Rep. Prog. Phys. **50** (1987) 1087–1170. Printed in the UK

Kaluza–Klein theories

David Bailin† and Alex Love‡

† School of Mathematical and Physical Sciences, University of Sussex, Brighton, UK
‡ Department of Physics, R H B New College, University of London, Egham, UK

Abstract

Kaluza–Klein theory is developed starting from the simplest example in which a single extra spatial dimension is compactified to a circle, and a single Abelian gauge field emerges in four dimensions from the higher-dimensional metric. This is generalised to greater dimensionality whence non-Abelian gauge groups may be obtained, and possible mechanisms for achieving the compactification of the extra spatial dimensions are discussed. The spectrum of particles appearing in four dimensions is discussed with particular emphasis on the spectrum of light fermions, and the constraints arising from cancellation of anomalies when explicit higher-dimensional gauge fields are present are studied. Cosmological aspects of these theories are described, including possible mechanisms for cosmological inflation, and relic heavy particles. Finally, an introductory account of Kaluza–Klein supergravity is given leading towards superstring theory.

This review was received in July 1986.

0034-4885/87/091087+84$09.00

Contents

1. Five-dimensional Kaluza–Klein theory

1.1. Introduction

A theory unifying gravitation and electromagnetism using five-dimensional Riemannian geometry (Kaluza 1921, Klein 1926) and its higher-dimensional generalisations to include weak and strong interactions (DeWitt 1964, Kerner 1968, Trautman 1970, Cho 1975, Cho and Freund 1975) have become a focus of attention for many particle physicists in the past few years. This revival of interest in Kaluza-Klein theory stemmed in the first instance from work in string theories (Scherk and Schwarz 1975, Cremmer and Scherk 1977), and then from the usefulness of extra spatial dimensions in the construction of $N=8$ supergravity theory (Cremmer *et al* 1978, Cremmer and Julia 1978). In these contexts, it would have been possible to regard the extra spatial dimensions as a mathematical device. However, the above authors took the potentially more fruitful approach of considering them to be genuine physical dimensions which we do not normally observe because they have compactified down to a very small scale (spontaneous compactification). Extra *temporal* dimensions are undesirable for several reasons. Firstly, there would be tachyons observed in four dimensions. (See the analysis given in § 1.5.) Secondly, there would be closed timelike loops leading to world lines which violate causality. Thirdly, the sign of the Maxwell action would be incorrect. (See the analysis in § 1.4.)

Though the renaissance of Kaluza-Klein theory has received a considerable impetus from the possible relevance to supergravity, many theorists have taken the view that extra *spatial* dimensions may be an ingredient in the unification of all interactions even if supergravity should eventually have to be relinquished. In this review, we shall avoid, as far as possible, reliance on supergravity as a framework for Kaluza-Klein theory. Instead we shall try to emphasise those aspects of such theories which might be of importance for unification of interactions with or without supergravity. Our approach is pedagogical and directed primarily towards readers without an extensive background in supergravity theories. Accordingly, we have been selective in the material included, and consequently in that section of the literature to which we refer. The reader may restore the imbalance against work deeply rooted in supergravity theory by a reading of one of the recent excellent review papers which have approached the subject from that standpoint (Duff *et al* 1986, Englert and Nicolai 1983).

1.2. The five-dimensional theory

Five-dimensional Kaluza-Klein theory (Kaluza 1921, Klein 1926) unifies electromagnetism with gravitation by starting from a theory of Einstein gravity in five dimensions. Thus, the initial theory has five-dimensional general coordinate invariance. However, it is assumed that one of the spatial dimensions compactifies so as to have the geometry of a circle S^1 of very small radius. Then, there is a residual four-dimensional general coordinate invariance, and, as we see in § 1.3, an Abelian gauge invariance associated with transformations of the coordinate of the compact manifold, S^1. Put another way, the original five-dimensional general coordinate invariance is spontaneously broken in the ground state. In this way, we arrive at an ordinary theory of gravity in four dimensions, together with a theory of an Abelian gauge field, with connections between the parameters of the two theories because they both derive from the same initial five-dimensional Einstein gravity theory.

We adopt coordinates $\bar{x}^A$, $A=1,\ldots,5$ with

$$\bar{x}^\mu = x^\mu \qquad \mu=0,1,2,3 \tag{1.1}$$

being coordinates for ordinary four-dimensional spacetime, and

$$\bar{x}^5=\theta \tag{1.2}$$

being an angle to parametrise the compact dimension with the geometry of a circle. The ground-state metric after compactification is

$$\bar{g}^{(0)}_{AB}=\mathrm{diag}\{\eta_{\mu\nu}-\tilde{g}_{55}\} \tag{1.3}$$

where

$$\eta_{\mu\nu}=(1,-1,-1,-1) \tag{1.4}$$

is the metric of Minkowski space, M_4, and

$$\tilde{g}_{55}=\tilde{R}^2 \tag{1.5}$$

is the metric of the compact manifold S^1, where $\tilde{R}$ is the radius of the circle.

The identification of the gauge field arises from an expansion of the metric about the ground state. Quite generally, we may parametrise the metric in the form

$$\bar{g}_{AB}(x,\theta)=\begin{pmatrix} g_{\mu\nu}(x,\theta)-B_\mu(x,\theta)B_\nu(x,\theta)\Phi(x,\theta) & B_\mu(x,\theta)\Phi(x,\theta) \\ B_\nu(x,\theta)\Phi(x,\theta) & -\Phi(x,\theta)\end{pmatrix}. \tag{1.6}$$

To extract the graviton and the Abelian gauge field it proves sufficient to replace $\Phi(x,\theta)$ by its ground-state value $\tilde{g}_{55}$, and to use the ansatz without θ dependence:

$$\bar{g}_{AB}(x)=\begin{pmatrix} g_{\mu\nu}(x)-B_\mu(x)B_\nu(x)\tilde{g}_{55} & B_\mu(x)\tilde{g}_{55} \\ B_\nu(x)\tilde{g}_{55} & -\tilde{g}_{55}\end{pmatrix}. \tag{1.7}$$

We write

$$B_\mu(x)=\xi A_\mu(x) \tag{1.8}$$

where ξ is a scale factor we shall choose later so that $A_\mu(x)$ is a conventionally normalised gauge field.

1.3. Abelian gauge transformations

Coordinate transformations associated with the coordinate θ of the compact manifold may be interpreted as gauge transformations, as we now show. Consider the transformation

$$\theta\to\theta'=\theta+\xi\varepsilon(x). \tag{1.9}$$

For a general coordinate transformation

$$\bar{g}_{AB}=\bar{g}'_{A'B'}\frac{\partial\bar{x}'^{A'}}{\partial\bar{x}^A}\frac{\partial\bar{x}'^{B'}}{\partial\bar{x}^B}. \tag{1.10}$$

For the particular transformation (1.9), the off-diagonal elements of the metric give

$$A_\mu \to A'_\mu = A_\mu + \partial_\mu \varepsilon. \tag{1.11}$$

Thus the transformation (1.9) of the coordinates of the compact manifold induces an Abelian gauge transformation on A_μ. This means that the compact manifold is providing the internal symmetry space for the (Abelian) gauge group, and internal symmetry has now to be interpreted as just another spacetime symmetry, but associated with the extra spatial dimension.

1.4. Effective four-dimensional action

An effective action for the four-dimensional theory may be derived from the action for five-dimensional Einstein gravity

$$\bar{I} = -\frac{1}{16\pi\bar{G}} \int d^5\bar{x} |\det \bar{g}|^{1/2} \bar{\mathbb{R}} \tag{1.12}$$

where $\bar{\mathbb{R}}$ is the five-dimensional curvature scalar, and $\bar{G}$ is the gravitational constant for five dimensions. Substituting the ansatz (1.7) for $\bar{g}_{AB}$, and integrating over the extra spatial coordinate θ, gives an effective four-dimensional action

$$I = -\frac{2\pi\tilde{R}}{16\pi\bar{G}} \int d^4x |\det g|^{1/2} \mathbb{R} - \frac{\xi^2 \tilde{g}_{55}}{4} \frac{2\pi\tilde{R}}{16\pi\bar{G}} \int d^4x |\det g|^{1/2} F_{\mu\nu} F^{\mu\nu} \tag{1.13}$$

where $\tilde{R}$ is the radius of the compact manifold as in (1.5), $\mathbb{R}$ is the four-dimensional curvature scalar, and

$$F_{\mu\nu} \equiv \partial_\mu A_\nu - \partial_\nu A_\mu . \tag{1.14}$$

The four-dimensional gravitational constant G is thus identified as

$$G = \bar{G}/2\pi\tilde{R}. \tag{1.15}$$

To obtain standard normalisation for the gauge field we must then choose

$$\xi^2 = \frac{16\pi G}{\tilde{g}_{55}} = \frac{\kappa^2}{\tilde{R}^2} \tag{1.16}$$

where

$$\kappa^2 \equiv 16\pi G. \tag{1.17}$$

Then the effective four-dimensional action is

$$I = -\frac{1}{16\pi G} \int d^4x |\det g|^{1/2} \mathbb{R} - \tfrac{1}{4} \int d^4x \, F_{\mu\nu} F^{\mu\nu}. \tag{1.18}$$

(Had the extra dimension been timelike, we would have obtained the opposite (wrong) sign for the Maxwell action in (1.13).)

Whether retaining just the massless states in four dimensions in this way is an adequate approximation is discussed in § 2.3.

1.5. Mass eigenstates

The natural scale of mass for these theories is the Planck mass and massive fields in five dimensions will naturally lead in four dimensions to particles with masses on the Planck scale. Suppose instead we start with a massless field in five dimensions. For a five-dimensional scalar field $\phi(x, \theta)$ we may make the Fourier expansion on the compact manifold

$$\phi(x, \theta) = \sum_{n=-\infty}^{\infty} \phi^n(x)\, e^{in\theta}. \tag{1.19}$$

The Klein–Gordon equation

$$\left(\Box_x - \tilde{R}^{-2}\frac{\partial}{\partial\theta^2}\right)\phi(x, \theta) = 0 \tag{1.20}$$

then gives the equations for the Fourier components:

$$(\Box_x + m_n^2)\phi^n(x) = 0 \tag{1.21}$$

where

$$m_n^2 = n^2/\tilde{R}^2. \tag{1.22}$$

The fields $\phi^n(x)$ are thus the mass eigenstates in four dimensions, and the field $\phi^0(x)$ is the only massless one (or perhaps light, after allowing for radiative corrections). The other fields $\phi^n(x)$ have masses of order $\tilde{R}^{-1}$, which we would expect to be comparable to the Planck mass. If the extra dimension had had a timelike signature (positive in our convention), we would have obtained a negative mass squared in (1.22), i.e. tachyons.

1.6. Charge quantisation

If we apply the coordinate transformation

$$\theta \to \theta' = \theta + \xi\varepsilon(x) \tag{1.23}$$

to the field $\phi(x, \theta)$ of (1.19), we have

$$\phi^n(x) \to \exp(in\xi\varepsilon(x))\phi^n(x). \tag{1.24}$$

Since the Abelian gauge field transforms (according to (1.11)) in the manner

$$A_\mu \to A'_\mu = A_\mu + \partial_\mu\varepsilon \tag{1.25}$$

this means that $\phi^n(x)$ has charge

$$q_n = -n\xi = -n\kappa/\tilde{R} \tag{1.26}$$

where we have used the normalisation condition (1.16). Thus, charge is quantised in units of $\kappa/\tilde{R}$. The radius of the compact manifold may now be estimated from

$$\tilde{R}^2 = \kappa^2/e^2 = 4G/(e^2/4\pi). \tag{1.27}$$

Thus, identifying e with the quantum of electric charge

$$\tilde{R} \sim 10 m_{\mathrm{P}}^{-1}. \tag{1.28}$$

There is however a flaw in the five-dimensional Kaluza–Klein theory, even if we do not wish to include weak and strong interactions as well as electromagnetism. From (1.26), we see that all charged particles have $n \neq 0$. But, from (1.22), this means that they all have masses on the Planck scale m_{P}, whereas the familiar charged particles have very small masses on that scale.

1.7. *Scalar field from metric*

Starting from the general parametrisation of the metric (1.6), it is possible to extract a massless scalar field (Freund 1982, Appelquist and Chodos 1983a, b). Deleting θ dependence, which is associated with massive degrees of freedom, the graviton–scalar sector may be obtained from

$$\bar{g}_{AB}(x)=\begin{pmatrix} g_{\mu\nu}(x) & 0 \\ 0 & -\Phi(x)\end{pmatrix}. \tag{1.29}$$

Substituting in the action (1.12) for five-dimensional gravity leads to the effective four-dimensional action

$$I=-\frac{2\pi}{16\pi\bar{G}}\int d^4x\,|\det g|^{1/2}\Phi^{1/2}\mathbb{R}. \tag{1.30}$$

The $\Phi^{1/2}$ multiplying the four-dimensional curvature scalar may be removed by a Weyl scaling,

$$\begin{aligned} g_{\mu\nu} &\to g_{\mu\nu}\Phi^{-1/3} \\ \Phi &\to \Phi^{2/3}. \end{aligned} \tag{1.31}$$

Then,

$$I=-\frac{1}{16\pi G}\int d^4x|\det g|^{1/2}\{\mathbb{R}-\tfrac{1}{6}\Phi^{-2}\partial^{\mu}\Phi\partial_{\mu}\Phi\} \tag{1.32}$$

or, with the further change of variables

$$x=\frac{\kappa^{-1}}{\sqrt{3}}\ln(\Phi/\Phi_0) \tag{1.33}$$

we have

$$I=-\frac{1}{16\pi G}\int d^4x|\det g|^{1/2}\mathbb{R}+\frac{1}{2}\int d^4x|\det g|^{1/2}\,\partial^{\mu}x\,\partial_{\mu}x. \tag{1.34}$$

This is a particular case of four-dimensional Brans–Dicke theory (Jordan 1959, Brans and Dicke 1961).

2. (4 + *D*)-dimensional Kaluza–Klein theories

2.1. *Isometry group of a manifold*

In order to unify gravitation, not just with electromagnetism but also with weak and strong interactions, it is necessary to generalise the five-dimensional theory of § 1 to a higher-dimensional theory (Klein 1926, DeWitt 1964, Kerner 1968, Trautman 1970, Cho 1975, Cho and Freund 1975, Scherk and Schwarz 1975, Cremmer and Scherk 1977) so as to obtain a non-Abelian gauge group. In the five-dimensional case, an Abelian gauge group arose from the coordinate transformation

$$\theta\to\theta'=\theta+\xi\varepsilon(x) \tag{2.1}$$

on the single coordinate θ of the compact manifold. In the $(4+D)$-dimensional case we must look for symmetries of the compact manifold which generalise (2.1). The appropriate transformations to study are the isometries of the manifold (an introductory discussion of isometries is to be found in Weinberg 1972, ch 13). Let us denote the coordinates of ordinary four-dimensional space by x^μ, and the coordinates of the compact manifold K by y^n. An isometry of K is a coordinate transformation $y \to y'$ which leaves the *form* of the metric $\tilde{g}_{mn}$ for K invariant:

$$y \to y' : \tilde{g}'_{mn}(y') = \tilde{g}_{mn}(y'). \tag{2.2}$$

Isometries form a group, with generators t_a and structure constants C_{abc}, in the following way. The general infinitesimal isometry is

$$I + \mathrm{i}\varepsilon^a t_a : y^n \to y^{n'} = y^n + \varepsilon^a \xi^n_a(y) \tag{2.3}$$

where the infinitesimal parameters ε^a are independent of y, and the Killing vectors ξ^n_a, which are associated with the *independent* infinitesimal isometries, obey the algebra

$$\xi^m_b \partial_m \xi^n_c - \xi^m_c \partial_m \xi^n_b = -C_{abc} \xi^n_a. \tag{2.4}$$

Correspondingly by considering the commutator of two infinitesimal isometries, we can show that

$$[t_a, t_b] = \mathrm{i} C_{abc} t_c. \tag{2.5}$$

For instance, the N-dimensional sphere S^N has isometry group SO$(N+1)$, and the $2N$(real)-dimensional complex projective plane CP^N has isometry group SU$(N+1)$. The isometry group for the compact manifold S^1 of the five-dimensional theory is just the SO(2) (or U(1)) group of transformations of (2.1). As we shall discuss later, it is possible to choose the compact manifold to obtain the isometry group SU(3) $\times$ SU(2) $\times$ U(1), which is the (observed) gauge group of electroweak and strong interactions.

2.2. *Non-Abelian gauge transformations*

The ground-state metric for the compactified $(4+D)$-dimensional theory may be written as

$$\bar{g}_{AB} = \mathrm{diag}^{(0)} \{\eta_{\mu\nu}, -\tilde{g}_{mn}(y)\} \tag{2.6}$$

where $\eta_{\mu\nu}$ is the metric of Minkowski space M_4 as in (1.4), and $\tilde{g}_{mn}(y)$ is the metric of the compact manifold. We return in § 3 to discuss whether such a ground state does exist. The non-Abelian gauge fields of the theory may be displayed by the expansion about the ground state

$$\bar{g}_{AB} = \begin{pmatrix} g_{\mu\nu}(x) - \tilde{g}_{mn}(y) B^m_\mu B^n_\nu & B^n_\mu \\ B^m_\nu & -\tilde{g}_{mn}(y) \end{pmatrix} \tag{2.7}$$

with

$$B^n_\mu \equiv \xi^n_a(y) A^a_\mu(x). \tag{2.8}$$

(A more general ansatz including x dependence for $\tilde{g}_{mn}$ and y dependence for $g_{\mu\nu}$ and A_μ is necessary to display any massless scalars arising from the metric and massive states.)

Non-Abelian gauge transformations arise by considering the effect on the components $\bar{g}_{\mu n}$ of the metric of the infinitesimal isometry with x-dependent parameters:

$$y^n \to y^n + \xi^n_a(y) \varepsilon^a(x). \tag{2.9}$$

We then find that

$$A_\mu^a \to A_\mu^{a'} = A_\mu^a + \partial_\mu \varepsilon^a(x) + C_{abc}\varepsilon^b(x) A_\mu^c \tag{2.10}$$

which is just the usual Yang-Mills gauge transformation if we display the gauge coupling constant g explicitly by writing

$$C_{abc} = g f_{abc} \tag{2.11}$$

and

$$t_a = g T_a \tag{2.12}$$

so that

$$[T_a, T_b] = \mathrm{i} f_{abc} T_c. \tag{2.13}$$

Thus, non-Abelian gauge transformations are generated by x-dependent infinitesimal isometries of the compact manifold K.

2.3. *Effective four-dimensional action*

The action for Einstein gravity in $4+D$ dimensions is

$$\bar{I} = -\frac{1}{16\pi\bar{G}} \int \mathrm{d}^{4+D}\,\bar{x} |\det \bar{g}|^{1/2} \bar{\mathbb{R}} \tag{2.14}$$

where $\bar{\mathbb{R}}$ is the $(4+D)$-dimensional curvature scalar, and $\bar{G}$ is the gravitational constant for $4+D$ dimensions. Substituting the ansatz (2.7) for $\bar{g}_{AB}$, and integrating over the compact degrees of freedom y gives an effective four-dimensional action

$$\begin{aligned} I = &-\left(\int \mathrm{d}^D y |\det \tilde{g}|^{1/2}\right)(16\pi\bar{G})^{-1} \int \mathrm{d}^4 x |\det g|^{1/2} \mathbb{R} \\ &-\left(\int \mathrm{d}^D y |\det \tilde{g}|^{1/2} \xi_a^m(y) \xi_b^n(y) \tilde{g}_{mn}(y)\right) \\ &\times (16\pi\bar{G})^{-1} \frac{1}{4} \int \mathrm{d}^4 x |\det g|^{1/2} F_{\mu\nu}{}^a (F^{\mu\nu})^b \end{aligned} \tag{2.15}$$

with

$$F_{\mu\nu}{}^a \equiv \partial_\mu A_\nu^a - \partial_\nu A_\mu^a - C_{abc} A_\mu{}^b A_\nu{}^c \tag{2.16}$$

and $\mathbb{R}$ denoting the four-dimensional curvature scalar. The four-dimensional gravitational constant G is thus identified by

$$\kappa^{-2} \equiv (16\pi G)^{-1} = (16\pi\bar{G})^{-1} \int \mathrm{d}^D y |\det \tilde{g}(y)|^{1/2} \tag{2.17}$$

and standard normalisation of the gauge fields requires the Killing vectors to be scaled so that

$$\langle \xi_a^m(y) \xi_b^n(y) \tilde{g}_{mn}(y) \rangle = \kappa^2 \delta_{ab} \tag{2.18}$$

where we have introduced the notation (of Weinberg 1983)

$$\langle f(y) \rangle \equiv \frac{\int \mathrm{d}^D y |\det \tilde{g}|^{1/2} f(y)}{\int \mathrm{d}^D y |\det \tilde{g}|^{1/2}}. \tag{2.19}$$

Then we have the standard action for Einstein gravity plus non-Abelian gauge fields in four dimensions:

$$I=-(16\pi G)^{-1}\int \mathrm{d}^4x|\det g|^{1/2}\mathbb{R}-\frac{1}{4}\int \mathrm{d}^4x|\det g|^{1/2}F_{\mu\nu}{}^a(F^{\mu\nu})^a. \tag{2.20}$$

The discussion given above has to be modified if the higher-dimensional theory one starts from is supergravity rather than ordinary Einstein gravity. For instance, there is present in eleven-dimensional supergravity a third-rank antisymmetric tensor field. (See, for example, the review articles of Englert and Nicolai (1983) and Duff *et al* 1986.) In an expansion of this field about its expectation value, terms proportional to the gauge fields appear (Duff *et al* 1983c, 1984a) and in four dimensions there is an additional contribution for the Yang-Mills Lagrangian. Then, the normalisation of the Killing vectors in terms of the gravitational constant differs from (2.18). This amounts to a modification of the relationship between the gauge coupling constant and the gravitational constant, as can be seen from § 5.

A further subtlety, which has been emphasised by Duff *et al* (1984b) and Duff and Pope (1984), is that simply retaining the massless states in the four-dimensional theory after compactification may give field equations whose solutions are not exact solutions of the full $(4+D)$-dimensional field equations. (This includes the five-dimensional case.) However, it has been argued by Witten and Weinberg (see Duff and Pope 1984) that the errors incurred are suppressed at energy scale E by powers of E/E_0, where E_0 is the compactification scale, *provided* the four-dimensional ground state is Minkowski space.

2.4. Graviton-scalar sector

The graviton-scalar action may be derived by making the ansatz

$$\bar{g}_{AB}(x,y)=\begin{pmatrix} g_{\mu\nu}(x) & 0 \\ 0 & -\Phi_{mn}(x,y)\end{pmatrix} \tag{2.21}$$

which is more general than (2.7), in that it allows x dependence for $\tilde{g}_{mn}$. Substitution of this ansatz in the $(4+D)$-dimensional gravitational action (2.14) yields (Cho and Freund 1975)

$$\begin{aligned}\bar{I}=-(16\pi G)^{-1}&\int \mathrm{d}^4x\int \mathrm{d}^Dy|\det g_{\mu\nu}|^{1/2}|\det\Phi_{mn}|^{1/2}\\ &\times\{\mathbb{R}+\tilde{\mathbb{R}}-\Phi^{mn}D_\mu D^\mu\Phi_{mn}-\tfrac{1}{2}D^\mu\Phi_{mn}D^\mu\Phi^{mn}\\ &-\tfrac{1}{4}\Phi^{mn}D^\mu\Phi_{mn}\Phi^{pq}D_\mu\Phi_{pq}+\tfrac{1}{4}\Phi^{mn}\Phi^{pq}D_\mu\Phi_{pm}D^\mu\Phi_{qn}\}\end{aligned} \tag{2.22}$$

where $\mathbb{R}$ and $\tilde{\mathbb{R}}$ are the curvature scalars for 4 and D dimensions, respectively. The integration over the compact manifold $\int \mathrm{d}^Dy$ may be performed given the metric.

2.5. Differential geometry

The modern formulation of differential geometry, in terms of differential forms, has great calculational and notational advantages for studying the compact manifold of Kaluza-Klein theory. Accordingly, we give here a brief resumé of the more important definitions and results. For a more extensive and thorough discussion see, for instance, Eguchi *et al* (1980) and Salam and Strathdee (1982). Differential forms are a convenient formalism for manipulating totally antisymmetric tensors. An antisymmetric product of coordinate differentials (the exterior product $\wedge$) is defined so that

$$\mathrm{d}y^m \wedge \mathrm{d}y^n = -(\mathrm{d}y^n \wedge \mathrm{d}y^m) \tag{2.23}$$

and a p-form ω is constructed from any pth-rank antisymmetric tensor $\omega_{m_1 \ldots m_p}$:

$$\omega \equiv \omega_{m_1 m_2 \ldots m_p} \mathrm{d}y^{m_1} \wedge \mathrm{d}y^{m_2} \wedge \ldots \wedge \mathrm{d}y^{m_p}. \tag{2.24}$$

Then for a p-form α_p and a q-form β_q

$$\alpha_p \wedge \beta_q = (-1)^{pq} \beta_q \wedge \alpha_p. \tag{2.25}$$

A totally antisymmetric differentiation, the exterior derivative d, is defined by

$$\mathrm{d}\omega \equiv \partial_{m_{p+1}} \omega_{m_1 \ldots m_p} \, \mathrm{d}y^{m_{p+1}} \wedge \mathrm{d}y^{m_1} \wedge \ldots \wedge \mathrm{d}y^{m_p}. \tag{2.26}$$

There follows immediately the important property

$$\mathrm{d}\mathrm{d}\omega = 0 \tag{2.27}$$

and it is easy to show that

$$\mathrm{d}(\alpha_p \wedge \beta_q) = \mathrm{d}\alpha_p \wedge \beta_q + (-1)^p (\alpha_p \wedge \mathrm{d}\beta_q). \tag{2.28}$$

An important theorem which generalises Gauss's Law and Stokes's theorem is that

$$\int_M \mathrm{d}\omega_{p-1} = \int_{\partial M} \omega_{p-1} \tag{2.29}$$

where M is a p-dimensional manifold and ∂M is its boundary.

