PHYSICS AND GEOMETRY*

Edward Witten†

In many past epochs, problems arising from theoretical physics influenced the development of mathematics, or structures that first arose in mathematics entered in the development of physics. Famous twentieth century examples would be the role of Riemannian geometry in facilitating the invention of general relativity or the influence of quantum mechanics in the development of functional analysis. The above-cited examples, however, involve innovations in physics that took place sixty or seventy years ago. In the last half century, mathematics and physics developed in very different directions, and really significant interaction between the two disciplines became comparatively rare.

In part this happened because mathematics progressed into abstract realms seemingly unrelated to the humdrum world of the theoretical physicist. In part, it resulted from the way that physics developed. The two basic theories in twentieth century physics are general relativity and quantum field theory. Their successes are in very different realms. General relativity – Einstein's theory of gravity – has its successful applications to large scale astronomical phenomena, while quantum field theory is the framework within which physicists have been able to understand many properties of the elementary particles. General relativity was put in its final form by Einstein in 1915, while quantum field theory has been an open frontier since its formulation in the late 1920's. For half a century, the really fundamental advances in physics have mainly been developments in

* Address at the International Congress of Mathematicians, Berkeley, August 1986.
† Research supported in part by NSF Grant PHY80-19754.

quantum field theory. In this period, it is quantum field theory which has been the central arena for possible interaction between mathematics and physics.

For some decades after the invention of quantum field theory, this theory was formulated in a rather technical and clumsy way, hard to work with even for physicists. It was not at all obvious that quantum field theory really exists as a sound mathematical theory. Most important, in the first few decades of quantum field theory, this subject did not give rise to very many interesting mathematical structures.

The outlook changed in the mid-1970's after non-abelian gauge theories emerged as the quantum field theories most relevant to physics. In the context of these theories, many significant physical problems lead to significant concepts in modern mathematics. For example, the study of magnetic monopoles and instantons involves the topological classification of vector bundles. The solution to the '$U(1)$ problem' of quantum chromodynamics turned out to involve the Atiyah-Singer index theorem. The proper understanding of local and global 'anomalies' involves fairly subtle properties of families of elliptic operators. Various other examples could be cited.

It certainly is charming to see 'practical' applications of some seemingly abstruse mathematics. In some of the cases I have mentioned, the solution of the physics problem has actually required the uncovering of new mathematical theorems. All the same, the mutual interaction of mathematics and physics would remain rather limited, I believe, if it were only a question of quantum field theory. The applications of modern mathematics to quantum field theory are fascinating but relatively specialized; and the same can be said for the role that quantum field theory has so far played in stimulating mathematical innovations. It is in trying to go beyond the limitations of quantum field theory that physicists have really begun to meet mathematical frontiers.

The basic limitation of quantum field theory is that, as we noted earlier, it is only one of two fundamental theories in twentieth century physics, the second being general relativity. Both of these theories play a role in describing the same natural world, so a more complete description of nature must encompass both of them. It has been rather clear, however, since the early days of quantum field theory that there are severe difficulties in trying to combine quantum field theory with general relativity. The formal attempt to quantize general relativity leads to nonsensical infinite formulas. In its early days, quantum field theory faced many difficulties, of which this was only one. As the other difficulties were overcome and quantum field theory emerged as an adequate framework for describing all of the natural forces except gravity, the inconsistency between general relativity and quantum field theory emerged clearly as *the* limitation of quantum field theory.

This problem is a theorists' problem *par excellence*. Experiment provides little guide except for the bare fact that quantum field theory and general relativity both play a role in the description of natural law. Unfortunately, gravitational effects are unmeasurably small in all feasible experiments in which quantum field theory plays an observeable role – and vice-versa. All the same, the inconsistency between the two central theories in physics is clearly an important problem on the logical plane. Indeed, the history of physics gives many examples showing how important such problems are. For example, general relativity was invented in Einstein's effort to resolve an inconsistency between two leading theories of that time, namely special relativity and Newtonian gravity. Quantum field theory was similarly born in an attempt to reconcile non-relativistic quantum mechanics with special relativity.

In the discovery of general relativity, the logical framework came first. Einstein first thought through the physical principles which the new theory should embody, then found in Riemannian geometry the correct mathematical framework, and finally formulated the theory. The development of quantum mechanics and quantum field theory was quite different. There was no *a priori* conceptual insight; experimental clues played an extensive role. As I have indicated, experiment is not likely to provide detailed guidance about the reconciliation of general relativity with quantum field theory. One might therefore believe that the only hope is to emulate the history of general relativity, inventing by sheer thought a new mathematical framework whch will generalize Riemannian geometry and will be capable of encompassing quantum field theory. Many ambitious theoretical physicists have aspired to do such a thing, but little has come of such efforts.

Progress seems to have come, instead, in a rather different way. In the course of attempting to understand the strong interactions, physicists were led in the late 1960's and early 1970's to investigate what came to be known as 'string theory.' String theory was originally discovered by accident, or at least in an exceedingly indirect way, starting with the 'Veneziano model' [30]. Surveys of string theory can be found in [17, 11,25,13]. As string theory was developed, a remarkably rich mathematical structure emerged, but one which bore increasingly little resemblance to strong interactions. By about 1973-4, a successful theory of strong interactions emerged in the context of non-abelian gauge theory. The mathematical structure of string theory retained its fascination, however. By around 1974, just as the original motivation for work on string theory was fading, it was suggested that string theory should be viewed not as a theory of strong interactions but as a framework for reconciling gravitation with quantum mechanics [24]. This idea has many bizarre implications. For instance, it is necessary to believe that (insofar as the conventional concepts of geometry are valid) space-time is ten dimensional rather than four dimensional. After some years of neglect, this idea has been revived in the 1980's, and there are many indications that this framework is close to the truth.

The roundabout path to the discovery of string theory has had a price. Despite learning much about this subject, we still do not know the logical framework in which it has its proper home. It is roughly as if general relativity had been invented, in some peculiar formulation, without knowing about Riemannian geometry; the task would then arise of reconstructing Riemannian geometry as the basic framework behind general relativity. The idea of knowing about general relativity without knowing about Riemannian geometry may sound outlandish, but we are in just such an outlandish situation in string theory. We do not know what the basic logical setting for string theory will turn out to be. We can say that some of the ingredients in string theory are Riemann surfaces, modular forms, and representation theory of infinite dimensional Lie algebras. These are preliminaries for thinking about string theory just as a modicum of elementary linear algebra is a prerequisite for Riemannian geometry and general relativity. While we do not know the proper logical setting for string theory, it seems rather clear that it will involve some fundamental generalization of the usual concepts of geometry. This generalization of geometry is bound to have widespread repercussions for mathematics as well as physics. The unearthing of it will entail a new golden age in the interaction of mathematics and physics.

Very probably, in some suitable sense, the number of fundamental mathematical problems is infinite. On the other hand, I personally believe that the number of really fundamental physics problems is finite. If this is so, then there will only be a finite number of episodes in the future in which mathematics and physics will interact in a really fundamental way. It seems likely that the next several decades will be one of those periods.

1 PARTICLE PHYSICS IN THE 1980'S

This article is written in four sections. In this section, I will review the basic ingredients in our present knowledge of fundamental physics. In the next section, I will try to explain why the idea that space-time is ten dimensional (this is one of the requirements of string theory) is not only compatible with everyday experience but even attractive. Surveys of the subjects treated in these two sections can be found in [35,5,13]. In the third section, I will sketch a very brief introduction to quantum field theory, emphasizing features that are relevant to string theory. The last section will be devoted to string theory.

We will begin our review of theoretical physics with general relativity. In this theory, space-time is a pseudo-Riemannian manifold M, of signature $(- + \dots +)$. In this section, M is four dimensional. I will denote local space-time

$R_{ij} - \frac{1}{2} g_{ij} R = 0$

coordinates as x^i, $i = 1 \ldots 4$. General relativity is governed by a variational principle associated with the Lagrangian

$$S_{GR} = \frac{1}{16\pi G} \int_M R \tag{1}$$

where R is the Ricci scalar of M and G is Newton's constant. The variational (Euler-Lagrange) equation derived from (1) is the equation

$$R_{ij} = 0 \tag{2}$$

for vanishing of the Ricci tensor R_{ij}. From the fundamental natural constants, G, $\hbar$ (Planck's constant), and c (the speed of light), we can form a quantity with dimensions of mass:

$$M_{Pl} = \sqrt{\frac{\hbar c}{G}}. \tag{3}$$

This mass, called the Planck mass, is the really natural mass scale in physics. In conventional units, its numerical value is roughly

$$M_{Pl} \approx 1.02 \times 10^{-5}\,\text{grams}. \tag{4}$$

From the fundamental constants we can likewise construct a fundamental length

$$R_{Pl} = \frac{\hbar}{M_{Pl} c} \approx 10^{-33}\,\text{centimeters} \tag{5}$$

which is known as the Planck length, and a time

$$t_{Pl} \approx 10^{-43}\,\text{second} \tag{6}$$

called the Planck time. The constants $\hbar, c$, and G are so fundamental in physics that it is most natural to work in units in which $\hbar = c = G = M_{Pl} = R_{Pl} = t_{Pl} = 1$. In such units, any physical quantity – any length, time, or mass – is simply a number.

Now, the actual values of the fundamental mass, length, and time are very strange. The Planck mass (4) is actually a macroscopic mass (the mass of a bacterium, perhaps), and is totally off the scale of masses of known elementary particles. The electron mass is 10^{22} times smaller than the Planck mass, and the heaviest elementary particles that we are able to produce in accelerators are still 10^{17} times lighter than the Planck mass. (5) and (6) are likewise completely off the usual scale of elementary particle physics. Everything that we know about quantum field theory comes from experiment probing length scales of at least 10^{-16} cm or times of at least 10^{-26} sec. Such lengths and times are very small by ordinary standards, of course, but by an appropriate yardstick determined by the fundamental quantities of physics they are very large. To make direct experimental probes of how nature reconciles quantum mechanics with general relativity would require experiments sensitive to processes that occur on times of order t_{Pl} or lengths of order R_{Pl}, or with individual elementary particles accelerated to kinetic energies of order M_{Pl}. This is regrettably out of reach for the forseeable future. We can hope for indirect clues from experiment, but progress with quantum gravity will require a great deal of theoretical luck and insight.

Actually, the large value of M_{Pl} has consequences visible in everyday life. Saying that M_{Pl} is very large compared to the mass m of an ordinary particle is the same as saying that Newton's constant G is very tiny on a scale determined by $\hbar$, c, and m:

$$G = \frac{\hbar c}{M_{Pl}^2} << \frac{\hbar c}{m^2} \tag{7}$$

This smallness of Newton's constant means that gravitational forces among individual particles of mass m are very tiny. For interactions among ordinary atoms, gravitation becomes significant only when one considers an aggregation of matter so gigantic that the cumulative effect of gravitational forces among many particles overpowers the extreme weakness of gravity at the atomic level. This means that bodies – such as planets or stars – that form gravitationally out of ordinary atoms must be very large. The fact that the length scale of astronomy is so large compared to the length scale of atoms has the same origin as the fact that the length scale of atoms is so large compared to the Planck length. Both facts are mysteries.

The ordinary masses are so small compared to the natural mass scale of physics that there must be a natural idealization in which they are zero. One of our goals in this section will be to elucidate this generalization.

I would now like to briefly discuss the physical content of general relativity. In two space-time dimensions, the integral (1) is a topological invariant (the Euler characteristic of space-time) and the theory determined by (1) alone does not have much content. In three space-time dimensions, the variational equation $R_{ij} = 0$ implies that space-time is flat (since on a three dimensional manifold the whole Riemann tensor can be written in terms of the Ricci tensor). The characteristic features of general relativity first appear in four dimensions. In four dimensions, the Einstein equation $R_{ij} = 0$ does not by any means imply that space-time is flat. On the contrary, this equation has wave-like solutions; if η_{ij} is the flat space Lorentz metric ($\eta = \mathrm{diag} - + \dots +$) then we can look for a nearly flat solution

$$g_{ij} = \eta_{ij} + h_{ij}, \tag{8}$$

with h considered small. To lowest order in h, the Einstein equations have plane wave solutions

$$h_{ij} = \epsilon_{ij} e^{ik\cdot x} + \text{complex conjugate} \tag{9}$$

where k_i and ϵ_{ij} are constants, obeying

$$k_i k^i = k^i \epsilon_{ij} = \epsilon^i_i = 0. \tag{10}$$

The solution (9) is rather analogous to the plane wave solutions of Maxwell's equations, which describe light waves. When the solution (8) of the linearized

Einstein equations was discovered, immediately after the formulation of general relativity, this was interpreted as a prediction of the existence of gravitational waves – which should travel at the same speed as light waves because they are both governed by $k_i k^i = 0$. At the time, particles (or 'matter') and waves were interpreted as two very different things, and it was definitely a new kind of wave, not a new kind of particle, that was predicted by general relativity.* Ten years later, however, quantum mechanics was developed, and it became clear that waves and particles are different sides of the same coin – the same basic entity will appear as a wave or as a particle depending on the circumstances. Thus general relativity is not only a theory of gravitational forces; it also describes a definite kind of 'matter.' On a conceptual plane this is a remarkable triumph. Merely in trying to invent a theory of gravitational forces based on Riemannian geometry, Einstein was forced to invent a unified theory of gravity and matter. A few things are missing, however. General relativity does not seem to make sense as a quantum theory, and the forms of matter observed in nature are richer than what is predicted by general relativity.

