatz for the five-dimensional metric ${}^{(5)}g$ in the form

$${}^{(5)}g = \phi^{-1/3}(g - \phi\omega \otimes \omega), \tag{23}$$

where $g = g_{\mu\nu}\, dx^\mu\, dx^\nu$ is the space-time metric and ω is a differential 1-form of the type

$$\omega = dx^5 + \kappa A_\mu dx^\mu. \tag{24}$$

Like ϕ, $A = A_\mu\, dx^\mu$ is independent of x^5; κ is a coupling constant to be determined. The convenience of the conformal factor $\phi^{-1/3}$ will become clear shortly.

Klein considers the subgroup of five-dimensional coordinate transformations which respect the form (23) of the $d=5$ metric:

$$x^\mu \to x^\mu, \quad x^5 \to x^5 + f(x^\mu). \tag{25}$$

Indeed, the pull-back of ${}^{(5)}g$ is again of the form (23) with

$$g \to g, \quad \phi \to \phi, \quad A \to A + \frac{1}{\kappa} df. \tag{26}$$

Thus $A = A_\mu\, dx^\mu$ transforms like a gauge potential under the Abelian gauge group (25) and is therefore interpreted as the electromagnetic potential. This is further justified by the most remarkable result derived by Kaluza and Klein, often called the Kaluza-Klein miracle. It turns out that the five-dimensional Ricci scalar ${}^{(5)}R$ splits as follows:

$${}^{(5)}R = \phi^{1/3}\left(R + \frac{1}{4}\kappa^2 \phi F_{\mu\nu}F^{\mu\nu} - \frac{1}{6\phi^2}(\nabla\phi)^2 + \frac{1}{3}\Delta \ln \phi \right). \tag{27}$$

For $\phi \equiv 1$ this becomes the Lagrangian of the coupled Einstein-Maxwell system. In view of the gauge group (25), this split is actually no miracle, because no other gauge-invariant quantities can be formed.

For the development of gauge theory this dimensional reduction was particularly important, because it revealed a close connection between coordinate transformations in higher-dimensional spaces and gauge transformations in space-time.

With Klein we consider the $d=5$ Einstein-Hilbert action

$${}^{(5)}S = \frac{-1}{\kappa^2 L}\int {}^{(5)}R \sqrt{|{}^{(5)}g|}\, d^5x, \tag{28}$$

assuming that the higher-dimensional space is a cylinder with $0 \leq x^5 \leq L = 2\pi R_5$. Since

$$\sqrt{|{}^{(5)}g|}\, dx^5 = \sqrt{-g}\, \phi^{-1/3}\, d^4x\, dx^5 \tag{29}$$

we obtain

$${}^{(5)}S = -\int \left(\frac{1}{\kappa^2} R + \frac{1}{4}\phi F_{\mu\nu}F^{\mu\nu} - \frac{1}{6\kappa^2\phi^2}(\nabla\phi)^2 \right) \sqrt{-g} d^4x. \tag{30}$$

[5] For instance, Einstein extensively discussed Nordström's second version in his famous Vienna lecture "On the Foundations of the Problem of Gravitation" (23 September 1913) and made it clear that Nordström's theory was a viable alternative to his own attempt with Grossmann. [See Doc. 17 of Vol. 4 of the *Collected Papers of Albert Einstein* (Einstein, 1987)].

Our choice of the conformal factor $\phi^{-1/3}$ in Eq. (23) was made so that the gravitational part in Eq. (30) is just the Einstein-Hilbert action, if we choose

$$\kappa^2 = 16\pi G. \tag{31}$$

For $\phi \equiv 1$ a beautiful geometrical unification of gravitation and electromagnetism is obtained.

We pause by noting that nobody in the early history of Kaluza-Klein theory seems to have noticed the following inconsistency in putting $\phi \equiv 1$ [see, however, Lichnerowicz (1995)]: The field equations for the dimensionally reduced action (30) are just the five-dimensional equations ${}^{(5)}R_{ab}=0$ for the Kaluza-Klein ansatz (23). Among these, the ϕ equation, which is equivalent to ${}^{(5)}R_{55}=0$, becomes

$$\Box(\ln\phi) = \frac{3}{4}\kappa^2\phi F_{\mu\nu}F^{\mu\nu}. \tag{32}$$

For $\phi \equiv 1$ this implies the unphysical result $F_{\mu\nu}F^{\mu\nu}=0$. This conclusion is avoided if one proceeds in the reverse order, i.e., by putting $\phi \equiv 1$ in the action (30) and varying afterwards. However, if the extra dimension is treated as physical—a viewpoint adopted by Klein (as we shall see)—it is clearly essential that one maintain consistency with the $d=5$ field equations. This is an example of the crucial importance of scalar fields in Kaluza-Klein theories.

Kaluza and Klein both studied the $d=5$ geodesic equation. For the metric (23) this is just the Euler-Lagrange equation for the Lagrangian

$$L = \frac{1}{2}g_{\mu\nu}\dot{x}^\mu\dot{x}^\nu - \frac{1}{2}\phi(\dot{x}^5 + \kappa A_\mu\dot{x}^\mu)^2. \tag{33}$$

Since x^5 is cyclic, we have the conservation law (m = mass of the particle)

$$p_5/m := \frac{\partial L}{\partial \dot{x}^5} = \phi(\dot{x}^5 + \kappa A_\mu\dot{x}^\mu) = \text{const.} \tag{34}$$

If use of this is made in the other equations, we obtain

$$\ddot{x}^\mu + \Gamma^\mu_{\alpha\beta}\dot{x}^\alpha\dot{x}^\beta = -\frac{p_5}{m}\kappa F^\mu_{\nu}\dot{x}^\nu - \frac{1}{2}\left(\frac{p_5}{m}\right)^2\phi^{-2}\nabla^\mu\phi. \tag{35}$$

Clearly, p_5 has to be interpreted as q/κ, where q is the charge of the particle,

$$p_5 = q/\kappa. \tag{36}$$

The physical significance of the last term in Eq. (35) remained obscure. Much later, Jordan (1949, 1955) and Thiry (1948, 1951) tried to make use of the new scalar field to obtain a theory in which the gravitational constant is replaced by a dynamical field. Further work by Jordan (1949, 1955), Fierz (1956), and Brans and Dicke (1961) led to a much studied theory, which has been for many years a serious competitor of general relativity. Generalized versions (Bergmann, 1968) have recently played a role in models of inflation (see, for example, Steinhardt, 1993). The question of whether the low-energy effective theory of string theories, say, has Brans-Dicke-type interactions has lately been investigated for instance by Damour and collaborators (Damour and Polyakov, 1994).

Since the work of Fierz (published in German, Fierz, 1956) is not widely known, we briefly describe its main point. Quoting Pauli, Fierz emphasizes that, in theories containing both tensor and scalar fields, the tensor field appearing most naturally in the action of the theory can differ from the "physical" metric by some conformal factor depending on the scalar fields. In order to decide which is the "atomic-unit" metric and thus the gravitational constant, one has to look at the coupling to matter. The "physical" metric $g_{\mu\nu}$ is the one to which matter is universally coupled (in accordance with the principle of equivalence). For instance, the action for a spin-0 massive matter field ψ should take the form

$$S_\psi = \frac{1}{2}\int (g^{\mu\nu}\,\partial_\mu\psi\,\partial_\nu\psi - m^2\psi^2)\sqrt{-g}d^4x. \tag{37}$$

A unit of length is then provided by the Compton wavelength $1/m$, and test particles fall along geodesics of $g_{\mu\nu}$. Fierz specializes Jordan's theory (with two free constants) such that the Maxwell density, expressed in terms of the physical metric, is not multiplied with a spacetime-dependent function. Otherwise the vacuum would behave like a variable dielectric and this would have unwanted consequences, although the refraction is 1: The fine structure constant would become a function of spacetime, changing the spectra of galaxies over cosmological distances.

With these arguments Fierz arrives at a theory which was later called the Brans-Dicke theory. He did not, however, confront the theory with observations, because he did not believe in its physical relevance. [The intention of Fierz's publication was mainly pedagogical (Fierz, 1999, private communication).]

Equation (36) brings us to the part of Klein's first paper that is related to quantum theory. There he interprets the five-dimensional geodesic equation as the geometrical optical limit of the wave equation ${}^{(5)}\Box\Psi=0$ on the higher-dimensional space and establishes for special situations a close relation of the dimensionally reduced wave equation with Schrödinger's equation, which had been discovered in the same year. His ideas are more clearly spelled out shortly afterwards in a brief *Nature* note entitled "The Atomicity of Electricity as a Quantum Theory Law" (Klein, 1926b). There Klein says in connection with Eq. (36):

> The charge q, so far as our knowledge goes, is always a multiple of the electronic charge e, so that we may write
>
> $$p_5 = n\frac{e}{\kappa} \quad [n \in \mathbf{Z}]. \tag{38}$$
>
> This formula suggests that the atomicity of electricity may be interpreted as a quantum theory law. In fact, if the five-dimensional space is assumed to be closed in the direction of x^5 with period L, and if

> we apply the formalism of quantum mechanics to our geodesics, we shall expect p_5 to be governed by the following rule:
>
> $$p_5 = n\frac{h}{L}, \tag{39}$$
>
> n being a quantum number, which may be positive or negative according to the sense of motion in the direction of the fifth dimension, and h the constant of Planck.

Comparing Eqs. (38) and (39), Klein finds the value of the period L,

$$L = \frac{h}{ec}\sqrt{16\pi G} = 0.8\times 10^{-30}\ \text{cm}, \tag{40}$$

and adds:

> The small value of this length together with the periodicity in the fifth dimension may perhaps be taken as a support of the theory of Kaluza in the sense that they may explain the non-appearance of the fifth dimension in ordinary experiments as the result of averaging over the fifth dimension.