It is often convenient to work with the vielbein $e_m^\alpha(y)$ for the compact manifold K rather than its metric $\tilde{g}_{mn}(y)$. If the tangent space to the manifold is a Euclidean space so that we may choose its metric to be $\delta_{\alpha\beta}$, then the vielbein is defined by

$$\tilde{g}_{mn}(y) = \delta_{\alpha\beta} e_m^\alpha(y) e_n^\beta(y). \tag{2.30}$$

(We shall refer to α, β, ... as flat indices, or tangent space indices, and m, n, ... as curved indices.) The vielbein has inverse $e_\alpha^m(y)$ given by

$$e_\alpha^m(y) = \delta_{\alpha\beta} \tilde{g}^{mn}(y) e_n^\beta(y) \tag{2.31}$$

so that

$$e_\alpha^m e_m^\beta = \delta_\alpha^\beta \tag{2.32}$$

and

$$\delta^{\alpha\beta} e_\alpha^m e_\beta^n = \tilde{g}^{mn}. \tag{2.33}$$

Flat indices may be converted into curved indices, and vice versa, using the vielbein and its inverse.

The vielbein 1-form

$$e^\alpha(y) = e_m^\alpha(y) \, \mathrm{d}y^m \tag{2.34}$$

may be used to construct the spin-connection 1-form $\omega^\alpha{}_\beta$, and the torsion and curvature 2-forms T^α and $R^\alpha{}_\beta$:

$$\mathrm{d}e^\alpha + \omega^\alpha{}_\beta \wedge e^\beta = T^\alpha \tag{2.35}$$

and

$$\begin{aligned} R^\alpha{}_\beta &= \mathrm{d}\omega^\alpha{}_\beta + \omega^\alpha{}_\gamma \wedge \omega^\gamma{}_\beta \\ &\equiv \tfrac{1}{2} R^\alpha{}_{\beta\gamma\delta} (e^\gamma \wedge e^\delta). \end{aligned} \tag{2.36}$$

(We shall normally deal with manifolds which admit a torsion-free metric, $T^\alpha = 0$.) The usual curvature tensor $R^m{}_{npq}$ is obtained by using the vielbein to convert flat indices into curved indices. The relationship of the spin connection as defined by (2.35) to the usual definition (as, for example, in Weinberg 1972, § 12.5) may be read off from the covariant derivative $D_m\psi$ of a field ψ transforming as the representation $M^{\alpha\beta}$ of the SO(D) tangent space group of the compact manifold K:

$$D_m\psi = (\partial_m - \mathrm{i}\omega_m)\psi \tag{2.37}$$

where

$$\omega^\alpha{}_\beta \equiv (\omega^\alpha{}_\beta)_m \mathrm{d}y^m \tag{2.38}$$

and

$$\omega_m \equiv \tfrac{1}{2}(\omega_{\alpha\beta})_m M^{\alpha\beta}. \tag{2.39}$$

As a 1-form,

$$D\psi = (\mathrm{d} - \mathrm{i}\omega)\psi. \tag{2.40}$$

When the compact manifold K is a coset space G/H, the vielbein may be constructed as follows. The points y^m of K may be represented by chosen elements $L(y)$ of G, one from each coset. Let the Hermitian generators of G be $Q_{\hat\alpha}$ with commutation relations

$$[Q_{\hat\alpha}, Q_{\hat\beta}) = \mathrm{i}f_{\hat\alpha\hat\beta\hat\gamma}Q_{\hat\gamma}. \tag{2.41}$$

The object

$$e(y) \equiv L^{-1}(y)\,\mathrm{d}L(y) \tag{2.42}$$

may be written in the form

$$\begin{aligned} e(y) &= e^{\hat\alpha}(y)\mathrm{i}Q_{\hat\alpha} \\ &= e^\alpha(y)\mathrm{i}Q_\alpha + e^{\bar\alpha}(y)\mathrm{i}Q_{\bar\alpha} \end{aligned} \tag{2.43}$$

where the generators $Q_{\hat\alpha}$ have been divided into those, $Q_{\bar\alpha}$, which belong to H, and the others Q_α which are associated with the tangent space. Then, $e^\alpha(y)$ is the vielbein 1-form for G/H as in (2.34).

The spin connection for a coset space is conveniently calculated from the structure constants of the group G using the Maurer–Cartan formula

$$\mathrm{d}e(y) + e(y) \wedge e(y) = 0 \tag{2.44}$$

which follows from (2.42) and

$$\mathrm{dd}L(y). \tag{2.45}$$

In terms of the structure constants of G this is

$$\mathrm{d}e^{\hat\alpha}(y) = \tfrac{1}{2}f_{\hat\alpha\hat\beta\hat\gamma}(e^{\hat\beta}(y) \wedge e^{\hat\gamma}(y)). \tag{2.46}$$

In particular.

$$\mathrm{d}e^\alpha(y) = \tfrac{1}{2}f_{\alpha\gamma\beta}(e^\gamma \wedge e^\beta) + f_{\alpha\bar\gamma\beta}(e^{\bar\gamma} \wedge e^\beta). \tag{2.47}$$

Comparing with (2.35) *for zero torsion* we identify the spin connection

$$\omega^\alpha{}_\beta = \tfrac{1}{2}f_{\alpha\gamma\beta}e^\gamma + f_{\alpha\bar\gamma\beta}e^{\bar\gamma}. \tag{2.48}$$

In particular for the important case of a symmetric coset space where

$$f_{\alpha\beta\gamma}=0 \tag{2.49}$$

we have the simple form

$$\omega^{\alpha}{}_{\beta}=f_{\alpha\bar{\gamma}\beta}e^{\bar{\gamma}}. \tag{2.50}$$

It is important in applications to know that the generators $Q_{\bar{\alpha}}$ of H may be embedded in the tangent space group SO(D) for K. The reasoning is as follows. A left translation $y\to y'$ may be defined on K by

$$gL(y)=L(y')h \tag{2.51}$$

where g is an arbitrary element of G, and h is a suitable element of H. When g is independent of position on the manifold, it may be deduced from (2.42) (Salam and Strathdee 1982) that under left translations

$$e^{\alpha}(y')=e^{\beta}(y)D_{\beta}{}^{\alpha}(h^{-1}) \tag{2.52}$$

where $D_{\alpha}{}^{\beta}$ is a matrix of the adjoint representation of G, defined by

$$g^{-1}Q_{\hat{\alpha}}g=D_{\hat{\alpha}}{}^{\hat{\beta}}(g)Q_{\hat{\beta}}. \tag{2.53}$$

Equation (2.52) specifies an embedding of H in the tangent space group SO(D). In terms of the matrix elements of the adjoint representation of G,

$$(Q_{\bar{\gamma}})_{\alpha\beta}=-\mathrm{i}C_{\bar{\gamma}\alpha\beta}. \tag{2.54}$$

But the standard generators of SO(D) are

$$(M^{\alpha\beta})_{\gamma\delta}=-\mathrm{i}(\delta^{\alpha}_{\gamma}\delta^{\beta}_{\delta}-\delta^{\alpha}_{\delta}\delta^{\beta}_{\gamma}). \tag{2.55}$$

Thus the embedding is

$$Q_{\bar{\gamma}}=-\tfrac{1}{2}C_{\alpha\bar{\gamma}\beta}M^{\alpha\beta}. \tag{2.56}$$

2.6. The manifolds M^{pqr}

To obtain the known gauge group SU(3)×SU(2)×U(1) of strong and electroweak interactions as (a subgroup of) the isometry group of the compact manifold K, it is necessary to have K at least seven dimensional. This is because (Witten 1981) the lowest-dimensional manifold with isometry group G is a coset space G/H with H a maximal subgroup of G, but with none of the factors in H identical to any factor in G. Thus, for

$$\mathrm{G}=\mathrm{SU}(3)\times\mathrm{SU}(2)\times\mathrm{U}(1) \tag{2.57}$$

we must choose the maximal subgroup

$$\mathrm{H}=\mathrm{SU}(2)\times\mathrm{U}'(1)\times\mathrm{U}''(1) \tag{2.58}$$

leading to the manifold

$$K=\mathrm{G}/\mathrm{H} \tag{2.59}$$

of dimensionality $12-5=7$.

Almost the most general coset space of this type, denoted M^{pqr} by Witten, may be constructed as follows. Denote the generators of SU(3), SU(2) and U(1) by $\frac{1}{2}\lambda_a$, $a=1,\ldots,8$, $\frac{1}{2}\sigma_{\alpha}$, $\alpha=1,2,3$, and Y. It is necessary (without loss of generality) to select two U(1) factors commuting with the SU(2) isospin subgroup of SU(3) to be in H. In other words, it is necessary to select one combination of $\frac{1}{2}\lambda_8$, $\frac{1}{2}\sigma_3$ and Y which does *not* occur in H (and so, as in § 2.5, is associated with the tangent space). Let this combination be

$$Z \equiv \tfrac{1}{2}(\sqrt{3}p\lambda_8 + q\sigma_3 + 2rY) \tag{2.60}$$

with p, q and r arbitrary integers to give a compact U(1). Then, the two combinations which lie in H may be taken to be the orthogonal combinations

$$Z' = \tfrac{1}{2}[2\sqrt{3}pr\lambda_8 + 2qr\sigma_3 - 2(3p^2 + q^2)\,Y] \tag{2.61}$$

and

$$Z'' = \tfrac{1}{2}(-\sqrt{3}q\lambda_8 + 3p\sigma_3). \tag{2.62}$$

This completes the construction of K.

Generally, the isometry group for M^{pqr} is SU(3) × SU(2) × U(1). However, for the exceptional cases ($p = 1$, $q = 0$, $r = 1$) and ($p = 0$, $q = 1$, $r = 1$) the manifolds are $S^5 \times S^2$ and $CP^2 \times S^3$ with the larger isometry groups SO(6) × SO(3) and SU(3) × SO(4), respectively (see § 2.1).

The manifolds M^{pqr} are not quite the most general (orientable) seven-dimensional manifolds with isometry group at least SU(3) × SU(2) × U(1), because it is *not* necessary to chose Z, Z' and Z'' orthogonal. All that is required is that they should be independent (Randjbar-Daemi *et al* 1984a). Thus, we may choose Z as in (2.60), but take the two U(1) factors in H to be

$$X' = \tfrac{1}{2}\sqrt{3}\,\lambda_8 + sY \tag{2.63}$$

and

$$X'' = \tfrac{1}{2}\,\sigma_3 + tY \tag{2.64}$$

where s and t are free parameters, subject only to the constraint

$$ps + qt - r \neq 0. \tag{2.65}$$

This slightly more general class of manifolds with isometry group SU(3) × SU(2) × U(1) is labelled M^{pqrst} by Randjbar-Daemi *et al* (1984a).

3. Compactification mechanisms

3.1. General considerations

For a Kaluza–Klein theory to be able to describe the observed four-dimensional world it is necessary for the extra spatial dimensions to be compactified (except possibly in the early universe) down to a size which we do not probe in particle physics experiments (e.g. the Planck length). As has been emphasised by Appelquist and Chodos (1983a, b), a basic difference between five-dimensional Kaluza–Klein theory and $(4+D)$-dimensional theory (with $D > 1$) is that the five-dimensional gravitational field equations have a compactified classical solution of the form (1.3), but, in general, the $(4+D)$-dimensional equations do *not* have a compactified classical solution of the form (2.6). Thus, in the $(4+D)$-dimensional theory, compactification of the D extra spatial dimensions requires that either matter fields are introduced to provide an energy-momentum tensor, or that the $(4+D)$-dimensional gravitational action differs from the minimal Einstein action. We shall mostly discuss the former possibility, but return to the latter possibility in § 3.8.

When compactification is due to matter fields, an energy-momentum tensor $\bar{T}_{AB}$ must be introduced on the right-hand side of the $(4+D)$-dimensional gravitational field equations:

$$\bar{\mathbb{R}}_{AB}-\tfrac{1}{2}(\bar{\mathbb{R}}+\bar{\Lambda})\bar{g}_{AB}=8\pi\bar{G}\bar{T}_{AB} \tag{3.1}$$

where for generality a $(4+D)$-dimensional cosmological constant $\bar{\Lambda}$ has been included. (The following argument follows closely Randjbar-Daemi *et al* (1983b).) If we demand compactification into $M_4\times K$, where M_4 is four-dimensional Minkowski space, and K is the D-dimensional compact manifold, then

$$\bar{\mathbb{R}}_{\mu\nu}=0 \qquad \mu,\nu=0,1,2,3 \tag{3.2}$$

and because of Lorentz invariance the 4-space components of the energy-momentum tensor are of the form

$$\bar{T}_{\mu\nu}=\frac{c}{8\pi\bar{G}}\,\eta_{\mu\nu} \tag{3.3}$$

where $\eta_{\mu\nu}$ is the Minkowski metric of (1.4). If the compact manifold is an Einstein space we may also write

$$\bar{\mathbb{R}}_{mn}=-2\tilde{k}\bar{g}_{mn} \qquad \tilde{k}>0 \tag{3.4}$$

and the components of the energy-momentum tensor on the compact manifold are of the form

$$\bar{T}_{mn}=\frac{\tilde{c}}{8\pi\bar{G}}\,\bar{g}_{mn}. \tag{3.5}$$

It then follows from the field equations (3.1) that

$$\bar{\mathbb{R}}_{mn}=(\tilde{c}-c)\bar{g}_{mn} \tag{3.6}$$

$$\bar{\mathbb{R}}=\bar{g}^{mn}\bar{\mathbb{R}}_{mn}=D(\tilde{c}-c) \tag{3.7}$$

and

$$\bar{\Lambda}=-D\tilde{c}+(D-2)c. \tag{3.8}$$

Equation (3.7) shows that compactification occurs provided

$$\tilde{c}-c<0. \tag{3.9}$$

3.2. Freund-Rubin compactification

A mechanism for compactification which arises naturally in eleven-dimensional supergravity (see § 7.2) has been particularly explored by Freund and Rubin (1980). This mechanism depends on a third-rank antisymmetric tensor field A_{BCD} with field strength

$$F_{ABCD}\equiv\partial_A A_{BCD}-\partial_B A_{ACD}+\partial_C A_{ABD}-\partial_D A_{ABC} \tag{3.10}$$

and action

$$\bar{I}=\int \mathrm{d}^{4+D}x\,|\det\bar{g}|^{1/2}(-\tfrac{1}{48}F_{ABCD}F^{ABCD}) \tag{3.11}$$

(apart from couplings to fermions, and trilinear self-couplings which do not contribute for the type of solution considered here). Then, the field equation for F_{ABCD} is

$$|\det \bar{g}|^{-1/2}\bar{\partial}_A(|\det \bar{g}|^{1/2}F^{ABCD})=0 \tag{3.12}$$

and the energy-momentum tensor is

$$\bar{T}_{AB}=-\tfrac{1}{6}(F_{CDEA}F^{CDE}{}_B-\tfrac{1}{8}F_{CDEF}F^{CDEF}\bar{g}_{AB}). \tag{3.13}$$

The field equation (3.12) has a solution of the type

$$F^{\mu\nu\rho\sigma}=|\det g|^{-1/2}\varepsilon^{\mu\nu\rho\sigma}F \qquad \mu,\ \nu,\ \rho,\ \sigma=0,1,2,3 \tag{3.14}$$

and all other entries zero, where det g refers to four-dimensional space, F is a constant and the Levi-Civita symbol is defined such that

$$\varepsilon^{0123}=1. \tag{3.15}$$

For such a solution the energy-momentum tensor takes the form

$$\bar{T}_{\mu\nu}=-\frac{F^2}{2}\frac{\det g}{|\det g|}\bar{g}_{\mu\nu} \tag{3.16}$$

and

$$\bar{T}_{mn}=\frac{F^2}{2}\frac{\det g}{|\det g|}\bar{g}_{mn} \tag{3.17}$$

where m, n run over the compact manifold. If we look for a compactification with Minkowski four-dimensional space, then in the notation of (3.3) and (3.5),

$$\frac{c}{8\pi\bar{G}}=-\frac{\tilde{c}}{8\pi\bar{G}}=-\frac{F^2}{2}\frac{\det g}{|\det g|}=\frac{F^2}{2}. \tag{3.18}$$

Thus (3.9) is satisfied and compactification occurs. From (3.8), consistency requires a $(4+D)$-dimensional cosmological constant

$$\bar{\Lambda}=8\pi\bar{G}(D-1)F^2. \tag{3.19}$$

However, in the case of eleven-dimensional supergravity, an eleven-dimensional cosmological constant in the theory destroys the supersymmetry. In that case, one must set $\bar{\Lambda}$ to zero, and one then finds

$$\mathbb{R}=8\pi\bar{G}F^2 4(D-1)/(D+2) \tag{3.20}$$

and

$$\tilde{\mathbb{R}}=-8\pi\bar{G}F^2 3D/(D+2) \tag{3.21}$$

where $\mathbb{R}$ and $\tilde{\mathbb{R}}$ denote the four- and D-dimensional curvature scalars, respectively. Thus, for eleven-dimensional supergravity, there is the unwelcome feature of an anti-de Sitter four-dimensional space!

3.3. Quantum fluctuations in massless higher-dimensional fields

An alternative compactification mechanism (Candelas and Weinberg 1984, Weinberg 1983) is through quantum fluctuations of massless fields in the $(4+D)$-dimensional theory. As these authors observe, for it to make sense to consider quantum fluctuations in light matter fields, while neglecting gravitational quantum connections, it is necessary for the number of light matter fields to be large. By dimensional analysis, when the matter fields are massless in $(4+D)$-dimensions, we may write the matter field action as

$$\bar{I} = -\int d^4x |g|^{1/2} V(\tilde{R}) \tag{3.22}$$

where $\tilde{R}$ is the radius of the compact manifold and

$$V(\tilde{R}) = C_D \tilde{R}^{-4} \tag{3.23}$$

with C_D a constant dependent on the number of matter fields. The energy-momentum tensor may be computed using

$$\delta\bar{I} = -\frac{1}{2}\int d^{4+D}x |\det \bar{g}|^{1/2} \bar{T}^{AB} \delta\bar{g}_{AB}. \tag{3.24}$$

Thus,

$$\bar{T}_{\mu\nu} = \Omega_D^{-1} V(\tilde{R}) \eta_{\mu\nu} = \Omega_D^{-1} C_D \tilde{R}^{-4} \eta_{\mu\nu} \tag{3.25}$$

and

$$\bar{T}_{mn} = \Omega_D^{-1} D^{-1} \tilde{R} (dV/d\tilde{R}) \bar{g}_{mn} = -4\Omega_D^{-1} D^{-1} C_D \tilde{R}^{-4} \bar{g}_{mn} \tag{3.26}$$

with

$$\Omega_D \equiv \int d^D y |\det \tilde{g}|^{1/2}. \tag{3.27}$$

Also, for an Einstein space,

$$\bar{R}_{mn} = -2\tilde{k}\tilde{R}^{-2} \bar{g}_{mn} \tag{3.28}$$

with $\tilde{k} > 0$ for a compact manifold. Demanding that four-dimensional space be Minkowski, the field equations (3.1) yield

$$\frac{2\tilde{k}}{8\pi G} \tilde{R}^{-2} = V(\tilde{R}) - D^{-1} \tilde{R} \frac{dV}{d\tilde{R}} \tag{3.29}$$

and with $V(\tilde{R})$ as in (3.23),

$$\tilde{R}^2 = \frac{8\pi G(D+4)}{2\tilde{k}D} C_D \tag{3.30}$$

and

$$\bar{\Lambda} = 8\pi G(D+2) C_D \tilde{R}^{-4} \tag{3.31}$$

where

$$G = \bar{G}/\Omega_D \tag{3.32}$$

as in (2.17).

It can be seen from (3.30) that for compactification to occur C_D must be positive. (Alternatively, this follows from (3.25), (3.26) and (3.9).) For any given manifold, C_D may be evaluated by calculating the eigenvalues of the Klein–Gordon or Dirac operator on the compact manifold (for massless scalar or spinor fields in $4+D$ dimensions, respectively) to obtain the masses of the 'tower' of four-dimensional fields (see §§ 1.5

and 5.1), and then calculating the effective potential in the usual way (Coleman and Weinberg 1973).

A similar discussion has been given earlier for five-dimensional Kaluza-Klein theory (Appelquist and Chodos 1983a, b). In this case $\tilde{k}$ is zero, because S^1 is flat, and *no* solution for $\tilde{R}$ arises. The coefficient C_D is found to be negative, and the energy $V(\tilde{R})$ decreases without limit as $\tilde{R}$ approaches zero. Thus, it has to be assumed that as $\tilde{R}$ approaches the Planck length, and the loop expansion of the effective potential fails, the dynamics stabilise $\tilde{R}$ at that scale. This is in contrast to the situation in the absence of quantum corrections where any value of R satisfies the classical field equations (neutral stability).

3.4. Compactification due to explicit gauge fields

It is contrary to the spirit of the original Kaluza-Klein theory to introduce gauge fields explicitly in $4+D$ dimensions, since the hope was that *all* gauge fields might arise from the isometry group of the manifold, upon dimensional reduction. However, there are three advantages to be derived from doing so. Firstly, the explicit gauge fields provide a compactification mechanism. Secondly, the D 'extra' components of the $(4+D)$-dimensional gauge fields provide scalar fields in four dimensions which may prove useful as Higgs scalars. The $F_{\mu\nu}F^{\mu\nu}$ terms in the $(4+D)$-dimensional gauge field kinetic term provide in four dimensions the gauge field kinetic term, the $F_{\mu m}F^{\mu m}$ terms give the Higgs scalar kinetic terms, and the $F_{mn}F^{mn}$ terms provide the Higgs scalar potential, after integration over the compact manifold. (This possibility has been advocated by Manton (1979), Forgacs and Manton (1980) and Chapline and Manton (1981), though with a mathematical dimensional reduction rather than a physical compactification.) Thirdly, an explicit gauge field expectation value in a topologically non-trivial configuration can overcome the difficulty endemic to pure Kaluza-Klein theories of obtaining chiral fermions in four dimensions (see § 5). Kaluza-Klein theories with explicit gauge fields (higher-dimensional Einstein-Yang-Mills theories or supergravity Yang-Mills theories) have been explored by Cremmer and Scherk (1976, 1977), Horvath *et al* (1977), Horvath and Palla (1978), Luciani (1978), Randjbar-Daemi and Percacci (1982), Randjbar-Daemi *et al* (1983a, b, c), Witten (1983), and many subsequent authors, the most recent theories of this type to excite interest being the ten-dimensional supergravity Yang-Mills theories derived from superstring theory (Green and Schwarz 1984, Gross *et al* 1985).

For *explicit* gauge fields with field strength F_a^{AB}, where A and B run over all $4+D$ dimensions, the action is

$$\bar{I} = -\frac{1}{4}\int \mathrm{d}^{4+D}\bar{x}|\det \bar{g}|^{1/2}(F_a)_{AB}(F_a)^{AB} \tag{3.33}$$

and the energy-momentum tensor is

$$\bar{T}_{AB} = -((F_a)_{AC}(F_a)_B{}^C - \tfrac{1}{4}(F_a)_{CD}F_a{}^{CD}\bar{g}_{AB}). \tag{3.34}$$

Assuming that the field strength has a non-zero expectation value only on the compact manifold, i.e. for A, B, C, D, ... taking values m, n, p, q, ..., and that the compact manifold is an Einstein space, then in the notation of § 3.1 we have

$$\frac{c}{8\pi\bar{G}} = \tfrac{1}{4}(F_a)_{CD}(F_a)^{CD} \tag{3.35}$$

and

$$\frac{\tilde{c}}{8\pi\bar{G}} = \tfrac{1}{4}(F_a)_{CD}(F_a)^{CD} - \frac{2\tilde{k}}{8\pi\bar{G}}. \tag{3.36}$$

Thus, (3.9) is satisfied consistently with compactification occurring.

3.5. *Monopole solutions*

The simplest example of compactification using a topologically non-trivial explicit gauge field configuration (Randjbar-Daemi *et al* 1983a) is obtained for $D=2$ with compact manifold the surface of an ordinary sphere:

$$S^2 \equiv \mathrm{G/H} = \mathrm{SU}(2)/\mathrm{U}(1). \tag{3.37}$$

Then an explicit U(1) gauge field may be introduced with a monopole expectation value on the compact manifold:

$$A_m(y)\,\mathrm{d}y^m = a(1 \mp \cos\theta)\,\mathrm{d}\phi \tag{3.38}$$

where a is a constant to be determined in terms of the charges of the matter fields, and the plus and minus signs refer to the upper and lower hemispheres, respectively. The monopole solution is SU(2) invariant (up to a U(1) gauge transformation) (Randjbar-Daemi *et al* 1983a).