Our next task, then, is to discuss some of the other forms of matter (or equivalently, some of the other types of wave) observed in nature. First of all, we have non-abelian gauge forces. Thus, the space-time manifold M, apart from a Riemannian metric, is endowed with additional structure. Over M we have a principal bundle X

$$\begin{array}{c} X \\ \Big\downarrow G \\ M \end{array} \qquad (11)$$

with a structure group G that is known by physicists as the 'gauge group.' Given any representation R of G, there is an associated vector bundle V_R, which will

* At the time this prediction was made, and for many decades thereafter, the prospect of actually testing this prediction experimentally seemed hopelessly remote – because of the extreme weakness of the gravitational force. However, the invention of radio astronomy and the discovery of radio pulsars has in the last few years made possible an indirect but compelling experimental test of the theory of gravitational waves.

play an important role in our story later. About G, we know experimentally only that it contains $SU(3) \times SU(2) \times U(1)$ as a subgroup, corresponding to the strong, weak, and electromagnetic interactions, respectively.† Let A be a connection on the bundle V and let F be the corresponding curvature two form. The Yang-Mills action (Lagrangian) is then

$$S_{YM} = -\frac{1}{4e^2}\int_M |F|^2, \tag{12}$$

where

$$|F|^2 = g^{ii'}g^{jj'}\langle F_{ij}|F_{i'j'}\rangle. \tag{13}$$

Here g_{ij} is the space-time metric, and $\langle\ |\ \rangle$ is the Cartan-Killing form on the Lie algebra of G. The constant e in (12) is called the Yang-Mills coupling constant. If the group G is not simple, it is possible to generalize (12), introducing a separate coupling constant for each simple factor in the gauge group.

Of course, the metric g that appears in (12) is supposed to be the same as the one in (1), since all this is happening on a single space-time manifold M. We should properly add the Einstein and Yang-Mills Lagrangians and study the combined theory

$$S = S_{GR} + S_{YM}. \tag{14}$$

Upon deriving the Euler-Lagrange variational equations, we will find coupled equations for the Yang-Mills and gravitational fields. This is the proper framework for describing the deflection of light by the sun, and various more exotic processes that unfortunately are undetectably weak.

† Experiment tells us more directly about the Lie algebra of G than about G itself. When I say that G contains the subgroup $SU(3) \times SU(2) \times U(1)$, I really mean only that the Lie algebra of G contains that of $SU(3) \times SU(2) \times U(1)$; there is no claim about the global form of G. For the same reason, in later comments I will not be very precise in distinguishing different groups that have the same Lie algebra.

The next major step is to incorporate what physicists call 'fermions,' as opposed to the Yang-Mills and gravitational fields which are 'bosons.' To this end, we introduce a Clifford algebra, that is, we introduce the 'Dirac matrices' Γ^i, $i = 1 \dots n$, obeying

$$\Gamma^i\Gamma^j + \Gamma^j\Gamma^i = -2g^{ij}, \quad i, j = 1 \dots n. \tag{15}$$

For even n, the irreducible representation S of the Clifford algebra has dimension $2^{n/2}$, while for odd n it is $2^{(n-1)/2}$. The Clifford module S automatically furnishes a representation of the Lorentz group $SO(1, n-1)$. The representation in question is called the spinor representation, the Lorentz generators in this representation being

$$\Sigma^{ij} = -\Sigma^{ji} = \frac{1}{4}[\Gamma^i, \Gamma^j]. \tag{16}$$

For odd n, S is an irreducible representation of the Lorentz group, but for even n – the case that will interest us more – S decomposes as

$$S = S_+ \oplus S_-. \tag{17}$$

S_+ and S_- are the eigenspaces of the involution

$$\overline{\Gamma} = i^{(n+2)/4}\Gamma^1\Gamma^2 \dots \Gamma^n. \tag{18}$$

In four dimensions, the representations S_+ and S_- are complex conjugates of each other.

If the second Stiefel-Whitney class of M vanishes, we can define the 'spin bundle' of M, which I will call $\hat{S}$. It is essentially the vector bundle whose bundle at a fiber $p \in M$ is the Clifford module specified in (15). A physical field ψ which is a section of $\hat{S}$ is called a 'spin one half fermi field.' Leptons (such as electrons) and quarks (from which protons and neutrons are made) are important examples. The phrase 'spin one half' refers to the fact that the weights of the spinor representation of $SO(1, N-1)$ are half-integral.

In even dimensions, the spin bundle decomposes as

$$\hat{S} = \hat{S}_+ \oplus \hat{S}_- \tag{19}$$

by analogy with (17). A physical field which is a section of $\hat{S}_+$ or $\hat{S}_-$ is called a fermi field of positive or negative chirality. Positive and negative chirality fermions are often described as being right-handed or left-handed, respectively; if one shines a beam of positive chirality fermions (particles described mathematically as sections of $\hat{S}_+$) into a block of matter, it will begin to spin in a right handed sense.

Various principles of quantum field theory, such as the CPT theorem, require that fermi fields should be real. In four dimensions, since S_+ and S_- are complex conjugates of one another, if we introduce a fermi field ψ which is a section of $\hat{S}_+$, we must also introduce a fermi field $\tilde{\psi}$ which is a section of $\hat{S}_-$. The same would be true, for the same reason, in $4k$ dimensions, for any k. Matters would be very different in $4k+2$ dimensions, since in that case S_+ and S_- are both real (or pseudoreal), but we will reserve that discussion for the next section.

If we do have in a given theory a spin one half fermi field, what equation should it obey? There is a very natural first order elliptic operator, the Dirac operator D, which maps sections of $\hat{S}$ to sections of $\hat{S}$. It is defined as

$$D = \Gamma^i D_i, \tag{20}$$

where D_i is the covariant derivative. For the time being, we consider the Dirac equation for fermions that interact with the gravitational field only, so we consider the covariant derivative constructed from the Levi-Civita connection. In even dimensions, D can be decomposed as

$$D = D_+ \oplus D_-, \tag{21}$$

with

$$D_- : \hat{S}_+ \to \hat{S}_-, \quad D_+ : \hat{S}_- \to \hat{S}_+. \tag{22}$$

(In other words, D_+ maps sections of $\hat{S}_+$ to sections of $\hat{S}_-$, and vice-versa.) The Dirac equation is thus

$$D_-\psi_+ = 0, \quad D_+\psi_- = 0. \tag{23}$$

This actually is what we would usually call the 'massless' Dirac equation. It makes sense to introduce an arbitrary complex constant λ and write

$$\begin{aligned} D_-\psi_+ + \lambda\psi_- &= 0 \\ D_+\psi_- + \lambda^*\psi_+ &= 0 \end{aligned} \tag{24}$$

(In the second line, λ^* is simply the complex conjugate of λ, so that the second equation is the complex conjugate of the first.) While (23) describes massless waves which travel at the speed of light just like electromagnetic or gravitational waves, (24) describes massive fermions, in fact fermions of mass λ. Now, as I have explained, there are many important cases (like leptons and quarks) in which the mass term is very tiny, less than 10^{-17} in Planck units. More exactly, the cases in which the mass term is absent to such enormous precision are the only cases which we know about, since our technology does not enable us to discover fermions which have masses of order one, if there are any.

In the precise framework that we have been discussing – fermions that couple to the gravitational field only – there is no natural explanation for why the λ term should be absent. To obtain such an explanation – and describe one of the really central observations in physics – we must reintroduce the nonabelian gauge fields that we have temporarily been suppressing. Letting R be any representation of the gauge group G, there is an associated vector bundle V_R. If $\bar{R}$ is the dual or complex conjugate representation of R, then $V_{\bar{R}}$ is the dual bundle to V_R. V_R is canonically isomorphic to $V_{\bar{R}}$ if R is real or pseudo-real, but not if R is complex.

It makes sense now to consider fermi fields which are sections not of S_+ but of

$$W_+ = \hat{S}_+ \otimes V_R. \tag{25}$$

If we do introduce 'right-handed' fermions ψ_+ which are sections of W_+, then in four dimensions the CPT theorem requires us to also introduce 'left-handed' fermions which are sections of the complex conjugate bundle

$$\tilde{W}_- = \hat{S}_- \otimes V_{\bar{R}}. \tag{26}$$

Now, since V_R and $V_{\bar{R}}$ are endowed with their Yang-Mills connections, we can write Dirac equations,

$$\begin{aligned} D_- &: W_+ \to W_- \\ D_+ &: \tilde{W}_- \to \tilde{W}_+ \end{aligned}. \tag{27}$$

Here

$$W_- = \hat{S}_- \otimes V_R, \;\; \tilde{W}_+ = \hat{S}_+ \otimes V_{\bar{R}}. \tag{28}$$

Formally, the Dirac equation takes the same form as before,

$$0 = D_-\psi_+ = D_+\psi_- \tag{29}$$

with now the combined Levi-Civita plus Yang-Mills connection. There is a big difference, though, between (29) and the analogous equation (23) in the absence of Yang-Mills fields. The difference is that if the representation R is complex, then we cannot add a mass term to (29). The equation

$$???\;\; 0 = D_-\psi_+ + \lambda\psi_- \tag{30}$$

only makes sense if the representation R is isomorphic to its dual, since $D_-\psi_+$ is a section of $\hat{S}_- \otimes V_R$, while ψ_+ is a section of $\hat{S}_- \otimes V_{\bar{R}}$. Therefore, *if* fermions transform in a complex representation of the gauge group G, then we can naturally understand why they are so extremely light compared to the natural scale, the Planck scale.

This is precisely what is observed, and it is believed by most physicists to be the proper explanation for why the observed fermions are so very light. Actually, a more precise statement is necessary, since in fact in nature the observed representation R is far from being irreducible. The decomposition of R into irreducible representations of the gauge group $SU(3) \times SU(2) \times U(1)$ seems to contain at least fifteen pieces:

$$R = \oplus_{i=1}^{15} R_i. \tag{31}$$

I will not give an explicit description of the R_i now, since this can be done more economically when we discuss 'grand unification' presently. It follows, of course, from (31) that

$$\tilde{R} = \oplus_{j=1}^{15} \tilde{R}_j. \tag{32}$$

Now, if there were i and j with $R_i \approx \tilde{R}_j$, then the corresponding fermions ψ_{i+} and ψ_{j-} could have masses, since (30) would make sense. However, the observed fermions all have $R_i \neq \tilde{R}_j$, and this, together with the other facts sketched above, explains why they are so light.

Now, let us shift our point of view slightly. There is no reason at all to believe that the fermions that have been discovered experimentally are all of those which exist. Very probably there are many, perhaps even infinitely many, fermions with masses 'of order one,' that is, of order M_{Pl}. Such fermions, of course, must transform in real representations of G. And conversely, fermions in real representations will very plausibly have masses of order one. Let U and $\tilde{U}$ be the G representations of fundamental right and left handed fermions in some underlying theory of nature. Of course, they are duals of one another. In the spirit of K theory, form the formal difference of representations:

$$\Delta = U \ominus \tilde{U}. \tag{33}$$

In addition to observed right handed fermions which transform as R, U may contain additional right handed fermions transforming in some representation U_0

$$U = U_0 \oplus R \tag{34}$$

which must be real since the U_0 particles have gotten mass. (34) implies

$$\tilde{U} = U_0 \oplus \tilde{R}. \tag{35}$$

Therefore, when we form the representation difference (33), U_0 will cancel out:

$$\Delta = U \ominus \tilde{U} = R \ominus \tilde{R}. \tag{36}$$

We conclude, then, that $R \ominus \tilde{R}$, which we can determine at accelerators, is the same as the underlying representation difference (33) which is a property of the fundamental theory. This is why the quantum numbers of the low energy fermions are so important.