Klein concludes this note with the daring speculation that the fifth dimension might have something to do with Planck's constant:

> In a former paper the writer has shown that the differential equation underlying the new quantum mechanics of Schrödinger can be derived from a wave equation of a five-dimensional space, in which h does not appear originally, but is introduced in connection with the periodicity in x^5. Although incomplete, this result, together with the considerations given here, suggests that the origin of Planck's quantum may be sought just in this periodicity in the fifth dimension.

This was not the last time that such speculations have been put forward. The revival of (supersymmetric) Kaluza-Klein theories in the eighties (Appelquist, Chados, and Freund, 1987; Kubyshin *et al.*, 1989) led to the idea that the compact dimensions would necessarily give rise to an enormous quantum vacuum energy via the Casimir effect. There were attempts to exploit this vacuum energy in a self-consistent approach to compactification, with the hope that the size of the extra dimensions would be calculable as a pure number times the Planck length. Consequently the gauge-coupling constant would then be calculable.

Coming back to Klein we note that he would also have arrived at Eq. (39) by the dimensional reduction of his five-dimensional equation. Indeed, if the wave field $\psi(x,x^5)$ is Fourier decomposed with respect to the periodic fifth coordinate,

$$\psi(x,x^5) = \frac{1}{\sqrt{L}}\sum_{n\in Z}\psi_n(x)e^{inx^5/R_5}, \tag{41}$$

one obtains for each amplitude $\psi_n(x)$ [for the metric (23) with $\phi\equiv 1$] the following four-dimensional wave equation:

$$\left(D^\mu D_\mu - \frac{n^2}{R_5^2}\right)\psi_n = 0, \tag{42}$$

where D_μ is the doubly covariant derivative (with respect to $g_{\mu\nu}$ and A_μ) with the charge

$$q_n = n\frac{\kappa}{R_5}. \tag{43}$$

This shows that the mass of the nth mode is

$$m_n = |n|\frac{1}{R_5}. \tag{44}$$

Combining Eq. (43) with $q_n = ne$, we obtain, as before, Eq. (40) or

$$R_5 = \frac{2}{\sqrt{\alpha}}l_{\text{Pl}}, \tag{45}$$

where α is the fine-structure constant and l_{Pl} is the Planck length.

Equations (43) and (44) imply a serious defect of the five-dimensional theory: The (bare) masses of all charged particles ($|n|\geqslant 1$) are of the order of the Planck mass

$$m_n = n\frac{\sqrt{\alpha}}{2}m_{\text{Pl}}. \tag{46}$$

The pioneering papers of Kaluza and Klein were taken up by many authors. For some time the "projective" theories of Veblen (1933), Hoffmann (1933), and Pauli (1933) played a prominent role. These are, however, just equivalent formulations of Kaluza's and Klein's unification of the gravitational and the electromagnetic field (Bergmann, 1942; Ludwig, 1951).

Einstein's repeated interest in five-dimensional generalizations of general relativity has been described by Bergmann (1942) and Pais (1982) and will not be discussed here.

V. KLEIN'S 1938 THEORY

The first attempt to go beyond electromagnetism and gravitation and apply Weyl's gauge principle to the nuclear forces occurred in a remarkable paper by Oskar Klein, presented at the Kazimierz Conference on New Theories in Physics (Klein, 1938). Assuming that the mesons proposed by Yukawa were vectorial, Klein proceeded to construct a Kaluza-Klein-like theory which would incorporate them. As in the original Kaluza-Klein theory he introduced only one extra dimension but his theory differed from the original in two respects:

(i) The fields were not assumed to be completely independent of the fifth coordinate x^5 but to depend on it through a factor e^{-iex^5} where e is the electric charge.

(ii) The five-dimensional metric tensor was assumed to be of the form

$$g_{\mu\nu}(x), \quad g_{55}=1, \quad g_{\mu 5} = \beta\chi_\mu(x), \tag{47}$$

where $g_{\mu\nu}$ was the usual four-dimensional Einstein metric, β was a constant, and $\chi_\mu(x)$ was a *matrix-valued field* of the form

$$\chi_\mu(x)=\begin{pmatrix} A_\mu(x) & \tilde{B}_\mu(x) \\ B_\mu(x) & A_\mu(x)\end{pmatrix}=\sigma_3[\vec{\sigma}\cdot\vec{A}_\mu(x)], \qquad (48)$$

where the σ's are the usual Pauli matrices and $\vec{A}_\mu(x)$ is what we would now call an $SU(2)$ gauge potential. This was a most remarkable ansatz considering that it implies a matrix-valued metric, and it is not clear what motivated Klein to make it. The reason that he multiplied the present-day $SU(2)$ matrix by σ_3 is that σ_3 represented the charge matrix for the fields.

Having made this ansatz, Klein proceeded in the standard Kaluza-Klein manner and obtained, instead of the Einstein-Maxwell equations, a set of equations that we would now call the Einstein-Yang-Mills equations. This is a little surprising because Klein inserted only electromagnetic gauge invariance. However, one can see how the $U(1)$ gauge invariance of electromagnetism could generalize to $SU(2)$ gauge invariance by considering the field strengths. The $SU(2)$ form of the field strengths corresponding to the $\tilde{B}_\mu$ and B_μ fields, namely,

$$F^{\tilde{B}}_{\mu\nu}=\partial_\mu\tilde{B}_\nu-\partial_\nu\tilde{B}_\mu+ie(A_\mu\tilde{B}_\nu-A_\nu\tilde{B}_\mu), \qquad (49)$$

$$F^{B}_{\mu\nu}=\partial_\mu B_\nu-\partial_\nu B_\mu-ie(A_\mu B_\nu-A_\nu B_\mu), \qquad (50)$$

actually follows from the electromagnetic gauge principle $\partial_\mu\rightarrow D_\mu=\partial_\mu+ie(1-\sigma_3)A_\mu$, given that the three vector fields belong to the same 2×2 matrix. The more difficult question is why the expression

$$F^{A}_{\mu\nu}=\partial_\mu A_\nu-\partial_\nu A_\mu-ie(\tilde{B}_\mu B_\nu-\tilde{B}_\nu B_\mu) \qquad (51)$$

for the field strength corresponding to A_μ contained a bilinear term when most other vector-field theories, such as the Proca theory, contained only the linear term. The reason is that the geometrical nature of the dimensional reduction meant that the usual space-time derivative ∂_μ was replaced by the covariant space-time derivative $\partial_\mu+ie(1-\sigma_3)\chi_\mu/2$, with the result that the usual curl $\partial\wedge\chi$ was replaced by $\partial_\mu\chi_\nu-\partial_\nu\chi_\mu+ie/2[\chi_\mu,\chi_\nu]$, whose third component is just the expression for $F^A_{\mu\nu}$.

Being primarily interested in the application of his theory to nuclear physics, Klein immediately introduced the nucleons, treating them as an isodoublet $\psi(x)$ on which the matrix ξ_μ acted by multiplication. In this way he was led to field equations of the familiar $SU(2)$ form, namely,

$$(\gamma\cdot D+M)\psi(x)=0, \quad D_\mu=\partial_\mu+\frac{ie}{2}(1-\sigma_3)\chi_\mu. \qquad (52)$$

However, although the equations of motion for the vector fields A_μ, $\tilde{B}_\mu$, and B_μ would be immediately recognized today as those of an $SU(2)$ gauge-invariant theory, this was not at all obvious at the time and Klein does not seem to have been aware of it. Indeed, he immediately proceeded to break the $SU(2)$ gauge invariance by assigning *ad hoc* mass terms to the $\tilde{B}_\mu$ and B_μ fields.

An obvious weakness of Klein's theory is that there is only one coupling constant β, which implies that the nuclear and electromagnetic forces would be of approximately the same strength, in contradiction with experiment. Furthermore, the nuclear forces would not be charge independent, as they were known to be at the time. These weaknesses were noticed by Møller, who, at the end of the talk, objected to the theory on these grounds. Klein's answer to these objections was astonishing: this problem could easily be solved he said, because the strong interactions could be made charge independent (and the electromagnetic field separated) by introducing one more vector field C_μ and generalizing the 2×2 matrix χ_μ

$$\text{from}\quad \chi_\mu=\sigma_3(\vec{\sigma}\cdot\vec{A}_\mu)\quad\text{to}\quad \sigma_3(C_\mu+\vec{\sigma}\cdot\vec{A}_\mu). \qquad (53)$$

In other words, he there and then generalized what was effectively a (broken) $SU(2)$ gauge theory to a broken $SU(2)\times U(1)$ gauge theory. In this way, he anticipated the mathematical structure of the standard electroweak theory by 21 years!

Klein has certainly not forgotten his ambitious proposal of 1938, in contrast to what has been suspected by Gross (1995). In his invited lecture at the Berne Congress in 1955 (Klein, 1956) he came back to some main aspects of his early attempt and concluded with the statement:

> On the whole, the relation of the theory to the five-dimensional representation of gravitation and electromagnetism on the one hand and to symmetric meson theory on the other hand—through the appearance of the charge invariance group—may perhaps justify the confidence in its essential soundness.

VI. THE PAULI LETTERS TO PAIS

The next attempt to write down a gauge theory for the nuclear interactions was due to Pauli. During a discussion following a talk by Pais at the 1953 Lorentz Conference in Leiden (Pais, 1953), Pauli said:

> . . . I would like to ask in this connection whether the transformation group with constant phases can be amplified in way analogous to the gauge group for electromagnetic potentials in such a way that the meson-nucleon interaction is connected with the amplified group . . .