The action for this theory is

$$\bar{I} = -\int \mathrm{d}^6\bar{x}|\det \bar{g}|^{1/2}\left(\frac{1}{16\pi\bar{G}}(\bar{\mathbb{R}} + \bar{\Lambda}) + \tfrac{1}{4}F_{AB}F^{AB}\right) \tag{3.39}$$

and the requirement that (3.38) is a solution of the field equations, for Minkowski four-dimensional space, determines the radius of the compact manifold in terms of the six-dimensional gravitational constant and the field strength a:

$$\tilde{R}^2 = 8\pi\bar{G}a^2. \tag{3.40}$$

When both the metric and the U(1) gauge field are expanded about the ground state the situation is more complicated than for pure Kaluza-Klein theory. In the pure Kaluza-Klein theory of § 2, the SU(2) isometry group of S^2 leads to SU(2) gauge fields in four dimensions arising from the off-diagonal components of the metric. In the present case, the gauge group of the effective four-dimensional action is SU(2) × U(1), with the SU(2) gauge fields being a superposition of the gauge fields from the metric and the original explicit U(1) gauge field. Because the radius of the compact manifold is related to the gravitational constant and the Yang-Mills field strength by (3.40), the gauge coupling constant g for the four-dimensional SU(2) gauge group is related to the monopole strength

$$g^2 = 3/2a^2. \tag{3.41}$$

3.6. *Instanton solutions*

If, instead, the compact manifold is

$$S^4 \equiv \mathrm{G/H} = \mathrm{SO}(5)/\mathrm{SO}(4) \tag{3.42}$$

then explicit SU(2) gauge fields may be introduced and an instanton solution on the compact manifold may be used (Randjbar-Daemi *et al* 1983c) instead of a monopole solution. (This solution is only formally an instanton. It is constructed using four spatial coordinates, rather than three spatial coordinates and a time coordinate. It is invariant, up to a gauge transformation under the action of SO(5) on the manifold.) Then, an expansion of the metric and the explicit gauge fields about the ground state leads to SO(5) Yang–Mills fields in four dimensions (rather than SO(5)×SU(2)), because the original SU(2) gauge symmetry is spontaneously broken by the instanton solution.

3.7. Generalised monopole and instanton solutions

If the compact manifold is the coset space

$$K \equiv \mathrm{G}/\mathrm{H} \tag{3.43}$$

then, in the notation of § 2.5, a G-invariant solution of the Einstein–Yang–Mills field equations may always be obtained (Randjbar-Daemi and Percacci 1982) by taking

$$A^{\bar{\alpha}} \equiv A^{\bar{\alpha}}_m \, \mathrm{d}y^m = e^{\bar{\alpha}} \qquad \mathrm{H} \neq \mathrm{U}(1) \tag{3.44}$$

or

$$A^{\bar{\alpha}} = a e^{\bar{\alpha}} \qquad \mathrm{H} = \mathrm{U}(1) \tag{3.45}$$

where a is a constant, and $e^{\bar{\alpha}}$ are the components of the covariant basis (2.43) associated with the generators of H. In (3.44) and (3.45), it is understood that an embedding of H in the explicit Yang–Mills group G_{YM} has been specified (which of course requires that G_{YM} is large enough to contain H). The gauge field configuration of (3.44) or (3.45) generalises the monopole and instanton solutions of the last two sections to an arbitrary coset space. If H is of the form

$$\mathrm{H} = \mathrm{H}' \otimes \mathrm{U}(1) \tag{3.46}$$

then a monopole solution may be constructed by applying the ansatz (3.45) to the U(1) factor. For instance, monopole solutions have been employed by Watamura (1983, 1984) for the complex projective planes

$$CP_N \equiv \frac{\mathrm{SU}(N+1)}{\mathrm{SU}(N) \times \mathrm{U}(1)}. \tag{3.47}$$

If H is of the form

$$\mathrm{H} = \mathrm{H}' \otimes \mathrm{SU}(2)$$

then an instanton solution may be constructed by applying the ansatz (3.44) to the SU(2) factor displayed. (In § 3.6, H′ was SU(2).)

Such gauge field configurations are in general topologically non-trivial. For instance, if for a monopole solution we write

$$F \equiv \tfrac{1}{2} F_{mn} \mathrm{d}y^m \wedge \mathrm{d}y^n \tag{3.48}$$

(with the gauge coupling absorbed in the definition of the gauge field) then the first Chern number (Eguchi *et al* 1980)

$$C_1 = -\frac{1}{2\pi} \int_K F = \text{integer} \tag{3.49}$$

is the monopole number, e.g.

$$C_1 = -2a \qquad \text{for } S^2 \tag{3.50}$$

with a as in (3.38).

For an instanton solution on S^4, if we write

$$F \equiv -\frac{t_a}{2}\frac{F^a_{mn}}{2}(\mathrm{d}y^m \wedge \mathrm{d}y^n) \tag{3.51}$$

where $(t_a/2)$ are the matrices representing the SU(2) factor of H in the fundamental representation of G_{YM}, the second Chern number (Eguchi *et al* 1980)

$$C_2 = \frac{1}{8\pi^2}\int_K \mathrm{Tr}(F \wedge F) = \text{integer} \tag{3.52}$$

is the instanton number, and for an instanton solution

$$|C_2| = 1. \tag{3.53}$$

(There is *no* contribution from the curvature to C_2 for S^4.)

Another very natural choice of gauge field configuration (Charap and Duff 1977, Wilczek 1977, Witten 1983) is

$$A^{\alpha\beta} \equiv A_m{}^{\alpha\beta}\,\mathrm{d}y^m = \omega_m{}^{\alpha\beta}\,\mathrm{d}y^m \equiv \omega^{\alpha\beta} \tag{3.54}$$

where $\omega_m{}^{\alpha\beta}$ is the spin connection for K of § 2.5, and the SO(D) tangent space group has been embedded in G_{YM}. ($A^{\alpha\beta}$ is a labelling of the gauge fields corresponding to the standard antisymmetric generators $M^{\alpha\beta}$ of SO(D).) Then the topology of the gauge field configuration is directly related to the topology of the manifold (the Euler characteristic).

3.8. *Non-minimal gravitational actions*

A possible compactification mechanism which does *not* require the introduction of matter fields is through non-minimal terms added to the $(4+D)$-dimensional Einstein action. Wetterich (1982a) has considered an action of the form

$$\bar{I} = -(16\pi\bar{G})^{-1}\int \mathrm{d}^{4+D}\bar{x}|\det \bar{g}|^{1/2}\{\bar{\mathbb{R}} + \alpha\bar{\mathbb{R}}^2 + \beta\bar{\mathbb{R}}_{AB}\bar{\mathbb{R}}^{AB} + \gamma\bar{\mathbb{R}}_{ABCD}\bar{\mathbb{R}}^{ABCD}\}. \tag{3.55}$$

He finds that the field equations admit a compactified solution of the type $M^4 \times S^D$, where M^4 is Minkowski space, though to obtain other compact manifolds (e.g. a product of spheres) requires higher curvature invariants to stabilise the effective action.

3.9. *Stability of compactification of pure Kaluza-Klein theories*

The simplest perturbation under which to consider stability (Candelas and Weinberg 1984) is a change in the overall scale $\tilde{R}$ of the manifold. Suppose that the action can be written in the form

$$\bar{I} = -(16\pi\bar{G})^{-1}\int \mathrm{d}^{4+D}\bar{x}|\bar{g}|^{1/2}(\tilde{\mathbb{R}}(y) + \bar{\Lambda}) - \int \mathrm{d}^4x|g|^{1/2}V(\tilde{R}) \tag{3.56}$$

where $\tilde{\mathbb{R}}(y)$ is the curvature scalar for the compact manifold, and $V(\tilde{R})$ is a compactifying action. Using (3.28) for an Einstein space,

$$\tilde{\mathbb{R}} = -2D\tilde{k}\tilde{R}^{-2}. \tag{3.57}$$

Integrating over the coordinates y of the compact manifold, and using (3.27) and (3.32), we may write (for Minkowski four-dimensional space)

$$\bar{I} = -\int \mathrm{d}^4x \, V_{\mathrm{eff}} \tag{3.58}$$

with

$$V_{\mathrm{eff}} = (16\pi G)^{-1}(\bar{\Lambda} - 2\tilde{k}D\tilde{R}^{-2}) + V(\tilde{R}). \tag{3.59}$$

When $V(\tilde{R})$ has the simple form

$$V(\tilde{R}) = c_D\tilde{R}^{-q} \qquad \text{with} \qquad c_D > 0 \tag{3.60}$$

the effective potential always has a *stable* minimum, when $\tilde{k} \neq 0$, provided

$$q + D - 2 > 0. \tag{3.61}$$

This includes the case of compactification by quantum fluctuations in matter fields (Candelas and Weinberg 1984), when $q = 4$. It also includes (Bailin *et al* 1984) the case of Freund–Rubin compactification because the field equation (3.12) implies that F of (3.14) has the property

$$F \sim \tilde{R}^{-D} \tag{3.62}$$

so that (3.16) and (3.17) imply that $V(\tilde{R})$ is of the form (3.60) with

$$q = D \qquad \text{Freund–Rubin compactification.} \tag{3.63}$$

One subtlety which arises for compactification by matter field quantum fluctuations is that the $(\mathrm{d}\tilde{R}/\mathrm{d}t)$-dependent terms in the effective action may as a result of quantum effects (Gilbert *et al* 1984) have a different sign from the tree approximation value deriving from $\tilde{\mathbb{R}}(y)$. This occurs (Gilbert and McClain 1984) when the ratio of the number of massless scalar fields to the number of massless spinor fields in the higher-dimensional theory is sufficiently small. Then, instability would arise at the apparently stable minimum of the effective potential!

More general perturbations may be considered than just an overall change of scale of the compact manifold. For instance, for a product compact manifold and Freund–Rubin compactification instability occurs under independent changes of scale of the two spaces in the product (Duff *et al* 1984b, Bailin and Love 1985a) with the radius of one growing at the expense of the other. A similar phenomenon occurs as a result of one-loop quantum corrections for a toroidal compact manifold T^D, with one of the D dimensions contracting while the others expand (Appelquist *et al* 1983, Inami and Yasuda 1983). Stability under more general perturbations including squashing has also been studied for spheres (Page 1984) and other manifolds.

All the above considerations of stability are in terms of classical perturbations of an effective action. However, in the five-dimensional theory there is an instability of the Kaluza–Klein ground state (Witten 1982, Kogan *et al* 1983) due to quantum tunnelling. Fortunately, there does *not* appear to be a (simple) generalisation of this phenomenon to higher dimensions (Young 1984).

3.10. Stability of compactification for Einstein–Yang–Mills theories

When the theory contains explicit gauge fields, there is a further danger of instability of the compactification, because of classical perturbations in the gauge fields. For the case of an explicit U(1) gauge field in a monopole configuration on S^2 it has been shown by Randjbar-Daemi *et al* (1983a) that *no* instability arises in this way. However, for any *non*-Abelian gauge group G_{YM} with an SU(2) invariant vacuum configuration on S^2 instability occurs (Randjbar-Daemi *et al* 1983b) as can be seen by expanding to quadratic order in the classical perturbation, in the action, and looking for tachyons.

In general (Randjbar-Daemi *et al* 1983b, Schellekens 1985a, b), for a compact manifold G/H, only the perturbations in the gauge bosons associated with H mix with the perturbations in the metric. This sector has been calculated (Schellekens 1984) for a general hypersphere S^D, for a general explicit gauge group G_{YM}. Except for the case $D=3$, there are never any tachyons in this sector, and the compactification is stable in this respect.

However, even when the perturbations in the H gauge bosons do *not* lead to instability, the perturbations in the gauge bosons in G but not in H often do (Randjbar-Daemi 1983c, 1984c, Frampton *et al* 1984a, Schellekens 1985a), as in the case discussed by Randjbar-Daemi *et al* (1983b). A general discussion of this type of instability has been given by Schellekens (1985b) for a general hypersphere S^D with the explicit gauge fields of gauge group G_{YM} in a generalised monopole configuration (as in § 3.7).

4. Gauge coupling constants

4.1. Geometric interpretation

In the five-dimensional case of § 1, where the gauge group was Abelian, charge was quantised and the quantum of charge was related to the radius of the compact manifold S^1 (and the four-dimensional gravitational constant) as in (1.26). In higher-dimensional theories in six or more dimensions where the gauge group is non-Abelian we would like to be able, in a similar fashion, to relate the gauge coupling constant (or coupling constants) to the geometry of the compact manifold. The connection can be made (Weinberg 1983) by considering isometric curves on the manifold (i.e. curves traced out when an infinitesimal isometry is exponentiated). Consider the infinitesimal isometry with parameter $d\sigma$ associated with a particular generator t_a of the isometry group. From (2.3) this is

$$I+\mathrm{i}\,d\sigma t_a: \qquad y^n \rightarrow y^{n'} = y^n + d\sigma \xi_a^n(y) \tag{4.1}$$

where ξ_a^n is the corresponding Killing vector. The exponentiated curve traced out as σ varies is

$$e^{\mathrm{i}\sigma t_a}: \qquad y^n = Y^n(\sigma, y_0) \tag{4.2}$$

where Y^n is the solution of

$$dY^n/d\sigma = \xi_a^n(y) \qquad Y^n(\sigma=0) = y_0^n. \tag{4.3}$$

For any representation of the non-Abelian isometry group, the eigenvalues of a diagonal generator t_a will be integral multiples of some lowest (positive) eigenvalue $g_{\min}$ (with the gauge coupling constant g absorbed in the definition of the generators). As σ increases from 0 to $2\pi/g_{\min}$, $e^{\mathrm{i}\sigma t_a}$ returns to its starting value, and, if the representation t_a is N valued, we go exactly N times round the manifold. The circumference $S(y_0)$

of the manifold along the isometric curve with starting point y_0 is

$$S(y_0)=\frac{1}{N}\int_0^{2\pi/g_{\min}} d\sigma[\tilde{g}_{mn}(Y)\xi_a^m(Y)\xi_a^n(Y)]^{1/2}$$

$$=\frac{2\pi}{Ng_{\min}}[\tilde{g}_{mn}(y_0)(y_0)\xi_a^m(y_0)\xi_a^n(y_0)]^{1/2} \tag{4.4}$$

where the metric $\tilde{g}_{mn}(y)$ is as in (2.6), and infinitesimal distance ds along the isometric curve is given by

$$ds^2=\tilde{g}_{mn}(y)\xi_a^m\xi_a^n\, d\sigma^2. \tag{4.5}$$

Averaging over the starting points y_0, and using the notation (2.19),

$$\langle s^2\rangle=(2\pi/Ng_{\min})^2\langle\tilde{g}_{mn}(y)\xi_a^m(y)\xi_a^n(y)\rangle. \tag{4.6}$$

For standard normalisation of the gauge fields (2.18),

$$\langle s^2\rangle=(2\pi/Ng_{\min})^2\kappa^2 \tag{4.7}$$

where κ is the four-dimensional gravitation constant of (2.17). Thus, the gauge coupling $g_{\min}$ is related to the geometry of the compact manifold by

$$g_{\min}=2\pi\kappa/N\langle s^2\rangle^{1/2} \tag{4.8}$$

where $\langle s^2\rangle^{1/2}$ is the root-mean-square circumference of the manifold along the isometric curve associated with the generator t_a averaged over starting points. The result is general enough to handle situations where the isometry group is a product of non-Abelian factors, so that there might be different gauge coupling constants associated with different diagonal generators t_a. Clearly, the above reasoning is not directly applicable to cases where the gauge group has an Abelian factor, since Abelian gauge fields do not have any self-coupling. In that case, to interpret the result (Weinberg 1983) it is necessary to introduce a matter field (e.g. a complex scalar field).

For an Einstein space, where the Ricci tensor is proportional to the metric, a 'radius' a for the compact manifold may be introduced by writing

$$\tilde{R}_{mn}=a^{-2}\tilde{g}_{mn}. \tag{4.9}$$

For the hypersphere S^D, (4.8) yields (Weinberg 1983)

$$g^2=\frac{\kappa^2}{a^2}\frac{(D+1)}{2(D-1)} \tag{4.10}$$

where the gauge coupling constant g for the $SO(D+1)$ isometry group has been normalised so that in the single-valued defining representation the eigenvalues of, say, M^{12} are $(g, -g, 0, 0, \ldots)$, with the normalisation of generators

$$[M^{ij}, M^{kl}]=i(\delta^{ik}M^{jl}+\delta^{jl}M^{ik}-\delta^{il}M^{jk}-\delta^{jk}M^{il}) \tag{4.11}$$

and couplings $M^{12}A_{12}^{\mu}+\ldots$. For the four-dimensional manifold CP^2, the corresponding result (Weinberg 1983) is

$$g^2=\tfrac{4}{3}\kappa^2/a^2 \tag{4.12}$$

with a standard normalisation of generators for the isometry group SU(3), such that the eigenvalues of t_3 in the **3** are $(g/2, -g/2, 0)$.

4.2. Absolute values of gauge coupling constants

When compactification is by quantum fluctuations in massless higher-dimensional fields, it is possible, as discussed in § 3.3, to determine the constant C_D, in the effective potential of (3.23), by calculating with the tower of four-dimensional fields deriving from the harmonic expansion of the higher-dimensional fields. Then, the radius of the compact manifold is determined by (3.30) where, in the present notation,

$$a^{-2}=2\tilde{k}\tilde{R}^{-2}. \tag{4.13}$$

If for instance, the compact manifold is a sphere S^D, then

$$2\tilde{k}=D-1 \qquad \text{for} \qquad S^D. \tag{4.14}$$

For this case, the absolute value of the gauge coupling constant (at the compactification scale) is then determined by (4.10). Unfortunately (Candelas and Weinberg 1984), very large numbers ($>10^3$) of massless higher-dimensional fields seem to be required to obtain $g^2/4\pi\leqslant 1$, as we might expect if g^2 varies under the renormalisation group by less than an order of magnitude between the electroweak scale and the compactification scale.

4.3. Ratios of gauge coupling constants

When the compactification is be some mechanism other than quantum fluctuations, it may not be possible to calculate absolute values of gauge coupling constants, e.g., for Freund–Rubin compactification (Freund and Rubin 1980) the expectation value (3.14) of the antisymmetric tensor field strength is unknown, and consequently the 'radius' of the compact manifold is unknown. However, it is still possible to calculate *ratios* of gauge coupling constants provided the geometry of the compact manifold is completely determined apart from the overall scale. This is the case for Freund–Rubin compactification, because (3.17) (together with the Einstein field equation) implies that the compact manifold is an Einstein space, so that there is only a *single* overall scale, and the various circumferences of the manifold are all related. (A subtlety arises for the Kaluza–Klein supergravity case to which we return shortly.)

For the manifolds M^{pqr} of § 2.6, with isometry group SU(3) × SU(2) × U(1) (except in the exceptional cases mentioned in § 2.6), the geometry is specified by three scales, a, b and c (Castellani *et al* 1984a). However, when the condition is imposed that M^{pqr} should be an Einstein space for Freund–Rubin compactification, these scales are related (Castellani *et al* 1984a). Thus, the coupling constants g_3, g_2 and g_1 for the SU(3), SU(2) and U(1) factors of the isometry group are related by using (4.8), even though the absolute values cannot be calculated. It is found (Bailin and Love 1984b, Ezawa and Koh 1984a) that for all non-zero values of p, q and r

$$1<g_2^2/g_3^2<\tfrac{3}{2} \tag{4.15}$$

and

$$\tfrac{27}{2}(p^2n^2/r^2)<g_1^2/g_3^2<\infty \tag{4.16}$$

where n is an integer. Equation (4.15) gives results which can be in reasonable agreement with extrapolations (Bailin and Love 1984a) of the known gauge coupling constants to the compactification scale. However, (4.16) gives values of g_1^2/g_3^2 that are far too large unless r is bigger than 1, corresponding to non-simply connected manifolds (Witten 1981). The situation is a little better for the exceptional case $S^5\times S^2$ (Bailin and Love 1984a), but this product manifold is probably unstable, as discussed in § 3.9.

Up to this point, we have been relating gauge coupling constants to the geometry by assuming that the gauge field terms in the effective four-dimensional theory arise entirely from the $(4+D)$-dimensional curvature scalar, as in § 2.3. However, as we have already mentioned in § 2.3, some refinement is necessary in the supergravity context where there are other terms in the $(4+D)$-dimensional action which may contribute to the gauge field action in four dimensions. For instance, in eleven-dimensional supergravity (see § 7.1) contributions to the Yang-Mills Lagrangian in four dimensions can come from the fourth-rank antisymmetric field strength (employed in the mechanism of Freund and Rubin (1980)). These contributions (Duff *et al* 1983c) arise by taking an expectation value for the tensor field strength of the form (3.14), and writing, in the expansion about the ground state,

$$F_{\mu\nu mn} \propto \varepsilon_{\mu\nu\rho\sigma} D^{\rho} A^{\sigma}_{a} D_{m}(\tilde{g}_{np}(y)\xi^{p}_{a}(y)) \qquad \mu, \nu = 0, 1, 2, 3, m, n = 4, \ldots, 10 \qquad (4.17)$$

in the notation of § 2.2. When (4.17) is substituted in the action (7.11), extra contributions to the gauge field kinetic term arise from $D^{\rho}A^{\sigma}_{a}$ in (4.17), resulting in this term being multiplied by a factor of four relative to the pure eleven-dimensional Einstein gravity case. The only effect is on the overall normalisation of the gauge field term in four dimensions compared with the four-dimensional Einstein gravity term, so that *ratios* of gauge coupling constants are unmodified.

The situation can be different in other supergravity theories. For instance, in $N=2$ (two supersymmetry generators) supergravity in ten dimensions, the extra contributions to the Yang-Mills Lagrangian in four dimensions come from both a fourth-rank tensor field strength and a second-rank tensor field strength belonging to the supergravity multiplet, and these contributions have a *different structure* to the contribution due to the metric of § 2. Then the *ratios* of gauge coupling constants differ from the ratios in pure $(4+D)$-dimensional Einstein gravity. As we have seen in § 3.5, a similar phenomenon can occur in Einstein-Yang-Mills theories, with mixing between explicit higher-dimensional gauge fields and gauge fields arising from the metric on dimensional reduction.

In eleven-dimensional supergravity, the ratio of the gauge coupling constants g_5 and g_3 for the isometry group $SO(5)\times SO(3)$ of the squashed 7-sphere (see § 7.2) has been calculated by de Alwis *et al* (1985) to be

$$g_5^2/g_3^2 = \tfrac{15}{13}. \qquad (4.18)$$

5. Particle spectrum of Kaluza-Klein theories

5.1. Boson spectrum

We have seen in § 1.5 that fields which depend upon the coordinate(s) associated with the compactified dimension(s) have a mass whose scale is set by the size of the compact space; the fields $\phi^{n}(x)$ in (1.19), for example, have a mass

$$m_n = n/\tilde{R} \qquad (n = 0, 1, 2, \ldots) \qquad (5.1)$$

where $\tilde{R}$ is the 'radius' of the circle formed by the compact dimension. This scale is likely to be gigantic compared with the energies available to present or forseeable particle accelerators, since we saw in (1.28) that $\tilde{R}^{-1}$ is of order of the Planck mass:

$$\tilde{R}^{-1} \sim m_{\mathrm{p}} \equiv G^{-1/2} \sim 10^{19}\,\mathrm{GeV}. \qquad (5.2)$$

Nevertheless these massive states may still affect the low-energy sector (Duff 1984) because of quantum effects. This low-energy dependence upon the high-mass states derives from the propagators of the massive states

$$\Delta(p, m_n) \propto \frac{1}{p^2 - m_n^2} = \mathrm{O}(m_n^{-2}) \qquad \text{as} \quad m \to \infty. \tag{5.3}$$

Since $m_n \to \infty$ as $n \to \infty$ in (5.1), we might suppose that at most it would be necessary to retain only the lowest excitations, say $n = 0, 1, 2, 3$, at least in *finite* diagrams. (There is also the possibility of logarithmic dependence upon m, which arises on renormalisation of divergence diagrams.) Quite apart from the question of the renormalisability of Kaluza-Klein theories, the above argument for retaining at most the lowest excitations is complicated by the fact that the charge q_n of the particle of mass m_n also increases with n, as is apparent from (1.26). Since

$$q_n = n\kappa/\tilde{R}$$

$$q_n^2 \Delta(p, m_n) \propto n^2 \frac{\kappa^2}{\tilde{R}^2} \frac{1}{p^2 - n^2/\tilde{R}^2} = \mathrm{O}(\kappa^2) \tag{5.4}$$

$$= \mathrm{O}(16\pi m_{\mathrm{p}}^{-2}) \qquad \text{as} \quad n \to \infty.$$

Thus the contribution from each component of the whole 'tower' of massive states is equally important (or unimportant). Our chief preoccupation in the succeeding parts of this section will be the characterisation of the *zero* modes of Kaluza-Klein theories, since it is these modes, which develop non-zero masses at a scale well below the Planck scale, which are presumed to constitute the particles actually observed in the (low-energy) world which is accessible to experiment. Even so, for the reasons already given, it is of some interest to characterise the massive modes, and in any case these modes are as important as the zero modes in determining the cosmological evolution of the very early universe, near the compactification scale. (We shall discuss this cosmological role of the massive modes in the following section.)