A few questions arise here. First of all, one might feel that we have done too good a job of explaining why the light fermions are light. The quark and lepton masses are not zero; they range from about 10^{-22} to perhaps 10^{-17} in the natural units I described earlier. The framework we have sketched might appear to force the quark and lepton masses to be strictly zero, since a mass term in (30) is strictly impossible. The answer to this involves something called 'symmetry breaking.' If we are presented with a vector bundle V with structure group G, it might happen that under some conditions the structure group of V can be reduced to a subgroup G_0. This phenomenon has an analogue in particle physics; it is called gauge symmetry breaking and plays a central role in the Weinberg-Salam-Glashow model of weak interactions. For the moment, I will simply assume that to a mathematical audience it is plausible that there can be a 'physical' counterpart of reducing the structure group of a vector bundle to a subgroup.

Suppose that at some low mass scale m, the gauge group G is effectively reduced to a subgroup G_0. Even if the representations R and $\tilde{R}$ are inequivalent as representations of G, they may be equivalent as representations of G_0. In this case, the fermions that were kept massless by the inequivalence of R and $\tilde{R}$ will be able to gain masses of order m. This is precisely what seems to happen in nature. At a mass scale of order $10^{-17} M_{Pl}$, the gauge group $SU(3) \times SU(2) \times U(1)$ is reduced to $SU(3) \times U(1)$.* At this point, some of the gauge fields become massive (these being the W and Z particles which were discovered at CERN several years ago – the heaviest elementary particles that have been discovered). At the same time, the representations R and $\tilde{R}$ are isomorphic as representations of $SU(3) \times U(1)$, so the light fermions can and do gain mass. In all of this, we do not understand why the mass scale associated with symmetry breaking is so tiny compared to the natural mass scale M_{Pl}. It is, however, pretty clear from our discussion that the idealization in which the masses of the particles are all zero is the situation in which the gauge group $SU(3) \times SU(2) \times U(1)$ is not broken to a subgroup.

In this brief survey of contemporary physics, the idea that the representation R might be complex probably did not seem very shocking. Yet historically, when this was discovered in the 1950's (or more precisely when certain facts were discovered which nowadays have the interpretation I have indicated) it came as a shocking surprise. Why is this? In the basic decomposition $S = S_+ \oplus S_-$ of the spinor representation S into spinors $S_\pm$ of positive and negative chirality, the distinction between S_+ and S_- is a matter of convention. Under a change of the orientation of space-time, called a parity transformation by physicists, S_+ and S_- are exchanged. The representations R and $\tilde{R}$ are therefore exchanged by parity. If we assume that the laws of nature are invariant under parity, then R and $\tilde{R}$ must be isomorphic. Our explanation of the lightness of the fermions therefore rests on parity violation. It was considered startling when in the 1950's it was

---

* These are strong and electromagnetic interactions, respectively. The $U(1)$ in question is a sort of skewed embedding in the original $SU(2) \times U(1)$.

discovered that the weak interactions violate parity. On the other hand, parity is conserved by strong and electromagnetic interactions; this is the statement that R and $\tilde{R}$ are isomorphic as representations of $SU(3) \times U(1)$.

One we realize that symmetry breaking plays an important role in explaining why the light fermions are not strictly massless, it is natural to carry this thought a step further. We know that the gauge group of nature contains at least $G = SU(3) \times SU(2) \times U(1)$. We also know that at very low energies this is reduced to a subgroup $G_0 = SU(3) \times U(1)$. Looking up to higher energies, might the gauge group G which we observe at accessible energies be itself a reduction of a larger group $\overline{G}$ which is relevant at higher energies? It would be far more satisfying to describe nature not by the ungainly product $SU(3) \times SU(2) \times U(1)$ but by a more simple structure. This is the goal of so-called 'grand unified theories.' The most obvious example of a simple group that contains $SU(3) \times SU(2) \times U(1)$ is $SU(5)$. The embedding of $SU(3) \times SU(2)$ in 5×5 matrices of $SU(5)$ can be indicated as follows:

$$\begin{pmatrix} SU(3) & 0 \\ 0 & SU(2) \end{pmatrix}. \tag{37}$$

We then take the $U(1)$ generator to be the unique traceless 5×5 matrix which commutes with the above embedding of $SU(3) \times SU(2)$, namely

$$\mathrm{diag}(1, 1, 1, -3/2, -3/2). \tag{38}$$

If we are to embed the observed gauge group $SU(3) \times SU(2) \times U(1)$ in $SU(5)$, then the representation difference $\Delta = R \ominus \tilde{R}$ must have a natural interpretation as an difference of $SU(5)$ representations. Indeed it does, and here I will write down for the first time an explicit formula for the representation difference observed in nature; this would have been rather clumsy at the $SU(3) \times SU(2) \times U(1)$ level. Let $\mathbf{5}$ be the fundamental five dimensional representation of $SU(5)$, and let $\overline{\mathbf{5}}$ be its dual. Let $\mathbf{10}$ be the antisymmetric part of the tensor product $\mathbf{5} \otimes \mathbf{5}$, and let $\overline{\mathbf{10}}$ be its dual. Then the fermions that we observe in nature come in three copies of a basic structure usually called a 'generation.' The left-handed fermions of a given generation are observed to transform as $\overline{\mathbf{5}} \oplus \mathbf{10}$, while the right-handed fermions transform as the dual of this or $\mathbf{5} \oplus \overline{\mathbf{10}}$. In nature we observe three copies of this, so in other words the representation difference which we observe is

$$\Delta = 3(\overline{\mathbf{5}} \oplus \mathbf{10} \ominus \mathbf{5} \ominus \overline{\mathbf{10}}). \tag{39}$$

When one contemplates extending the observed gauge group G to a larger group $\overline{G}$, it is important to understand that this would entail predicting new forces. In the case of $SU(5)$, and most other grand unified theories, the new forces have a dramatic consequence: they cause the proton (which is otherwise absolutely stable), to be unstable, with a lifetime typically in the range of $10^{32} - 10^{45}$ years, depending on which grand unified theory one considers. In 10^4 tons of water, roughly the largest quantity in a feasible experiment, there are about 10^{34} protons, so assuming that an event rate of one proton decay per year is detectable, the lower part of the interesting range can be probed experimentally. So far proton decay has not been observed.

(39) is much nicer than the corresponding formula at the $SU(3) \times SU(2) \times U(1)$ level, and this is one of the main reasons to believe that the $SU(5)$ model probably has some truth in it. However, it is natural to wonder whether there are groups beyond $SU(5)$ that would be as good or better. One group that works very nicely is $SO(10)$, with the unique non-trivial embedding of $SU(5)$ in $SO(10)$. (The fundamental vector of $SO(10)$, which we will call $\mathbf{10}$, decomposes under $SU(5)$ as $5 \oplus \overline{5}$.) $SO(10)$ has two complex conjugate spinor representations of 16 dimensions each; let us call them $\mathbf{16}$ and $\overline{\mathbf{16}}$. Under $SU(5)$ the $\mathbf{16}$ decomposes as $\mathbf{1} \oplus \overline{5} \oplus \mathbf{10}$, where $\mathbf{1}$ is the trivial representation of $SU(5)$. This is, of course, a real representation. The $\overline{\mathbf{16}}$ decomposes, of course, as the complex conjugate of the 16, or $\mathbf{1} \oplus \mathbf{5} \oplus \overline{\mathbf{10}}$. Therefore, at the $SO(10)$ level, we can rewrite (39) as

$$\Delta = 3(\mathbf{16} \ominus \overline{\mathbf{16}}). \tag{40}$$

This is about as simple as a non-trivial representation difference can be, except that the factor of three is rather peculiar; we will discuss some attempts to explain it in the next section. It is natural, though, to ask whether there are any examples of groups beyond $SO(10)$ that work just as nicely. It turns out that there is one, namely the exceptional group E_6. The smallest non-trivial representation of E_6 is a complex representation which we will call the $\mathbf{27}$, this being its dimension; its dual will be denoted as the $\overline{\mathbf{27}}$. E_6 has an $SO(10)$ subgroup, the decomposition of the $\mathbf{27}$ under $SO(10)$ being $\mathbf{27} = \mathbf{16} \oplus \mathbf{10} \oplus \mathbf{1}$, with $\mathbf{16}$ and $\mathbf{10}$ being as before the spinor and vector of $SO(10)$. At the E_6 level, we can then write

$$\Delta = 3(\mathbf{27} \ominus \overline{\mathbf{27}}). \tag{41}$$

This is the third and last example of a group that seems suitable for grand unification in four dimensions. Having found a hint that E_6 can play a role, it is natural to wonder whether one can go further and base a grand unified theory on the biggest exceptional group E_8. This runs into the problem that E_8 only has real representations, and so would automatically lead to $\Delta = 0$. It is possible to try to use E_8 for grand unification, but this requires some additional ingredients, which we will discuss in the next section.

If one wants to summarize our knowledge of physics in the briefest possible terms, there are three really fundamental observations: (i) Space-time is a pseudo-Riemannian manifold M, endowed with a metric tensor and governed by geometrical laws. (ii) Over M is a vector bundle X with a nonabelian gauge group G. (iii) Fermions are sections of $(\hat{S}_+ \otimes V_R) \oplus (\hat{S}_- \otimes V_{\tilde{R}})$. R and $\tilde{R}$ are not isomorphic; their failure to be isomorphic explains why the light fermions are light and presumably has its origins in a representation difference Δ in some underlying theory. All of this must be supplemented with the understanding that the geometrical laws obeyed by the metric tensor, the gauge fields, and the fermions are to be interpreted in quantum mechanical terms.

2 Physics in Ten Dimensions

The standard model of physics which we have just surveyed has many successes, but leaves many questions open. Surely we would like to find a natural explanation for the peculiar factor of three in (40) and (41). We do not want to merely assume that nature is based on three copies of $\mathbf{16}\ominus\overline{\mathbf{16}}$ of $SO(10)$ or three copies of $\mathbf{27}\ominus\overline{\mathbf{27}}$ of E_6. There must be some more economical structure at a more elementary level. Also, it is very peculiar to study a chain that passes from $SU(5)$ to $SO(10)$ to E_6 without finding some way to extend this to E_8. We would like some natural understanding of the symmetry breaking steps along this chain. Finally, the fact that all we observe in experiment is the character *difference* (33) or (36) is a warning that the true underlying structure may be far richer than is apparent. The underlying representation U may even be infinite dimensional; of course, in this case there will have to be some suitable regularization in the definition of the character difference $\Delta = U \ominus \tilde{U}$.

I will now describe an approach to some of these questions. Let us assume that space-time is not a *four* dimensional pseudo-Riemannian manifold, but has a higher dimension; in fact, the case of ten dimensions is favored in string theory. We consider thus a *ten* dimensional manifold $\overline{M}$, oriented, with a spin structure, and with a metric of signature $(-+++\ldots+)$. One ingredient in the ten dimensional theory will be the Einstein action,

$$S_{GR} = -\frac{1}{16\pi\overline{G}}\int_{\overline{M}} R. \tag{42}$$

At first sight, one might believe that it is preposterous to imagine that the world is ten dimensional. The world seems to be 'obviously' four dimensional (if one includes time along with the three obvious space dimensions). The following is crucial, however. We tend to take for granted that there is some notion of 'empty space' and that any other physical state is obtained by adding particles to 'empty space.' In reality, though, what we interpret as empty space is just some solution of the underlying physical equations which plays a special role in our experience. The reason that there is a solution playing such a special role is largely that the energy scales of our experiments are so small compared to the Planck mass M_{Pl}. The tools at our disposal are so feeble that we can bring about only minor disturbances in whatever solution of the underlying equations we happen to have been born into. Just to make this discussion more concrete, suppose that the ten dimensional Einstein equations derived from (42) are the fundamental equations of physics. Let M^4 be flat four dimensional Minkowski space, and let K be some compact Ricci-flat six manifold. Suppose that the 'vacuum state' of the ten dimensional world is

$$\overline{M} = M^4 \times K. \tag{43}$$

Suppose finally that the radius of K (by which I mean any characteristic measure of the size of K) is comparable to the typical length scale of physics, namely R_{Pl}. This is so tiny compared to the length scales of our observations that K will be indistinguishable from a point; the existence of extra dimensions will not be obvious in everyday life, or even in accelerator experiments. Thus, $\overline{M}$ will be indistinguishable from four dimensional Minkowski space; likewise, we could imitate a four dimensional cosmological model.