Stimulated by this discussion, Pauli worked on this problem and drafted a manuscript to Pais that begins with the heading (Pauli, 1999).

> Written down July 22–25, 1953, in order to see how it looks. Meson-Nucleon Interaction and Differential Geometry.

In this manuscript, Pauli generalizes the Kaluza-Klein theory to a six-dimensional space and arrives through dimensional reduction at the essentials of an $SU(2)$ gauge theory. The extra dimensions form a two-sphere S^2 with space-time-dependent metrics on which $SU(2)$

operates in a space-time-dependent manner. Pauli develops first in "local language" the geometry of what we now call a fiber bundle with a homogeneous space as typical fiber [in his case $S^2 \cong SU(2)/U(1)$]. Studying the curvature of the higher-dimensional space, Pauli automatically finds, for the first time, the correct expression for the non-Abelian field strength.

Since it is somewhat difficult to understand exactly what Pauli did, we give some details, using more familiar formulations and notations.

Pauli considers the six-dimensional total space $M \times S^2$, where S^2 is the two-sphere on which $SO(3)$ acts in the canonical manner. He distinguishes among the diffeomorphisms (coordinate transformations) those which leave M pointwise fixed and induce space-time-dependent rotations on S^2:

$$(x,y) \rightarrow [x, R(x) \cdot y]. \tag{54}$$

Then Pauli postulates a metric on $M \times S^2$ that is supposed to satisfy three assumptions. These lead him to what is now called the non-Abelian Kaluza-Klein ansatz: The metric $\hat{g}$ on the total space is constructed from a space-time metric g, the standard metric γ on S^2, and a Lie-algebra-valued 1-form,

$$A = A^a T_a, \quad A^a = A^a_\mu \, dx^\mu, \tag{55}$$

on M [T_a, $a = 1, 2, 3$, are the standard generators of the Lie algebra of $SO(3)$] as follows: If $K^i_a \partial/\partial y^i$ are the three Killing fields on S^2, then

$$\hat{g} = g - \gamma_{ij}[dy^i + K^i_a(y) A^a] \otimes [dy^j + K^j_a(y) A^a]. \tag{56}$$

In particular, the nondiagonal metric components are

$$\hat{g}_{\mu i} = A^a_\mu(x) \gamma_{ij} K^j_a. \tag{57}$$

Pauli does not say that the coefficients of A^a_μ in Eq. (57) are the components of the three independent Killing fields. This is, however, his result, which he formulates in terms of homogeneous coordinates for S^2. He determines the transformation behavior of A^a_μ under the group (54) and finds in matrix notation what he calls "the generalization of the gauge group":

$$A_\mu \rightarrow R A_\mu R^{-1} + R^{-1} \partial_\mu R. \tag{58}$$

With the help of A_μ, he defines a covariant derivative, which is used to derive "field strengths" by applying a generalized curl to A_μ. This is exactly the field strength that was later introduced by Yang and Mills. To our knowledge, apart from Klein's 1938 paper, it appears here for the first time. Pauli says that "this is the *true* physical field, the analog of the *field strength*" and he formulates what he considers to be his "main result":

> The vanishing of the field strength is necessary and sufficient for the $A^a_\mu(x)$ in the whole space to be transformable to zero.

It is somewhat astonishing that Pauli did not work out the Ricci scalar for $\hat{g}$ as for the Kaluza-Klein theory. One reason may be connected with his remark on the Kaluza-Klein theory in Note 23 of his relativity article (Pauli, 1958) concerning the five-dimensional curvature scalar (p. 230):

> There is, however, no justification for the particular choice of the five-dimensional curvature scalar P as integrand of the action integral, from the standpoint of the restricted group of the cylindrical metric [gauge group]. The open problem of finding such a justification seems to point to an amplification of the transformation group.

In a second letter (Pauli, 1999), Pauli also studies the dimensionally reduced Dirac equation and arrives at a mass operator that is closely related to the Dirac operator in internal space (S^2, γ). The eigenvalues of the latter operator had been determined by him long before (Pauli, 1939). Pauli concludes with the statement: "So this leads to some rather unphysical 'shadow particles.' "

VII. YANG-MILLS THEORY

In his Hermann Weyl Centenary Lecture at the ETH (Yang, 1980), C. N. Yang commented on Weyl's remark "The principle of gauge-invariance has the character of general relativity since it contains an arbitrary function λ, and can certainly only be understood in terms of it" (Weyl, 1968) as follows:

> The quote above from Weyl's paper also contains something which is very revealing, namely, his strong association of gauge invariance with general relativity. That was, of course, natural since the idea had originated in the first place with Weyl's attempt in 1918 to unify electromagnetism with gravity. Twenty years later, when Mills and I worked on non-Abelian gauge fields, our motivation was completely divorced from general relativity and we did not appreciate that gauge fields and general relativity are somehow related. Only in the late 1960s did I recognize the structural similarity mathematically of non-Abelian gauge fields with general relativity and understand that they both were connections mathematically.

Later, in connection with Weyl's strong emphasis of the relation between gauge invariance and conservation of electric charge, Yang continues with the following instructive remarks:

> Weyl's reason, it turns out, was also one of the melodies of gauge theory that had very much appealed to me when as a graduate student I studied field theory by reading Pauli's articles. I made a number of unsuccessful attempts to generalize gauge theory beyond electromagnetism, leading finally in 1954 to a collaboration with Mills in which we developed a non-Abelian gauge theory. In [$\cdots$] we stated our motivation as follows:
>
> The conservation of isotopic spin points to the existence of a fundamental invariance law similar to the conservation of electric charge. In the latter case, the electric charge serves as a source of electromagnetic field; an important concept in this case is gauge invariance, which is closely connected with (1) the equation of motion of the electro-

> magnetic field, (2) the existence of a current density, and (3) the possible interactions between a charged field and the electromagnetic field. We have tried to generalize this concept of gauge invariance to apply to isotopic spin conservation. It turns out that a very natural generalization is possible.
>
> Item (2) is the melody referred to above. The other two melodies, (1) and (3), where what had become pressing in the early 1950s when so many new particles had been discovered and physicists had to understand how they interact with each other.
>
> I had met Weyl in 1949 when I went to the Institute for Advanced Study in Princeton as a young "member." I saw him from time to time in the next years, 1949–1955. He was very approachable, but I don't remember having discussed physics or mathematics with him at any time. His continued interest in the idea of gauge fields was not known among the physicists. Neither Oppenheimer nor Pauli ever mentioned it. I suspect they also did not tell Weyl of the 1954 papers of Mills' and mine. Had they done that, or had Weyl somehow came across our paper, I imagine he would have been pleased and excited, for we had put together two things that were very close to his heart: gauge invariance and non-Abelian Lie groups.

It is indeed astonishing that during those late years neither Pauli nor Yang ever talked with Weyl about non-Abelian generalizations of gauge invariance.

With the background of Sec. VI, the following story of spring 1954 becomes more understandable. In late February, Yang was invited by Oppenheimer to return to Princeton for a few days and to give a seminar on his joint work with Mills. Here is Yang's report (Yang, 1983):

> Pauli was spending the year in Princeton, and was deeply interested in symmetries and interactions. (He had written in German a rough outline of some thoughts, which he had sent to A. Pais. Years later F. J. Dyson translated this outline into English. It started with the remark, "Written down July 22–25, 1953, in order to see how it looks," and had the title "Meson-Nucleon Interaction and Differential Geometry.") Soon after my seminar began, when I had written down on the blackboard,
>
> $(\partial_\mu - i\epsilon B_\mu)\psi,$
>
> Pauli asked, "What is the mass of this field B_μ?" I said we did not know. Then I resumed my presentation, but soon Pauli asked the same question again. I said something to the effect that that was a very complicated problem, we had worked on it and had come to no definite conclusions. I still remember his repartee: "That is not sufficient excuse." I was so taken aback that I decided, after a few moments' hesitation to sit down. There was general embarrassment. Finally Oppenheimer said, "We should let Frank proceed." I then resumed, and Pauli did not ask any more questions during the seminar.
>
> I don't remember what happened at the end of the seminar. But the next day I found the following message:
>
> February 24, Dear Yang, I regret that you made it almost impossible for me to talk with you after the seminar. All good wishes. Sincerely yours, W. Pauli.
>
> I went to talk to Pauli. He said I should look up a paper by E. Schrödinger, in which there were similar mathematics.[6] After I went back to Brookhaven, I looked for the paper and finally obtained a copy. It was a discussion of spacetime-dependent representations of the γ_μ matrices for a Dirac electron in a gravitational field. Equations in it were, on the one hand, related to equations in Riemannian geometry and, on the other, similar to the equations that Mills and I were working on. But it was many years later when I understood that these were all different cases of the mathematical theory of connections on fiber bundles.

Later Yang adds:

> I often wondered what he [Pauli] would say about the subject if he had lived into the sixties and seventies.

At another occasion (Yang, 1980) he remarked:

> I venture to say that if Weyl were to come back today, he would find that amidst the very exciting, complicated and detailed developments in both physics and mathematics, there are fundamental things that he would feel very much at home with. He had helped to create them.

Having quoted earlier letters from Pauli to Weyl, we add what Weyl said about Pauli in 1946 (Weyl, 1980):

> The mathematicians feel near to Pauli since he is distinguished among physicists by his highly developed organ for mathematics. Even so, he is a physicist; for he has to a high degree what makes the physicist; the genuine interest in the experimental facts in all their puzzling complexity. His accurate, instructive estimate of the relative weight of relevant experimental facts has been an unfailing guide for him in his theoretical investigations. Pauli combines in an exemplary way physical insight and mathematical skill.