We start by determining the four-dimensional classical mass spectrum of the original five-dimensional Kaluza-Klein theory. In the absence of any matter fields, the equations of motion are

$$\bar{\mathbb{R}}_{AB} - \tfrac{1}{2}\bar{g}_{AB}\bar{\mathbb{R}} = 0 \tag{5.5a}$$

or equivalently

$$\bar{\mathbb{R}}_{AB} = 0 \tag{5.5b}$$

and the ground-state solution is

$$\langle 0|\bar{g}_{AB}|0\rangle = \eta_{AB} \equiv \mathrm{diag}(1, -1, -1, -1, -\tilde{R}^2) \tag{5.6}$$

as in (1.3). To find the mass spectrum we vary the field $\bar{g}_{AB}$ around its ground-state value. Thus we write

$$\bar{g}_{AB}(x, \theta) = \eta_{AB} + h_{AB}(x, \theta) \tag{5.7}$$

and expand (5.5b) to lowest- (first-) order in h. This gives

$$\partial_B\partial_A h^C{}_C - \partial_C\partial_A h^C{}_B - \partial_C\partial_B h^C{}_A + \partial^C\partial_C h_{AB} = 0. \tag{5.8}$$

Note that the connection coefficients vanish in the ground state, so ordinary partial derivatives are sufficient. Equation (5.8) is invariant with respect to the gauge transformation

$$h_{AB} \to h_{AB} + \partial_A \zeta_B + \partial_B \zeta_A \tag{5.9}$$

so we may choose a convenient gauge in which to extract the physical content of the theory. The choice is (Dolan 1983)

$$\partial^\mu h_{\mu 5} = 0 \tag{5.10a}$$

$$\partial^5 h_{\mu 5} = 0 \tag{5.10b}$$

$$\partial^5 h_{55} = 0. \tag{5.10c}$$

Since the compactified manifold is a circle we may write

$$h_{AB}(x, \theta) = \sum_{n=-\infty}^{\infty} h_{AB}^{(n)}(x) e^{in\theta} \tag{5.11}$$

as in (1.19), so that h is well defined on S^1. The gauge choice (5.10) then implies that

$$\partial^\mu h_{\mu 5}^{(0)}(x) = 0 \tag{5.12a}$$

$$h_{\mu 5}^{(n)}(x) = 0 \qquad (n \neq 0) \tag{5.12b}$$

$$h_{55}^{(n)}(x) = 0 \qquad (n \neq 0). \tag{5.12c}$$

In other words, by an appropriate choice of gauge the $n \neq 0$ vector potentials $h_{\mu 5}^{(n)}(x)$ and the $n \neq 0$ scalar fields $h_{55}^{(n)}(x)$ may be transformed to zero (i.e. 'gauged away'). This sounds reminiscent of the situation in electroweak theory, for example, where the local $SU(2) \times U(1)$ gauge invariance is spontaneously broken. The would-be Goldstone boson modes (which can also be gauged away) associated with spontaneous breakdown of the global symmetry are 'eaten' by the hitherto massless gauge bosons. In the process some of the gauge bosons become massive. We might wonder, therefore, whether the $n \neq 0$ vector potentials and scalar fields are 'eaten' by the $n \neq 0$ tensor modes $h_{\mu\nu}^{(n)}(x)$, which thereby become massive. (Note that, if the $n \neq 0$ vector potentials and scalar fields really are (massless) Goldstone modes, then there are a total of *three* modes for each n available for eating, since a massless vector field has two degrees of freedom, and a massless scalar only one. Thus there would be a total of five degrees of freedom available, just the number required by a massive spin-2 field.) In fact, the scenario envisaged above is what actually happens, although it is not apparent at this stage precisely what symmetry it is which will yield a (three-times) infinite number (all $n \neq 0$) of Goldstone modes. This is the subject of the following section. For the present we shall content ourselves with verifying that the $n \neq 0$ tensor fields $h_{\mu\nu}^{(n)}(x)$ are indeed massive, as we have claimed, and that all the other modes are massless (Salam and Strathdee 1982). The derivation which we present follows that of Dolan (1983).

First we substitute (5.11) into (5.8), and use the gauge choice (5.10), or equivalently (5.11). From the 55 component of (5.8) we find

$$\partial^\lambda \partial_\lambda h_{55}^{(0)} = 0 \tag{5.13a}$$

and

$$h_\lambda^{\lambda(n)} = 0 \qquad (n \neq 0). \tag{5.13b}$$

The $\mu 5$ component gives

$$\partial^\lambda \partial_\lambda h^{(0)}_{\mu 5} = 0 \tag{5.14a}$$

and, using (5.13*b*),

$$\partial^\lambda h^{(n)}_{\mu\lambda} = 0 \qquad (n \neq 0). \tag{5.14b}$$

Finally, from the $\mu\nu$ component of (5.8), we find

$$\partial^\lambda \partial_\lambda h^{(0)}_{\mu\nu} + \partial_\mu \partial_\nu (h^{\lambda(0)}_{\lambda} + h^{5(0)}_{5}) - \partial_\lambda \partial_\mu h^{\lambda(0)}_{\nu} - \partial_\lambda \partial_\nu h^{\lambda(0)}_{\mu} = 0 \tag{5.15a}$$

and using (5.13*b*) and (5.14*b*)

$$(\partial^\lambda \partial_\lambda + n^2/\tilde{R}^2) h^{(n)}_{\mu\nu} = 0 \qquad (n \neq 0). \tag{5.15b}$$

Thus, as claimed, the $n \neq 0$ tensor modes $h^{(n)}_{\mu\nu}$ are indeed massive, and turn out to have masses m_n as in (5.1). Equations (5.13*a*) and (5.14*a*) show that the surviving scalar field $h^{(0)}_{55}$ and vector potential $h^{(0)}_{\mu 5}$ are both massless. Also, by defining

$$\bar{h}^{(0)}_{\mu\nu} \equiv h^{(0)}_{\mu\nu} + \tfrac{1}{2}\, \eta_{\mu\nu} h^{5(0)}_{5} \tag{5.16}$$

we find that we can recast (5.15*a*) in the form

$$\partial^\lambda \partial_\lambda \bar{h}^{(0)}_{\mu\nu} + \partial_\nu \partial_\mu \bar{h}^{\lambda(0)}_{\lambda} - \partial_\lambda \partial_\mu \bar{h}^{\lambda(0)}_{\nu} - \partial_\lambda \partial_\nu \bar{h}^{\lambda(0)}_{\mu} = 0 \tag{5.17}$$

which is the equation of a massless spin-2 (graviton) field (Weinberg 1972).

The masslessness of the $n = 0$ modes was anticipated in §§ 1.4 and 1.7, where we made the ansatz of discarding the θ dependence of all fields, as in equation (1.7) for example. In fact if we truncate the theory by retaining only the $n = 0$ modes, perform a Weyl rescaling (1.31), and carry out the θ integration in (1.12), we obtain an effective four-dimensional action

$$I = -\int \mathrm{d}^4 x |\det g^{(0)}|^{1/2} \left\{ \frac{1}{\kappa^2} \mathbb{R}(g^{(0)}) + \tfrac{1}{4}\, \phi^{(0)} F^{(0)}_{\mu\nu} F^{\mu\nu(0)} - \frac{1}{6\kappa^2}\, \phi^{(0)-2} (\partial_\mu \phi^{(0)}) \partial^\mu \phi^{(0)} \right\} \tag{5.18}$$

where κ^2 is defined in (1.16) and (1.17), $\phi^{(0)}$ is defined in (1.19), $g^{(0)}_{\mu\nu}$, $A^{(0)}_\mu$ and hence $F^{(0)}_{\mu\nu}$ are defined analogously to $\phi^{(0)}$, as in (5.11), and indices are raised and lowered with $g^{(0)}_{\mu\nu}$. This (truncated) action is derived in Appelquist and Chodos (1983a, b), for example. The action (5.18) is consistent with the previous results (1.18) and (1.32). The masslessness of A_μ derives from the invariance of (5.18) under the gauge transformation (1.11):

$$A_\mu \to A'_\mu = A_\mu + \partial_\mu \varepsilon \tag{5.19}$$

which in turn derives from the invariance of the original five-dimensional action under the (particular) coordinate transformation (1.9):

$$\theta \to \theta' = \theta + \xi \varepsilon(x) \tag{5.20a}$$

$$x^\mu \to x'^\mu = x^\mu. \tag{5.20b}$$

In the same way the masslessness of $g^{(0)}_{\mu\nu}$ derives from the invariance of the action under the generalised (four-dimensional) coordinate transformation

$$x^\mu \to x'^\mu = x^\mu + \zeta^\mu(x) \tag{5.21a}$$

$$\theta \to \theta' = \theta. \tag{5.21b}$$

However (5.18) is *also* invariant under a global scale transformation, in which

$$g_{\mu\nu}^{(0)} \to g_{\mu\nu}'^{(0)} = g_{\mu\nu}^{(0)} \tag{5.22a}$$

$$A_\mu^{(0)} \to A_\mu'^{(0)} = A_\mu^{(0)} + \lambda A_\mu^{(0)} \tag{5.22b}$$

$$\phi^{(0)} \to \phi'^{(0)} = \phi^{(0)} - 2\lambda\phi^{(0)} \tag{5.22c}$$

with λ infinitesimal and constant. The ground state of the system has

$$\langle g_{\mu\nu}^{(0)} \rangle = \eta_{\mu\nu} \qquad \langle A_\mu^{(0)} \rangle = 0 \qquad \langle \phi^{(0)} \rangle = \tilde{R}^2 \tag{5.23}$$

where $\eta_{\mu\nu}$ is the Minkowski metric. Thus it has symmetry $P^4 \times R^1$, where P^4 is the four-dimensional Poincaré group and R^1 is the gauge symmetry, which is *not* a compact U(1) because the truncated theory has no memory of the periodicity in θ. Since $\langle \phi^{(0)} \rangle$ is non-zero, the ground state is *not* invariant under the global scale transformation. Thus the masslessness of $\phi^{(0)}$ is because it is the Goldstone boson associated with the spontaneous breakdown of the global scale invariance (Dolan and Duff 1984).

It is important to remember that the results for the spectrum of the five-dimensional Kaluza–Klein theory derived so far determine only the *classical* mass values. Quantum effects can modify these values. In particular, we should expect that the massless modes will only be 'truly' massless if their masslessness is protected by some sort of symmetry. In the case of the graviton field $g_{\mu\nu}^{(0)}$ and of the gauge field $A_\mu^{(0)}$ this is indeed the case, since the gauge symmetries from which they derive are symmetries of the *full* theory. However the (Brans–Dicke) scalar field $\phi^{(0)}$ is not so protected. The global scale invariance, of which it is the Goldstone boson, is a symmetry only of the *truncated* theory, in which the $n \neq 0$ modes are discarded. We shall see in the following section that the full theory does not have this symmetry. So $\phi^{(0)}$ is only a pseudo-Goldstone boson (Dolan and Duff 1984). Its classical mass is zero, but we would expect that radiative corrections will shift the mass to a non-zero value. Evidently, an understanding of the symmetry of the full (untruncated) theory is necessary if we are to accurately determine the zero modes in a general Kaluza–Klein theory. However, before addressing that task (in the next section) we shall describe briefly the work which has been carried out to determine the classical mass spectrum in more realistic, and therefore higher-dimensional, theories.

The procedure is essentially a generalisation of that which we have described for the five-dimensional case. Firstly, one has to find a ground-state solution of the field equations, as in (5.6). Secondly, one considers arbitrary fluctuations of all fields about this solution, as in (5.7), and expands these fluctuations, as in (5.11), in a complete set of harmonics on the compact manifold. The coefficients of these harmonics are the physical (four-dimensional) fields, and substitution into the full higher-dimensional field equations yields the four-dimensional field equations, and hence the spectrum. However, there are a number of technical complications, which stem from the higher dimensionality, which deserve comment.

Firstly, the ground-state metric (2.6)

$$\langle 0 | \tilde{g}_{AB} | 0 \rangle = \mathrm{diag}(\eta_{\mu\nu}, -\tilde{g}_{mn}(y)) \tag{5.24}$$

where $\eta_{\mu\nu}$ is the metric of flat Minkowski space M_4, given in (1.4), and $\tilde{g}_{mn}(y)$ is the metric of the compact manifold, is in general *not* a solution of the $(4+D)$-dimensional pure gravity field equations. This was discussed in § 3. To achieve compactification it is necessary to introduce additional (matter) fields (or to allow a non-minimal gravitational action). The additional fields, of course, also have fluctuations about their ground-state (background field) values, and these fluctuations too must be expanded in terms of a complete set of harmonics on the compact manifold. Thus there are additional contributions to the spectrum, besides those originating in the purely gravitational sector. In general the final spectrum depends in an essential way upon the precise compactification mechanism actually utilised.

The second complication occurs because the 'harmonic expansion' is in general considerably more complicated than in the five-dimensional case (5.11), and the quantum numbers of the mass eigenstates are complicated by the general non-Abelian nature of the isometry group of the compact manifold. The harmonic expansion (5.11) expands the fields in terms of the harmonics $e^{in\theta}$ which form a complete set of representations of the isometry group U(1) on the compact manifold S^1. In general the fields of the theory transform as representations $|R\rangle$ of the tangent space group G_T. For the case of a D-dimensional compact manifold K, the tangent space group is

$$G_T = SO(1, 3+D). \tag{5.25}$$

The 'harmonic expansion' is then an expansion of the representations $|R\rangle$ in terms of the representations $|G_\alpha\rangle$ of the isometry group G_I of the compact manifold K:

$$|R\rangle = \sum_\alpha |G_\alpha\rangle\langle G_\alpha|R\rangle. \tag{5.26}$$

We shall assume that K is a coset space

$$K = G/H \tag{5.27}$$

where G is a Lie group and H is a subgroup of G. In the case that H is a maximal subgroup of G, the isometry group is G itself; but if H is a non-maximal subgroup, then the isometry group is (Castellani *et al* 1984c)

$$G_I = G \times N'(H) \tag{5.28}$$

where $N'(H)$ is the normaliser of H in G, but with any U(1) factors common with G deleted. Thus in any event H is always a subgroup of G_I so we may expand the representations $|G_\alpha\rangle$ of G_I in terms of the representations $|h_i\rangle$ of H:

$$|G_\alpha\rangle = \sum_i |h_i\rangle\langle h_i|G_\alpha\rangle. \tag{5.29}$$

Also, we showed in (2.56) that H is a subgroup of the tangent space SO(D) of the compact manifold K. Thus, since

$$\begin{aligned} G_T &\supset SO(1,3) \times SO(D) \\ &\supset SO(1,3) \times H \end{aligned} \tag{5.30}$$

we may also expand the representations $|R\rangle$ of G_T in terms of the representations of H:

$$|R\rangle = \sum_j |h_j'\rangle\langle h_j'|R\rangle. \tag{5.31}$$

Then the harmonic expansion (5.26) is

$$|R\rangle = \sum_{\alpha,i,j} |G_\alpha\rangle\langle G_\alpha|h_i\rangle\langle h_i|h_j'\rangle\langle h_j'|R\rangle. \tag{5.32}$$

This shows that the isometry group representations which actually occur in the harmonic expansion of $|R\rangle$ are those for which there is at least one overlap between the expansions (5.29) and (5.31) in representations of H (de Alwis and Koh 1984). The degeneracy of $|G_\alpha\rangle$ is given by the number of overlaps.

As an illustration consider the $(D=2)$ case in which

$$K=S^2=\mathrm{SO}(3)/\mathrm{SO}(2)\simeq \mathrm{SU}(2)/\mathrm{U}(1). \tag{5.33}$$

Since $\mathrm{H}=\mathrm{SO}(2)$ *is* the maximal subgroup of $\mathrm{G}=\mathrm{SO}(3)$, the isometry group is the 'rotation' group

$$\mathrm{G}_{\mathrm{I}}=\mathrm{SO}(3). \tag{5.34}$$

Suppose also that we are concerned with the vector representation $|\text{vector}_6\rangle$ of $\mathrm{SO}(1,5)$. Under the decomposition (5.30) the expansion (5.31) becomes

$$|\text{vector}_6\rangle=|\text{vector}_4,0\rangle+|\text{scalar}_4,+1\rangle+|\text{scalar}_4,-1\rangle \tag{5.35}$$

using the notation $|$Lorentz group representation, U(1) charge$\rangle$. Consider first the scalar component with charge +1. Then the representations of the isometry group SO(3) which actually occur in the harmonic expansion are those which contain an element with U(1) charge +1. Denoting the representations of SO(3) by their 'angular momentum' J, so that the dimensionality is $2J+1$, this means that the required representations are those with J integral and non-zero:

$$J=n \qquad (n=1,2,3,\ldots). \tag{5.36}$$

The inclusion of the vector component with charge 0 extends this to include also the $J=0$ representations of G_{I}.

The 'realistic' cases studied have for the most part been in the context of eleven-dimensional supergravity, in which the compactification is achieved using a background 'three-index photon' field A_{BCD} with field strength F_{ABCD}, defined in (3.10). As explained in § 3.2, since supersymmetry forbids an eleven-dimensional cosmological constant, these supergravity theories typically have a classical ground state which is a four-dimensional anti-de Sitter space times a compact manifold K. The known solutions of the field equations fall into two classes: the Freund–Rubin (1978, 1980) solution and the Englert (1982) solution. In both classes the gravitino field vanishes; the four-dimensional Riemann tensor is maximally symmetric:

$$R_{\mu\nu\rho\sigma}=-k[g_{\mu\rho}(x)g_{\nu\sigma}(x)-g_{\mu\sigma}(x)g_{\nu\rho}(x)] \tag{5.37}$$

where $g_{\mu\nu}(x)$ is the anti-de Sitter space metric, and the 'photon' field strength in the 4-space is given by

$$F_{\mu\nu\rho\sigma}=|\det g|^{-1/2}\varepsilon^{\mu\nu\rho\sigma}F \tag{5.38}$$

as in (3.14). Also, in both classes the metric and the field strength with mixed indices vanish:

$$\tilde{g}_{\mu m}=F_{\mu\nu\rho m}=F_{\mu\nu mn}=F_{\mu mnp}=0. \tag{5.39}$$

However, in the Freund–Rubin solution,

$$F_{mnpq}(y)=0 \qquad \mathbb{R}_{mn}=\tfrac{1}{3}F^2\tilde{g}_{mn} \qquad k=\tfrac{2}{3}F^2 \tag{5.40}$$

whereas the Englert solution has

$$F_{mnpq}(y)=\pm\tfrac{1}{2}F\bar{\eta}\tau_{mnpq}\eta \qquad \mathbb{R}_{mn}=\tfrac{3}{4}F^2\tilde{g}_{mn} \qquad k=\tfrac{5}{12}F^2 \tag{5.41}$$

where η is a Killing spinor and τ_{mnpq} is the totally antisymmetric product of four seven-dimensional Dirac matrices. All fields are then fluctuations about these ground-state values and these fluctuations may be decomposed into irreducible representations of the tangent space group SO(7) of the compact manifold K. For example, the fluctuations $h_{AB}(x, y)$ of the metric may be decomposed into 1-, 7- and 27-dimensional representations, the last one being the symmetric traceless representation, and the fluctuations $a_{ABC}(x, y)$ in the 'photon' field may be decomposed into 1-, 7-, 21- and 35-dimensional representations. We may then perform the harmonic expansion (5.26) for each of these tangent space group representations. This expresses the various fluctuations in terms of a basis of these irreducible representations of G satisfying the criteria described after (5.32). We call these basis elements $Y^{N_i}(y)$ 'spherical' harmonics, where N_1, N_7, N_{27}, etc, identify representations of the tangent space group from which the harmonic arose; the harmonics are eigenfunctions of the H-invariant d'Alembertian operator

$$\Delta_y Y^{N_i}(y) = -\mu^{N_i} Y^{N_i}(y) \tag{5.42}$$

where Δ_y is the Hodge-de Rham operator. The various field equations are linearised in the fluctuations, as in (5.8), and after fixing the gauge the mass eigenstates are identified. The mass eigenvalues are specified in terms of the eigenvalues μ^{N_i} of the invariant operators on the internal space. The general solution in the case of the Freund-Rubin solution has been derived by Castellani *et al* (1984b), but in the case of the Englert solution a calculation of the bosonic spectrum is still awaited. (It is not known whether these are the *only* solutions.) To fix the numerical values of the mass eigenstates it is necessary to specify precisely which coset space G/H is being considered. In the case of the round 7-sphere the general treatment reproduces the previously found results (Biran *et al* 1983, 1984, Duff and Pope 1983). In addition to the zero modes expected in an $N=8$ supergravity theory (namely, one massless graviton, 28 massless SO(7) gauge vector bosons, 35 scalars and 35 pseudoscalars) there are an additional 294 massless scalars. It is not known whether these scalars will remain massless when quantum effects are included, since their masslessness is not obviously protected by a gauge symmetry. The results for the squashed 7-sphere were derived by Awada *et al* (1983), Bais *et al* (1983) and by Nilsson and Pope (1984). The complete bosonic spectrum for the M^{pqr} solutions (Witten 1981) of eleven-dimensional supergravity, with Freund-Rubin compactification, has been found by D'Auria and Fré (1984a, b).

5.2. *Kac-Moody symmetries*

We saw in the previous section that the Brans-Dicke scalar field $\phi^{(0)}$ of the five-dimensional Kaluza-Klein theory was massless (only) at the classical level because of the scale invariance of the *truncated* theory. This masslessness is not expected to survive beyond the classical approximation, since the scale invariance is not an invariance of the complete theory. We now wish to analyse the symmetry of the theory in four dimensions when the complete tower of massive states is retained. This derives from the *general* five-dimensional coordinate invariance. Under a general (infinitesimal) coordinate transformation

$$x^\mu \to x^\mu + \zeta^\mu(x, \theta) \tag{5.43a}$$

$$\theta \to \theta + \zeta^5(x, \theta) \tag{5.43b}$$

where

$$\zeta^A(x, \theta) = \sum_{n=-\infty}^{\infty} \zeta^{(n)A}(x) \mathrm{e}^{\mathrm{i}n\theta} \tag{5.43c}$$

with

$$\zeta^{(n)A}(x)^* = \zeta^{(-n)A}(x). \tag{5.43d}$$

As in (1.19), and (5.11), we are restricted to *periodic* variation of the coordinates, because of the topology of the ground state. Then it is easy to see that the global scale transformation (5.22) is not a symmetry of the complete theory, since to effect the transformation it is required to rescale the metric

$$\bar{g}_{AB} \to (1+2\lambda/3)\bar{g}_{AB} \tag{5.44}$$

and combine this with the general coordinate transformation in which

$$\zeta^A = \delta_5^A(-\lambda\tilde{R}\theta). \tag{5.45}$$

Clearly this coordinate transformation does *not* satisfy the periodicity requirement (5.43c), which means that $\phi^{(0)}$ is only a pseudo-Goldstone boson, as claimed in the previous section.

In ordinary four-dimensional relativity the Poincaré invariance may be regarded as a special case of the general covariance in which the infinitesimal coordinate transformations $\zeta^\mu(x)$ are restricted to the linear form

$$\zeta^\mu(x) = a^\mu + \omega^\mu{}_\nu x^\nu \tag{5.46}$$

where a^μ and $\omega_{\mu\nu} = -\omega_{\nu\mu}$ are constant. To find the analogue in our five-dimensional theory we restrict the quantities $\zeta^{(n)A}(x)$ in an analogous manner

$$\zeta^{(n)\mu}(x) = a^{(n)\mu} + \omega^{(n)\mu}{}_\nu x^\nu \tag{5.47a}$$

$$\zeta^{(n)5}(x) = c^{(n)} \tag{5.47b}$$

where $a^{(n)}$, $\omega^{(n)\mu}{}_\nu$ and $c^{(n)}$ are constants. We may determine the generators associated with these constants in the usual way. Consider, for example, a translation with all $\omega^{(n)}$, $c^{(n)}$ zero and a single non-zero $a^{(n)\mu}$, i.e.

$$x^\mu \to x^\mu + \zeta^{(n)\mu} e^{in\theta} \tag{5.48a}$$

$$\theta \to \theta. \tag{5.48b}$$

Then

$$\phi(x, \theta) \to \phi(x, \theta) + e^{in\theta}\zeta^{(n)\mu}\partial_\mu\phi \equiv (1 - i\zeta^{(n)\mu}P_\mu^{(n)})\phi. \tag{5.49}$$

Hence

$$P_\mu^{(n)} = i e^{in\theta}\partial_\mu. \tag{5.50}$$

Similarly

$$M_{\mu\nu}^{(n)} = i e^{in\theta}(x_\mu\partial_\nu - x_\nu\partial_\mu) \tag{5.51}$$

generates the Lorentz transformation parametrised by $\omega_{\mu\nu}^{(n)}$ and

$$Q^{(n)} = i e^{in\theta}\partial_\theta \tag{5.52}$$

generates translations in S^1. These generate a (non-compact) infinite parameter Lie algebra (Dolan and Duff 1984) containing the usual Poincaré algebra:

$$[P_\mu^{(n)}, P_\nu^{(m)}] = 0 \tag{5.53a}$$

$$[M_{\mu\nu}^{(m)}, P_\lambda^{(n)}] = i(\eta_{\lambda\nu}P_\mu^{(m+n)} - \eta_{\lambda\mu}P_\nu^{(m+n)}) \tag{5.53b}$$

$$[M_{\mu\nu}^{(n)}, M_{\rho\sigma}^{(m)}] = i(\eta_{\nu\rho}M_{\mu\sigma}^{(n+m)} + \eta_{\mu\sigma}M_{\nu\rho}^{(m+n)} - \eta_{\mu\rho}M_{\nu\sigma}^{(m+n)} - \eta_{\nu\sigma}M_{\mu\rho}^{(m+n)}) \tag{5.53c}$$

$$[Q^{(n)}, Q^{(m)}] = (n-m)Q^{(n+m)} \tag{5.53d}$$

$$[Q^{(n)}, P_\mu^{(m)}] = -mP_\mu^{(n+m)} \tag{5.53e}$$

$$[Q^{(n)}, M_{\mu\nu}^{(m)}] = -mM_{\mu\nu}^{(n+m)}. \tag{5.53f}$$

If we restrict our attention to the subalgebra with $n = m = 0$, we of course obtain the usual Poincaré $\otimes$U(1) algebra. In fact this finite-dimensional subalgebra may be enlarged to Poincaré $\otimes$SO(1, 2) since the generators $P_\mu^{(0)}$, $M_{\mu\nu}^{(0)}$, $Q^{(1)}Q^{(0)}$, $Q^{(-1)}$ also close. The last three give

$$[Q^{(1)}, Q^{(-1)}] = 2Q^{(0)} \tag{5.54a}$$

$$[Q^{(0)}, Q^{(1)}] = -Q^{(1)} \tag{5.54b}$$

$$[Q^{(0)}, Q^{(-1)}] = +Q^{(1)} \tag{5.54c}$$

which is the SO(1, 2) algebra (*not* SO(3)). This was noted by Salam and Strathdee (1982).