This may seem reasonable intuitively. To make things more precise, we will now discuss an important aspect of ten dimensional physics, namely the ten dimensional Dirac equation. We thus introduce the ten dimensional Dirac operator

$$D_{(10)} = \sum_{i=1}^{10} \Gamma^i D_i. \tag{44}$$

We can write it as

$$D_{(10)} = D_{(4)} + D_K \tag{45}$$

where $D_{(4)}$ is the four dimensional Dirac operator

$$D_{(4)} = \sum_{i=1}^{4} \Gamma^i D_i \tag{46}$$

and $D_{(K)}$ is the Dirac operator of K,

$$D_K = \sum_{j=5}^{10} \Gamma^j D_j. \tag{47}$$

We would now like to analyze the ten dimensional Dirac equation

$$0 = D_{(10)} \Psi(x^i, y^j) \tag{48}$$

where x^i and y^j, $i = 1...4$, $j = 5...10$ are coordinates of M^4, and Ψ is a spinor field in ten dimensions.

Let us first briefly discuss just what kind of spinor Ψ will be. The irreducible representation of the Clifford algebra in ten dimensions decomposes as the sum of two irreducible representations of the Lorentz group $SO(1,9)$. They are distinguished by the eigenvalue of

$$\Gamma^{(10)} = \Gamma^1 \Gamma^2 \dots \Gamma^{10}. \tag{49}$$

In ten dimensions, unlike four dimensions, the two irreducible spin representations of $SO(1,9)$, which I will call $S_+^{(10)}$ and $S_-^{(10)}$, are both real. So the CPT theorem permits us to consider a theory in which the spinor field Ψ transforms according to one definite spin representation, say $S_+^{(10)}$. In other words

$$\Gamma^{(10)} \Psi = +\Psi. \tag{50}$$

Actually, ten dimensional supersymmetry and string theory force us to consider this case. If we are interested in thinking about four dimensional physics, we should decompose the spinor representation of $SO(1,9)$ under $SO(1,3) \times SO(6)$, where $SO(1,3)$ is the Lorentz group of M^4 and $SO(6)$ is the structure group of the tangent bundle of K. In this decomposition, only spinor representations of $SO(1,3)$ and $SO(6)$ will appear, since the representation space for a ten dimensional Clifford algebra certainly represents the four dimensional and six dimensional Clifford subalgebras. The precise decomposition is easily worked out by introducing the operators analogous to (49):

$$\Gamma^{(4)} = i\Gamma^1\Gamma^2 \ldots \Gamma^4, \; \Gamma^{(K)} = -i\Gamma^5\Gamma^6 \ldots \Gamma^{10}. \tag{51}$$

(The factors of $\pm i$ are conventional and ensure that the eigenvalues of $\Gamma^{(4)}$ and Γ^K are ± 1.) These matrices obey the obvious relation

$$\Gamma^{(10)} = \Gamma^{(4)} \cdot \Gamma^K. \tag{52}$$

Therefore, in an irreducible representation of $SO(1,9)$ which has $\Gamma^{(10)} = +1$, $\Gamma^{(4)}$ and Γ^K are equal,

$$\Gamma^{(4)} = \Gamma^K. \tag{53}$$

This means that the decomposition of the $SO(1,9)$ spinor representation $S_+^{(10)}$ of positive 'chirality' under $SO(1,3) \times SO(6)$ is

$$S_+^{(10)} = (S_+^{(4)} \otimes S_+^K) \; \oplus \; (S_-^{(4)} \otimes S_-^K). \tag{54}$$

Here $S_\pm^{(4)}$ and $S_\pm^K$ are the positive and negative chirality spin representations of $SO(1,3)$ and $SO(6)$, respectively.

Going back to our problem, we want to solve (48) by separation of variables, using the decomposition (45). Since $D_{(4)}$ and D_K do not commute, but anticommute, the standard procedure of separation of variables needs slight modification.

It is convenient to introduce

$$D'_K = \Gamma^{(4)} D_K \tag{55}$$

which does commute with $D_{(4)}$. D'_K is unitarily equivalent to D_K, and so has the same spectrum (since the matrices $\Gamma^{(4)}\Gamma^j$, $j = 5 \ldots 10$ obey the same Clifford algebra as Γ^j, and the irreducible representation of this algebra is known to be unique). Introduce a complete set of eigenfunctions χ_m of D'_K,

$$D'_K \chi_m = \lambda_m \chi_m. \tag{56}$$

We then write

$$\Psi(x^i, y^j) = \sum_m \phi_m(x^i) \otimes \chi_m(y^j) \tag{57}$$

whereupon (48) reduces to

$$0 = (D_{(4)} + \Gamma^{(4)} \lambda_m) \psi_m \tag{58}$$

or equivalently

$$0 = (D'_{(4)} + \lambda_m) \psi_m \tag{59}$$

where we have introduced

$$D'_{(4)} = \Gamma^{(4)} D_{(4)}. \tag{60}$$

$D'_{(4)}$ is unitarily equivalent to $D_{(4)}$ (since $\Gamma^{(4)}\Gamma^i$ generate a Clifford algebra). (59) is equivalent to the Dirac equation for a massive fermion introduced in (24).

We have thus learned the following important lesson. The eigenvalues λ_m of the Dirac operator D_K on the compact manifold K correspond in four dimensional terms to the masses of the fermions ψ_m. Of course, there are an infinite number of such eigenvalues. But as D_K is an elliptic operator on a compact manifold, there are only a finite number of zero eigenvalues – the eigenvalue zero only appears with a finite multiplicity. The non-zero eigenvalues of D_K will be of order $1/R$, with R being the radius of K. Since we are assuming that the radius of K is of order the Planck length (5), the non-zero eigenvalues of D_K will correspond to fermions with Planckian masses – fermions that we would certainly not be able to discover experimentally. Experimentally accessible four dimensional physics will be determined by the zero eigenvalues of D_K, corresponding to massless particles in four dimensions. As there are only finitely many of these, the ten dimensional theory will look for all practical purposes like a four dimensional theory with a finite number of fermi fields. This is just the sort of structure that we discussed in the last section, so we have achieved our goal of showing that a theory that is really ten dimensional can look four dimensional to an observer whose experiments are limited to low energies.

What is more, it is very interesting that in this framework the observed fermions in four dimensions originate as zero modes of the Dirac operator D_K. Zero eigenvalues of elliptic operators such as the Dirac operator do not arise by accident; they arise when they are related to suitable topological invariants. So here (and in many other instances) basic physical questions lead to questions about the topology of K.

The simplest topological invariant that can predict the existence of zero modes of an elliptic operator is the *index* of the operator. Let n_+ and n_- be the number of zero eigenvalues of D_K of positive and negative chirality, respectively (i.e., $\Gamma_K = \pm 1$). Then the difference $n_+ - n_-$ is called the *index* of the Dirac operator, and is easily shown to be a topological invariant. We have so far been tacitly considering a ten dimensional Dirac equation for a spinor field coupled to the space-time geometry only – no Yang-Mills vector bundle. In this case, it is easily seen from the Atiyah-Singer index theorem, or on various more elementary grounds, that in six dimensions (and more generally in $4k+2$ dimensions) the index of the Dirac operator vanishes. This should come as no surprise; in the last section we found a rationale for the existence of massless fermions in four dimensions only in the case in which there is a Yang-Mills group. Let us therefore generalize our discussion and reintroduce the Yang-Mills gauge fields.

The preceding considerations are rather general. The case favored by developments of the last few years [12,14] in string theory is a ten dimensional theory with gauge group $E_8 \times E_8$. For brevity I will consider only a single E_8. The fermions will be in the adjoint representation of E_8. In a ten dimensional theory which has Yang-Mills fields as well as the gravitational field, to describe the vacuum state it is not enough to specify the vacuum manifold. It is also necessary to specify an E_8 vector bundle X over space-time, endowed with a connection A.

What would be a natural choice of X?* Whenever we discuss a six manifold K, there is one vector bundle that is always present – the tangent bundle T. This of course is endowed with the Levi-Civita connection. The structure group of T is – in the generical case – $SO(6)$. Any embedding of $SO(6)$ in E_8 gives a canonical way to construct an E_8 bundle with connection from the tangent bundle T with its Riemannian connection. There is one embedding of $SO(6)$ in E_8 which is in a sense minimal among such embeddings. It comes from the chain

$$SO(6) \times SO(10) \in SO(16) \in E_8. \tag{61}$$

Here $SO(16)$ is a maximal subgroup of of E_8, and $SO(6) \times SO(10)$ is a maximal subgroup of $SO(16)$.† The embedding (61) turns out to lead to an interesting picture of four dimensional physics.

At this point we encounter the notion of gauge symmetry breaking, alluded to in the last section. What will a low energy physicist interpret as the gauge group? A low energy physicist is 'trapped' in a world with an E_8 bundle X and some particular connection A, and is not able to disturb this world very much. An E_8 gauge transformation that does not leave A invariant is not a symmetry of this particular world but relates it to another world with some other connection

* The following construction was considered in [31,4].

† Recall that we are really working at the Lie algebra level, and not specifying the global structure of the various groups. $SO(16)$ in (61) is really $spin(16)/Z_2$, etc.

A'. Probing such a gauge transformation would involve exciting the very massive degrees of freedom whose inaccessibility to the low energy observer is the reason that the world is apparently four dimensional. What the low energy physicist interprets as the gauge group is the subgroup of E_8 that acts on the bundle X, leaving the connection A invariant. For a generical choice of X and A, this group would be trivial. If, however, X is constructed from the tangent bundle *via* the embedding (61) of $SO(6)$ in E_8, then the subgroup of gauge transformations that leaves the connection invariant is the group $G = SO(10)$ that commutes with $SO(6)$.

This is a promising development, because although E_8 is a not a suitable gauge group for a four dimensional theory (it only has real representations and so would lead to $\Delta = 0$), $SO(10)$ is one of the natural candidates, as we learned in the last section. Let us now compute the character difference Δ which will emerge in the present framework. To do so, it is necessary to decompose the adjoint representation of E_8 under $SO(6) \times SO(10)$. The adjoint representation of E_8, which we will call the **248**, decomposes under $SO(6) \times SO(10)$ in the general form

$$\mathbf{248} = \sum_i L_i \otimes R_i \tag{62}$$

where L_i and R_i are certain representations of $SO(6)$ and $SO(10)$, respectively. For each L_i, there is a corresponding Dirac operator $D_K^{(i)} = D_K^{L_i}$ acting on fermions that transform as L_i under $SO(6)$. Massless fermions in four dimensions that transform as R_i under $SO(10)$ originate as zero modes of $D_K^{(i)}$. In view of (53), zero eigenvalues of $D_K^{(i)}$ with $\Gamma^K = +1$ (or -1) have $\Gamma^{(4)} = +1$ (or -1). Let δ_i be the index of $D_K^{(i)}$. It is the difference between the number of zero eigenvalues of $D_K^{(i)}$ with $\Gamma^K = \pm 1$ or equivalently the difference between the number of zero eigenvalues of $D_K^{(i)}$ with $\Gamma^{(4)} = \pm 1$. Therefore, in the basic character difference Δ of the massless fermions, the $SO(10)$ representation R_i will appear with the coefficient δ_i. Altogether, the character difference Δ will be

$$\Delta = \sum_i \delta_i R_i. \tag{63}$$

We will now evaluate (63) in detail. Let us adopt some conventions. If M is an $SO(6)$ representation and N is an $SO(10)$ representation, the tensor product $M \otimes N$ will be denoted (M, N). Representations of $SO(6)$ and $SO(10)$ will be labeled by their dimension. The relevant representations of $SO(6)$ are the adjoint, **15**, the vector, **6**, and the two spinor representations, **4** and $\overline{\mathbf{4}}$. The relevant representations of $SO(10)$ are the adjoint, **45**, the vector **10**, and the two spinor representations **16** and $\overline{\mathbf{16}}$. The adjoint representation of E_8 decomposes under $SO(6) \times SO(10)$ as

$$\mathbf{248} = (\mathbf{15},\mathbf{1}) \oplus (\mathbf{1},\mathbf{45}) \oplus (\mathbf{6},\mathbf{10}) \oplus (\mathbf{4},\mathbf{16}) \oplus (\overline{\mathbf{4}},\overline{\mathbf{16}}). \tag{64}$$

If L_i is a real representation of $SO(6)$, then (from the Atiyah-Singer index theorem or various more elementary considerations) the index δ_i is zero. What is more, if L_i and $L_{\bar{i}}$ are complex conjugate representations, then $\delta_i = -\delta_{\bar{i}}$. The only complex representations of $SO(6)$ in (64) are the **4** and $\overline{\mathbf{4}}$. So (63) reduces to

$$\Delta = \delta_4(\mathbf{16} - \overline{\mathbf{16}}). \tag{65}$$

Comparing to (40), we see that this is of the correct form to agree with observation, with δ_4 being ± 3 in nature as far as we can see.*

As for the actual value of δ_4, it is one of the most fundamental topological invariants of a six manifold. The de Rham complex of differential forms can be built from the tensor product of two spin bundles. Since the **4** of $SO(6)$ is

* We cannot determine the sign, since the difference between **16** and $\overline{\mathbf{16}}$ of $SO(10)$ is a matter of convention.

one of the two spinor representations, a spinor on K with values in the 4 of $SO(6)$ is equivalent to a certain collection of differential forms. δ_4 can therefore be related to the Euler characteristic χ and the Hirzebruch signature σ, which are the topological invariants that can be made from an index problem in the de Rham complex. Actually, $\sigma = 0$ in six (or $4k+2$) dimensions, and

$$\delta_4 = \chi/2. \tag{66}$$

Thus, we have seen how the observed character difference Δ can be related to something more fundamental, namely the topology of K. Although we have not succeeded in explaining the peculiar number 3 in (40), we have perhaps removed some of the mystery from this. The reason that nature seems to repeat herself, with several 'fermion generations,' is simply that there is no reason for a six dimensional manifold to have Euler characteristic ± 2. A suitable K, starting as we have done with a single spinor field Ψ in ten dimensions, can give rise to any desired number of fermion generations in four dimensions.