To conclude this section, let us emphasize the main differences between general relativity and Yang-Mills theories. Mathematically, the $so(1,3)$-valued connection forms ω in Sec. III A and the Lie-algebra-valued gauge potential A are on the same footing; they are both representatives of connections in (principle) fiber bundles

[6] E. Schrödinger, Sitzungsberichte der Preussischen (Akademie der Wissenschaften, 1932), p. 105.

$$\begin{tikzcd}
g & e^{\alpha} \arrow[l] \arrow[d] & \\
& \omega \arrow[r, leftrightarrow] \arrow[d] & A \arrow[d] \\
& \Omega \arrow[r, leftrightarrow] & F
\end{tikzcd}$$

FIG. 4. General relativity vs Yang-Mills theory.

over the space-time manifold. Equation (17) translates into the formula for the Yang-Mills field strength F,

$$F = dA + A \wedge A. \tag{59}$$

In general relativity one has, however, additional geometric structure, since the connection derives from a metric, or the tetrad fields $e^{\alpha}(x)$, through the first structure equation (15). This is shown schematically in Fig. 4. [In bundle theoretical language one can express this as follows: The principle bundle of general relativity, i.e., the orthonormal frame bundle, is constructed from the base manifold and its metric, and has therefore additional structure, implying, in particular, the existence of a canonical 1-form (soldering form), whose local representatives are the tetrad fields; see, for example, Bleecker (1981).]

Another important difference is that the gravitational Lagrangian $*R = 1/2\Omega_{\alpha\beta} \wedge *(e^{\alpha} \wedge e^{\beta})$ is linear in the field strengths, whereas the Yang-Mills Lagrangian $F \wedge *F$ is quadratic.

VIII. RECENT DEVELOPMENTS

The developments after 1958 consisted in the gradual recognition that—contrary to phenomenological appearances—Yang-Mills gauge theory could describe weak and strong interactions. This important step was again very difficult, with many hurdles to overcome.

One of them was the mass problem, which was solved, probably in a preliminary way, through spontaneous symmetry breaking. Of critical significance was the recognition that spontaneously broken gauge theories are renormalizable. On the experimental side the discovery and intensive investigation of the neutral current was, of course, crucial. For the gauge description of the strong interactions, the discovery of asymptotic freedom was decisive. That the $SU(3)$ color group should be gauged was also not at all obvious. And then there was the confinement idea, which explains why quarks and gluons do not exist as free particles. All this is described in numerous modern textbooks and does not have to be repeated.

The next step in creating a more unified theory of the basic interactions will probably be much more difficult. All major theoretical developments of the last twenty years, such as grand unification, supergravity, and supersymmetric string theory, are almost completely separated from experience. There is a great danger that theoreticians may get lost in pure speculations. As in the first unification proposal of Hermann Weyl, they may create beautiful and highly relevant mathematics, which do not, however, describe Nature. In the latter case history shows, however, that such ideas can one day also become fruitful for physics. It may, therefore, be appropriate to conclude with some remarks on current attempts in string theory and noncommutative geometry.

A. Gauge theory and strings

1. Introduction

So far we have considered gravitation and gauge theory only within the context of local field theory. However, gravitation and gauge theory also occur naturally in string theory (Green, Schwarz, and Witten, 1987; Polchinski, 1998). Indeed, whereas in field theory they are optional extras that are introduced on phenomenological grounds (equality of inertial and gravitational mass, divergence-free character of the magnetic field, etc.) in string theory they occur as an intrinsic part of the structure. Thus string theory is a very natural setting for gravitation and gauge fields. One might go so far as to say that, had string theory preceded field theory historically, the gravitational and gauge fields might have emerged in a completely different manner. An interesting feature of string gauge theories is that the choice of gauge group is quite limited.

String theory is actually a natural setting not only for gravitational and gauge fields but also for the Kaluza-Klein mechanism. Historically, the most obvious difficulty with Kaluza-Klein reductions was that there was no experimental evidence and no theoretical need for any extra dimensions. String theory changes this situation dramatically. As is well known, string theory is conformally invariant only if the dimension d of the target space is $d = 10$ or $d = 26$, according to whether the string is supersymmetric or not. Thus, in contrast to field theory, string theory points to the existence of extra dimensions and even specifies their number.

We shall treat an important specific case of dimensional reduction within string theory, namely, the toroidal reduction from 26 to 10 dimensions, in Sec. VIII. A. 7. However, since no phenomenologically satisfactory reduction from 26 or 10 to 4 dimensions has yet emerged, and the dimensional reduction in string theory is rather similar to that in ordinary field theory (Appelquist, Chodos, and Freund, 1987; Kubyshin *et al.*, 1989), we shall not consider any other case, but refer the reader to the literature.

Instead we shall concentrate on the manner in which gauge theory and gravitation occur in the context of dimensionally unreduced string theory. We shall rely heavily on the result (Yau, 1985) that if a massless vector field theory with polarization vector ξ_{μ} and on-shell momentum p_{μ} $(p^2 = 0)$ is invariant with respect to the transformation

$$\xi_{\mu}(p) \rightarrow \xi_{\mu}(p) + \eta(p) p_{\mu}, \tag{60}$$

where $\eta(p)$ is an arbitrary scalar, then it must be a gauge theory. Similarly, we shall rely on the result that a second-rank symmetric-tensor theory must be a gravitational theory if the polarization vector $\xi_{\mu\nu}$ satisfies

$$\xi^{\mu}_{\mu}(p)=0, \quad p^{\mu}\xi_{\mu\nu}(p)=0, \tag{61}$$

and if the theory is invariant with respect to

$$\xi_{\mu\nu}(p)\rightarrow\xi_{\mu\nu}(p)+\eta_{\mu}p_{\nu}+\eta_{\nu}p_{\mu} \tag{62}$$

for arbitrary $\eta_{\mu}(p)$, and $p^2=0$ (Weinberg, 1965; Wald, 1986, and references therein; Feynman, 1995).

2. Gauge properties of open bosonic strings

To fix our ideas we first recall the form of the path integral for the bosonic string (Green, Schwarz, and Witten, 1987; Polchinski, 1998), namely,

$$\int dh\, dX\, e^{\int d^2\sigma h^{\alpha\beta}\eta_{\mu\nu}\,\partial_{\alpha}X^{\mu}(\sigma)\partial_{\beta}X^{\nu}(\sigma)}, \tag{63}$$

where σ are the coordinates, $d^2\sigma$ is the diffeomorphic-invariant measure, and $h^{\alpha\beta}$ is the metric on the two-dimensional world sheet of the string, while $\eta_{\mu\nu}$ is the Lorentz metric for the 26-dimensional target space in which the string, with coordinates $X^{\mu}(\sigma)$, moves. Thus the $X^{\mu}(\sigma)$ may be regarded as fields in a two-dimensional quantum field theory. The action in Eq. (63), and hence the classical two-dimensional field theory, is conformally invariant, but, as is well known, the quantum theory is conformally invariant only if $N=26$ or $N=10$ in the supersymmetric version.

The *open* bosonic string is the one in which gauge fields occur. Indeed, one might go so far as to say that the open string is a natural nonlocal generalization of a gauge field. The ends of the open strings are usually assumed to be attached to quarks, and thus there is a certain qualitative resemblance between the open bosonic strings and the gluon flux lines that link the quarks in theories of quark confinement. We wish to make the relationship between gauge fields and open bosonic strings more precise.

As is well known (Green, *et al.* 1987; Polchinski, 1998), the vacuum state $|0\rangle$ of the open string is a scalar tachyon and the first excited states are $X^{\mu}(\tau)|0\rangle=\int d\,(\sin s)X^{\mu}_{+}(\tau,s)|0\rangle$, where τ and s are the time and space components of σ and $X^{\nu}_{+}(\tau,s)$ is the positive-frequency part of $X^{\mu}(\tau,s)$. For a suitable standard value of the normal-ordering parameter for the Noether generators of the conformal (Virasoro) group, these states are massless, i.e., $p^2=0$, where p_{μ} is the 26-dimensional momentum. Furthermore, they are the only massless states. Thus if there are gauge fields in the theory, these are the states that must be identified with them. On the other hand, since all the other (massive) states in the Fock space of the string are formed by acting on $|0\rangle$ with higher moments of $X_{-}(\sigma)$ we see that the operators $X^{\mu}_{+}(\tau)$ are the prototypes of the operators that create the whole string. It is in this sense that the string can be regarded as a nonlocal generalization of a gauge field.

The question is: how is the identification of the massless states $X_{-}(\tau)|0\rangle$ to be justified? The justification comes about through the so-called vertex operators for the emission of the on-shell massless states. The vertices are the analogs of the ordinary Feynmann diagrams in quantum field theory and take the form

$$V(\xi,p)=\int ds e^{ip\cdot X(\sigma)}\xi\cdot\partial_{s}X(\sigma), \tag{64}$$

where ξ_{μ} is the polarization vector, p_{μ} is the momentum $(p^2=0)$ of the emitted particle, and ∂_{η} is a spacelike derivative. This vertex operator is to be inserted in the functional integral (63). Although the form of this vertex is not deduced from a second-quantized theory of strings (which does not yet exist) the vertex operator (64) is generally accepted as the correct one, because it is the only vertex that is compatible with the two-dimensional structure and conformal invariance of the string, and that reduces to the usual vertex in the point-particle limit. Suppose now that we make the transformation $\xi_{\nu}\rightarrow\xi_{\nu}+\alpha(p)p_{\nu}$ in Eq. (60). Then the vertex $V(\xi,p)$ acquires the additional term

$$\eta(p)\int ds e^{ip\cdot X(\sigma)}p\cdot\partial_{s}X(\sigma)=-i\,\eta(p)\int ds\,\partial_{s}(e^{1p\cdot X(\sigma)}), \tag{65}$$

which is an integrable factor that attaches itself to the two ends of the string and thus displays the gauge-covariant character of $V(\xi,p)$. The important point is that this gauge covariance is not imposed from outside, but is an intrinsic property of the string. It is a consequence of the fact that the string is conformally invariant (which dictates the form of the vertex operator) and has an internal structure (manifested by the fact that it has an internal two-dimensional integration).