The algebra (5.53) gives the symmetry of the full four-dimensional Lagrangian, when all of the $n \neq 0$ modes are retained. It is the natural extension of the ordinary Poincaré invariance with which we are familiar. However the ground state, the particle physics vacuum, described by (5.23), only has symmetry Poincaré $\otimes$U(1), so the full (Kac-Moody) symmetry (Kac 1968, Moody 1968) is spontaneously broken. Since the restriction (5.47) makes the symmetry a global one, rather than a local one, we expect there to be Goldstone bosons generated by the symmetry breaking. To identify the Goldstone bosons we need to determine which fields are transformed inhomogeneously by the broken generators. For example in the case of the Abelian Higgs model with a complex scalar field $\phi(x)$

$$\phi(x) = (1/\sqrt{2})(\phi_1(x) + i\phi_2(x)) \tag{5.55}$$

suppose that

$$\langle\phi_1(x)\rangle = v \tag{5.56}$$

$$\langle\phi_2(x)\rangle = 0. \tag{5.57}$$

Then we write

$$\phi(x) = (1/\sqrt{2})[v + \hat{\phi}_1(x) + i\hat{\phi}_2(x)] \tag{5.58}$$

where

$$\langle\hat{\phi}_i(x)\rangle = 0 \qquad (i = 1, 2). \tag{5.59}$$

Under an infinitesimal U(1) gauge transformation

$$\phi(x) \to (1 + i\Lambda)\phi(x) = (1/\sqrt{2})[v + \hat{\phi}_1 + i\hat{\phi}_2 - \Lambda\hat{\phi}_2 + i\Lambda(v + \hat{\phi}_1)]. \tag{5.60}$$

Thus the Higgs field

$$\hat{\phi}_1 \to \hat{\phi}_1 - \Lambda\hat{\phi}_2 \tag{5.61}$$

whereas the Goldstone boson field

$$\hat{\phi}_2 \to \hat{\phi}_2 + i\Lambda(v + \hat{\phi}_1) \tag{5.62}$$

which is an inhomogeneous transformation.

In the same way we may determine the properties of the fields $h^{(n)}_{AB}$ (defined in (5.11)) appearing in the metric, under the transformations (5.47), and hence find the Goldstone modes.

To do this the metric is first written in the rescaled form

$$\bar{g}_{AB} = \phi^{-1/3}\begin{pmatrix} g_{\mu\nu} + \kappa A_\mu A_\nu & \kappa\phi A_\mu \\ \kappa\phi A_\nu & \phi \end{pmatrix} \tag{5.63}$$

and then the fields are shifted so that

$$g_{\mu\nu} = \langle g_{\mu\nu}\rangle + \hat{g}_{\mu\nu} \tag{5.64a}$$

$$A_\mu = \langle A_\mu\rangle + \hat{A}_\mu \tag{5.64b}$$

$$\phi = \langle\phi\rangle + \hat{\phi}. \tag{5.64c}$$

Under a general infinitesimal coordinate transformation the metric transforms according to

$$\bar{g}_{AB} \to \bar{g}_{AB} + \bar{g}_{CB}\partial_A\zeta^C + \bar{g}_{AC}\partial_B\zeta^C + \zeta^C\partial_C\bar{g}_{AB} \tag{5.65}$$

and from this we may determine how the fields $\hat{g}_{\mu\nu}$, $\hat{A}_\mu$ and $\hat{\phi}$ transform under the (global) transformations (5.47). In this way we can find how the component fields $\hat{g}^{(n)}_{\mu\nu}(x)$, $\hat{A}^{(n)}_\mu(x)$, $\hat{\phi}^{(n)}(x)$ transform, and hence identify the Goldstone modes. The conclusion (Dolan 1984a, b) is that the fields $A^{(n)}_\mu$ and $\phi^{(n)}$ with $n \neq 0$ *are* Goldstone modes, as anticipated in the previous section. This is why the full theory, which is invariant under *local* five-dimensional coordinate transformations, has an infinite tower of massive tensor and vector modes, as found in the last section.

There will, of course, be generalisations of the infinite parameter symmetry (5.53) which occur in the various higher-dimensional ($N > 5$) Kaluza-Klein theories which we have considered. For example in the case (5.33) that the compactified manifold K is a sphere S^2, the expansion (5.43) will be replaced by an expansion in spherical harmonics. Evidently the derivation of the analogue of (5.53) will involve products of these harmonics, and the structure constants of the infinite parameter Lie algebra will be Clebsch-Gordan coefficients. It is expected that a full understanding of these Kac-Moody symmetries will be important in determining the ultraviolet properties of the non-Abelian Kaluza-Klein theories (Dolan and Duff 1984).

Four-dimensional Kac-Moody algebras have been studied in a different context by Dolan (1984c). In this case they have been used to find exact non-perturbative solutions for the (special) theories in which they are manifest.

5.3. Fermions (Witten 1983)

We have seen in § 3 that in order to achieve compactification of the extra dimensions to a manifold K having isometry group

$$G_I \supset SU(3)\times SU(2)\times U(1)$$

it is likely that matter fields, not arising from the metric, have to be input from the start (unless we are prepared to believe that the compactification derives entirely from quantum effects). We shall see in this section that there are additional reasons why *fundamental* gauge fields, as opposed to those emerging from G_I, are necessary if we are to understand how light (on the Planck scale) fermions can arise in a Kaluza-Klein scenario.

Consider, for example, a massless spinor particle ψ in $4+D$ dimensions. The Dirac equation can be written as

$$\mathrm{i}\not{D}\psi \equiv (\mathrm{i}\gamma^{\mu}\partial_{\mu} + \mathrm{i}\Gamma^{\alpha}e_{\alpha}^{m}\nabla_{m})\psi = 0 \tag{5.66}$$

where γ^{μ} are $(2^{2+D/2}\times 2^{2+D/2})$ gamma matrices satisfying

$$\{\gamma^{\mu}, \gamma^{\nu}\} = 2\eta^{\mu\nu} \qquad (\mu, \nu = 0, 1, 2, 3) \tag{5.67}$$

with $\eta^{\mu\nu}$ as in (1.4) and the gamma matrices of the compact space satisfying

$$\{\Gamma^{\alpha}, \Gamma^{\beta}\} = 2\delta^{\alpha\beta} \qquad (\alpha, \beta = 4, \ldots, 3+D)$$

and (5.68)

$$\{\Gamma^{\alpha}, \gamma^{\mu}\} = 0.$$

∇_m is the covariant derivative defined in (2.37), so

$$\nabla_m\psi = (\partial_m - \tfrac{1}{2}\mathrm{i}\omega_m^{\alpha\beta}M_{\alpha\beta})\psi \tag{5.69}$$

with

$$M_{\alpha\beta} = \tfrac{1}{4}\mathrm{i}[\Gamma^{\alpha}, \Gamma^{\beta}] \tag{5.70}$$

being the spinor representation of the tangent space group SO(D) of the compact manifold K. Clearly

$$M \equiv -\mathrm{i}\Gamma^{\alpha}e_{\alpha}^{m}\nabla_{m} \tag{5.71}$$

plays the role of the mass operator, since its eigenvalues give the particle masses in four dimensions. However, it can be shown that M has no zero eigenvalues (Lichnerowitz 1963, Schrödinger 1932). We outline the derivation given by Zee (1981). First we write

$$M = -\mathrm{i}\Gamma^{m}\nabla_{m} \tag{5.72}$$

where

$$\Gamma^{m} = \Gamma^{\alpha}e_{\alpha}^{m} \tag{5.73}$$

so that

$$\{\Gamma^{m}, \Gamma^{n}\} = 2\delta^{\alpha\beta}e_{\alpha}^{m}e_{\beta}^{n} = 2\tilde{g}^{mn}. \tag{5.74}$$

Squaring M gives

$$\begin{aligned}
-M^2\psi &= \Gamma^{m}\nabla_{m}\Gamma^{n}\nabla_{n}\psi \\
&= \Gamma^{m}\Gamma^{n}\nabla_{m}\nabla_{n}\psi + \Gamma^{m}[\nabla_{m}, \Gamma^{n}]\nabla_{n}\psi \\
&= (\tfrac{1}{2}\{\Gamma^{m}, \Gamma^{n}\} + \tfrac{1}{2}[\Gamma^{m}, \Gamma^{n}])\nabla_{m}\nabla_{n}\psi \\
&= \tilde{g}^{mn}\nabla_{m}\nabla_{n}\psi + \tfrac{1}{4}[\Gamma^{m}, \Gamma^{n}][\nabla_{m}, \nabla_{n}]\psi \\
&= \nabla^{m}\nabla_{m}\psi - \tfrac{1}{2}M^{mn}M^{\alpha\beta}R_{\alpha\beta mn}\psi \\
&= \nabla^{m}\nabla_{m}\psi - \tfrac{1}{4}\tilde{\mathbb{R}}\psi.
\end{aligned} \tag{5.75}$$

In deriving this we have used the identities

$$[\nabla_{m}, \Gamma^{n}]\nabla_{n}\psi = 0$$

$$[\nabla_{m}, \nabla_{n}]\psi = -\tfrac{1}{2}\mathrm{i}M^{\alpha\beta}R_{\alpha\beta mn}\psi \tag{5.76}$$

as well as the well known cyclic properties of the Riemann tensor. $\tilde{\mathbb{R}}$ is the curvature scalar of K. The first term of (5.75) is not positive, since $\nabla_m \tilde{g}^{mn} = 0$ so

$$\int \mathrm{d}^D y \sqrt{\tilde{g}}\, \psi^\dagger \tilde{g}^{mn} \nabla_m \nabla_n \psi = -\int \mathrm{d}^D y \sqrt{\tilde{g}} (\nabla_m \psi^\dagger) \tilde{g}^{mn} (\nabla_n \psi). \tag{5.77}$$

Also $\tilde{\mathbb{R}}$ is positive for a (closed) compact space, so it follows that M^2 has only positive eigenvalues. Furthermore, we expect these non-zero eigenvalues to be of order $\tilde{R}^{-2}$, with $\tilde{R}$ the characteristic size of the compact manifold, and from (5.2) this means that only fermions with masses of order m_{P} are allowed. Of course we could instead have considered a *massive* spinor in $4+D$ dimensions, and then adjusted its mass to give any desired mass in four dimensions. This requires a fine tuning to at least one part in 10^{19} to account for the masses of the fermions which we observe in nature, and this is generally regarded as a most unattractive explanation. A much preferred scenario is that the observed fermions are in fact zero modes of the Dirac operator, and that their small non-zero masses arise from physics at a much lower (possibly electroweak) energy scale. (There may still be a measure of fine tuning, but nothing like so much as is needed at the Planck scale.)

If the observed fermions really are zero modes, it follows that we must change some of the (tacit) assumptions made in deriving Lichnerowitz's theorem. One possibility, which has been explored by Destri *et al* (1983) and by Wu and Zee (1984), is to introduce torsion on the internal manifold K. This modifies (5.75) so that

$$-M^2 \psi = \{\nabla^m \nabla_m - \tfrac{1}{2} M^{mn} M^{\alpha\beta} R_{\alpha\beta mn} + \mathrm{i} M^{mn} T^k_{mn} \nabla_k\} \psi \tag{5.78}$$

where

$$T^k_{mn} \equiv \Gamma^k_{mn} - \Gamma^k_{nm} \tag{5.79}$$

is non-zero when the Christoffel connection is not symmetric in its lower indices; in the form language of § 2.5 the torsion 2-form is defined by

$$T^\alpha \equiv \mathrm{d}e^\alpha + \omega^\alpha{}_\beta \wedge e^\beta. \tag{2.35}$$

The first term of (5.78) is still not positive, but the remaining terms can have either sign, so zero modes are possible. Incidentally, the cyclic properties of the Riemann tensor are not valid in the presence of torsion so we cannot reduce the second term to the scalar curvature. Unfortunately the explorations of Wu and Zee have shown that this possible escape route is also unattractive. First there is the arbitrariness of precisely how the torsion should be introduced. These authors study only the case when K is a group manifold, so the torsion is parallelisable, as proposed by Cartan and Schouten (1926). (Parallelisable means that the curvature tensor, constructed from g_{mn} and Γ^k_{mn}, vanishes.) They find that there are large numbers of zero modes (1024 in the case of the SU(5) manifold, for example) but these do not appear in the representations which appear to be inhabited by the known fermions. This high dimensionality is related to the high dimensionality of the manifold, so a better bet might be to look at homogeneous coset spaces G/H, instead of G. In any case there is no compelling reason to insist upon a parallelisable torsion. It is therefore possible that the unsatisfactory results so far obtained could be improved upon given sufficient ingenuity. However even if this could be achieved, the massless fermions so obtained would most likely *not* have the quantum numbers of the fermions actually observed in nature.

Suppose, for example, we take the manifold K to be a homogeneous coset space

$$K = \mathrm{G}/\mathrm{H} \tag{5.80}$$

as in (5.27), with G and H chosen so that K is one of the M^{pqrst} spaces discussed in § 2.6. Then

$$G = G_1 = SU(3)_C \times SU(2)_L \times U(1)_Y \tag{5.81a}$$

and

$$H = SU(2)_{C'} \times U(1)_{X'} \times U(1)_{X''} \tag{5.81b}$$

with X' and X'' as in (2.63) and (2.64). Then we may decompose the various representations of G which are observed in nature in terms of representations of H, as (5.29). Thus

$$e_R = (\underline{1}, -s, -t) \tag{5.82a}$$

$$E_L = (\underline{1}, -\tfrac{1}{2}s, \tfrac{1}{2} - \tfrac{1}{2}t) + (\underline{1}, -\tfrac{1}{2}s, -\tfrac{1}{2} - \tfrac{1}{2}t) \tag{5.82b}$$

$$d_R = (\underline{2}, \tfrac{1}{2} - \tfrac{1}{3}s, -\tfrac{1}{3}t) + (\underline{1}, -1 - \tfrac{1}{3}s, -\tfrac{1}{3}t) \tag{5.82c}$$

with similar expansions for u_R and the quark doublet Q_L. The embedding of H in the tangent space group SO(7) of K is given by (2.56). So, in an obvious notation, the diagonal generators of H are

$$Q(\tfrac{1}{2}\lambda^3) = \tfrac{1}{2}(M^{12} - M^{34}) \tag{5.83a}$$

$$Q(X') = \tfrac{3}{2}(M^{12} + M^{34}) \tag{5.83b}$$

$$Q(X'') = M^{56} \tag{5.83c}$$

where $M^{\alpha\beta}$ ($\alpha, \beta = 1, \ldots, 7$) are the generators of SO(7). The representations of the full tangent space group SO(1, 10) may be expanded in terms of the decomposition (5.30). In particular it is clear that the spinor representation decomposes into a product of spinors of SO(1, 3) and of SO(7). Similarly a vector-spinor of SO(1, 10) (which transforms as a product of vector and spinor representations) decomposes into products of an SO(1, 3) vector-spinor with an SO(7) spinor, and an SO(1, 3) spinor with an SO(7) vector-spinor. If we restrict ourselves to particles having spin less than 2, this means that (the observed) SO(1, 3) spinors must be associated with spinors or vector-spinors of SO(7) (de Alwis and Koh (1984). Now, the spinor representation of SO(7) is generated by the 8×8 gamma matrices (see, for example, Rajpoot 1980)

$$\begin{aligned} &\Gamma^1 = \sigma_1 \times 1 \times 1 && \Gamma^2 = \sigma_2 \times 1 \times 1 && \\ &\Gamma^3 = \sigma_3 \times \sigma_1 \times 1 && \Gamma^4 = \sigma_3 \times \sigma_2 \times 1 && \\ &\Gamma^5 = \sigma_3 \times \sigma_3 \times \sigma_1 && \Gamma^6 = \sigma_3 \times \sigma_3 \times \sigma_2 && \Gamma^7 = \sigma_3 \times \sigma_3 \times \sigma_3. \end{aligned} \tag{5.84}$$

Then

$$M^{\alpha\beta} \equiv \tfrac{1}{4}\mathrm{i}[\Gamma^\alpha, \Gamma^\beta] \tag{5.85}$$

and it is easy to see, using (5.83), that this spinor representation may be expanded in terms of representations of H as

$$|8-\text{spinor}\rangle = |\underline{2}, 0, \tfrac{1}{2}\rangle + |\underline{2}, 0, -\tfrac{1}{2}\rangle + |\underline{1}, \tfrac{3}{2}, \tfrac{1}{2}\rangle + |\underline{1}, \tfrac{3}{2}, -\tfrac{1}{2}\rangle + |\underline{1}, -\tfrac{3}{2}, \tfrac{1}{2}\rangle + |\underline{1}, -\tfrac{3}{2}, -\tfrac{1}{2}\rangle. \tag{5.86}$$

As explained in (5.32), in order for a particular (fermion) representation (5.82) of G to occur in the harmonic expansion of the spinor representation of SO(7), there must be at least one overlap between the expansion (5.82) and (5.86). Thus if we demand that e_R arises in the harmonic expansion then

$$|t| = \tfrac{1}{2} \tag{5.87a}$$

$$|s| = \tfrac{3}{2}. \tag{5.87b}$$

But then E_L does *not* occur in the harmonic expansion, and neither does d_R, or indeed any of the other fermion representations. Including the vector–spinor representation of SO(7) does not alleviate the problem (Randjbar-Daemi *et al* 1984b). It *is* possible to meet this objection by enlarging the manifold K. For example, taking (Bailin and Love 1985a)

$$\mathrm{H} = \mathrm{SU}(2) \times \mathrm{U}(1)_Z \tag{5.88}$$

with Z as in (2.60), does allow the observed fermion quantum numbers to arise in the expansion of the SO(8) spinor, provided

$$-3p = \pm q = r \tag{5.89}$$

or in the vector–spinor if

$$3p = \pm q = r \qquad \text{or} \qquad 3p = \pm 3q = \pm r. \tag{5.90}$$

However the choice (5.88) gives an isometry group $\mathrm{G_I}$ larger than G, as in (5.28); in fact

$$\mathrm{G_I} = \mathrm{SU}(3) \times \mathrm{SU}(2) \times \mathrm{U}(1) \times \mathrm{U}(1) \tag{5.91}$$

in this case, so there is an additional (hitherto) unobserved neutral gauge boson associated with the extra U(1).

However, there are further objections to the programme we have pursued, quite apart from the difficulty of arranging for masslessness and the observed fermion quantum numbers. Even when we *are* able to satisfy these criteria there remains the objection that the resulting fermions are not *chiral* fermions. By this we mean that the left-chiral components of the (observed) fermion fields transform differently, with respect to SU(3) × SU(2) × U(1), from the right-chiral components. In other words the quantum numbers do not appear to be vector-like; for example the left-handed quark fields transform as

$$Q_L = (\underline{3}, \underline{2}, \tfrac{1}{6}) \tag{5.92}$$

whereas the right-handed quarks transform as

$$u_R = (\underline{3}, \underline{1}, \tfrac{2}{3}) \tag{5.93a}$$

$$d_R = (\underline{3}, \underline{1}, -\tfrac{1}{3}). \tag{5.93b}$$

Equivalently the left components of the antiquarks transform as

$$u_L^C = (\underline{\bar{3}}, \underline{1}, -\tfrac{2}{3}) \tag{5.94a}$$

$$d_L^C = (\underline{\bar{3}}, \underline{1}, \tfrac{1}{3}). \tag{5.94b}$$

Thus the fermions of a given helicity transform as a *complex* representation of SU(3) × SU(2) × U(1); the right-chiral components transform according to the complex conjugate of the left-chiral representation, and the two representations are inequivalent. Of course, it is always possible that future high-energy experiments will discover new 'mirror' fermion states $\tilde{u}$ and $\tilde{d}$ transforming as

$$\tilde{Q}_L = (\underline{\bar{3}}, \underline{2}, -\tfrac{1}{6}) \tag{5.95a}$$

and

$$\tilde{u}_L^C = (\underline{3}, \underline{1}, \tfrac{2}{3}) \tag{5.95b}$$

$$\tilde{d}_L^C = (\underline{3}, \underline{1}, -\tfrac{1}{3}) \tag{5.95c}$$

and that the quantum numbers of the complete fermion generation are vector-like after all. This seems unlikely for several reasons (Witten 1983). First, it is remarkable that none of the (fourteen multiplets of) observed fermions are the mirror partners of others. Second, the well known cancellation of the Adler-Bardeen γ_5 anomalies within each of the three known fermion generations would then be pure coincidence, since vector-like theories are guaranteed to be anomaly free; the mirror states would automatically have cancelled any non-zero anomaly generated by the known fermions.

We shall see that the requirement that the left-chiral zero modes form a *complex* representation of the symmetry group is extremely restrictive. For instance, the dimensionality of spacetime cannot be odd. This is because in odd dimensionality there is a *unique* spinor representation. (This is illustrated for SO(7) in (5.84). The product of all of the Γ matrices $\mathrm{i}\Gamma^1 \ldots \Gamma^7 = I_8$, whereas in four dimensions the product $\mathrm{i}\gamma^0\gamma^1\gamma^2\gamma^3 \equiv \gamma^5 \neq I_4$, and permits the decomposition of the spinor into left- and right-chiral pieces.) This means that $SO(1, 3+D)$ has a unique spinor representation if D is odd, which transforms under $SO(1,3)\times SO(D)$ as the product of a (four-component) spinor of SO(1, 3) with the unique spinor of $SO(D)$. Thus when we perform the harmonic expansion (5.26) fermions which are left- or right-handed in four dimensions transform in the same way with respect to the isometry group, and so furnish a *real* representation.

Next, suppose that D is even and

$$D = 2N. \tag{5.96}$$

Then defining Γ matrices as in (5.84), we find that

$$\Gamma \equiv \Gamma^1\Gamma^2 \ldots \Gamma^D = \mathrm{i}^N \sigma_3 \times \sigma_3 \times \ldots \times \sigma_3 \tag{5.97}$$

anticommutes with all of the matrices $\Gamma^1, \Gamma^2, \ldots, \Gamma^D$. The gamma matrices for the full tangent space group $SO(1, 3+D)$ are given by

$$\bar{\Gamma}^A = \gamma^\mu \times I \qquad (A = \mu = 0, 1, 2, 3) \tag{5.98a}$$

$$= \mathrm{i}\gamma^5 \times \Gamma^{A-3} \qquad (A = 4, \ldots, D+3). \tag{5.98b}$$

(The reason for the γ^5 is so that the gamma matrices associated with M_4 *anticommute* with those of K.) In the full $(4+D)$-dimensional space the 'chirality' χ is defined by

$$\begin{aligned} \chi &= \bar{\Gamma}^0\bar{\Gamma}^1\bar{\Gamma}^2 \ldots \bar{\Gamma}^{D+3} \\ &= -\mathrm{i}(-1)^N \gamma^5 \times \Gamma \end{aligned} \tag{5.99}$$

and χ anticommutes with all of the matrices $\bar{\Gamma}^0, \ldots, \bar{\Gamma}^{D+3}$, since γ^5 and Γ do in SO(1, 3) and $SO(D)$ respectively. Thus, χ commutes with all of the generators M^{AB} of $SO(1, 3+D)$, and can be used to label inequivalent spinor representations. Since

$$\begin{aligned} (\chi)^2 &= -(\gamma^5)^2 \times (\Gamma)^2 \\ &= -1 \qquad (N = \text{even}) \\ &= +1 \qquad (N = \text{odd}) \end{aligned} \tag{5.100}$$

the eigenvalues of χ are

$$\chi = \pm \mathrm{i} \qquad (N = \text{even}) \tag{5.101a}$$

$$= \pm 1 \qquad (N = \text{odd}). \tag{5.101b}$$

Either way, the important thing is that for fixed χ the four-dimensional chirality γ_5 is correlated with the internal chirality Γ; so for N odd, for example, and $\chi = \pm 1$, we have spinor representations with

$$\gamma_5 = +1 \qquad \Gamma = -\mathrm{i} \tag{5.102a}$$

or

$$\gamma_5 = -1 \qquad \Gamma = +\mathrm{i} \tag{5.102b}$$

using (5.99). Thus fermions having left-handed physical chirality have 'internal' chirality $+\mathrm{i}$, and satisfy a different Dirac equation from the right-handed fermions. Thus the zero modes *might* have different quantum numbers. Actually the only possibility of this happening is if N is *odd*. This is because for N even the complex conjugate of the spinor representation with $\chi = +\mathrm{i}$ is equivalent to the spinor representation with $\chi = -\mathrm{i}$.

The situation is as follows. (5.99) shows that for N even

$$\chi = +\mathrm{i} \Rightarrow \gamma_5 = 1 \qquad \Gamma = -1 \tag{5.103a}$$

or

$$\gamma_5 = -1 \qquad \Gamma = +1 \tag{5.103b}$$

while

$$\chi = -\mathrm{i} \Rightarrow \gamma_5 = 1 \qquad \Gamma = 1 \tag{5.103c}$$

or

$$\gamma_5 = -1 \qquad \Gamma = -1. \tag{5.103d}$$

The complex conjugate of a $\gamma_5 = 1$ spinor is a $\gamma_5 = -1$ spinor, as is usual in SO(1, 3). However the complex conjugate of the $\Gamma = -1$ spinor of SO($2N$) is equivalent to itself, for *N even*. (The difference arises because of the signature of SO(1, 3). For a very clear discussion of this and other details see Chadha and Daniel (1985a, b) or Gourdin (1982).) Thus the complex conjugate of (5.103a) is equivalent to (5.103d), and similarly for (5.103b, c). Evidently there is no net correlation between the physical chirality and the internal chirality, since the $\chi = +\mathrm{i}$ field is equivalent to the complex conjugate of the $\chi = -\mathrm{i}$ field, and there will be equal numbers of left- and right-chiral fields having given internal helicity.