I have tried to give the flavor of how properties of four dimensional physics can be extracted from geometric and topological properties of K. There are various other examples, but these should suffice as illustrations. I would like to emphasize, though, that while we have supposed the *vacuum* to be a product $M^4 \times K$, the general physical disturbances do not preserve this product structure. The basic laws are supposed to be ten dimensional laws, governed by (42) or (more likely) some much more refined ten dimensional theory, and an approximate four dimensional picture arises only because the 'vacuum' admits four dimensional but not ten dimensional Poincaré symmetry. Why this is, and what K should be, remain mysteries.

3 QUANTUM FIELD THEORY ON A RIEMANN SURFACE

In the last two sections, we have sketched some of the key ingredients in physics in *classical* terms. Buried in the fine print was the proviso that Yang-Mills theory, etc., should actually be reinterpreted quantum mechanically. This in fact leads to a vastly richer and more formidable structure.*

Quantum field theory has not yet emerged as an important tool in pure mathematics. But there are indications that this will change in the coming period. Both in the theory of affine Lie algebras and in algebraic geometry, structures that are familiar in quantum field theory have recently come to play a major role. (We will hear about the latter subject from G. Faltings in the next lecture.) In trying here to give a very brief introduction to quantum field theory, I will emphasize those aspects of this subject which are necessary for understanding string theory, and which are likely to be related to the areas of mathematics which I have just mentioned.

Let Σ be a Riemann surface, perhaps with boundary, and let ϕ be a real valued function on Σ, that is, a map from Σ to R (the real numbers). Let

$$I(\phi) = \frac{1}{2}\int_{\Sigma} d\phi \wedge *d\phi. \tag{67}$$

This is called 'the action functional of free boson field theory.' Let f be a real valued function on $\partial\Sigma$ (the boundary of Σ), and let $\Omega_f(\Sigma)$ be the space of continuous real-valued functions

$$\phi : \Sigma \rightarrow R \tag{68}$$

whose restriction to $\partial\Sigma$ coincides with f.

If we are given an actual metric on Σ (and not just a conformal class of metrics), then the affine space $\Omega_f(\Sigma)$ acquires a natural Riemannian metric. For

* See, e.g., [16, 10] for introductions.

ϕ, $\phi' \in \Omega_f(\Sigma)$, one defines

$$|\phi - \phi'|^2 = \int_\Sigma (\phi - \phi')^2. \tag{69}$$

In finite dimensions, a Riemannian metric always induces a measure ('the square root of the determinant of the metric'). Hoping that this still works in infinite dimensions, we can try to define

$$Z_f(\Sigma) = \int_{\Omega_f(\Sigma)} e^{-I(\phi)}. \tag{70}$$

Let us see what is involved in trying to define (70). $I(\phi)$ is a quadratic functional of ϕ, so the integral that we are trying to do is rather similar to a finite dimensional Gaussian integral

$$Z(A) = \int_{-\infty}^{\infty} d\phi_1 d\phi_2 \ldots d\phi_n \; e^{-\frac{1}{2}(\phi, A\phi)}. \tag{71}$$

Here ϕ_i are coordinates in an n dimensional Euclidean space, and A is a positive definite quadratic form in n variables. It is well known that the value of (71) is

$$Z(A) = (2\pi)^{\frac{n}{2}} \frac{1}{\det A} = (2\pi)^{\frac{n}{2}} \prod_i \lambda_i^{-1}, \tag{72}$$

with λ_i being the eigenvalues of A. Trying to generalize (72) to infinite dimensions, the factor of $(2\pi)^{\frac{n}{2}}$ does not look very promising for $n \to \infty$. For this reason, and because of the difficulty in defining the determinant $\det A$ in infinite dimensions, we will not be able to define the integral $Z_f(\Sigma)$ for given f and Σ, but ratios such as $Z_{f_1}(\Sigma_1)/Z_{f_2}(\Sigma_2)$ will make sense.

What is the analogue of A in (70)? Evidently, $I(\phi)$ as defined in (67) is equivalent to

$$I(\phi) = \frac{1}{2} \int_{\Sigma} \phi \Delta \phi, \tag{73}$$

where $\Delta = *d*d$ is the usual Laplacian. Let us consider first the case in which Σ has no boundary. Then Δ is a positive definite operator with a discrete spectrum

$$\Delta \phi_i = \lambda_i \phi_i, \; i = 1, 2, 3 \ldots. \tag{74}$$

There is one zero eigenvalue, corresponding to the constant function one, and the others are positive. The integral in (70) should then be precisely an infinite dimensional analogue of (71), so we have to define a regularized version of

$$\prod_i \lambda_i^{-1}. \tag{75}$$

Before trying to define this infinite product, it is necessary to remove the zero eigenvalue. We try to define

$$\prod_{\lambda_i \neq 0} \lambda_i^{-1}. \tag{76}$$

This may be done with zeta function regularization. Let

$$\varsigma(s) = \sum_{\lambda_i \neq 0} \lambda_i^{-s}. \tag{77}$$

The series converges if the real part of s is large enough. It defines a meromorphic function of s which can be shown to be regular at $s = 0$. Then we define

$$\prod_{\lambda_i \neq 0} \lambda_i^{-1} = \exp(\varsigma'(0)). \tag{78}$$

For $\partial\Sigma = 0$, we define then

$$Z(\Sigma) = \exp(\varsigma'(0)). \tag{79}$$

Now, let us work out a formula analogous to (79) for the more general case $\partial\Sigma \neq 0$. The first case to consider is $f = 0$. This hardly presents any novelty. If

$f = 0$, then (70) involves an integral over functions ϕ that vanish on $\partial\Sigma$. With the boundary condition $\phi|_{\partial\Sigma} = 0$, the Laplacian Δ has a discrete spectrum of positive eigenvalues, so we define as before

$$Z_{f=0}(\Sigma) = \exp(\varsigma'(0)). \tag{80}$$

With $\partial\Sigma \neq 0$, and boundary conditions that $\phi = 0$ on $\partial\Sigma$, the operator Δ is strictly positive definite, so there is no need to remove zero eigenvalues in defining (80). It remains to generalize (80) to $f \neq 0$. To accomplish this, we recall the classical theorem that the functional $I(\phi)$ has a unique extremum subject to the boundary condition $\phi|_{\partial\Sigma} = f$. Denote the extremizing function as ϕ_0, and write $\phi = \phi_0 + \phi'$. Since $I(\phi)$ is a quadratic functional of ϕ, and stationary at $\phi = \phi_0$, we have

$$I(\phi) = I(\phi_0) + I(\phi'). \tag{81}$$

If ϕ and ϕ_0 are in $\Omega_f(\Sigma)$, then $\phi' = \phi - \phi_0$ is in $\Omega_0(\Sigma)$. So looking back to (70), the change of variables from ϕ to ϕ' converts an integral over $\Omega_f(\Sigma)$ into an integral over $\Omega_0(\Sigma)$, giving

$$Z_f(\Sigma) = e^{-I(\phi_0)} \cdot Z_0(\Sigma) = e^{-I(\phi_0)} e^{\varsigma'(0)}. \tag{82}$$

Of course, as we have formulated things, (79) and (82) are just definitions. They become theorems only in the context of a suitable infinite dimensional integration theory.

Now, a variety of comments are in order here. First of all, in the above we have been discussing *real-valued* functions on Σ - that is, maps $\Sigma \to R$. More generally, we could pick a Riemannian manifold X, with metric tensor γ, and consider maps from Σ to X. Fixing a map $f : \partial\Sigma \to X$, let $\Omega_f(\Sigma; X)$ be the space of continuous maps from Σ to X whose restriction to $\partial\Sigma$ coincides with f.

Picking local coordinates ϕ_i on X, a map $\Phi : \Sigma \to X$ can be described locally in terms of real-valued functions ϕ^i on Σ. We generalize (67) to

$$I(\Phi) = \int_\Sigma \gamma_{ij}(\Phi) d\phi^i \wedge * d\phi^j. \tag{83}$$

Instead of (70), we want to define

$$Z_f(\Sigma; X) = \int_{\Omega_f(\Sigma;X)} e^{-I}. \tag{84}$$

Now, (70) was an integral over an affine space, and could be defined explicitly as in (79) and (82). (70) is one way of defining what physicists call (bosonic) free field theory. If X is not flat, (84) is an integral over a nonlinear space, and is rather difficult to deal with. (84) is an example of what physicists would call an interacting or nonlinear quantum field theory. The quantum field theory defined in (84) is known as the nonlinear sigma model, and is particularly important in string theory. To define (84) requires much more than the ζ function regularization that we used in free field theory. In fact, (84) cannot be satisfactorily defined for arbitrary X. Generally speaking, (84) can be defined for manifolds X of positive or zero Ricci tensor but not for manifolds of negative Ricci tensor. Much is known about the nonlinear sigma model, though not much of this has been proved. Unfortunately, to explain here what is known about (84) would require a lengthy explanation of the renormalization group, the $1/N$ expansion, factorizable S matrices, etc.

Now, the action function (67) and the integrand in (70) depend only on the conformal class of the metric tensor of Σ – in other words, they depend only on the complex structure of Σ. However, to define even formally an integration measure in (70) we needed to give Σ an actual Riemannian structure, not just a conformal class of Riemannian structures. Therefore, our eventual answer (79) is not a function on the moduli space of Riemann surfaces Σ – it is not invariant under a conformal rescaling of the metric of Σ. However, (79) can be shown to transform rather simply under a conformal rescaling of the metric tensor of Σ. It has been interpreted by Quillen as defining a metric on a certain holomorphic line bundle over moduli space [23]; this structure plays a significant role in the developments we will hear about from Faltings in the next lecture.

The next major step in explaining what quantum field theory is as understood by physicists is to explain the relation between the Feynman path integrals which we have been describing and the older Hilbert space interpretation of quantum field theory. If the boundary of Σ is not empty, it consists of several disjoint circles S_i. Let us consider the case of a single circle S. Let $\Omega(S)$ be the space of continuous real-valued functions on S. As in our discussion of Σ, a Riemannian structure on S induces a Riemannian structure and therefore a measure on $\Omega(S)$. Given such a measure, we can speak of square integrable complex valued functions on $\Omega(S)$, that is, functions ψ with

$$\int_{\Omega(S)} |\psi|^2 < \infty. \tag{85}$$

Of course, it is necessary here to define precisely the measure on $\Omega(S)$, just as we needed zeta function (or some other) regularization in discussing integrals on $\Omega(\Sigma)$. While this can easily be made precise (at least in the case of free field theory), I will just proceed formally. Let us denote the Hilbert space of square integrable functions on $\Omega(S)$ as H_S. It is known as the Hilbert space of the quantum field theory that we are discussing.