3. Gravitational properties of closed bosonic strings

Just as the open bosonic string is the one in which gauge fields naturally occur, the closed bosonic string is the one in which gravitational fields naturally occur. It turns out, in fact, that a gravitational field, a dilaton field, and an antisymmetric tensor field occur in the closed string in the same way that the gauge field appears in the open string. The ground state $|0\rangle$ of the closed string is again a tachyon but the new feature is that, for the standard value of the normal-ordering constant, the first excited states are massless states of the form $X^{\mu}_{+}(\sigma)X^{\nu}_{+}(\sigma)|0\rangle$ and it is the symmetric, trace, and antisymmetric parts of the two-tensor formed by the X's that are identified with the gravitational, dilaton, and antisymmetric tensor fields, respectively. The question is how the identification with the gravitational field is to be justified, and again the answer is by means of a vertex operator, this time for the emission of an on-shell graviton. The vertex operator that describes the emission of an on-shell massless spin-2 field (graviton) of polarization $\xi_{\mu\nu}$ and momentum p_{λ}, where $p^2=0$, is

$$V(\xi,p)=\int d^2\sigma\, e^{ip\cdot X(\sigma)}\sqrt{h}h^{\alpha\beta}\xi_{\mu\nu}\,\partial_{\alpha}X^{\mu}(\sigma)\partial_{\beta}X^{\nu}(\sigma). \tag{66}$$

Already at this stage there is a feature that does not arise in the gauge-field case: Since the vertex operator is bilinear in the field X^μ it has to be normal ordered, and it turns out that the normal ordering destroys the classical conformal invariance, unless

$$\xi^\mu_\mu = 0 \quad \text{and} \quad p^\mu \xi_{\mu\nu} = 0. \tag{67}$$

We next make the momentum-space version of an infinitesimal coordinate transformation, namely,

$$\xi_{\mu\nu} \to \xi_{\mu\nu} + \eta_\mu(p) p_\nu + \eta_\nu(p) p_\mu . \tag{68}$$

Under this transformation the vertex V_ξ picks up an additional term of the form

$$2\eta_\nu \int d^2\sigma\, h^{\alpha\beta} e^{ip\cdot X(\sigma)} p_\mu \partial_\alpha X^\mu(\sigma) \partial_\beta X^\nu(\sigma)$$

$$= -2i\eta_\nu \int d^2\sigma\, h^{\alpha\beta} (\partial_\alpha e^{ip\cdot X(\sigma)}) \partial_\beta X^\nu(\sigma). \tag{69}$$

In analogy with the electromagnetic case we can carry out a partial integration. However, this time the expression does not vanish completely but reduces to

$$2i\eta_\nu \int d^2\sigma e^{ip\cdot X(\sigma)} \Delta X^\nu(\sigma), \tag{70}$$

where Δ denotes the two-dimensional Laplacian. On the other hand, $\Delta X^\nu(\sigma) = 0$ is just the classical field equation for $X^\nu(\sigma)$, and it can be shown that even in the quantized version it is effectively zero (Green, Schwarz, and Witten, 1987; Polchinski, 1998). Thus, thanks to the dynamics, we have invariance with respect to the transformations (68). But Eq. (67) and invariance with respect to Eq. (68) are just the conditions (61) and (62) discussed earlier for the vertex to be a gravitational field. As in the gauge-field case, the important point is that the general coordinate invariance is not imposed, but is a consequence of the conformal invariance and internal structure of the string.

The appearance of a scalar field in this context is not too surprising, since a scalar also appeared in the Kaluza-Klein reduction. What is more surprising is the appearance of an antisymmetric tensor. From the point of view of traditional local gravitational and gauge-field theory the presence of an additional antisymmetric tensor field seems at first sight to be an embarrassment. But it turns out to play an essential role in maintaining conformal invariance (cancellation of anomalies), so its presence is to be welcomed.

4. The presence of matter

Of course, the open bosonic string is not the whole story any more than pure gauge fields are the whole story in quantum field theory. One still has to introduce quantities that correspond to fermions (and possibly scalars) at the zero-mass level. There are essentially two ways to do this. The first is the Chan-Paton mechanism (Green, Schwarz and Witten, 1987; Polchinski, 1998), which dates from the days of strong-interaction string theory. In this mechanism one simply attaches charged particles to the open ends of the string. These charged particles are not otherwise associated with any string and thus the mechanism is rather *ad hoc* and leads to a hybrid of string and field theories. But it has the merit of introducing charged particles directly and thus emphasizing the relationship between strings and gauge fields.

The Chan-Paton mechanism has the further merit of allowing a simple generalization to the non-Abelian case. This is done by replacing the charged particles by particles belonging to the fundamental representations of compact internal symmetry groups G, typically quarks $q_a(x)$ and antiquarks $\overline{q_b}(x)$. The vertex operator then generalizes to one with double labels (a,b) and represents non-Abelian gauge fields in much the same way that the simple bosonic string represents an Abelian gauge field.

An interesting restriction arises from the fact that since the string represents gauge fields, and gauge fields belong to the adjoint representation of the gauge group, the vertex function must belong to the adjoint representation. This implies that even at the tree level the tensor product of the fundamental group representation with itself must produce only the adjoint representation, and this restricts G to be one of the classical groups $SO(n)$, $Sp(2n)$, and $U(n)$. Furthermore, it is found that $U(n)$ violates unitarity at the one-loop level, which leaves only $SO(n)$ and $SP(2n)$. Finally, these groups require symmetrization and antisymmetrization in the indices a and b to produce only the adjoint representation, and this implies that the string is oriented (symmetric with respect to its end points). When all these conditions are satisfied it can be shown that the non-Abelian vertex corresponding to Eq. (64) is covariant with respect to non-Abelian gauge transformations corresponding to $\xi_\mu \to \xi_\mu + \eta(p) p_\mu$ above. But since these transformations are nonlinear the proof is more difficult than in the Abelian case.

5. Fermionic and heterotic strings: Supergravity and non-abelian gauge theory

The Chan-Paton version of gauge string has the obvious disadvantage that the charged fields (quarks) are not an intrinsic part of the theory. A second method of introducing fermions is to place them in the string itself. This is done by replacing the kinetic term $(\partial X)^2$ by a Dirac term $\overline{\Psi} \partial\!\!\!/ \Psi$ in the Lagrangian density. Interesting cases are those in which the number of fermion components just matches the number of bosons, so that the Lagrangian is supersymmetric. In that case the condition for quantum conformal invariance reduces from $d=26$ to $d=10$. An interesting case from the point of view of gauge theory and dimensional reduction is the heterotic string, in which the left-handed part forms a superstring and the right-handed part forms a bosonic string in which 16 of the bosons are fermionized. For the heterotic string the Lagrangian in the bosonic-string path integral (63) is replaced by the Lagrangian

$$\sum_{\mu=1}^{\mu=10} \partial_\alpha X^\mu \partial^\alpha X_\mu - 2\sum_{\mu=1}^{\mu=10} \psi_+^\mu \partial_- \psi_{\mu+} - 2\sum_{A=1}^{A=32} \lambda_-^A \partial_+ \lambda_-^A , \tag{71}$$

where the ψ's and λ's are Majorana-Weyl fermions and the λ's belong to a representation (labeled with A) of an internal symmetry group G. It is only through the λ's that the internal symmetry group enters. The left- and right-handed parts of the theory are conformally invariant for quite different reasons. The left-handed part of the X's and the left-handed fermions ψ are conformally invariant, because together they form the left-handed part of a superstring (this is why the summation over the X's is only from 1 up to 10). The right-handed part of the X's and the right-handed fermions λ^A are conformally invariant because, from the point of view of anomalies, two Majorana-Weyl fermions are equivalent to one boson and thus the system is equivalent to the right-handed part of a 26-dimensional bosonic string. (This is why the index A runs from 1 to 32.) The fact that there are 32 fermions obviously puts strong restrictions on the choice of the internal symmetry group G.

We now examine the particle content of the theory, using the light-cone gauge, where there are no redundant fields. There are no tachyons for the left-moving fields; the first excited states are massless and take the form

$$|i\rangle_L \quad \text{and} \quad |\alpha\rangle_L , \tag{72}$$

where the $|i\rangle_L$ for $i=1\cdots 8$ are the left-handed components of a massless space-time vector in the eight transverse directions in the light-cone gauge and $|\alpha\rangle_L$ are the components of a massless fermion in one of the two fundamental spinor representations of the same space-time $SO(8)$ group. These states are all G invariant.

The first excited states for the right-moving sector are

$$|i\rangle_R \quad \text{and} \quad \lambda^A \lambda^B |0\rangle, \tag{73}$$

where the $|i\rangle_R$ are the right-handed analogs of the $|i\rangle_L$ and the $\lambda\lambda|0\rangle$ states are massless space-time scalars. The states $|i\rangle_R$ are G invariant but the states $\lambda\lambda|0\rangle$ belong to the adjoint representation of G and thus it is only through these states that the internal symmetry enters at the massless level.