In the case when N is odd the situation is quite different. Equation (5.99) shows that for N odd

$$\chi = +1 \Rightarrow \gamma_5 = 1 \qquad \Gamma = -\mathrm{i} \tag{5.104a}$$

or

$$\gamma_5 = -1 \qquad \Gamma = +\mathrm{i} \tag{5.104b}$$

while

$$\chi = -1 \Rightarrow \gamma_5 = 1 \qquad \Gamma = \mathrm{i} \tag{5.104c}$$

or

$$\gamma_5 = -1 \qquad \Gamma = -\mathrm{i}. \tag{5.104d}$$

The complex conjugate of the $\Gamma=-\mathrm{i}$ spinor is equivalent to the $\Gamma=+\mathrm{i}$ spinor when N is *odd*. So the complex conjugate of (5.104*a*) is equivalent to (5.104*b*). This is just what we need; a zero mode of the internal Dirac operator with $\Gamma=+\mathrm{i}$ corresponds to a left-handed massless fermion in four dimensions, and its complex conjugate, having $\Gamma=-\mathrm{i}$, corresponds to a right-handed massless fermion, provided we restrict ourselves to a theory in which only $\chi=+1$ representations appear. In any case, the $\chi=-1$ representations are unrelated to the $\chi=+1$ representations. The special importance for fermions of theories with $4n+2$ dimensions has been emphasised by Brink *et al* (1977), Wetterich (1981, 1982a, b, c, d, 1983) and by Randjbar-Daemi *et al* (1983a, b).

This is all very well, but we still have to arrange that the zero modes of the Dirac operator having $\Gamma=+\mathrm{i}$ form a *complex* representation of the symmetry group. Equivalently, we must arrange that the $\Gamma=+\mathrm{i}$ zero modes transform differently from the $\Gamma=-\mathrm{i}$ zero modes. Unfortunately, the possibility of achieving this, even in $4n+2$ dimensions, is severely constrained by a theorem of Atiyah and Hirzebruch (1970). One consequence of this theorem is that, in the absence of elementary gauge fields, the zero modes of the Dirac operator form a *real* representation (in any even number of dimensions). Thus we are forced to introduce gauge fields if there is to be any chance of arranging for the zero modes to form a complex representation.

Following Witten (1983) we define

$$\hat{\Gamma}\equiv\mathrm{i}\Gamma \tag{5.105}$$

so that $\hat{\Gamma}$ has eigenvalues ±1 when N is odd. Then using the definition of M in (5.71), it follows that $\hat{\Gamma}$ anticommutes with M and

$$[\hat{\Gamma}, M^2]=0. \tag{5.106}$$

Thus eigenstates of M^2 can be chosen to be eigenstates of $\hat{\Gamma}$. Now suppose ψ is an eigenstate of M^2 with eigenvalue E:

$$M^2\psi=E\psi. \tag{5.107}$$

Then

$$\begin{aligned} M^3\psi &= M^2(M\psi)\\ &= M(E\psi)=E(M\psi) \end{aligned} \tag{5.108}$$

which shows that the states ψ and $M\psi$ are degenerate, unless $M\psi=0$. Furthermore, since $\hat{\Gamma}$ anticommutes with M, ψ and $M\psi$ have opposite eigenvalues of $\hat{\Gamma}$. Thus the eigenfunctions with non-zero eigenvalues ($E\neq0$) are paired: for every ψ with an eigenvalue of $\hat{\Gamma}=+1$ there is another state with $\hat{\Gamma}=-1$. The zero modes, satisfying

$$M\psi=0 \tag{5.109}$$

are not necessarily paired in this way, and this leads to the notion of an 'index'. The index of M is defined as the number of zero modes of M with $\hat{\Gamma}=+1$ minus the number with $\hat{\Gamma}=-1$. Evidently the index of M is a topological invariant; since it is integer valued, no smooth deformation of the manifold can alter it. In the case of $4n+2$ dimensions (i.e. $D=2N$ with $N=2n+1$) the index of M is zero, since the $\Gamma=-\mathrm{i}$ ($\hat{\Gamma}=+1$) and then $\Gamma=+\mathrm{i}$ ($\hat{\Gamma}=-1$) zero modes are just complex conjugate representations, as we have just observed. What we require to be non-zero is the 'character-valued' index of M. This is defined when the compact manifold K has an isometry group $\mathrm{G_I}$, so that the eigenvectors of M^2 belong to representations of $\mathrm{G_I}$. If R is a representation of $\mathrm{G_I}$, then

$$\operatorname{index}_R(M)\equiv n_+(R)-n_-(R) \tag{5.110}$$

where $n_{\pm}(R)$ is the number of $\hat{\Gamma}=\pm 1$ zero modes, belonging to the representation R. For a complex representation the index of M in the complex conjugate representation $\bar{R}$ is

$$\begin{aligned}\text{index}_{\bar{R}}(M) &= n_+(\bar{R})-n_-(\bar{R}) \\ &= n_-(R)-n_+(R) \\ &= -\text{index}_R(M)\end{aligned} \tag{5.111}$$

in $4n+2$ dimensions. (This is consistent with the fact that the number of left-handed quarks (for example) in the representation R is equal to the number of right-handed antiquarks in $\bar{R}$.)

The Atiyah–Hirzebruch theorem states that in the absence of gauge fields $\text{index}_R(M)$ is zero, for any manifold with a continuous isometry group (in any even number of dimensions). We shall not prove the theorem, but we can verify its validity in a particularly simple example (Witten 1983). We consider the $D=2$ case that K is a 2-sphere. So, as in (5.33)

$$K=S^2=\text{SO}(3)/\text{SO}(2)\simeq \text{SU}(2)/\text{U}(1) \tag{5.112}$$

and

$$\text{G}_\text{I}=\text{SO}(3). \tag{5.113}$$

Suppose we have a spinor representation of the tangent space of K:

$$\text{G}_\text{T}=\text{SO}(2). \tag{5.114}$$

The representations of G_I are labelled by the eigenvalues of the operator $\boldsymbol{J}^2$, where $\boldsymbol{J}$ is the angular momentum operator. We want to calculate the number of zero modes of M having $\hat{\Gamma}=\pm 1$, and belonging to a particular representation $|\boldsymbol{J}_1\rangle$ of G_I where

$$\boldsymbol{J}^2|\boldsymbol{J}_1\rangle = J_1(J_1+1)|\boldsymbol{J}_1\rangle. \tag{5.115}$$

We have already derived the technology needed to answer this question in (5.26) and the following equations, since K is indeed a (homogeneous) coset space. It amounts to calculating how many times the representation $|\boldsymbol{J}_1\rangle$ occurs in the harmonic expansion of the SO(2) 2-spinor. We need to expand both the 2-spinor and $|\boldsymbol{J}_1\rangle$ in terms of the representations of the $\text{SO}(2)\simeq\text{U}(1)$ group. Clearly

$$|\text{2-spinor}\rangle = |+\tfrac{1}{2}\rangle + |-\tfrac{1}{2}\rangle \tag{5.116}$$

and

$$|\boldsymbol{J}_1\rangle = |J_1\rangle + |J_1-1\rangle + \ldots + |-J_1\rangle. \tag{5.117}$$

The condition for overlap between these two expansions is simply that J_1 is half-odd integral and, if so,

$$n_+(J_1)=1=n_-(J_1) \tag{5.118}$$

since $|\pm\tfrac{1}{2}\rangle$ has $\hat{\Gamma}=\text{i}\Gamma=-\sigma_3=\mp 1$. Thus

$$\text{index}_{J_1}(M)=0 \tag{5.119}$$

for all J_1, since for integral J_1 $n_{\pm}(J_1)=0$ in any case. Although this is consistent with the theorem of Atiyah and Hirzebruch, it was of course a foregone conclusion, since for a sphere $\tilde{\mathbb{R}}$ is positive and M^2 has only positive eigenvalues by the Lichnerowitz theorem, as we have already demonstrated.

We can see how the existence of background gauge fields *might* permit us to evade the Lichnerowitz theorem. Suppose that the massless spinor particle ψ discussed in (5.66) onwards belongs to some non-trivial representation of a local gauge symmetry group G_{YM}. Then the Dirac equation (5.11) becomes

$$[\gamma^{\mu}(\mathrm{i}\partial_{\mu} - gt^{a}A^{a}_{\mu}) + \mathrm{i}\Gamma^{\alpha}e_{\alpha}{}^{m}D_{m}]\psi = 0 \tag{5.120a}$$

where

$$D_{m}\psi = (\partial_{m} - \tfrac{1}{2}\mathrm{i}\omega^{\alpha\beta}_{m}M_{\alpha\beta} - \mathrm{i}gt^{a}A^{a}_{m})\psi \tag{5.120b}$$

is the gauge and gravitational covariant derivative. The matrices t^{a} represent G_{YM} in the representation to which ψ belongs, and A^{a}_{m} are the gauge fields. (In the case of a U(1) gauge symmetry t^{a} is just the 'charge' of ψ in units of g.) If we choose

$$A^{a}_{\mu} = 0 \tag{5.121}$$

then we are guaranteed to preserve the (four-dimensional) Poincaré invariance of the vacuum, but we must choose some of the A^{a}_{m} to be non-zero so as to arrange some sort of cancellation between the spin connection and the gauge connection contributions to D_{m}, thereby (it is to be hoped) evading Lichnerowitz's theorem. However the choice of the non-zero components of A^{a}_{m} is severely constrained by the requirement that the gauge field configuration preserves the symmetry group G_{I} associated with the isometries of the compact space K. In other words, we require the background gauge fields to be solutions of the (Yang-Mills) field equations which are invariant under G_{I}. Further, since we wish to have chiral zero modes, the background gauge fields must be in a topologically non-trivial configuration, so as to evade the Atiyah-Hirzebruch theorem; we have already observed that the (character-valued) $\mathrm{index}_{R}(M)$ is a topological quantity, so if it is to be non-zero because of the existence of A^{a}_{m}, it is clear that these gauge fields must not be continuously deformable to zero.

We have already discussed gauge fields with such non-trivial topology in §§ 3.5, 3.6 and 3.7. The best known example is provided by the Dirac monopole and its use in the previous simple example (5.112):

$$K = S^{2} = \mathrm{SO}(3)/\mathrm{SO}(2) \simeq \mathrm{SU}(2)/\mathrm{U}(1)$$

was discussed first by Randjbar-Daemi *et al* (1983a, d). The gauge field 1-form is given in (3.38) or (3.45). In this case it follows from (5.70) that

$$M^{12} = -\tfrac{1}{2}\sigma_{3} \tag{5.122}$$

and from (2.50) that

$$\omega^{12} = -e^{\bar{3}} \tag{5.123}$$

so the gauge and gravitational covariant derivative of a spinor ψ with charge q is

$$D_{m}\psi = (\partial_{m} - \tfrac{1}{2}\mathrm{i}\sigma_{3}e^{\bar{3}}_{m} - \mathrm{i}qae^{\bar{3}}_{m})\psi. \tag{5.124}$$

Evidently the effect of the coupling to the U(1) gauge field is to change the *effective* H = U(1) content of the 2-spinor. Instead of (5.116), the presence of the monopole now means that we have

$$|\text{2-spinor}\rangle = |\tfrac{1}{2} + qa\rangle + |-\tfrac{1}{2} + qa\rangle. \tag{5.125}$$

The well known charge quantisation condition for the monopole gives

$$qa = \tfrac{1}{2}n \tag{5.126}$$

for some integer n, and this can be proved by parallel transporting the spinor around the 2-sphere. Thus we again have overlap with (5.117) for suitable values of J_1. Assuming $qa > 0$ we have

$$n_+(J_1) = 1 \qquad \text{for } J_1 = qa - \tfrac{1}{2}, qa + \tfrac{1}{2}, qa + \tfrac{3}{2}, \ldots \tag{5.127a}$$

$$n_-(J_1) = 1 \qquad \text{for } J_1 = \qquad qa + \tfrac{1}{2}, qa + \tfrac{3}{2}, \ldots . \tag{5.127b}$$

In other words there is a state with $J_1 = qa - \frac{1}{2}$ for the *positive* chirality spinor, but *not* for the negative chirality spinor, and the character-valued index is non-zero for $J_1 = qa - \frac{1}{2}$:

$$n_+(qa - \tfrac{1}{2}) - n_-(qa - \tfrac{1}{2}) = 1 \tag{5.128}$$

$$n_+(J_1) - n_-(J_1) = 0 \qquad J_1 \neq qa - \tfrac{1}{2}. \tag{5.129}$$

Thus in the harmonic expansion of a spinor field on the 2-sphere there appears a $\hat{\Gamma} = +1$ zero mode ψ with 'isospin' $qa - \frac{1}{2}$. This state is a zero mode because if $i\not{D}\psi$ is non-zero it has $\hat{\Gamma} = -1$ *and* $J_1 = qa - \frac{1}{2}$, and no such state exists.

In summary, then, the general technique is to choose a topologically non-trivial configuration of the background gauge fields, such as those discussed in § 3, which is invariant (up to a gauge transformation) under the action of the isometry group G_I. In the case of a coset space

$$K = G/H \tag{5.130}$$

such G_I-invariant solutions exist if either $H \subseteq G_{YM}$, or if $H \supseteq G_{YM}$ but both H and G_{YM} have a common non-trivial normal subgroup (Randjbar-Daemi 1983). This changes the effective H content of the SO($2N$) spinor (more generally of the representations R of the tangent space group G_T) in its expansion (5.31), and allows for non-zero values of the character-valued index. By a judicious choice of the parameters (e.g. monopole strengths and instanton numbers) it is then possible to arrange for chiral zero modes in the desired representations of the isometry group G_I.

The most comprehensive study of the complete fermion spectrum, not just the zero modes, has been carried out by Schellekens (1985a, b). He has expressed the eigenvalues of the fermion mass-squared operator M^2, in the presence of a general instanton background gauge field configuration on a symmetric coset space G/H, in terms of the Casimir invariants of G and H. For massless fermions the problem is that given a fermion transforming according to some representation Γ of $H \subset G_{YM}$ (this determines the coupling to the generalised instanton background field) to determine in which representation Γ of G the zero modes occur. This has been done in general for the hyperspheres (S^D) and the complex projective planes (CP^N). In the former case when D is even, and the fermion is in an irreducible representation of $H = SO(D)$, then the zero modes form a (known) irreducible representation of $G = SO(D+1)$. This generalises a previously known result (Randjbar-Daemi *et al* 1983c) for the case $D = 4$.

Watamura (1983, 1984) has also provided a general treatment of compactification in the case $K = CP^N$ in the presence of a monopole U(1) gauge field, and has shown how for a particular choice of the monopole charge, namely $ec = (N+3)/4$, the fermionic zero modes belong to the fundamental representation of $G = SU(N+1)$. We have already observed that SO(1, 3) spinors must be associated with spinors or vector-spinors of the tangent space SO(D) of K, so it is worth studying the zero modes associated with vector-spinors on CP^N, as well as spinors. This has been done by Bailin and Love (1985d) (in the presence of a background monopole field). In particular we have addressed the problem of obtaining the observed fermion quantum numbers. We find that the twelve-dimensional ($D=8$) theories envisaged in (5.88) do *not* admit chiral fermions, so these models will require the existence of mirror fermions, if they are to have any chance of being realistic. For the case $K = CP^4$ which has $G = SU(5)$, we can obtain the 'observed' $\underline{\bar{5}}$ as a zero mode from either a left-handed spinor or a left-handed vector-spinor, while the $\underline{10}$ is a zero mode only of a left-handed vector-spinor.

The use of the third possible configuration of gauge fields (3.45), in which the gauge connection and spin connection contribute to the covariant derivative (5.120) of the spinor field, evidently requires the embedding

$$SO(D) \subset G_{YM}. \tag{5.131}$$

We shall discuss the use of this mechanism in connection with anomaly cancellation in the next section. It has been used recently in the context of superstring theories, in which the compact manifold has no isometry group. The Euler characteristic of K determines the number of fermion generations (Candelas *et al* 1985). Also Schellekens (1985a, b) has studied the boson spectrum in the case $K = CP^N$ and $G_{YM} = SU(N) \times U(1)$.

5.4. *Anomalies*

We have discussed at length in the previous section the difficulty of arranging that the fermions in our Kaluza-Klein theory are *chiral* fermions, as seems to be required by the fermions observed in nature. The essence of the chiral property is that the gauge bosons couple differently to left- and right-helicity fermions, since the fermions of a given helicity transform as a *complex* representation of the gauge group. It is well known that in general this chiral feature generates 'anomalies', at least in four-dimensional gauge theories (Adler 1969, Bell and Jackiw 1969, Bardeen 1969). The effect of anomalies is that quantum effects lead to a breakdown in the local gauge invariance which is imposed on the classical (tree-level) theory. This is most easily appreciated in the functional approach to quantum field theory (Fujikawa 1979, 1980a, b). In this, the generating function for Green functions is obtained by a path integral over all classical field configurations, and for fermions these fields are Grassmann variables ψ. Now consider, for example, an infinitesimal local U(1) gauge transformation in which the left- and right-chiral components of ψ transform differently:

$$\psi_{\substack{L\\R}}(x) \to [1 + i\theta_{\substack{L\\R}}(x)]\psi_{\substack{L\\R}}(x) \tag{5.132a}$$

where

$$\theta_R - \theta_L \neq 0. \tag{5.132b}$$

Then evidently

$$\psi(x) \to \psi + i(\theta_L\psi_L + \theta_R\psi_R) = [1 + \tfrac{1}{2}i(\theta_L + \theta_R) + \tfrac{1}{2}i(\theta_R - \theta_L)\gamma_5]\psi(x) \tag{5.133a}$$

and

$$\bar{\psi}(x) \to \bar{\psi}(x)[1 - \tfrac{1}{2}i(\theta_L + \theta_R) + \tfrac{1}{2}i(\theta_R - \theta_L)\gamma_5]. \tag{5.133b}$$

It follows that the fermionic functional integration measure transforms as

$$\mathscr{D}\psi \to \{1 - \mathrm{Tr}[\tfrac{1}{2}\mathrm{i}(\theta_\mathrm{L} + \theta_\mathrm{R}) + \tfrac{1}{2}\mathrm{i}(\theta_\mathrm{R} - \theta_\mathrm{L})\gamma_5]\}\mathscr{D}\psi$$
$$\mathscr{D}\bar{\psi} \to \{1 - \mathrm{Tr}[-\tfrac{1}{2}\mathrm{i}(\theta_\mathrm{R} + \theta_\mathrm{L}) + \tfrac{1}{2}\mathrm{i}(\theta_\mathrm{R} - \theta_\mathrm{L})\gamma_5]\}\mathscr{D}\bar{\psi}. \tag{5.134}$$

(The *inverse* of the determinant appears because ψ is a Grassmann variable.) Hence

$$\mathscr{D}\psi\mathscr{D}\bar{\psi} \to \{1 - \mathrm{Tr}[\mathrm{i}(\theta_\mathrm{R} - \theta_\mathrm{L})\gamma_5]\}\mathscr{D}\psi\mathscr{D}\bar{\psi} \tag{5.135}$$

which means that the measure is *not* gauge invariant since $\theta_\mathrm{R} \neq \theta_\mathrm{L}$. In consequence the *quantum* field theory of what was a perfectly well defined classical theory is meaningless; S-matrix elements will depend on the gauge-fixing parameters through the appearance of unphysical poles in physical amplitudes, and unitarity will be violated. Thus theories with anomalies are not even quantisable, and cannot be entertained as candidates to describe reality. It is therefore of paramount importance to construct theories in which the anomalies cancel. Since the appearance of anomalies is due specifically to fermions it may be possible to arrange that anomalies arising from different fermions cancel against each other. This is what happens in the standard electroweak theory, for example. The anomaly arises only in the fermion triangle diagram, with an odd number (one or three) of axial vertices. Then the total anomaly is multiplied by a factor

$$A^{abc} \equiv \mathrm{Tr}(t^a\{t^b t^c + t^c t^b\}) \tag{5.136}$$

where the t^a are the appropriate fermionic representations of electroweak theory, and the trace adds together the contribution from different fermions in the representation. For the standard assignments of weak isospin and hypercharge the above trace is always zero, but other choices of representations would yield a non-zero anomaly, and thereby an unphysical theory. Evidently the requirement that the theory be anomaly free imposes important constraints on the choice of fermion representations.

The anomalies discussed so far are the 'pure' gauge anomalies; that is why the vertices of the anomalous fermion triangle diagram were all interactions of the fermion with the gauge bosons. However it is easy to see that the triangle diagram remains anomalous when there is one axial current, and two energy-momentum tensor vertices (Delbourgo and Salam 1972, Eguchi and Freund 1976) representing the *gravitational* interaction of the fermions. Thus there are *mixed* gauge and gravitational anomalies. (The vanishing of this anomaly in standard electroweak theory is ensured by

$$\mathrm{Tr}(Y) = 0 \tag{5.137}$$

where Y is the weak hypercharge.) In fact in $4n+2$ dimensions there are in general *purely* gravitational anomalies (Brink *et al* 1977, Manton 1981, Chapline and Manton 1981, Chapline and Slansky 1982). The reason for this is essentially that already given in § 5.3. Only in $4n+2$ dimensions does the tangent space $\mathrm{SO}(1, 4n+1)$ admit *complex* representations. This is why we were able to construct a theory with only a $\chi = +1$ state (see (5.104)) for example. In such a chirally asymmetric theory the (Euclidean space) effective action is complex and there are anomalies (Alvarez-Gaumé and Witten 1984). Just as the gauge anomalies imply a violation of gauge invariance generated by quantum effects, so gravitational anomalies indicate a violation of the general coordinate invariance by fermion-loop quantum effects. Thus we are constrained to construct a theory which is free of *all* anomalies generated by fermions: gauge, gravitational and mixed.

It has been shown by Alvarez-Gaumé and Witten (1984) that, in an even number of dimensions ($4+D=4+2N$), fermion loops with $N+3$ external gauge bosons are anomalous, as are those with $N+1$ gauge bosons and two gravitons, etc; the general anomalous diagram has $r=N+3-2k$ external gauge bosons and $2k$ external graviton lines with $0\leqslant k\leqslant\frac{1}{2}(N+3)$. Note that when N is odd, there is a purely gravitational anomaly, as already observed. We denote by $t^a_\pm$ the (possibly reducible) representation of the gauge group G_{YM} formed by the $\chi=\pm1$ (spin-$\frac{1}{2}$) fermions in the case that N is odd; for N even we use the same notation to distinguish the $\chi=\pm\mathrm{i}$ (spin-$\frac{1}{2}$) fermions. Evidently there is a group theory factor associated with the anomaly arising from the (spin-$\frac{1}{2}$) fermion loop with r external gauge bosons:

$$A^{a_1\ldots a_r}\equiv S\,\mathrm{tr}(t^{a_1}_+t^{a_2}_+\ldots t^{a_r}_+)-S\,\mathrm{tr}(t^{a_1}_-t^{a_2}_-\ldots t^{a_r}_-) \tag{5.138}$$

where the 'S' denotes the (Bose) symmetrisation with respect to all indices $a_1\ldots a_r$. To make the theory anomaly free it is sufficient to require that $A^{a_1\ldots a_r}$ is zero for all allowed values of r. The simplest way to do this is, of course, to choose $t^a_+=t^a_-$, but this will not give a *chiral* theory as we have already observed; such vector-like theories are trivially anomaly free. For the case that N is odd, the allowed values of r are all even. In particular, in the case that r is zero, corresponding to the purely gravitational anomaly, we require that the number of (spin-$\frac{1}{2}$) fermions with $\chi=+1$ is equal to the number with $\chi=-1$. The only known method of satisfying (5.138) is to choose the gauge group

$$G_{YM}=O(2N+10) \tag{5.139a}$$

or

$$G_{YM}=O(2N+4k+10) \tag{5.139b}$$

and (for odd N) to choose the $\chi=\pm1$ spinors of $SO(1,3+2N)$ to transform as $\Gamma=\pm1$ spinors of the gauge group G_{YM} (Witten 1983). Thus

$$t^a_\pm=M^{\alpha\beta}a_\pm \tag{5.140a}$$

where

$$M^{\alpha\beta}=\tfrac{1}{4}\mathrm{i}[\Gamma^\alpha,\Gamma^\beta] \qquad (\alpha,\beta=1,\ldots,2N+10) \tag{5.140b}$$

analogously to (5.85) and

$$a_\pm=\tfrac{1}{2}(1\pm\Gamma) \tag{5.140c}$$

with

$$\Gamma\equiv\Gamma^1\Gamma^2\ldots\Gamma^{2N+10} \tag{5.140d}$$

as in (5.97). Then

$$\begin{aligned}\mathrm{Tr}(t^{a_1}_+t^{a_2}_+\ldots t^{a_r}_+)-\mathrm{Tr}(t^{a_1}_-t^{a_2}_-\ldots t^{a_r}_-)&=\mathrm{Tr}(M^{\alpha_1\beta_1}M^{\alpha_2\beta_2}\ldots M^{\alpha_r\beta_r}\Gamma)\\&=0 \qquad \text{for} \qquad r\leqslant N+4.\end{aligned} \tag{5.141}$$

(This is the higher-dimensional version of the familiar four-dimensional trace theorems $\mathrm{Tr}\,\gamma_5=\mathrm{Tr}\,\gamma_\mu\gamma_\nu\gamma_5=0$.) It follows that the group theory factor $A^{a_1\ldots a_r}$ defined in (5.138) vanishes for *all* of the anomalous diagrams. Another possibility might be to use some of the complex anomaly-free representations of $G_{YM}=SU(n)$. However it is not known if (5.138) can be satisfied for *all* r, and in any case the dimensionality of these representations is colossal; the smallest is the 374 556-dimensional representation of SU(6), and even this can accommodate only one generation of quarks and leptons, leaving a (very) large number of unwanted (unobserved) exotic particles (Eichten *et al* 1982). (See also Frampton (1983), Frampton and Kephart (1983a, b, c) who have studied the problem using the antisymmetric tensor representations.) The groups (5.139) also give anomaly-free theories in the case that N is even with an analogous choice of representations.