Now I would like to define a certain linear operator on H_S known as the 'Hamiltonian' of the theory in question. Let us consider a Feynman path integral (70) on a very special Riemann surface Σ with boundary – a cylinder with a standard, flat metric (figure (1)). The boundary consists of two components, S_1 and S_2. We will think of S_1 and S_2 as two copies of the 'same' circle, displaced in the vertical direction through a distance ('proper time') β. When desired we will feel free therefore to identify the function spaces $\Omega(S_1)$ and $\Omega(S_2)$ as well as the Hilbert spaces H_{S_1} and H_{S_2}. So for $\psi_1 \in H_{S_1}$ and $\psi_2 \in H_{S_2}$ we have an inner product

$$\langle\psi_2|\psi_1\rangle = \int\limits_{\Omega(S)} \psi_2^* \psi_1. \tag{86}$$

A real-valued function f on the boundary of Σ is the same as a pair of functions f_1 and f_2 on the boundary components; we write $f = (f_2, f_1)$. We have already discussed in (70) the Feynman path integral on Σ with given boundary values:

$$Z_{(f_2,f_1)}(\Sigma) = \int\limits_{\Omega_{(f_2,f_1)}} e^{-I}. \tag{87}$$

Now we will go a step further and define a bilinear functional on the Hilbert spaces H_{S_1} and H_{S_2}. Given $\psi_1 \in H_{S_1}$ and $\psi_2 \in H_{S_2}$, we define

$$R(\psi_2, \psi_1) = \int\limits_{\Omega(S_2)} \psi_2{}^*(f_2) \int\limits_{\Omega(S_1)} \psi_1(f_1) \cdot Z_{(f_1,f_2)}(\Sigma). \tag{88}$$

$R(\psi_2, \psi_1)$ is obviously linear in ψ_1 and antilinear in ψ_2, so it has the general form

$$R(\psi_2, \psi_1) = \langle\psi_2|T_\beta|\psi_1\rangle \tag{89}$$

for some linear operator T_β. (Recall that β, indicated in figure (1), is the 'height' of the cylinder Σ.)

T_β is simply that linear operator on $\Omega(S)$ whose kernel is

$$T_\beta(f_2, f_1) = Z_{(f_2,f_1)}(\Sigma). \tag{90}$$

Now, suppose that, as in figure (2), we glue together two cylinders Σ_1 and Σ_2 of thickness β_1 and β_2 along a common boundary component S to make a cylinder

Σ of thickness $\beta = \beta_1 + \beta_2$. The boundary of Σ_1 consists of two components, say S_1 and S, and the boundary of Σ_2 consists of two components, say S and S_2. Suppose we fix real-valued functions f_1, f_2, and f on S_1, S_2, and S, respectively, and look at the product

$$T_{\beta_2}(f_2, f)T_{\beta_1}(f, f_1) = Z_{(f_2,f)}(\Sigma_2)Z_{(f,f_1)}(\Sigma_1). \tag{91}$$

The product (91) is an integral over real valued functions on Σ that have prescribed values on the three circles S_1, S_2, and S (functions, in other words, whose restriction to S_1, S_2, or S coincides with f_1, f_2, or f). If we do *not* wish to specify the values on S, we can avoid this by integrating (91) over $\Omega(S)$. We thus consider

$$\int_{\Omega(S)} T_{\beta_2}(f_2, f)T_{\beta_1}(f, f_1) = \int_{\Omega(S)} Z_{(f_2,f)}(\Sigma_2)Z_{(f,f_1)}(\Sigma_1). \tag{92}$$

Here we are integrating over real-valued functions on Σ whose values are specified only on $\partial\Sigma$ – in other words, we are considering our basic integral (70). Thus, (92) coincides with the kernel $T_\beta(f_2, f_1)$ associated according to (90) with integration over real-valued functions on the cylinder Σ of height $\beta = \beta_1 + \beta_2$:

$$T_{\beta_1+\beta_2}(f_2, f_1) = \int_{\Omega(S)} T_{\beta_2}(f_2, f)T_{\beta_1}(f, f_1). \tag{93}$$

This equation is a semigroup law:

$$T_{\beta_1+\beta_2} = T_{\beta_2}T_{\beta_1}. \tag{94}$$

(94) means that

$$T_\beta = e^{-\beta H} \tag{95}$$

for some linear operator H. H is known as the 'Hamiltonian' of the quantum field theory. In the free field theory that we have been discussing, an explicit formula for H is easily written. The reasoning by which we have given a Hilbert space interpretation to the Feynman path integral (70) and extracted a Hamiltonian is, however, far more general.

We would now like to compute the trace of the operator $e^{-\beta H} = T_\beta$. This is done, of course, by integrating the kernel $T_\beta(f_2, f_1)$ along the diagonal:

$$\mathrm{Tr}\, e^{-\beta H} = \int_{\Omega(S)} T_\beta(f, f). \tag{96}$$

In fact, (96) is simply an integral over real valued functions on the cylinder Σ of figure (1) with the boundary values on the two components identified but otherwise unrestricted. By identifying the boundary components of Σ one forms a Riemann surface (without boundary) of genus one, which we may call $\overline{\Sigma}$. (96) is simply the path integral of (70) carried out over $\overline{\Sigma}$:

$$\mathrm{Tr}\, e^{-\beta H} = \int_{\Omega(\overline{\Sigma})} e^{-I}. \tag{97}$$

Before discussing the significance of (97), we need a generalization. We return to the Hilbert space H_S associated with a circle S. The operation of rotating the circle by an angle θ obviously acts in a natural way on real-valued functions $\Omega(S)$ and therefore also on the Hilbert space H_S. We thus have a linear operator R_θ representing the action of the rotation on H_S. (We will use the same symbol R_θ to denote the action of this rotation on $\Omega(S)$.) We again obviously have a semigroup law, $R_{\theta_1+\theta_2} = R_{\theta_2} R_{\theta_1}$, so

$$R_\theta = e^{i\theta P} \tag{98}$$

where P is some linear operator, known as the momentum operator of the quantum field theory. Again, there is no difficulty in writing down explicit formulas for P. It is easy to see that it commutes with H. We now wish to generalize

(97) and calculate the trace of the operator $T_{\beta,\theta} = e^{-\beta H + i\theta P}$. The kernel of this operator is simply

$$T_{\beta,\theta}(f_2, f_1) = T_\beta(f_2, R_\theta f_1). \tag{99}$$

Here $T_\beta(f_2, f_1)$ is the kernel of $e^{-\beta H}$, which we have discussed at length already, and the formula (99) holds because in $e^{-\beta H + i\theta P} = e^{-\beta H} \cdot e^{i\theta P}$, the effect of $e^{i\theta P}$ is just to replace f_1 with $R_\theta f_1$. Again constructing the trace by integrating the kernel along the diagonal, we have

$$\mathrm{Tr}\, e^{-\beta H + i\theta P} = \int\limits_{\Omega(S)} T_\beta(f, R_\theta f). \tag{100}$$

As in our previous discussion of the case $\theta = 0$, (100) has a simple interpretation. The right hand side of (100) is, in the terminology of (70),

$$\int\limits_{\Omega(S)} Z_{(f,R_\theta f)}(\Sigma) = \int\limits_{\Omega(S)} \int\limits_{\Omega_{(f,R_\theta f)}(\Sigma)} e^{-I}. \tag{101}$$

This is easily described in words. The right hand side of (101) is a path integral on the cylinder of figure (1). The integral ranges over all real valued functions whose values on the boundary components S_1 and S_2 coincide after rotating S_1 through an angle θ. Thus, if we take our cylinder Σ of height β, and glue the two components together after rotating through an angle θ, we form a Riemann surface of genus one without boundary. We will call this surface $\Sigma(\beta, \theta)$. Our conclusion is

$$\mathrm{Tr}\, e^{-\beta H + i\theta P} = \int\limits_{\Omega(\Sigma(\beta,\theta))} e^{-I}. \tag{102}$$

(102) is a very general formula in quantum field theory. However, it is particularly interesting in the case in which the action functional I is conformally invariant – independent of a conformal rescaling of the metric of Σ. Such theories

are said to be conformal field theories. For example, free boson field theory is a conformal field theory. In the case of a conformal field theory, the right hand side of (102) might appear at first sight to define a conformal invariant, depending only on the conformal structure of the surface $\Sigma(\beta,\theta)$. This is not quite true, because there is no conformally invariant way to define the measure in the integral over $\Omega(\Sigma)$. Nevertheless, it is possible to construct simple formulas for the deviation from conformal invariance of the right hand side of (102). For mathematicians, these formulas involve the theory of the determinant line bundle; for physicists, they have involved the theory of anomalies.

It is well known that every Riemann surface of genus one is isomorphic to $\Sigma(\beta,\theta)$ for some values of β,θ in the range

$$0 \leq \beta < \infty, \; 0 \leq \theta < 2\pi \tag{103}$$

The correspondence is not one-to-one. If we introduce the complex variable $\tau = (\theta + i\beta)/2\pi$, then the group $SL(2,Z)$ acts on τ by

$$\tau \rightarrow \frac{a\tau + b}{c\tau + d}, \tag{104}$$

where a, b, c, and d are integers with $ad - bc = 1$. Riemann surfaces with values of τ that are related by (104) are isomorphic. In the theory of modular forms it is standard to introduce $q = e^{2\pi i\tau}$. If we define

$$H_{\pm} = \frac{H \pm P}{2} \tag{105}$$

then (102) amounts to

$$\mathrm{Tr}\, q^{H_+}\bar{q}^{H_-} = \int\limits_{\Omega(\Sigma(\tau))} e^{-I}. \tag{106}$$

Conformal field theories in which $H_- = 0$ are said to be chiral theories. The free boson theory (67) is not chiral. The simplest example of a chiral theory is the theory of free chiral fermions, but I will not explain this here. In a chiral theory, the right hand side of (106) defines a holomorphic function of q or in other words a modular form. In any case, in conformal field theories, whether chiral or not, one can obtain simple formulas for the transformation of (106) under conformal transformations, by using the theory of the determinant line bundle or the theory of anomalies [32, 9].

For example, in the theory of affine Lie algebras, it is known that the characters of the integrable highest weight modules transform in a simple way under $SL(2,Z)$. (See [18] for a survey.) At first sight this comes as something of a surprise; $SL(2,Z)$ does not enter the structure of these algebras in any obvious way; it is certainly not a group of automorphisms of them. However, it is known that the highest weight modules of affine Lie algebras have a natural description in terms of quantum field theory. For the level one modules of most of the affine Lie algebras this can be done using free bosons or free fermions [6,20,26, 7]. In general, arbitrary integrable highest weight modules of affine Lie algebras have quantum field theory realizations, albeit more complicated [33]. The quantum field theories in question are conformal field theories, so the appearance of $SL(2,Z)$ in the theory of affine Lie algebras is a special case of our assertion that $SL(2,Z)$ has a simple action on (106).

Most of the work on affine Lie algebras has been done in the context of the Hamiltonian formulation of quantum field theory. As we have just seen, the role of $SL(2,Z)$ becomes clear only in the path integral formulation. The discussion has made it clear that the Hamiltonian perspective is related to path integrals on a cylinder. In and of itself the Hamiltonian perspective seems self-contained. The path integral approach suggests the obvious generalization to other Riemann surfaces, which would scarcely occur to us if we think in Hamiltonian terms only. Given a quantum field theory whose path integral on a cylinder constructs a highest weight module of an affine Lie algebra, what is the mathematical significance of the same path integral on, say, a Riemann surface of genus g? I cannot propose an answer to this question. One reason I hope that this question will attract interest is that a proper answer might well shed some light on string theory.

The path integrals that we have discussed until now are by no means the most general ones that are usually considered in quantum field theory. In fact, we have only considered very special cases of the usual structures. Usually one defines what are known as local operators. Consider, for instance, the free boson field theory which we have given as a simple example of quantum field theory. Let Σ be a Riemann surface. Let P be a point on Σ, and let x^i be local coordinates near P. In studying free boson field theory, we are integrating over real-valued functions $\phi : \Sigma \rightarrow R$. By a local operator at P we mean a functional O of ϕ which depends only on $\phi(P)$ and the derivatives of ϕ at P. More specifically, O is required to have only a polynomial dependence on the derivatives of $\phi(P)$. Thus, examples of local operators would be

$$F(\phi(P)),\ \ F^i(\phi(P))\frac{\partial\phi}{\partial x^i},\ \ F^{ij}(\phi(P))\frac{\partial^2\phi}{\partial x^i\partial x^j}, \tag{107}$$

where F, F^i, F^{ij} are arbitrary real valued functions of $\phi(P)$. (In string theory the case $F = e^{i\lambda\phi}$, $\lambda \in R$, is particularly important.) What do we do with such local operators? Let P_i be some points on Σ, and let O_i be local operators at P_i. Then we generalize (70) to consider

$$Z_f(O_i;\Sigma) = \int_{\Omega_f(\Sigma)} e^{-I}\prod_i O_i(P_i). \tag{108}$$

This would be described as a path integral on the surface Σ with insertion of the operators $O_i(P_i)$. One usually defines the 'correlation function'

$$\langle O_1(P_1)O_2(P_2)\dots O_n(P_n)\rangle = \frac{Z_f(O_i;\Sigma)}{Z_f(\Sigma)}. \tag{109}$$

For the significant choices of the O_i (for example, the operators (107) with $F = e^{i\lambda\phi}$), it is possible to work out quite explicit formulas for such correlation functions, which depend only on the Green function of the Laplacian Δ. I will not enter into this here.