The physical states are obtained by tensoring the left- and right-moving states (72) and (73). On tensoring the right-handed states with the vectors in Eq. (73) we obviously obtain states that are G invariant, and they turn out to be just the states that would occur in $N=1$ supergravity. An analysis of the vertex operators, similar to that carried out above for closed bosonic strings, confirms that these fields do indeed correspond to supergravity.

6. The internal symmetry group G

From the point of view of non-Abelian gauge theory the interesting states are those belonging to the nontrivial representations of G, and these are the ones obtained from the tensor products of Eq. (73) with the space-time scalars $\lambda\lambda|0\rangle$. At this point one must make a choice about the internal symmetry group G. The simplest choice is evidently $G=SO(32)$, and it is obtained by assigning antiperiodic boundary conditions to *all* the fermion fields λ. (Assigning periodic boundary conditions to all of them violates the masslessness condition.) Since the product states continue to belong to the adjoint representation of $SO(32)$, they are the natural candidates for states associated with non-Abelian gauge-fields, and an analysis of the vertex operators associated with these states confirms that they do indeed correspond to $SO(32)$ gauge fields.

In sum, the heterotic string produces both supergravity and non-Abelian gauge theory.

7. Dimensional reduction and the heterotic symmetry group $E_8\times E_8$

A variety of other left-handed internal symmetry groups $G\subset SO(32)$ can be obtained by assigning periodic and antiperiodic boundary conditions to the fermions λ^A of the heterotic string in a nonuniform manner. However, apart from the $SO(32)$ case just discussed, the only assignment that satisfies unitarity at the one-loop level is an equipartition of the 32 fermions into two sets of 16, with mixed boundary conditions. This would appear, at first sight, to lead to an $SO(16)\times SO(16)$ internal symmetry and gauge group, on the same grounds as $SO(32)$ above, but a closer analysis shows that it actually leads to a larger group, namely $E_8\times E_8$, which actually has the same dimension (496) as $SO(32)$. This group is quite attractive for grand unification theory, as it breaks naturally to E_6, which is one of the favorite grand unified theory groups.

Once we accept that $SO(16)\times SO(16)$ is a gauge group and that a rigid internal symmetry group $E_8\times E_8$ exists, it follows immediately that $E_8\times E_8$ must be a gauge group, because the action of the rigid generators of $E_8\times E_8$ on the $SO(16)\times SO(16)$ gauge fields produces $E_8\times E_8$ gauge fields.

This reduces the problem to the existence of a rigid $E_8\times E_8$ symmetry, but, within the context of our present methods, this is a rather convoluted process. One must introduce special representations of $SO(16)\times SO(16)$, project out some of the resulting states, and construct vertices that represent the elements of the coset $(E_8\times E_8)/[SO(16)\times SO(16)]$. Luckily there is a much more intuitive way to establish the existence of the $E_8\times E_8$ symmetry, and, as this way provides a very nice example of dimensional reduction within string theory, we shall now sketch it.

We have already remarked that, from the point of view of Virasoro anomalies, the 32 right-handed Majorana-Weyl fermions λ are equivalent to the right-handed parts of 16 bosons. This relationship can be carried farther by bosonizing the fermions according to $\lambda_\pm(\sigma)= :\exp[\pm\phi_i^R(\sigma)]:$, where $\phi^R(\sigma)$ is a right-moving bosonic field, compactified so that $0\leqslant\phi(\sigma)<2\pi$. In that case we may regard the right-handed part of the heterotic string as originating in the right-handed part of an ordinary 26-dimensional bosonic string, in which 16 of the 26 right-moving bosonic fields $X^R(\sigma)$ have been fermionized by letting $X_a^R(\sigma)\rightarrow\phi_a^R(\sigma)$ for $0\leqslant\phi_a(\sigma)<2\pi$ and $a=1\cdots16$. Since the X's correspond to coordinates in the target space of the string, this is equivalent to a toroidal compactification of 16 of the target-space dimensions and thus is equivalent to a Kaluza-Klein-type dimensional reduction from 26 to 10 dimensions. It turns out that the toroidal compactification and conversion to fermions is consistent only if the lattice that defines the 16-dimensional torus is even and self-dual. But it is well known that there are only two such lattices, called D_{16}^{+} and E_8+E_8, and since these have automorphism groups $SO(32)$ and $E_8\times E_8$, respectively, one sees at once where the origin of these symmetry groups lies. The further reduction from ten to four dimensions is, of course, another question. One of the more attractive proposals is that the quotient, six-dimensional space, be a Calabi-Yau space (Green, Schwarz, and Witten, 1987; Yau, 1985), but we do not wish to pursue this question further here.

B. Gauge theory and noncommutative geometry

The recent development of noncommutative geometry by Connes (1994) has permitted the generalization of gauge-theory ideas to the case in which the standard differential manifolds (Minkowskian, Euclidean, Riemannian) become mixtures of differential and discrete manifolds. The differential operators then become mixtures of ordinary differential operators and matrices. From the point of view of the fundamental physical interactions, the interest in such a generalization of gauge theory is that the Higgs field and its potential, which are normally introduced in an *ad hoc* manner, appear as part of the gauge-field structure. Indeed the Higgs field emerges as the component of the gauge potential in the "discrete direction" and the Higgs potential, like the self-interaction of the gauge field, emerges from the square of the curvature. The theory also relates to Kaluza-Klein theory because the Higgs field and potential can also be regarded as coming from a dimensional reduction in which the discrete direction in the gauge group is reduced to an internal direction.

1. Simple example

To explain the idea in its simplest form we follow Connes (1994) and use as an example the simplest nontrivial case, namely, when the continuous manifold is a four-dimensional compact Riemannian manifold with gauge group $U(1)$ and the discrete manifold consists of just two points. With respect to the new discrete (two-point) direction the zero-forms (functions) $\omega_0(x)$ are taken to be diagonal 2×2 matrices with ordinary scalar functions as entries:

$$\omega_0(x)=\begin{pmatrix} f_a(x) & 0 \\ 0 & f_b(x)\end{pmatrix}\in\Omega_0, \tag{74}$$

where Ω_0 denotes the space of zero-forms. The essential new feature is the introduction of a discrete component $\mathbf{d}$ of the outer derivative d. This is defined as a self-adjoint off-diagonal matrix, i.e.,

$$\mathbf{d}=\begin{pmatrix} 0 & k \\ k & 0\end{pmatrix} \tag{75}$$

with constant entries k. (More generally one could take the off-diagonal elements in d to be complex-conjugate bounded operators, but that will not be necessary for our purpose.) The outer derivative of the zero-forms with respect to $\mathbf{d}$ is obtained by commutation,

$$\mathbf{d}\wedge\omega_0\equiv[\mathbf{d},\omega_0]=[f_b(x)-f_a(x)]\begin{pmatrix} 0 & k \\ -k & 0\end{pmatrix}. \tag{76}$$

The noncommutativity enters in the fact that $d\omega_0$ does not commute with the forms in Ω_0. The one-forms ω_1 are taken to be off-diagonal matrices,

$$\omega_1(x)=\begin{pmatrix} 0 & v_a(x) \\ v_b(x) & 0\end{pmatrix}\in\Omega_1, \tag{77}$$

where the $v(x)$'s are ordinary scalar functions and Ω_1 denotes the space of one-forms. Note that, according to Eq. (76), the discrete component of the outer derivative maps Ω_0 into Ω_1. The outer derivative of a one-form with respect to $\mathbf{d}$ is obtained by anticommutation. Thus

$$\mathbf{d}\wedge\omega_1\equiv\{\mathbf{d},\omega_1\}=[v_b(x)-v_a(x)]kI\in\Omega_0, \tag{78}$$

where I is the unit 2×2 matrix. It is easy to check that with this definition we have $\mathbf{d}\wedge\mathbf{d}\wedge=0$ on both Ω-spaces.

The $U(1)$ gauge group is a zero-form and is the direct sum of the $U(1)$ gauge groups on the two sectors of the zero-forms. Thus it has elements of the form

$$U(x)=\begin{pmatrix} e^{i\alpha(x)} & 0 \\ 0 & e^{i\beta(x)}\end{pmatrix}\in U(1). \tag{79}$$

Its action on both Ω_0 and Ω_1 is by conjugation. Thus under a gauge transformation the zero-forms are invariant and the one-forms transform according to

$$\omega_1(x)\rightarrow\omega_1'(x)=U^{-1}(x)\omega_1(x)U(x). \tag{80}$$

Explicitly,

$$\omega_1'(x)=\begin{pmatrix} 0 & e^{i\lambda(x)}g_a(x) \\ e^{-i\lambda(x)}g_b(x) & 0\end{pmatrix},\quad \text{where } \lambda(x)=\beta(x)-\alpha(x). \tag{81}$$

The discrete component of a connection takes the form

$$V(x)=\begin{pmatrix} 0 & v(x) \\ v^*(x) & \end{pmatrix} \tag{82}$$

and thus resembles a Hermitian one-form. But, being a connection, it is assumed to transform with respect to $U(x)$ as

$$V(x)\rightarrow V_u(x)=U^{-1}(x)V(x)U(x)+e^{-1}U^{-1}(x)dU(x), \tag{83}$$

where e is a constant. The transformation law (83) is the natural extension of the conventional transformation law for connection forms.