The price of achieving an anomaly-free theory is that we are forced to choose the group G_{YM} of the background gauge fields much larger than is required just to ensure the existence of chiral fermions with the observed quantum numbers. This in turn generates high-dimensional spinor representations containing fermions most of which are (so far) unobserved. We may illustrate this by considering the case (Bailin and Love 1985d) $K = CP^4$, discussed at the end of § 5.3. Since $N = 4$ in this case, (5.139a) shows that to obtain an anomaly-free theory we can take

$$G_{YM} = O(18). \tag{5.142}$$

Since the observed gauge bosons are assumed to come from the isometry group $G = SU(5)$ of K, this means that we have 153 additional gauge bosons for which we must presumably arrange large masses. The monopole U(1) gauge group must be embedded in G_{YM}, and one possibility is to identify it with the U(1) factor of the maximal subgroup

$$U(1) \times SU(9) \subset SO(18). \tag{5.143}$$

Then the effective gauge group is $SU(5) \times SU(9)$. We have already noted that to arrange for a fermionic zero mode belonging to the $\underline{5}$ (or $\bar{\underline{5}}$) representation of the isometry group, it is necessary to have a U(1) background monopole field satisfying $ec = \frac{7}{4}$ (Watamura 1983, 1984). That is to say, the U(1) charge of the $\bar{\underline{5}}$ is seven times the smallest possible unit. The $\Gamma = \pm i$ spinors of SO(18) are each 2^8-dimensional, and we have to arrange that the U(1) charge of the $\bar{\underline{5}}$ is associated with one of possible U(1) values arising in the expansion of the 2^8-spinors in terms of their $U(1) \times SU(9)$ content. It is easy to see that this requires us to associate the $\bar{\underline{5}}$ of SU(5) with the $\underline{9}$ of SU(9). This means that there are nine identical (zero-mode) fermion families with the standard $\bar{\underline{5}}$ quantum numbers, together with a large number of other fermions having non-standard quantum numbers. Spontaneous breaking of SU(9) might generate large masses for these unwanted states. To generate zero modes in the $\underline{10}$ of SU(5) requires the use of vector–spinors, as we showed in § 5.3, and these cause even more complexity when we attempt to remove the anomalies which they generate.

A rather more attractive scenario emerges if we take a non-Abelian subgroup of G_{YM} to have an expectation value, rather than the (monopole) U(1) gauge field considered so far. Equating the gauge field to the spin connection, as in (3.54), evidently gives a background $SO(2N)$ gauge field. This can be embedded in $G_{YM} = O(2N+10)$ in the obvious way, and this breaks G_{YM} to O(10). The effective gauge group is then $G_I \times O(10)$. However the fermion zero modes are singlets of G_I, so their quantum numbers derive entirely from the O(10) which came from removing the anomalies (Witten 1983). Consider first the $\Gamma = +1$ spinor of $O(2N+10)$. For N odd, it decomposes into products of spinors of $O(2N)$ and O(10) as

$$\begin{aligned} |2^{N+4}\text{-spinor}, \Gamma = 1\rangle = &|2^{N-1}\text{-spinor}, \Gamma_{2N} = i\rangle |2^4\text{-spinor}, \Gamma_{10} = -i\rangle \\ &+ |2^{N-1}\text{-spinor}, \Gamma_{2N} = -i\rangle |2^4\text{-spinor}, \Gamma_{10} = +i\rangle \end{aligned} \tag{5.144}$$

where $\Gamma_{2N,10}$, are the products of the Γ matrices in the $2N$- and ten-dimensional spaces respectively. The $\Gamma_{10} = -i$ spinor is (equivalent to) the complex conjugate of the $\Gamma_{10} = +i$

spinor, as we have previously observed. Evidently the fermion zero modes transform as $\underline{16}$ (or $\overline{\underline{16}}$) representations of O(10), which is well known to be an attractive way of accommodating all of the observed fermions in one generation (plus a ν_R state). The number of fermion generations is therefore

$$N_g = |n_L(\underline{16}) - n_R(\underline{16})| \tag{5.145}$$

where $n_{L,R}(\underline{16})$ is the number of left- or right-zero modes belonging to the $\underline{16}$ of O(10). These zero modes emerge from the tangent space SO(1, 2N+3) spinor, and the $\Gamma = +1$ spinor of (5.144) gives the G_{YM} behaviour of the $\chi = +1$ spinor of SO(1, 2N+3), according to the anomaly cancellation prescription. The $\chi = +1$ spinor can also be decomposed into spinors on M_4 and spinors on K:

$$\begin{aligned} |2^{N+1}\text{-spinor}, \chi = 1\rangle \\ &= |2\text{-spinor}, \gamma_5 = -1\rangle |2^{N-1}\text{-spinor}, \Gamma_K = \mathrm{i}\rangle \\ &\quad + |2\text{-spinor}, \gamma_5 = -1\rangle |2^{N-1}\text{-spinor}, \Gamma_K = -\mathrm{i}\rangle \end{aligned} \tag{5.146}$$

and we see that left ($\gamma_5 = -1$) or right ($\gamma_5 = +1$) chirality is correlated with the $\Gamma_K = \mathrm{i}$ or $\Gamma_K = -\mathrm{i}$ chirality on K. Further, from (5.144) fermions transforming as the $\underline{16}$ of O(10) transform as the 2^{N-1}-spinor of O(2N). Thus

$$\begin{aligned} N_g &= |n_+(2^{N-1}) - n_-(2^{N-1})| \\ &= |\mathrm{index}_{2^{N-1}}(M)| \end{aligned} \tag{5.147}$$

where $n_\pm(2^{N-1})$ is the number of zero modes transforming as the 2^{N-1} spinor of O(2N) with $\hat{\Gamma}_K \equiv \mathrm{i}\Gamma_K = \pm 1$, and we have used the definition (5.110) of the character-valued index. For odd N we may use (5.111) and obtain

$$\begin{aligned} N_g &= \tfrac{1}{2}|\mathrm{index}_{2^{N-1}}(M) - \mathrm{index}_{\overline{2^{N-1}}}(M)| \\ &= \tfrac{1}{2}(\text{Euler characteristic of } K). \end{aligned} \tag{5.148}$$

The contribution from the $\Gamma = -1$, $\chi = -1$ spinor is the same, so finally we have (Witten 1983)

$$N_g = (\text{Euler characteristic of } K). \tag{5.149}$$

For spheres (S^n), the Euler characteristic is 2 if n is even, and zero otherwise, while for CP^N the Euler characteristic is $N+1$. The strange thing about this scenario is that it is a *complete* negation of the Kaluza–Klein philosophy, since the isometry of K is irrelevant to the fermions actually observed.

It *can* be argued that the requirement that the anomaly $A^{a_1 \cdots a_r}$, defined in (5.138), vanishes for *all* allowed r is too restrictive. Kephart (1984) and Frampton and Yamamoto (1984) have proposed a weaker condition which can be applied in a theory containing only spin-$\frac{1}{2}$ fermions coupled to gravity. The constraint is that only the leading (fully connected) gauge and gravitational anomalies are required to cancel. (In such a theory it is always possible to remove the purely gravitational anomaly by simply adding to the theory the required number of gauge singlet spin-$\frac{1}{2}$ fields of the appropriate chirality.) The non-leading terms can be removed by introducing new massless fields which are totally antisymmetric and which transform under gauge transformations in a (peculiar) way designed to cancel the gauge non-invariance generated by the anomalies. (Precisely this mechanism is used in superstring theory (Green and Schwarz 1984) in which only the leading hexagon loop gauge anomaly in ten dimensions is cancelled. The non-leading anomalies are cancelled by gauge-dependent pieces of the action.) Whether this simplification leads to more attractive and realistic Kaluza–Klein models remains to be seen.

Another proposal along these lines was made by Alvarez-Gaumé and Witten (1984), who attempted to arrange that the purely gravitational anomalies cancel between fields of different spin. We have primarily had in mind the anomalies generated by spin-$\frac{1}{2}$ fermions, but there are also anomalies generated by spin-$\frac{3}{2}$ fermions, and we should certainly expect such fields if we are forced to deal with vector–spinors to solve the chiral fermion problem. However, Alvarez-Gaumé and Witten have demonstrated that antisymmetric tensor fields *also* generate gravitational anomalies, although they are covariant Bose fields, whenever they do not have a covariant Lagrangian (Frampton and Kephart 1983a, b, c, Townsend and Sierra 1983, Zumino *et al* 1983, Matsuki 1983, Matsuki and Hill 1983). This happens in $4+2N$ dimensions, when N is odd, and when the field A is antisymmetric with $N+1$ indices; its field strength F has $N+2$ indices and can be constrained to be self-dual ($^*F=F$). This together with the associated Bianchi identity serves as a covariant equation of motion. (The same effect is known to arise in two dimensions because of the bosonisation of fermions (Coleman 1975, Mandelstam 1975). The total cancellation of gravitational anomalies is possible only if there are all three types of field present (spin $\frac{1}{2}$, spin $\frac{3}{2}$ and antisymmetric tensor). In six dimensions this theory is (gravitational) anomaly free with 21 (positive chirality) spin-$\frac{1}{2}$ fields, one (negative chirality) spin-$\frac{3}{2}$ (gravitino) field and eight tensor fields. In ten dimensions the cancellation is achieved only by one (negative chirality) spin-$\frac{1}{2}$ field, a (positive chirality) spin-$\frac{3}{2}$ field and one real self-dual antisymmetric tensor field.

6. Kaluza–Klein cosmology

6.1. *Introduction*

The existence of extra (compactified) spatial dimensions has only indirect consequences for the (relatively) low energy particle physics which is open to experimental investigation either now or in the foreseeable future. However in the very early universe, i.e. at times t satisfying

$$tm_{\mathrm{P}} \gtrsim 1 \tag{6.1}$$

presumably *all* spatial dimensions were of the same scale and participating in the dynamical evolution of the universe. Thus we need to understand how the universe reached the present form $F^4\times K$, with F^4 the observed (flat) four-dimensional Friedmann solution, and K the (constant) compact manifold.

To address this problem it is natural to assume a form of the metric with *two* time-dependent scale factors $a(t)$ and $b(t)$:

$$\mathrm{d}s^2=\mathrm{d}t^2-a^2(t)\tilde{g}_{ij}(x)\,\mathrm{d}x^i\,\mathrm{d}x^j-b^2(t)\hat{g}_{mn}(y)\,\mathrm{d}y^m\,\mathrm{d}y^n \tag{6.2}$$

with $i, j=1, 2, 3$ and $m, n=4, \ldots, D+3$. $\tilde{g}_{ij}$ is the metric of the 3-space, usually taken to be maximally symmetric, and $\hat{g}_{mn}$ is the metric of the D-dimensional space, which is assumed to be an Einstein space. Thus we take

$$\tilde{R}_{ij}=2\tilde{k}\tilde{g}_{ij}(x) \tag{6.3a}$$

$$\hat{R}_{mn}=2\hat{k}\hat{g}_{mn}(y) \tag{6.3b}$$

with $\tilde{k}=+1,-1,0$ depending upon whether the 3-space is taken to be closed, open or flat and $\hat{k}=+1$, since K is closed. Aothough (6.2) is a natural ansatz with which to study the cosmological implications of having additional dimensions, it must be admitted from the outset that the implicit assumption of isotropy in the three spatial dimensions is unlikely to be correct, especially at early times. Analysis of the vacuum Einstein equations (Barrow and Stein-Schabes 1985) indicates that the generic behaviour is anisotropic, although without the chaotic unpredictability characteristic of having only three spatial dimensions. Similar conclusions have also been reached by Furusawa and Hosoya (1985). Insofar as (6.2) does allow the three spatial dimensions to evolve monotonically, we may at least hope that (6.2) will give a reasonable qualitative representation of the early-time behaviour.

The gravitational field equations are

$$\bar{\mathbb{R}}^{AB}-\tfrac{1}{2}\bar{\mathbb{R}}\bar{g}^{AB}=8\pi\bar{G}\bar{T}^{AB} \tag{6.4}$$

where $\bar{T}^{AB}$ is the energy-momentum tensor in the full $(4+D)$-dimensional space. Depending upon the era, $\bar{T}^{AB}$ receives contributions from incoherent radiation (massless particles) in 4 or $4+D$ dimensions and from any background fields necessary to ensure the compactification; this background may derive from classical fields, as in the case of the Freund–Rubin compactification, and/or from quantum fluctuations of the particle fields. Also, in order to arrange that there is no cosmological constant in the 4-space it is usually necessary to add a suitably tuned overall cosmological constant, $\bar{\Lambda}$, although its origin is often not specified. Such a cosmological constant breaks the supersymmetry, so cosmological studies of supergravity theories usually require only that the 4-space is anti-de Sitter.

The general form of $\bar{T}_{AB}$, consistent with the metric is

$$\bar{T}_{00}=\rho \qquad \bar{T}_{ij}=-\tilde{p}\bar{g}_{ij} \qquad \bar{T}_{mn}=-\hat{p}\bar{g}_{mn}. \tag{6.5}$$

Including the contribution from the overall cosmological constant $\bar{\Lambda}$, we may rewrite (6.4) in the form

$$\bar{\mathbb{R}}^{AB}=8\pi\bar{G}\left(\bar{T}^{AB}-\frac{\bar{T}}{2+D}\bar{g}^{AB}\right)-\frac{\bar{\Lambda}}{2+D}\bar{g}^{AB} \tag{6.6a}$$

where

$$\bar{T}\equiv\bar{T}^{A}{}_{A}=\rho-3\tilde{p}-D\hat{p} \tag{6.6b}$$

is the trace of the energy-momentum tensor. With the metric given in (6.2) this gives (Freund 1982)

$$\frac{3\ddot{a}}{a}+D\frac{\ddot{b}}{b}=\frac{\bar{\Lambda}}{D+2}-\frac{8\pi\bar{G}}{D+2}[(D+1)\rho+3\tilde{p}+D\hat{p}] \tag{6.7a}$$

$$\frac{\ddot{a}}{a}+2\left(\frac{\dot{a}}{a}\right)^2+\frac{2\tilde{k}}{a^2}+D\frac{\dot{a}\dot{b}}{ab}=\frac{\bar{\Lambda}}{D+2}+\frac{8\pi\bar{G}}{D+2}[\rho+(D-1)\tilde{p}-D\hat{p}] \tag{6.7b}$$

$$\frac{\ddot{b}}{b}+(D-1)\left(\frac{\dot{b}}{b}\right)^2+\frac{2\hat{k}}{b^2}+\frac{3\dot{a}\dot{b}}{ab}=\frac{\bar{\Lambda}}{D+2}+\frac{8\pi\bar{G}}{D+2}[\rho-3\tilde{p}+2\hat{p}]. \tag{6.7c}$$

The second derivative may be eliminated from (6.7a) using (6.7b, c) to yield the Friedmann equation

$$\left(3\frac{\dot{a}}{a}+D\frac{\dot{b}}{b}\right)^2-3\left(\frac{\dot{a}}{a}\right)^2-D\left(\frac{\dot{b}}{b}\right)^2+\frac{6\tilde{k}}{a^2}+2D\frac{\hat{k}}{b^2}=\bar{\Lambda}+16\pi\bar{G}\rho. \tag{6.8}$$

Of course, ρ, $\tilde{p}$ and $\hat{p}$ are constrained by energy-momentum conservation, as follows from (6.4),

$$D_B \bar{T}^{AB} = 0 \tag{6.9}$$

(and this remains true even when we include the cosmological constant $\bar{\Lambda}$). Then

$$\dot{\rho} + 3\frac{\dot{a}}{a}(\rho + \tilde{p}) + D\frac{\dot{b}}{b}(\rho + \hat{p}) = 0. \tag{6.10}$$

In the special case that T_{AB} is traceless, which is characteristic of *massless* particles, (6.6*b*) implies

$$\rho = 3\tilde{p} \qquad \text{if} \quad \hat{p} = 0 \tag{6.11a}$$

or

$$\rho = (D+3)p \qquad \text{if} \quad \tilde{p} = \hat{p} \equiv p. \tag{6.11b}$$

Then (6.10) is easily solved to give

$$\rho = 3\tilde{p} \propto a^{-4} b^{-D} \qquad \text{if} \quad \hat{p} = 0 \tag{6.12a}$$

or

$$\rho = (D+3)p \propto (a^3 b^D)^{-(D+4)(D+3)^{-1}} \qquad \text{if} \quad \tilde{p} = \hat{p} = p. \tag{6.12b}$$

6.2. *Late-time solutions (Bailin et al 1984)*

At later stages of the evolution, i.e. when

$$a \gg 1/T \gg b \tag{6.13}$$

we expect the dominant contributions to the energy-momentum tensor to be from the usual four-dimensional radiation, which fuels the observed Friedmann expansion of the 3-space, plus the contribution from the background fields necessary to achieve the (observed) compactification. Since the role of the background fields is to separate the two spaces their contribution to the energy-momentum tensor is typically of the form

$$\bar{T}^{\text{bg}}_{\mu\nu} = \tilde{H}(b)\bar{g}_{\mu\nu} \qquad (\mu, \nu = 0, 1, 2, 3) \tag{6.14a}$$

$$\bar{T}^{\text{bg}}_{mn} = \hat{H}(b)\bar{g}_{mn} \qquad (m, n = 4, \ldots, D+3). \tag{6.14b}$$

This is because $\bar{T}_{AB}$ derives from an action of the form

$$\bar{I}^{\text{bg}} = -\int d^{4+D}x\sqrt{\bar{g}}\, W(b) \tag{6.15}$$

so that the background energy density

$$\rho^{\text{bg}} = W(b) \tag{6.16}$$

depends only on the scale factor associated with the compact manifold K. In the case of Freund–Rubin compactification, for example, which was considered in § 3.2, (3.12) and (3.14) show that

$$b^D F = \text{constant} \tag{6.17}$$

from which it follows that $\bar{I}^{\mathrm{bg}}$ given in (3.11) has the form (6.15). A similar result holds also when the compactification is via quantum fluctuations, as is apparent from (3.22). The energy-momentum tensor following from (6.15) may be computed as in (3.25) and (3.26) with a result of the form (6.14) and

$$\tilde{H}(b) = W(b) \tag{6.18a}$$

$$\hat{H}(b) = W(b) + (b/D)\,W'(b). \tag{6.18b}$$

Thus the background field contribution is given by

$$\rho^{\mathrm{bg}} = -\tilde{p}^{\mathrm{bg}} = W(b) \tag{6.19a}$$

and

$$-\hat{p}^{\mathrm{bg}} = W(b) + (b/D)\,W'(b). \tag{6.19b}$$

At late times, characterised by (6.13), excitations in the compact space K are non-thermal, so the dominant thermal contribution is in the 4-space and given by (6.5) with

$$\rho^{\mathrm{th}} = 3\tilde{p}^{\mathrm{th}} \propto T^4 \tag{6.20a}$$

and

$$\hat{p}^{\mathrm{th}} = 0. \tag{6.20b}$$

The thermal energy density ρ^{th} is related to the four-dimensional density ρ_4^{th} by

$$\rho^{\mathrm{th}} = \rho_4^{\mathrm{th}}/V_D \tag{6.21}$$

where V_D is the volume of K. Also from (2.17) we find

$$V_D = \bar{G}/G \tag{6.22}$$

so

$$\rho^{\mathrm{th}} = \frac{G}{\bar{G}}\rho_4^{\mathrm{th}} = \frac{G}{\bar{G}}\sigma_4 T^4 \tag{6.23}$$

where σ_4 is determined by the number of massless helicity states.

We require that K is not evolving at late times. As seen in § 4 the observed gauge coupling constants are inversely proportional to the scale of K, so any time dependence in b would lead to time dependence of the gauge coupling constants. In any case, it is clear from (2.17) that if K were time dependent then the effective four-dimensional gravitational 'constant' would also be time dependent. There is by now a considerable body of evidence that none of these constants has any observable time dependence (Irvine and Humphreys 1984). We therefore require that the field equations admit a solution with

$$b = b_0 = \text{constant}. \tag{6.24}$$

Substituting (6.19) and (6.20) into (6.7c), (6.21) then requires

$$(D+2)(2\hat{k}/b_0^2) = \bar{\Lambda} + 16\pi\bar{G}[\,W(b_0) - (b_0/D)\,W'(b_0)]. \tag{6.25}$$

Using this and (6.23) the remaining equations can be cast in the standard cosmological form (Randjbar-Daemi *et al* 1984b, c, d):

$$\left(\frac{\dot{a}}{a}\right)^2+\frac{\tilde{k}}{a^2}=\tfrac{1}{6}\Lambda_4+\tfrac{8}{3}\pi G\rho_4^{\text{th}} \tag{6.26a}$$

$$\frac{\ddot{a}}{a}=\tfrac{1}{6}\Lambda_4-\tfrac{8}{3}\pi G\rho_4^{\text{th}} \tag{6.26b}$$

where

$$\tfrac{1}{2}\Lambda_4\equiv\frac{\bar{\Lambda}}{D+2}+\frac{8\pi\bar{G}}{D+2}[2W(b_0)+b_0W'(b_0)] \tag{6.26c}$$

is the effective four-dimensional cosmological constant. The contribution from ρ_4^{th} is non-negligible only in the case that $\Lambda_4\simeq 0$ which is the case of especial physical interest. In this special case we may eliminate $\bar{\Lambda}$ from (6.25) and (6.26). Then b_0 is a stationary point of an 'effective potential':

$$W'_{\text{eff}}(b_0)=0 \tag{6.27a}$$

where

$$W_{\text{eff}}(b)\equiv W(b)-\frac{D\hat{k}}{8\pi\bar{G}b^2} \tag{6.27b}$$

is obtained by modifying $W(b)$ with a 'curvature potential' (Maeda 1986). Although $W_{\text{eff}}(b)$ is a useful construction with which to discuss the theory, it is important to bear in mind that it only plays the role of an effective potential for values of $b(t)$ near the stationary point(s) b_0 at which $\Lambda_4\simeq 0$. At other possible constant values of $b(t)$, W_{eff} is *not* in general stationary, as can be seen from (6.61) for example. In general arranging that (6.27) is satisfied requires that $\bar{\Lambda}$ be non-zero. This is why in Kaluza–Klein supergravity theories which require $\bar{\Lambda}=0$, it is not usually possible to insist that $\Lambda_4=0$ (see, for example, Okada 1985 and Gleiser *et al* 1984).

Although (6.25) is sufficient to ensure a solution of the field equations of the required $F^4\times K$ form, this solution is unlikely to be reached unless it is stable against perturbations. To test for stability we write

$$a(t)=a_0(t)[1+\alpha(t)] \tag{6.28a}$$

$$b(t)=b_0[1+\beta(t)] \tag{6.28b}$$

where $a_0(t)$ is the (Friedmann) solution of (6.26). Substituting these into (6.7*c*) and linearising we find

$$\ddot{\beta}=-3(\dot{a}_0/a_0)\dot{\beta}+K_0\beta \tag{6.29a}$$

where

$$K_0=\frac{4\hat{k}}{b_0^2}+\frac{16\pi\bar{G}}{D(D+2)}[(D-1)b_0W'(b_0)-b_0^2W''(b_0)]. \tag{6.29b}$$

In the special case that $\Lambda_4\simeq 0$, (6.26*a*, *b*) are solved by

$$a_0(t)\propto t^{1/2} \tag{6.30}$$

when $\tilde{k}=0$, since $\rho_4^{\text{th}}\propto a^{-4}$. Then

$$(\dot{a}_0/a_0)=(1/2t) \tag{6.31}$$

is positive, which means that the second term of (6.29*a*) acts as a *damping* force. The third term acts as a *restoring* force if $K_0 < 0$, and then (6.29*a*) shows that the perturbation β executes damped simple harmonic motion. Thus a necessary condition of stability is that K_0 is negative. Using (6.27) we may express K_0 in terms of the effective potential:

$$K_0 = -\frac{16\pi\bar{G}b_0^2}{D(D+2)} W''_{\text{eff}}(b_0). \tag{6.32}$$

So the necessary condition for stability is that

$$W''_{\text{eff}}(b_0) > 0 \tag{6.33}$$

which shows that b_0 is a *minimum* of the effective potential W_{eff}, defined in (6.27*b*).

This conclusion applies more generally than the circumstances described above. For example, when $\Lambda_4 \simeq 0$ but $\tilde{k} = -1$

$$\dot{a}_0/a_0 = 1/t > 0 \tag{6.34}$$

and the qualitative behaviour of β is unaffected, although the precise solution of (6.29*a*) is different. In fact, in both of the above cases we may solve for β in terms of Bessel functions (Barrow and Stein-Schabes 1985, Maeda 1986):

$$\beta = t^{-p}[AJ_p(\sqrt{-K_0}t) + BN_p(\sqrt{-K_0}t)] \tag{6.35a}$$

where

$$p = \tfrac{1}{4} \qquad \text{for} \quad \tilde{k} = 0 \tag{6.35b}$$

$$p = 1 \qquad \text{for} \quad \tilde{k} = -1. \tag{6.35c}$$

For $K_0 < 0$ the Bessel functions are oscillatory and for large t

$$\beta \sim t^{-p-1/2} \tag{6.36}$$

showing that the oscillations are damped out and that the solution is stable against this perturbation. Similar conclusions apply also when Λ_4 is non-zero. For example, when $\Lambda_4 \gg \rho_4^{\text{th}} > 0$ and $\tilde{k} = -1$, $\dot{a}_0/a_0$ is positive and approaches a constant

$$\dot{a}_0/a_0 \to \sqrt{\Lambda_4/6} \qquad \text{as} \quad t \to \infty. \tag{6.37}$$

So asymptotically β again executes damped simple harmonic motion provided $K_0 < 0$.