It remains to explain why the local functionals $O(P)$ are called local operators. Indeed, they correspond to operators in the Hamiltonian formulation of quantum field theory. Consider again a path integral on our familiar cylinder Σ, but now with an insertion of the operator $O(P)$ at the point P. Let β_1 and β_2 be the distance from P to S_1 and S_2, respectively. Generalizing (88), we consider the integral

$$R(O(P);\psi_2,\psi_1) = \int\limits_{\Omega(S_2)} \psi_2{}^*(f_2) \int\limits_{\Omega(S_1)} \psi_1(f_1)\cdot Z_{(f_1,f_2)}(O;\Sigma). \tag{110}$$

As in our previous case, linearity in the ψ_i ensures that (110) is of the form $\langle\psi_2|U|\psi_1\rangle$ for some operator U in the Hilbert space H_S of the quantum field theory. In fact, one defines a linear operator $\hat{O}(P)$ by

$$R(O(P);\psi_2,\psi_1) = \langle\psi_2|e^{-\beta_2 H}\hat{O}(P)e^{-\beta_1 H}|\psi_1\rangle. \tag{111}$$

This a canonical correspondence between local functionals $O(P)$ that can be inserted in the path integral and local operators $\hat{O}(P)$ in the Hilbert space. (The precise sense in which the Hilbert space operator $\hat{O}(P)$ is 'local' is something I will not enter into here.) Local operators $O(P)$ are crucial in physical applications of quantum field theory, and suitable local operators are the natural tools for describing the highest weight modules of affine Lie algebras.

There is in addition an important correspondence between operators and vectors in Hilbert space. Consider the simplest Riemann surface with boundary, namely a disc D. Fix a point P in the interior of Σ. We have a Hilbert space H_S associated with real valued functions on the boundary S of D. We also have local operators at P; they form a vector space H_P. I will describe a natural map from H_P to H_S. Let $O(P)$ be a local operator at P. We wish to construct an element of H_S. Let f be a real-valued function on S, and let

$$\psi_O(f) = Z_f(O;D). \tag{112}$$

The correspondence $O \to \psi_O(f)$ gives the desired map $H_P \to H_S$. In conformal field theory, this correspondence is an isomorphism between H_S and H_P, and is of special importance.

One local operator of particular importance is the following. Under an infinitesimal change in the metric g_{ij} of Σ, the change in the action I of a quantum field theory is

$$\delta I = \int_{\Sigma} \delta g^{ij} T_{ij}, \tag{113}$$

where the symmetric tensor T_{ij} is known as the energy-momentum tensor. A conformal field theory is precisely one in which $g^{ij}T_{ij} = 0$. In a conformal field theory on a Riemann surface, T_{ij} has only two independent non-zero components, which transform as differentials of type $(2,0)$ and $(0,2)$, respectively. Let us call these T and $\tilde{T}$, respectively. Letting σ be an angular variable on our canonical circle S, we define

$$L_n = \int_S e^{in\sigma} T(\sigma), \quad \tilde{L}_n = \int_S e^{in\sigma} \tilde{T}(\sigma). \tag{114}$$

These can be shown to obey

$$\begin{aligned} [L_n, L_m] &= (m-n)L_{m+n} + c\delta_{m+n}(m^3 - m) \\ [\tilde{L}_n, \tilde{L}_m] &= (m-n)\tilde{L}_{m+n} + \tilde{c}\delta_{m+n}(m^3 - m) \\ [L_n, \tilde{L}_m] &= 0 \end{aligned} \tag{115}$$

where c and $\tilde{c}$ are constants; δ_n is one for $n = 0$ and otherwise zero. Restricting ourselves to the L's, (115) is the so-called Virasoro algebra, a central extension of the Lie algebra of diffeomorphisms of S. The latter algebra is generated by the vector fields

$$D_n = -ie^{in\sigma}\frac{d}{d\sigma}, \tag{116}$$

which are easily seen to obey $[D_n, D_m] = (m-n)D_{m+n}$; the first line in (115) is a central extension of (116). A representation of the Virasoro algebra is called a highest weight representation if L_0 is bounded above; a highest weight vector is one annihilated by the L_n for $n > 0$. Conformal field theories lead to highest weight representations of the Virasoro algebra, a special case of this being the assertion that the highest weight representations of affine Lie algebras can be extended to the semi-direct product with the Virasoro algebra.

The introduction to quantum field theory that I have given has been very sketchy to say the least. Quantum field theory is a rather rich and complex subject, and I have only explained a few generalities. Describing quantum field theory as a mathematical theory will become far easier and more natural once some of its characteristic mathematical applications begin to emerge. There are grounds for believing that – after sixty years – this time may be nearly at hand.

4 STRING THEORY

Finally, I would like to briefly describe what string theory is. String theory is a subject even more vast than those that we have considered previously, and my treatment will be even sketchier. Let us return to general relativity, described by the action functional

$$S = \frac{-1}{16\pi G} \int_M R, \tag{117}$$

with R being the Ricci scalar of a Riemannian metric g on a space-time manifold M. Treating (117) as a quantum theory would mean the following. Fixing the topology of M, let $\mathcal{A}$ be the space of metrics on M modulo diffeomorphisms. The quantization of general relativity would involve defining integrals such as

$$Z = \int_{\mathcal{A}} e^{-S} \tag{118}$$

as well as generalizations with insertions of operators, modifications of boundary conditions if $\partial M \neq 0$, etc. While there is no rigorous theorem here, all the indications are that there is no satisfactory way to make sense of the integration measure in (118). This is one way of expressing the inconsistency between general relativity and quantum mechanics which was cited in the introduction as a central problem in theoretical physics.

Before discussing the string theory generalization of (118), let us discuss the perturbative expansion of general relativity. Let η_{ij} be the metric tensor of flat Minkowski space, and expand the metric tensor g_{ij} as

$$g_{ij} = \eta_{ij} + h_{ij} \tag{119}$$

where h_{ij} is the metric disturbance. In an expansion near flat space, one can think of h_{ij} as taking values in a linear space Λ_0. If the action function I is written in terms of h, the linear term vanishes because flat Minkowski space is a solution of the Einstein equations. The quadratic term is of the general form

$$S_2 = \frac{-1}{16\pi G} \int_M h \Delta_L h, \tag{120}$$

where Δ_L is a certain linear operator known as the Lichnerowicz Laplacian. The solutions of $\Delta_L h = 0$ are the linearized gravitational waves discussed in section (1). The cubic term is

$$S_3 = -\frac{1}{16\pi G} \Phi_3(h), \tag{121}$$

with $\Phi_3(h)$ being a complicated cubic expression, the details of which need not concern us. On substituting in (118) the expansion $S = S_2 + S_3 + \ldots$ of the action, one can try to develop a perturbation expansion, viewing h as a small quantity. The starting point, discarding all terms beyond S_2, gives a Gaussian integral similar to those discussed in the last section. The correction terms can be computed but, in the case of general relativity, one runs into nonsensical infinite formulas. This is one major aspect of why we believe that general relativity does not make sense as a quantum theory.

In theories that do make sense as quantum theories, a major aspect of our understanding is the perturbative expansion analogous to the above. Such perturbative expansions are not beautiful. The beauty, if any, of a quantum field theory is to be found in the basic formulas, analogous to (117), not in the nitty-gritty of a perturbative expansion. Nevertheless, in the case of string theory, it is the perturbation expansion that we know. We do not know the basic formulas like (117), or the basic logical concepts that should play in string theory the role played by metrics, connections, and curvature in general relativity.

I will therefore explain in turn some of the ingredients which in string theory are analogous to h, Λ_0, I_2, and I_3. First of all, the linear space Λ_0 is replaced by the Hilbert space of a certain quantum field theory. Our discussion of quantum field theory in the last section was rather formal, and we did not discuss the physical interpretation of the Riemann surface Σ or the quantum field ϕ. In traditional applications of quantum field theory, Σ plays the role of space-time, and ϕ is a field (analogous to the electromagnetic field) propagating in space-time. In string theory, the interpretation is reversed; the Riemann surface Σ is an auxiliary object, and ϕ is replaced by a map $\Phi : \Sigma \to M$, with M interpreted as space-time. Thus, let M be a flat manifold of dimension D, with standard coordinates X^i, $i = 1 \dots D$. The map $\Phi : \Sigma \to M$ is specified by giving D real-valued functions X^i on Σ. The action functional is thus

$$I = \frac{1}{2\pi} \int_\Sigma \eta_{ij} dX^i \wedge *dX^j. \qquad (122)$$

Clearly, (122) is invariant under D dimensional Poincaré transformations of the X^i.

Now, in the linearized theory of general relativity, the main ingredients are the linear space Λ_0 of metric disturbances and the quadratic action functional S_2 on this space. What are the analogues of these in string theory? The analogue of Λ_0 is the Hilbert space H_S of the quantum field theory (122). The replacement of

Λ_0 by H_S is really a quite drastic step, since H_S is a quite infinite space compared to Λ_0. An element of Λ_0 is, concretely, a finite collection of functions

$$h_{ij}(x^k), \tag{123}$$

with x^k, $k = 1 \ldots D$, being the coordinates of the space-time manifold M. What is an element of H_S? Letting σ be an angular parameter on the circle S, we defined in the last section the space Ω_S of maps from S to M. Such a map is concretely given by specifying real-valued functions $X^k(\sigma)$ for which we will make a Fourier expansion

$$X^k(\sigma) = x^k + \sum_{n \neq 0} e^{in\sigma} x_n^k. \tag{124}$$

We have separated out the zeroth Fourier component, which is known as the 'center of mass of the string.' We would like to think of x^k in (124) as corresponding to the x^k in (123). Of course, x_n^k is the complex conjugate of x_{-n}^k. An element of H_S is a real-valued function on Ω_S, or concretely a function

$$\Phi(x^k; x_{\pm 1}^k, \ldots). \tag{125}$$

Thus, Φ depends on an infinity of variables besides the center of mass coordinates x^k which appear already in (123). The sense in which string theory has field theory as an approximation is that if the x_n^k in (125) can be considered as small, then Φ reduces, in the first instance, to a function of the x^k just like the gravitational field or any other field normally considered in physics. To be somewhat more precise about this, suppose that the x_n^k for $n \neq 0$ are small enough that it is appropriate to make a Taylor series expansion about $x_n^k = 0$. The form of the expansion would be*

$$\Phi(x^k; x_n^k) = \phi(x^k) + x_1^j B_j(x^k) + x_1^i x_{-1}^j h'_{ij}(x^k) + \ldots. \tag{126}$$

In this expansion, the $\phi(x^k)$, $B_i(x^k)$, h'_{ij}, etc., are ordinary functions of the center

* This is not precisely the right expansion to make, but it will do for our purposes here.

of mass coordinates, that is, ordinary functions on M. The first point here is that we can think of Φ as an infinite collection of functions on space-time. General relativity is based on certain nonlinear partial differential equations for h_{ij}. In string theory, we will likewise write down nonlinear partial differential equations for Φ (or a somewhat larger set of variables). Concretely, a nonlinear equation for Φ can be understood as a system of coupled nonlinear equations for an infinite set of variables in spacetime. The second major point about (126) is that there is a very definite sense in which Φ should be viewed as a generalization of (123). In (126), there appears the tensor field h'_{ij}, and this should be viewed as the analogue of h_{ij}. The nonlinear equation that one studies in string theory has the property that, when restricted to h'_{ij}, it reduces approximately to the Einstein equations obeyed by h_{ij}, on length scales large compared to the Planck length.

Having explained what is the string theoretic analogue of Λ_0, the next step is to explain what is the analogue in string theory of the quadratic action functional S_2 of the linearized theory of general relativity. Roughly speaking, the analogue of S_2 is just

$$S_2(\Phi) = (\Phi|H|\Phi) \tag{127}$$

where H is the Hamiltonian of the quantum field theory (122). This isn't quite the right definition, but must be supplemented by a restriction to highest weight vectors of the Virasoro algebra. However, it will do for the moment.

What is the analogue in string theory of the nonlinear terms S_3, etc., of general relativity? The right approach involves a simple elaboration of ideas from the last section. Consider a Riemann surface Σ whose boundary consists of three circles S_1, S_2, and S_3, as in figure (4). A simple variant of constructions in the last section is to consider a path integral on Σ in which the boundary condition on each boundary component is determined by Φ:

$$S_3(\Phi) = \int_{\Omega(S_1)} \Phi(f_1) \int_{\Omega(S_2)} \Phi(f_2) \int_{\Omega(S_3)} \Phi(f_3)\, Z_{(f_1,f_2,f_3)}(\Sigma). \tag{128}$$

This is then roughly the analogue of the cubic term in the Einstein action $S_3(h)$. It can obviously be generalized to a case with n boundary components.