The discrete component of the covariant outer derivative is defined to be

$$\mathbf{D}=\mathbf{d}+eV(x)=\begin{pmatrix} 0 & k+ev(x) \\ k+ev^*(x) & 0 \end{pmatrix}$$

$$=e\begin{pmatrix} 0 & \phi(x) \\ \phi^*(x) & 0 \end{pmatrix}, \tag{84}$$

where

$$\phi(x)=v(x)+c, \quad c=\frac{k}{e}. \tag{85}$$

The outer derivative with $\mathbf{D}$ is formed in the same way as with $\mathbf{d}$, namely, by commutation and anticommutation on the forms Ω_0 and Ω_1, respectively. From Eq. (83) it follows in the usual way that $\mathbf{D}$ transforms covariantly with respect to the $U(1)$ gauge group, i.e.,

$$\mathbf{D}[\phi(x)]\rightarrow\mathbf{D}[\phi_\lambda(x)]=U^{-1}(x)\mathbf{D}[\phi(x)]U(x), \tag{86}$$

where

$$\phi_\lambda(x)=e^{i\lambda(x)}\phi(x). \tag{87}$$

This is consistent with the fact that $\mathbf{D}$ acts on the gauge group by commutation.

Note that, although the component $v(x)$ of the connection does not transform covariantly with respect to $U(1)$, the field $\phi(x)$ does. Since $\phi(x)$ is also a spacetime scalar, it can therefore be identified as a Higgs field. As we shall see, the fact that $\phi(x)$ rather than $v(x)$ is identified as the Higgs field is of great importance for the Higgs potential.

Having defined the covariant derivative, we can proceed to construct the curvature. In an obvious notation this can be written as

$$F_{AB}=\begin{pmatrix} F_{\mu\nu} & F_{d\mu} \\ F_{d\mu} & F_{dd} \end{pmatrix}, \tag{88}$$

where $F_{\mu\nu}$ is the conventional curvature and

$$F_{d\mu}=\partial_\mu V-d\wedge A_\mu+[A_\mu,V]$$

$$=\begin{pmatrix} 0 & D_\mu\phi \\ D_\mu\phi^* & 0 \end{pmatrix}\equiv D_\mu\Phi, \tag{89}$$

where D_μ is the conventional covariant derivative. The interesting component is F_{dd}, which turns out to be

$$F_{dd}=\mathbf{d}\wedge V+eV^2. \tag{90}$$

The explicit form of Eq. (90) is easily computed to be

$$F_{dd}=(k(v+v^*)+evv^*)I=e(|\phi|^2-c^2)I. \tag{91}$$

Since it is $\phi(x)$ that must be identified as a Higgs field, the relationship between Eq. (91) and the standard $U(1)$ Higgs potential is obvious.

Before applying the above formalism to physics, however, we have to introduce fermion fields $\Psi(x)$. These are taken to be column vectors of ordinary fermions $\psi_a(x)$,

$$\Psi(x)=\begin{pmatrix} \psi_a(x) \\ \psi_b(x) \end{pmatrix}. \tag{92}$$

The action of the gauge group and the covariant derivative on them is by ordinary multiplication, i.e.,

$$U(x)\Psi(x)=\begin{pmatrix} e^{i\alpha(x)}\psi_a(x) \\ e^{i\beta(x)}\psi_b(x) \end{pmatrix} \tag{93}$$

and

$$\mathbf{D}\Psi(x)=e\begin{pmatrix} 0 & \phi(x) \\ \phi^*(x) & 0 \end{pmatrix}\begin{pmatrix} \psi_a(x) \\ \psi_b(x) \end{pmatrix}$$

$$=e\begin{pmatrix} \phi(x)\psi_b(x) \\ \phi^*(x)\psi_a(x) \end{pmatrix}, \tag{94}$$

respectively. As might be expected from the fact that the fermions are $U(1)$ covariant, it is the $U(1)$-covariant field $\phi(x)$, and not the component $v(x)$ of the connection, that couples to them in Eq. (94).

2. Application to the standard model

As has already been mentioned, the immediate physical interest of the noncommutative gauge theory lies in its application to the standard model of the fundamental interactions. The new feature is that it produces the Higgs field and its potential as natural consequences of gauge theory, in contrast to ordinary field theory in which they are introduced in an *ad hoc* phenomenological manner. The mechanism by which they are produced is very like that used in Kaluza-Klein reduction so, to put the noncommutative mechanism into perspective, let us first digress a little to recall the usual Kaluza-Klein mechanism.

a. The Kaluza-Klein mechanism

Consider the gauge-fermion Lagrangian density in $4+n$ dimensions, namely,

$$L=\frac{1}{4}\mathrm{Tr}(F_{AB})^2+\bar{\psi}\gamma^A D_A\psi, \tag{95}$$

where $A,B=1\ldots4+n$. If we let $\mu,\nu=0\ldots3$ and $r,s=4\ldots n$ and assume that the fields do not depend on the coordinates x_r, the Dirac operator and the curvature decompose into

$$F_{AB}=\begin{pmatrix} F_{\mu\nu} & D_\mu A_r \\ -D_\mu A_r & [A_r,A_s] \end{pmatrix}$$

and

$$\gamma^A D_A=\gamma^\mu D_\mu+\gamma^r A_r, \tag{96}$$

respectively, and hence the Lagrangian (95) decomposes into

$$L=\frac{1}{4}\mathrm{Tr}(F_{\mu\nu})^2-\frac{1}{2}(D_\nu A_r)^2+\bar{\psi}\gamma^\nu D_\mu\psi$$

$$+\bar{\psi}\gamma^s A_s\psi+\frac{1}{4}\mathrm{Tr}[A_s,A_r]^2. \tag{97}$$

The extra components A_r of the gauge potential are space-time scalars and may therefore be identified as Higgs fields. Thus the dimensional reduction produces a standard kinetic term, a standard Yukawa term, and a potential for the Higgs fields. The problem is that the Higgs potential is not the one required for the standard model. In particular, its minimum does not force $|A_r|$ to assume the fixed nonzero value that is necessary to produce the masses of the gauge fields and leptons.

b. The noncommutative mechanism

As we shall now see, the noncommutative mechanism is very similar to the Kaluza-Klein mechanism. But it eliminates the artificial assumption that the fields do not depend on the extra coordinates and it produces a Higgs potential that is of the same form as those used in standard models. As in the Kaluza-Klein case, the procedure is to start with the formal gauge-fermion Lagrangian (95) and expand around the conventional four-dimensional gauge and fermion fields.

From the discussion of the previous section we see that if we expand the Dirac operator and the Yang-Mills curvature in this way we obtain

$$F_{AB}=\begin{pmatrix} F_{\mu\nu} & D_\mu\Phi(x) \\ -D_\mu\Phi(x) & F_{dd} \end{pmatrix}$$

and

$$\gamma^A D_A=\gamma^\mu D_\mu+g\mathbf{D}, \tag{98}$$

respectively, where g is a constant whose value cannot be fixed as the theory does not relate the scales of D_μ and $\mathbf{D}$. The resemblance between Eq. (98) and the corresponding Kaluza-Klein expression (96) is striking.

It is clear from Eq. (98) that for the noncommutative case the formal Yang-Mills-fermion Lagrangian (95) decomposes to

$$L=\frac{1}{4}\mathrm{Tr}(F_{\mu\nu})^2-\frac{1}{2}(D_\nu\phi)^2+\overline{\Psi}\gamma^\nu D_\mu\Psi + G\overline{\Psi}\Phi\Psi+\frac{1}{4}\mathrm{Tr}[F_{dd}(\phi)]^2, \tag{99}$$

where

$$\Phi(x)=\begin{pmatrix} 0 & \phi(x) \\ \phi^*(x) & 0 \end{pmatrix} \quad \text{and} \quad G=eg. \tag{100}$$

Since the field $\phi(x)$ is a scalar that transforms covariantly with respect to the $U(1)$ gauge group it may be interpreted as a Higgs field. Hence, in analogy with the Kaluza-Klein mechanism, the noncommutative mechanism produces a standard kinetic term, a standard Yukawa term, and a potential for the Higgs field. The difference lies in the form of the potential, which is no longer the square of a commutator. From Eq. (91) we have

$$\frac{1}{4}\mathrm{Tr}(F_{dd})^2=\frac{1}{2}e^2[|\phi(x)|^2-c^2]^2. \tag{101}$$

But this is just the renormalizable potential that is used to produce the spontaneous breakdown of $U(1)$ invariance. Putting all the new contributions together, we see that the introduction of the discrete dimension and its associated gauge potential $\phi(x)$ produces exactly the extra terms

$$-\frac{1}{2}[D_\mu\phi(x)]^2+G\overline{\Psi}(x)\Phi(x)\Psi(x) +\frac{1}{2}e^2[|\phi(x)|^2-c^2]^2 \tag{102}$$

that describe the Higgs sector of the standard $U(1)$ model. Thus, when the concept of manifold is generalized in the manner dictated by noncommutative geometry, the standard Higgs sector emerges in a natural way. Note, however, that since there are three undetermined parameters in Eq. (99), the noncommutative approach does not achieve any new unification in the sense of reducing the number of parameters. However, it considerably reduces the ranges of the parameters, places strong restrictions on the matter-field representations, and even rules out the exceptional groups as gauge groups (Schucker, 1997).

Of course the above model is only a toy one, since it uses the gauge group $U(1)\times U(1)$ rather than the gauge groups $U(2)$ and $S[U(2)\times U(3)]$ of the standard electroweak and electroweak-strong models or the gauge groups of grand unified theory.

However, the general structure provided by noncommutative geometry can be applied to any gauge group. Connes himself (Connes, 1994) has applied it to the standard model. There is some difficulty in applying it to grand unified theories because of the restrictions on fermion representations, but a modified version has been applied to grand unified theories in the work of Chamseddine *et al.* (1993). As in the toy model, the noncommutative approach does not achieve any new unification in the sense of reducing the number of parameters, though, as already mentioned, it introduces some important restrictions. Most importantly, it provides a new and interesting interpretation.