There is, however, another language for thinking about this. At the end of the last section we noted that in conformal field theory there is a natural isomorphism $H_P \approx H_S$ between local operators $O(P)$ that can be inserted at a point P and vectors ψ_O in the Hilbert space H_S. The state Φ in (128) corresponds under this isomorphism to some operator V_Φ (called the vertex operator of Φ). Henceforth we will refer to V_Φ merely as V. According to (112), the relation between Φ and V is that if we introduce for $i = 1, 2, 3$ a disc D_i with boundary S_i, then

$$\Phi(f_i) = Z_{f_i}(V; D_i). \tag{129}$$

Suppose that we compactify the surface Σ of figure (4) by gluing in the disc D_i along each boundary component S_i, for $i = 1, 2, 3$. Let us call the resulting closed surface $\overline{\Sigma}$.

We want to reexpress (128) as a path integral on $\overline{\Sigma}$. In (128), Φ has been used to define the boundary condition on each S_i. According to (129), defining the boundary condition on S_i by Φ is the same as gluing in the disc D_i with the operator V inserted on a point P_i of D_i. Thus, (128) is really equivalent to a path integral on $\overline{\Sigma}$ with insertions of the vertex operator V at the three points P_1, P_2, and P_3:

$$S_3(\Phi) = Z(V(P_1), V(P_2), V(P_3); \overline{\Sigma}). \tag{130}$$

In (130) we have not been too particular about explaining which three points P_1, P_2, and P_3 have been chosen. There is a definite sense in which this does not matter. The Riemann surface $\overline{\Sigma}$ is isomorphic to the Riemann sphere. It is well known that the group $SL(2, C)$ acts transitively on the configuration of three disjoint points on the Riemann sphere, so there is no conformal invariant associated with the choice of the P_i. If, however, we want to define not $S_3(\Phi)$ but an analogous object $S_n(\Phi)$ for $n > 3$, we face the question of choosing points

P_i. The correct prescription, as articulated by Polyakov [22], is to integrate over the moduli of configurations of n points on $\bar{\Sigma}$:

$$S_n(\Phi) = \int_{\mathcal{M}_n} Z(\prod V(P_i); \bar{\Sigma}). \tag{131}$$

Here $\mathcal{M}_n$ is the space of moduli of n disjoint points on the Riemann sphere. This integral is indicated in figure (5) for the case of five points.

In fact, (131) is not just the string theoretic generalization of the Einstein action, but contains more information. (131) is the string theoretic analogue not just of (117) but of the perturbation series constructed from (117) in defining the 'tree approximation' to general relativity. The physical interpretation of (131) is that it gives the probability for scattering of n particles of type Φ. Unfortunately, to explain the latter remark would require a considerable enlargement of the brief introduction to quantum field theory in section (3).

Now, in (131) $\bar{\Sigma}$ is a Riemann surface of genus zero, because we are led to genus zero in the course of trying to find a string theory analogue of (117). However, in the mathematical sense, (131) has a very natural generalization (figure (6)) with $\bar{\Sigma}$ replaced by a Riemann surface of genus greater than zero. This generalization is of central importance in string theory. I have remarked that if we try to interpret (117) as the Lagrangian of a quantum theory, we run into severe difficulties. The attempt to calculate quantum corrections to the classical theory of general relativity gives rise to infinities. On the other hand, in string theory there is a meaningful and well-defined prescription for calculating quantum corrections to classical answers. One simply replaces $\bar{\Sigma}$ by a Riemann surface Σ_k of genus $k > 0$. Thus, if we want to calculate the probability of scattering of n particles of type Φ, then in string theory the classical answer, valid for $\hbar = 0$, is given by (131). If we want to calculate a quantum correction to (131) of order $\hbar^k$, then we calculate not (131) but

$$\int_{\mathcal{M}_{k,n}} Z(\prod V(P_i), \bar{\Sigma}). \tag{132}$$

Here $\tilde{\Sigma}$ has genus k, and $\mathcal{M}_{k,n}$ is the moduli space of Riemann surfaces of genus k with n marked points. In this way, physicists have become interested in the moduli space of Riemann surfaces and in path integrals over this space. The integrals in (132) actually have remarkably beautiful properties, some of which were described on Tuesday by Yu. Manin. But even if (132) were not beautiful, the fact that it is free of the ultraviolet divergences that plague the analogous formulas in the quantum theory of general relativity would be enough to give it a far-reaching importance in physics.

There are many major gaps in this exposition. A much nicer description can be given if one considers not the space H_S but a certain highest weight cohomology theory of the Virasoro algebra with values in H_S. The cohomology theory in question [3,29,19,8] has been presented at this conference by I. Frenkel, and I will not enter into it here. Also, I have avoided the question of what vertex operators V we are using in (131) and (132). These formulas are actually limited to vertex operators V that transform on the Riemann surface like differential forms of type $(1,1)$; they correspond, under our canonical correspondence between operators and vectors in H_S, to highest weight vectors of the Virasoro algebra. The highest weight cohomology theory that I just mentioned is the proper framework for formulating the quadratic action $S_2(\Phi)$ [27,1,28] and is also very useful in what little we can say about the nonlinear theory which has the perturbative expansion we have discussed [34,15,21].

I have tried to make it plausible that path integrals on Riemann surfaces can be used to formulate a generalization of general relativity. What is more, the resulting generalization is (especially in its supersymmetric forms) free of the ailments that plague quantum general relativity. If the logic has seemed a bit thin, it is at least in part because almost all we know in string theory is a trial and error construction of a perturbative expansion. (131) and (132) are probably the most beautiful formulas that we now know of in string theory, yet these formulas are merely a perturbative expansion (in powers of Φ and $\hbar$) of some underlying structure. Uncovering that structure is a vital problem if ever there was one.

REFERENCES

1. Banks, T., and Peskin, M., *Nucl. Phys.* **B264** (1986) 513.
2. Bismut, J.-P., and Freed, D., 'Analysis of Elliptic Families', preprint (1985).
3. Becchi, C., Rouet, A., and Stora, R., *Annals of Physics* **98** (1974) 287.
4. Candelas, P., Horowitz, G., Strominger, A., and Witten, E., *Nucl. Phys.* **B258** (1985) 46.
5. Appelquist, T., Chodos, A., Freund, P.G.O., eds., *Modern Kaluza-Klein Theory and Its Applications*, to be published by Benjamin-Cummings (1987).
6. Feingold, A., and Lepowsky, J, *Adv. Math.* **29** (1978) 271.
7. Frenkel, I., and Kac, V., *Inv. Math.* **62** (1980) 23.
8. Frenkel, I., Garland, H., and Zuckerman, G., Yale University preprint (1986).
9. Freed, D., MIT preprint (1986).
10. Glimm, J., and Jaffe, A., *Quantum Physics: A Functional Integral Point of View* (Springer-Verlag, 1981).
11. Green, M.B., and Gross, D.J., eds., *Unified String Theories* (World Scientific, Singapore, 1986).
12. Green, M.B., and Schwarz, J.H., *Phys. Lett.* (1984).
13. Green, M.B., Schwarz, J.H., and Witten, E., *Superstring Theory*, to be published by Cambridge University Press (1986).
14. Gross, D.J., Harvey, J.A., Martinec, E., and Rohm, R., *Phys. Rev. Lett.* 54 (1984) 502.
15. Hata, H., Itoh, K., Kugo, T., Kunitomo, H., and Ogawa, K., 'Covariant String Field Theory', Kyoto preprint (1986).
16. Itzykson, C., and Zuber, J.-B. *Quantum Field Theory* (McGraw-Hill, 1980).

17. Jacob, M., ed., *Dual Models* (North-Holland, Amsterdam, 1974).

18. Kac, V., *Infinite Dimensional Lie Algebras* (Birkhauser, 1983).

19. Kato, M., and Ogawa, K., *Nuclear Physics* B212 (1983) 443.

20. Lepowsky, J. and Wilson, R.L., *Comm. Math. Phys.* 62 (1978) 43.

21. Neveu, A., and West, P.C., *Phys. Lett.* 168B (1986) 192.

22. Polyakov, A.M., *Phys. Lett.* 203B (1981) 207.

23. Quillen, D., 'Determinants of Cauchy-Riemann Operators on a Riemann Surface,' (IHES preprint).

24. Scherk, J., and Schwarz, J.H., *Phys. Lett.* 52B (1974) 347.

25. Schwarz, J.H., ed., *Superstrings* (World Scientific, Singapore, 1986).

26. Segal, G., *Comm. Math. Phys.* 80 (1981) 301.

27. Siegel, W., *Phys. Lett.* 151 (1985) 391, 396.

28. Siegel, W., and Zweibach, B., *Nuclear Physics* B263 (1986) 105.

29. Tyupin, I.V., Lebedev preprint FIAN, No. 39 (1975), unpublished.

30. Veneziano, G., *Nuovo Cimento* 57A (1968) 190.

31. Witten, E., in Shelter Island II: Proceedings of the 1983 Shelter Island Conference on Quantum Field Theory and the Fundamental Problems of Physics, ed. N. Khuri et. al. (MIT Press, 1985.

32. Witten, E., in *Anomalies, Geometry, and Topology*, ed. A. White and W. Bardeen (World Scientific, 1985).

33. Witten, E., *Comm. Math. Phys.* 92 (1984) 495.

34. Witten, E., *Nucl. Phys.* B268 (1986) 253.

35. Zee, A., ed., Unity of Forces in the Universe (World Scientific, Singapore, 1982).

FIGURE CAPTIONS

1) A cylinder with a standard, flat metric. The boundary consists of two components S_1 and S_2 of circumference 2π. The 'height' of the cylinder is labeled as β.

2) A cylinder of height $\beta_1 + \beta_2$, made by gluing together cylinders Σ_1 and Σ_2 of height β_1 and β_2, respectively. The boundary of Σ_1 consists of the two circles S_1 and S, and the boundary of Σ_2 consists of the two circles S and S_2.

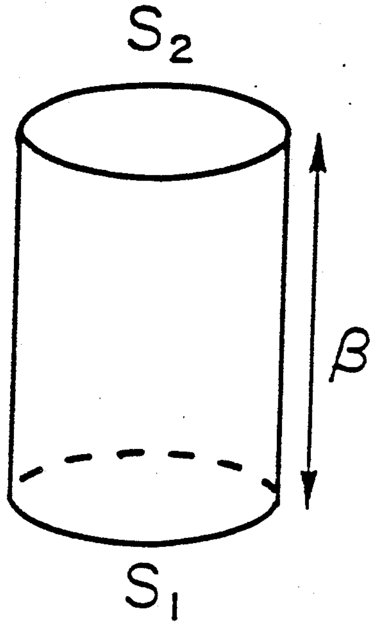
3) A path integral on a cylinder Σ with insertion of a local operator $O(P)$. The boundary of Σ again consists of the two circles S_1 and S_2.

4) A Riemann surface with three boundary components.

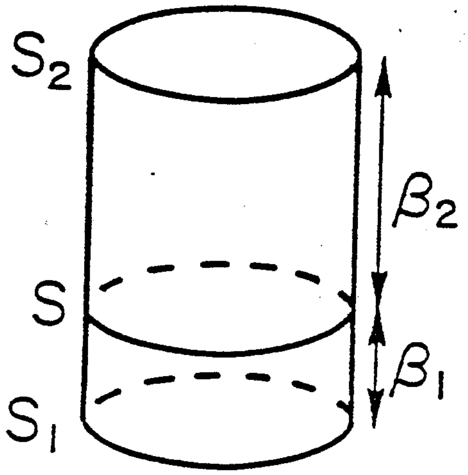
5) Integration over moduli of configurations of several points on the Riemann sphere.

6) A path integral on a Riemann surface of genus two, with insertions of several vertex operators.

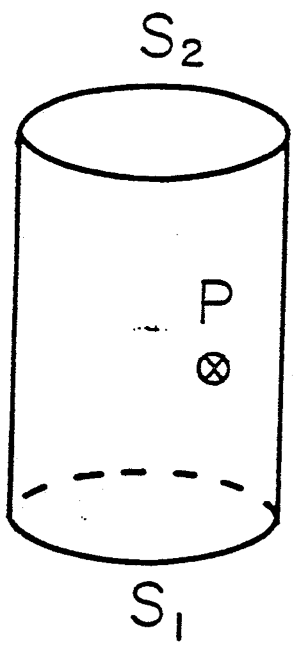
(1)



(2)



(3)



(4)