ACKNOWLEDGMENTS

We are indebted to C. N. Yang for important remarks which improved the paper. A number of people gave us positive reactions and welcome comments. We thank, in particular, D. Giulini, F. Hehl, U. Lindstrom, T. Schucker, and D. Vassilevich. Special thanks go to T. Damour for some pertinent remarks which led to several improvements of the printed version.

REFERENCES

Appelquist, T., A. Chodos, and P. G. O. Freund, 1987, *Modern Kaluza-Klein Theories* (Addison-Wesley, London).

Audretsch, J., F. Gähler, and N. Straumann, 1984, Commun. Math. Phys. **95**, 41.

Bergmann, V., 1932, Sitzungsber. K. Preuss. Akad. Wiss., Phys. Math. Kl. 346.

Bergmann, P. G., 1942, *An Introduction to the Theory of Relativity* (Prentice-Hall, New York), Chaps. XVII and XVIII.

Bergmann, P. G., 1968, Int. J. Theor. Phys. **1**, 25.

Bleecker, D., 1981, *Gauge Theory and Variational Principles* (Addison-Wesley, London).

Brans, C. H., and R. H. Dicke, 1961, Phys. Rev. **124**, 925.

Cartan, E., 1928, *Leçons sur la Géométrie des Espaces de Riemann*, 2nd ed. (Gauthier-Villars, Paris).

Case, C. M., 1957, Phys. Rev. **107**, 307.

Chamseddine, A. H., G. Felder, and J. Fröhlich, 1993, Nucl. Phys. B **395**, 672.

Chandrasekharan, K., 1986, Ed., *Hermann Weyl, 1885–1985* (Springer, New York).

Connes, A., 1994, *Noncommutative Geometry* (Academic, New York).

Damour, T., and A. M. Polyakov, 1994, Nucl. Phys. B **423**, 532.

Dicke, R. H., 1962, Phys. Rev. **125**, 2163.

Einstein, A., 1987, *The Collected Papers of Albert Einstein*, edited by J. Stachel, D. C. Cassidy, and R. Schulmann (Princeton University, Princeton, NJ).

Feynman, R. P., 1995, *Feynman Lectures on Gravitation*, edited by B. Hatfield (Addison-Wesley).

Fierz, M., 1956, Helv. Phys. Acta **29**, 128.

Fierz, M., 1999, private communication.

Fock, V., 1927, Z. Phys. **39**, 226.

Fock, V., 1929, Z. Phys. **57**, 261.

Green, M., J. Schwarz, and E. Witten, 1987, *Theory of Strings and Superstrings* (Cambridge University, Cambridge, England).

Gross, D., 1995, in *The Oskar Klein Centenary Symposium*, edited by U. Lindstrom (World Scientific, Singapore), p. 94.

Hoffmann, B., 1933, Phys. Rev. **37**, 88.

Infeld, L., and B. L. van der Waerden, 1932, Sitzungsber. K. Preuss. Akad. Wiss., Phys. Math. Kl. 380 and 474.

Jordan, P., 1949, Nature (London) **164**, 637.

Jordan, P., 1955, Schwerkraft und Weltall, 2nd ed. (Vieweg, Braunschweig).

Kaluza, Th., 1921, Sitzungsber. K. Preuss. Akad. Wiss., Phys. Math. Kl., 966; for an English translation see O'Raifeartaigh, 1997.

Klein, O., 1926a, Z. Phys. **37**, 895; for an English translation see O'Raifeartaigh, 1997.

Klein, O., 1926b, Nature (London) **118**, 516.

Klein, O., 1938, *1938 Conference on New Theories in Physics, held in Kasimierz, Poland*; reproduced in O'Raifeartaigh, 1997.

Klein, O., 1956, in *Proceedings of the Berne Congress*, Helv. Phys. Acta, Suppl. IV, 58.

Kubyshin, Y. A., J. M. Mourao, G. Rudolph, and I. P. Volobujev, 1989, *Dimensional Reduction of Gauge Theories, Spontaneous Compactification and Model Building*, Springer Lecture Notes in Physics No. 349 (Springer, New York).

Lichnerowicz, A., 1955, *Théories Relativiste de la Gravitation et de l'Electromagnétisme* (Masson, Paris), Chap. 4.

London, F., 1927, Z. Phys. **42**, 375.

Ludwig, C., 1951, *Fortschritte der Projektiven Relativitätstheorie* (Vieweg, Branunschweig).

Nordström, G., 1912, Phys. Z. **13**, 1126.

Nordström, G., 1913a, Ann. Phys. (Leipzig) **40**, 856.

Nordström, G., 1913b, Ann. Phys. (Leipzig) **42**, 533.

Nordström, G., 1914, Phys. Z. **15**, 504.

O'Raifeartaigh L., 1997, *The Dawning of Gauge Theory* (Princeton University, Princeton, NJ).

Pais, A., 1953, *Conference in Honour of H. A. Lorentz, Leiden 1953*, Physica (Amsterdam) **19**, 869.

Pais, A., 1982, *Subtle is the Lord*: *The Science and Life of Albert Einstein* (Oxford University, New York).

Pauli, W., 1919, Phys. Z. **20**, 457.

Pauli, W., 1921, "Relativitätstheorie," *Encyklopädie der Mathematischen Wissenschaften* (Leipzig, Teubner), Vol. 5.3, p. 539.

Pauli, W., 1933, Ann. Phys. (Leipzig) **18**, 305.

Pauli, W., 1939, Helv. Phys. Acta **12**, 147.

Pauli, W., 1958, *Theory of Relativity* (Pergamon, New York).

Pauli, W., 1979, *Wissenschaftlicher Briefwechsel*, Vol. I: 1919–1929 (Springer, Berlin), p. 505. (Translation of the letter by L. O'Raifeartaigh).

Pauli, W., 1999, *Wissenschaftlicher Briefwechsel*, Vol. IV, Part II (Springer, Berlin), Letters 1614 and 1682.

Polchinski, J., 1998, *String Theory*, Vols. I, II, Cambridge Monographs on Mathematical Physics (Cambridge University, Cambridge, England).

Raman, V., and P. Forman, 1969, Hist. Stud. Phys. Sci. **1**, 291.

Schrödinger, E., 1922, Z. Phys. **12**, 13.

Schrödinger, E., 1932, Sitzungsber. K. Preuss. Akad. Wiss., Phys. Math. Kl. 105.

Schrödinger, E., 1987, *Schrödinger*: *Centenary Celebration of a Polymath*, edited by C. Kilmister (Cambridge University, New York/Cambridge, England).

Schucker, T., 1997, "Geometry and Forces," in *Proceedings of the 1997 EMS Summer School on Noncommutative Geometry and Applications*, Monsaraz and Lisbon, edited by P. Almeida, to appear (hep-th/9712095)

Seelig, C., 1960, *Albert Einstein* (Europa, Zürich), p. 274.

Steinhardt, P. J., 1993, Class. Quantum Grav. **10**, 33.

Straumann, N., 1984, *General Relativity and Relativistic Astrophysics*, Texts and Monographs in Physics (Springer, Berlin).

Straumann, N., 1987, Phys. Bl. (Germany) **43** (11), 414.

Tetrode, H., 1928, Z. Phys. **50**, 336.

Thiry, Y. R., 1948, C. R. Acad. Sci. **226**, p. 216, 1881.

Thiry, Y. R., 1951, These (Université de Paris).

Veblen, O., 1933, *Projektive Relativitätstheorie* (Springer, Berlin).

Wald, R. M., 1986, Phys. Rev. D **33**, 3613.

Weinberg, S., 1965, Phys. Rev. **138**, A988.

Weyl, H., 1918, "Gravitation und Elektrizität," Sitzungsber. Deutsch. Akad. Wiss. Berlin, Klossefü... pp. 465–480. See also H. Weyl, 1968, *Gesammelten Abhandlungen*, edited by K. Chadrasekharan (Springer, Berlin). An English translation is given in O'Raifeartaigh, 1997.

Weyl, H., 1922, *Space, Time, Matter* (Methuen, London, and Dover, New York). Translated from the 4th German Edition. [Raum.· Zeit.· Materie, 8. Auflage (Springer, Berlin, 1993)].

Weyl, H., 1929, "Elektron und Gravitation. I" Z. Phys. **56**, 330.

Weyl, H., 1946, "Memorabilia," in *Hermann Weyl*, edited by K. Chandrasekharan (Springer, New York), p. 85.

Weyl, H., 1956, *Selecta* (Birkhäuser, Boston).

Weyl, H., 1968, *Gesammelte Abhandlungen*, edited by K. Chandrasekharan (Springer, Berlin), Vol. III, p. 229.

Weyl, H., 1980, in *Hermann Weyl*, edited by K. Chandrasekharan (Springer, Berlin), p. 85.

Weyl, H., 1981, *Gruppentheorie und Quantenmechanik* (Wissenschaftliche Buchgesellschaft, Darmstadt), Nachdruck der 2. Aufl., Leipzig 1931. [English translation: *Group Theory and Quantum Mechanics*, Dover, New York (1950)].

Wigner, E., 1929, Z. Phys. **53**, 592.

Yang, C. N., 1980, ''Hermann Weyl's Contribution to Physics,'' in *Hermann Weyl, 1885–1985*, edited by K. Chandrasekharan (Springer, New York), p. 7.

Yang, C. N., 1983, *Selected Papers 1945–1980 with Commentary* (Freeman, San Francisco), p. 525.

Yau, S. T., 1985, ''Compact three-dimensional Kahler Manifolds with zero Ricci curvature, in *Symposium on Anomalies, Geometry, and Topology*,'' edited by W. Bardeen and A. White (World Scientific, Singapore), p. 395. See also references therein.