It is also necessary to investigate the time development of the perturbation α, and this is most easily done using (6.8). Substituting (6.28) and linearising gives

$$\frac{\dot{a}_0}{a_0} 2\dot{\alpha} - \frac{2\tilde{k}}{a_0^2}\alpha + \frac{32\pi G}{3}\rho_4^{\text{th}}\alpha = -\frac{\dot{a}_0}{a_0} D\dot{\beta} - \frac{8\pi G}{3}\rho_4^{\text{th}} D\beta \tag{6.38}$$

where we have used (6.12*a*) to perturb ρ, and (6.27) to simplify the right-hand side. For large t it follows from (6.36) that the right-hand side of (6.38) decays as $t^{-p-5/2}$ when $\Lambda_4 \simeq 0$, from which we deduce that α also decays as $t^{-p-1/2}$. Thus $K_0 < 0$ also ensures that the perturbation α decays to zero. This conclusion extends to the case that $\Lambda_4 \gg \rho^{\text{th}} > 0$, $\tilde{k} = -1$ since (6.37) shows that the asymptotic behaviour of α is the same as that of β.

Thus the conclusion is that provided the constant value b_0 of b is a local minimum of the effective potential $W_{\rm eff}(b)$, defined in (6.27*b*), the $F^4 \times K$ solutions of the field equations are stable against infinitesimal perturbations of the scale factors $a(t)$, $b(t)$. In particular if

$$W(b) = w/b^n \qquad (w > 0) \tag{6.39}$$

then the stationary point $b_0 = (4\pi\bar{G}/nD)^{(n-1)^{-1}}$ *is* a minimum provided

$$n > 2. \tag{6.40}$$

Thus the Friedmann-like solutions are stable irrespective of whether compactification is achieved by the Freund–Rubin mechanism ($n = 2D > 2$), or by quantum fluctuations ($n = D + 4$) (Bailin *et al* 1984).

It is important to bear in mind that the above conclusions only apply to infinitesimal perturbations of the overall scale factors; the solutions may be unstable against finite perturbations of the scale factors, or against different infinitesimal perturbations. Maeda (1986) has investigated the former problem. He has observed that the term $(D\dot{b}/b + 3\dot{a}/a)\dot{b}/b$ in (6.7*c*) is *dissipative* when the proper volume V of the universe is *increasing*, and antidissipative when the proper volume is decreasing. Further, there is a region T in the $(b, \dot{b})$ plane, including the point $(b_0, 0)$, in which the sign of $\dot{V}$ does not change. By studying the classical turning points of the non-dissipative problem he then shows that if b_0 is a minimum of $W_{\rm eff}$, and if the universe enters the region T with V increasing, then the $F^4 \times K$ solution is an attractor; on the other hand if V is decreasing $F^4 \times K$ is a repulsor. The stability with respect to different infinitesimal perturbations has also been studied (Bailin *et al* 1985a). In the case when the compact manifold K is a product of two compact manifolds K_1 and K_2

$$K = K_1 \times K_2 \tag{6.41}$$

one can study perturbations with respect to the scale factors b_1 and b_2 separately. Then the $F^4 \times K_1 \times K_2$ solutions are always unstable.

6.3. *Early-time solutions*

Even if we understand how the observed $F^4 \times K$ solution can emerge from the field equations at late time, there remains the question of how, at an earlier stage, the separation of the two spaces arose. It is central to the Kaluza–Klein philosophy that (in some era) all of the dimensions were on the same footing. This is the motivation of the action (2.14), for example, which derives from the requirement of invariance under general $(4 + D)$-dimensional coordinate transformations. We should therefore like to understand how the vastly different scales associated with F^4 and K arose and, more ambitiously, how the *topological* separation between the spaces emerged.

The latter problem is beset by various no-go theorems which have been derived by Tipler (1985), generalising earlier work in four dimensions by Geroch (1967a, b, 1970) and Tipler (1977). The effect of these theorems is that in the context of a (causal) *classical* field theory the topology is invariant. Thus if we believe that the universe is well described at late times by the $F^4 \times K$ solutions, then their topology must also characterise the earlier-time solutions of the classical field equations from which they evolved. It follows that the topological separation must have occurred in the quantum era, in which not only is the classical field theory no longer an adequate description but also it is unclear as to how the question to be answered should be posed. We therefore concentrate in this section on the separation of the scale factors, assuming that the topological separation has already occurred. In particular, we shall see whether the Kaluza-Klein scenario affords novel opportunities for inflation, thereby solving the well known problems of the homogeneity and flatness of the observed universe (Guth 1981, Linde 1982). What is required is that there be an era in which the scale factor $a(t)$ associated with the observed 3-space grows very rapidly (exponentially), as if (or indeed because) there were a positive cosmological constant Λ_4. At the same time we shall require that the scale factor $b(t)$, associated with K, remains or becomes small, so as not to conflict with our observation of only three spatial dimensions.

One possibility is that initially both $a(t)$ and $b(t)$ expand, creating a large entropy per causal volume of the $(4+D)$-dimensional space, and that this entropy is then squeezed into the observed three dimensions as the manifold K is compactified. This scenario has been studied by Sahdev (1984), Abbott *et al* (1984, 1985) and by Kolb *et al* (1984). The assumption is that this era is dominated by $(4+D)$-dimensional radiation, so that the energy-momentum tensor is given by (6.5) with (6.11*b*) and (6.12*b*). To understand the general solution it is helpful (Yoshimura 1985) to consider two simple power-law solutions of the field equations. Thus we substitute

$$a(t) \propto t^{\tilde{\gamma}} \tag{6.42a}$$

$$b(t) \propto t^{\hat{\gamma}}. \tag{6.42b}$$

The (kinetic) terms involving time derivatives on the left of (6.7) are then all of order t^{-2}. Also the ideal-gas terms on the right of (6.7) are of order

$$\rho = (D+3)p \propto t^{-(3\tilde{\gamma}+D\hat{\gamma})(D+4)(D+3)^{-1}} \tag{6.43}$$

using (6.42) in (6.12*b*). If these ideal-gas terms are required to balance the kinetic terms then there is a generalised Friedmann solution with

$$\tilde{\gamma} = \hat{\gamma} = 2(D+4)^{-1} \tag{6.44}$$

in which the curvature terms on the left, proportional to a^{-2}, b^{-2}, are negligible since $\gamma < 1$. The other simple solution is the vacuum solution, in which the kinetic terms dominate both the curvature terms and the ideal-gas terms. Then there are (Kasner) solutions (i.e. $\tilde{\gamma} \neq \hat{\gamma}$) in which

$$3\tilde{\gamma} + D\hat{\gamma} = 1 \tag{6.45a}$$

$$3\tilde{\gamma}^2 + D\hat{\gamma}^2 = 1 \tag{6.45b}$$

so

$$3(3+D)\tilde{\gamma} = 3 \mp (3D(D+2))^{1/2} \tag{6.46a}$$

$$D(3+D)\hat{\gamma} = D \pm (3D(D+2))^{1/2}. \tag{6.46b}$$

Evidently $\tilde{\gamma}$ and $\hat{\gamma}$ have opposite signs (for $D \geqslant 2$) suggesting that we may find solutions, dominated by the Kasner singularity, in which $a(t)$ grows ($\tilde{\gamma} < 0$) and $b(t)$ decreases ($\hat{\gamma} > 0$). In fact this is precisely what the detailed numerical integration of the field equations show. This is illustrated in figure 1. The behaviour near $t = 0$ shows both scale factors controlled by the Friedmann behaviour (6.44), whereas near $t = t_0$ $b(t)$ is collapsing to zero while $a(t)$ is diverging according to the Kasner solution (6.46); for D large

$$\tilde{\gamma}_{\mathrm{KAS}} \sim -1/\sqrt{3} \tag{6.47a}$$

$$\hat{\gamma}_{\mathrm{KAS}} \sim (1+\sqrt{3})/D. \tag{6.47b}$$

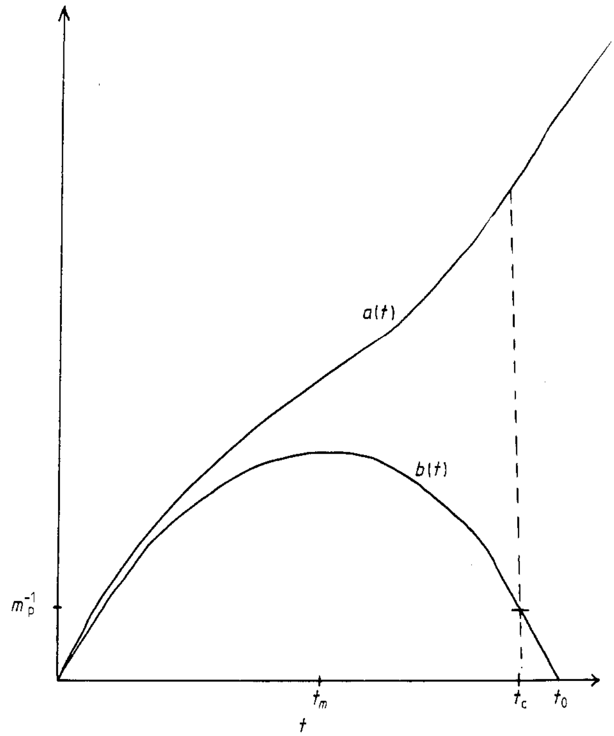

Figure 1. Time dependence of the scale factors. $a(t)$ scales the 3-space, $b(t)$ scales the compact space K.

The behaviour exhibited in figure 1 may also be understood qualitatively by studying the subleading terms in (6.44) and (6.46). Thus, writing the Friedmann solution as

$$a(t) = R_0 t^{\gamma}(1 + At^{\delta}) \qquad (6.48a)$$

$$b(t) = R_0 t^{\gamma}(1 + Bt^{\delta}) \qquad (6.48b)$$

with γ as in (6.44), and matching the next-to-leading-order terms one finds that (Yoshimura 1985)

$$\delta = 2(D+3)(D+4)^{-1} \qquad (6.49)$$

and that the curvature terms $\hat{k}/b^2$ tend to turn $b(t)$ over ($B<0$). When the 3-space is flat or open ($\tilde{k} \leqslant 0$) $a(t)$ increases monotonically (Abbott *et al* 1984) and $A>0$. Even when $\tilde{k}>0$, $A>B$ provided $D>3$, and for $D \gg 3$

$$A \sim D/12R_0^2 \qquad (6.50a)$$

$$B \sim -D/12R_0^2. \qquad (6.50b)$$

Thus we are able to understand qualitatively why the 3-space starts to expand more rapidly than the D-dimensional space K. To see whether this leads quantitatively to sufficient inflation we need to consider the thermodynamic evolution of the system. Following Abbott, *et al* (1985), Sahdev (1984) and Yoshimura (1984), we assume that the universe evolves slowly and adiabatically, so that the entropy per causal volume is conserved. In the early (expansion) stage the energy density $\rho = (D+3)p$ is that of $(D+4)$-dimensional thermal radiation. So

$$\rho = \sigma T^{D+4} \qquad (6.51)$$

and the energy density in the non-compact four dimensions is obtained, as in (6.21), by integrating over K:

$$\begin{aligned}\rho_{(4)} &= \sigma T^{D+4} V_D \\ &\propto \sigma b^D T^{D+4}. \end{aligned} \tag{6.52}$$

The assumption (6.51) is justified so long as the wavelength of the thermal excitations T^{-1} is small compared with the scale factor $b(t)$ of K. However when $bT \sim 1$ the excitations in K, of mass b^{-1}, become massive and the D extra dimensions are de-excited. At this point the (classical) equations we have used break down, and what happens next is slightly controversial.

The assumption of Kolb *et al* (1984) is that this 'freeze-out' of the extra dimensions coincides with the stabilisation of $b(t)$ which we know must occur, sooner or later. They assume that at freeze out

$$b_* T_* = 1 \tag{6.53}$$

and the energy density is instantly thermalised, so the four-dimensional entropy density

$$S_{(4)} \propto \rho_{(4)}^{3/4} \propto b_*^{-3} \tag{6.54}$$

using (6.53). The quantity of physical interest is the entropy S_h contained in a horizon volume after freeze out. The horizon length

$$l_h(t) \propto a(t) \tag{6.55}$$

so the required entropy is given by

$$S_h \propto a_*^3 b_*^{-3}. \tag{6.56}$$

For t close enough to t_0 the behaviour of $a(t)$ and $b(t)$ is dominated by the Kasner singularity, so

$$\begin{aligned} S_h &\propto (t_0 - t_*)^{3(\tilde{\gamma}_{KAS} - \hat{\gamma}_{KAS})} \\ &= (t_0 - t_*)^{-\sqrt{3}(1+2/D)^{1/2}}. \end{aligned} \tag{6.57}$$

Thus provided t_* is close enough to t_0 we might suppose that we can easily arrange that

$$S_h > 10^{88} \tag{6.58}$$

which is the inflation required (Guth 1981) to solve the horizon problem. However we should certainly *not* believe (6.58) if it required the freeze-out value b_* to be less than the Planck scale. In other words we cannot believe the values shown in figure 1 for $t > t_c$, which is when b falls below m_p^{-1}. It is clear from figure 1 that the largest value of S_h is when $t = t_c$ but then (6.58) is *not* satisfied. Since the only length scale in the problem is provided by

$$\bar{G} = G V_D \sim m_p^{-2} b_*^D \tag{6.59}$$

it is not surprising that $b_* \sim m_p^{-1}$ leads to an entropy S_h of order unity, as found by Kolb *et al* (1984). Typically, to ensure (6.58) we need $b_* \lll m_p^{-1}$ which is just when we cannot have any faith in classical field equations. A (slightly) more positive result is possible if one is prepared to believe that bT is of order 10–100, before compactification, say when $b \sim a$ at $t \sim t_m$ in figure 1 (Abbott *et al* 1984). However even then $D \sim$ 10–100 is also required to satisfy (6.58).

The deficiency of the foregoing scenario, which would have been its strength had it worked, is that there are no adjustable parameters and therefore no reason to suppose that large numbers, such as (6.58), will emerge. In any case, it seems unlikely that the compactification mechanism, which we know must participate eventually, merely plays the role of preventing the recollapse of $b(t)$. It is at least conceivable that the background fields responsible for $W(b)$ in (6.15) drive the inflation which we are seeking. One especially attractive scenario for inflation is that the universe underwent a 'slow roll-over' phase transition (for a review see Steinhardt 1983). If the order parameter of the phase transition is denoted by ϕ, we require that at very early times ϕ was trapped near ϕ_1 in a region of metastability. (Here, and elsewhere, we ignore the (possible) objection that the effective potential is known to be *convex* (Symanzik 1964, 1970, Iliopoulos *et al* 1975; see also Rivers 1984), and so can have at most one extremum; the 'potential' W_{eff} is in any case only an effective potential in the vicinity of $b=b_0$, as is apparent from (6.61), for example.) As the temperature decreased a more favoured value ϕ_0 developed, which was separated by an energy barrier from ϕ_1. The energy barrier prevented the phase transition occurring until there had been considerable supercooling. When eventually the barrier disappeared there was a first-order phase transition and the accompanying latent heat reheated the universe to a high temperature. During the metastable era the energy density was non-zero and almost constant (because of the '*slow*-roll-over'), so the universe underwent the exponential expansion called inflation. The high temperature achieved by reheating after the stable value ϕ_0 was attained ensures that the subsequent history was just as in the hot big bang theory.

In the present context the scale factor $b(t)$ is a natural vehicle for accomplishing such a transition. We have already seen that the 'observed' constant value b_0 of $b(t)$ minimises $W_{\text{eff}}(b)$, as shown in (6.27) and (6.33). (This required tuning the overall cosmological constant $\bar{\Lambda}$ so that the effective cosmological constant $\Lambda_4(b_0)$ in (6.26*c*) is zero.) Suppose that there is another constant value b_1 of $b(t)$ satisfying (6.7*c*). Then

$$(D+2)\frac{2\hat{k}}{b_1^2}=\bar{\Lambda}+16\pi\bar{G}\Big[W(b_1)-\frac{b_1}{D}W'(b_1)\Big]. \tag{6.60}$$

We can express this in terms of W_{eff}:

$$(b_1/D)\,W'_{\text{eff}}(b_1)=W_{\text{eff}}(b_1)-W_{\text{eff}}(b_0) \tag{6.61}$$

since from (6.25)

$$0=\bar{\Lambda}+16\pi\bar{G}W_{\text{eff}}(b_0). \tag{6.62}$$

When $b(t)\sim b_1$, the scale factor $a(t)$ inflates according to (6.8) with a Hubble constant H_1 given by

$$3H_1^2=8\pi\bar{G}[W_{\text{eff}}(b_1)-W_{\text{eff}}(b_0)]. \tag{6.63}$$

We can also calculate the rate at which $b(t)$ moves away from the stationary value b_1. As in (6.29) we linearise the departure from b_1 and then

$$\beta(t)\propto e^{\lambda_1 t} \tag{6.64a}$$

where

$$\lambda_1/H_1=\tfrac{3}{2}[-1+(1+\Delta)^{1/2}] \tag{6.64b}$$

and

$$\Delta \equiv 4K_1/9H_1^2$$
$$= \frac{8}{9H_1^2}\left(\frac{2\hat{k}}{b_1^2} + \frac{8\pi\bar{G}}{D(D+2)}[(D-1)b_1 W'(b_1) - b_1^2 W''(b_1)]\right). \tag{6.65}$$

To get sufficient inflation we require that λ_1/H_1, and hence Δ, is small so that $b(t)$ stays near b_1 long enough for $a(t)$ to inflate by many Hubble factors. In fact (6.58) requires

$$\lambda_1/H_1 \simeq \tfrac{3}{4}\Delta \lesssim \tfrac{1}{68}. \tag{6.66}$$

This too can be re-expressed in terms of W_{eff} to give

$$\frac{2}{D+2}\left(D-1-\frac{b_1^2}{D}\frac{W''_{\text{eff}}(b_1)}{W_{\text{eff}}(b_1)-W_{\text{eff}}(b_0)}\right) \lesssim \tfrac{1}{68} \tag{6.67}$$

and this imposes a severe constraint on possible compactifying forces. For example it *cannot* be satisfied if $W(b)$ is a single power, as in (6.39); this is because with only a single parameter w we cannot simultaneously satisfy the three conditions (6.27), (6.61) and (6.67). With *two* parameters, as in

$$W(b) = w_1 b^{-4-D} + w_2 b^{-2D} \tag{6.68}$$

it *is* possible, although *not* for $D=7$, provided w_1 and w_2 are tuned to critical values with an accuracy of 3% (Bailin *et al* 1985a).

It is also necessary to arrange that after the slow roll-over the universe is reheated to a temperature fairly close to its initial temperature. This reheating is caused by the emission of light particles as $b(t)$ oscillates about its stable value b_0. It can be seen that, *provided* b_0 is not (too) large compared with m_p^{-1}, then there *is* sufficient reheating to generate the observed entropy of the universe. However if this is the case it is likely that the model generates density fluctuations which are far too large to account for galaxy formation, unless there are a very large number of extra spatial dimensions (Bailin *et al* 1985a).

An alternative mechanism which realises an identical scenario has been proposed by Shafi and Wetterich (1983). They observe that in addition to the quantum effects considered by Candelas and Weinberg (1984) there are likely to be additional, non-minimal, gravitational terms in the effective action which are important at early times when $a(t) \sim b(t)$. Thus Shafi and Wetterich assume a background action of the form

$$\bar{I}^{\text{bg}} = -V_D^{-1}\int d^{4+D}x\sqrt{\bar{g}}[\alpha\bar{\mathbb{R}}^2 + \beta\bar{\mathbb{R}}_{AB}\bar{\mathbb{R}}^{AB} + \gamma\bar{R}_{ABCD}\bar{R}^{ABCD}] \tag{6.69}$$

instead of (6.15), which yields the most general form of the action involving up to four derivatives of $\bar{g}_{AB}$. (It is known that such terms arise naturally in the superstring theories which are currently the subject of intensive research.) It is easy to show that for rather general values of α, β, γ the field equations admit a solution of the form $M^4 \times S^D$, by fine tuning the cosmological constant $\bar{\Lambda}$, as in (6.26), and this is the presumed solution at late times. At early times there are (two) other solutions and, as in (6.63), these provide a non-zero effective cosmological constant with which to inflate $a(t)$. Thus with appropriate fine tuning it is again possible to arrange for sufficient inflation.

In principle a more sophisticated model than either of those so far advanced deserves analysis. We have already observed that it is unreasonable to suppose that the sole role of the compactifying fields at early times is to prevent the recollapse of b. Yoshimura (1985) has observed that the quantum Casimir effect, computed by Candelas and Weinberg (1984), for example, is relevant as it stands only at late times when $a(t) \gg b(t)$. The recalculation of these when $a(t) \sim b(t)$ has been performed (Yoshimura 1984, Koikawa and Yoshimura 1985) in the case when the 3-space and K are spherical, and has been generalised to finite temperature (see also Rubin and Roth 1983). The result is a free energy which interpolates between the pure Casimir and the thermal energy. One effect of this may be to generate a cosmological bounce in the space K. Then the present universe may be the outcome of a large number of such bounces in each of which a relatively moderate amount of entropy is gradually accumulated which in turn affects the size and duration of the succeeding bounce.

A similar calculation, using a free energy which interpolates between the pure Casimir and the pure thermal energy, has been carried out by Okada (1985). It has the particular feature that, like (6.68), it has *two* vacuum solutions. Therefore, by choosing the initial conditions carefully, i.e. by fine tuning, it can again be arranged that $a(t)$ inflates sufficiently while $b(t)$ remains near the (unstable) b_1.

6.4. *Survival of massive modes (Kolb and Slansky 1984)*

We have seen in §§ 1.5 and 5.1 that Kaluza-Klein theories typically require the existence of modes (particles) having masses of order $\tilde{R}^{-1}$, where $\tilde{R}$ is the radius of the compactified manifold K. If $\tilde{R}$ is of order $G^{1/2} \sim 10^{33}$ cm, then these particles have masses of the order of the Planck mass $m_p \sim 10^{19}$ GeV. (The particles observed in nature must then be zero modes which acquire their non-zero mass from some other (low-energy) mechanism, for example spontaneous symmetry breaking.) If the temperature T in the early universe was ever comparable with $\tilde{R}^{-1}$, these heavy modes, called 'pyrgons' by Kolb and Slansky (1984), would have been excited, and we are concerned with their fate in the subsequent cosmological evolution.

In the original five-dimensional theory of Kaluza and Klein the pyrgons are the $n \neq 0$ tensor modes $h_{\mu\nu}^{(n)}$ discussed in § 5.1, and these modes have non-zero charge $q_n = n(m_p \tilde{R})^{-1}$, as shown in (1.26). Evidently it is not possible for a pyrgon to decay solely into zero modes, because the zero modes all have $n = 0$ and therefore have zero charge. So the lightest pyrgons (i.e. the $|n| = 1$ modes in the five-dimensional theory) will be stable, and some of them might have survived to the present era. Whether or not they do so in observable numbers depends upon their annihilation rate Γ_A, since an $n = +1$ pyrgon can annihilate with an $n = -1$ pyrgon to form zero modes. The annihilation rate is given by

$$\Gamma_A = n_\psi \sigma_A |v| \tag{6.70}$$

where n_ψ is the pyrgon number density, σ_A is the cross section for the annihilation process and v is the relative velocity of the annihilating pyrgons. On dimensional grounds

$$\sigma_A |v| \sim \alpha^2 / m_\psi^2 \tag{6.71}$$

where α is the fine structure constant, and m_ψ is the mass of the $|n| = 1$ pyrgon. The pyrgon number density is

$$n_\psi \sim T^3 \tag{6.72}$$

and the expansion rate of the universe is

$$\Gamma_E \sim T^2/m_p. \tag{6.73}$$

Thus

$$\Gamma_A/\Gamma_E \sim \alpha^2 m_p T/m_\psi^2$$
$$\sim \alpha^2 m_p/m_\psi \qquad \text{at } T \sim m_\psi. \tag{6.74}$$

Now, because of the formula (1.26) for the charge of the pyrgons, which is characteristic of Kaluza–Klein theories,

$$\alpha \sim (\tilde{R} m_p)^{-2} \sim m_\psi^2/m_p^2 \tag{6.75}$$

and

$$\Gamma_A/\Gamma_E \sim m_\psi^3/m_p^3. \tag{6.76}$$

Thus if $m_\psi < m_p$ annihilation is ineffective at removing the pyrgons. But $m_\psi < m_p$ is just the condition

$$\tilde{R} > m_p^{-1} \tag{6.77}$$

which we have assumed throughout to justify our use of the classical field equations.

It is reasonable to assume that at the time of compactification (t_c) the abundances of pyrgons and of photons were comparable, since for $T \gg \tilde{R}^{-1}$ we should expect the $n \neq 0$ modes to be excited. Then it follows from (6.76) that the present pyrgon number density is of the order of the present photon number density n_γ:

$$n_\psi \sim n_\gamma \tag{6.78}$$

and this is supported by more careful numerical calculations (Kolb and Slansky 1984). (In fact the predicted pyrgon density is consistent with data on the present energy density only if m_ψ is less than about 100 eV, which is not consistent with regarding them as pyrgons.) Thus 10^{19} GeV pyrgons should be as abundant as photons are!

Such a prediction is likely to be characteristic of more realistic theories, as well as the five-dimensional model we have concentrated upon. One way to escape this disastrous prediction is if a large amount of entropy is created after compactification, for example by inflation as discussed in § 6.3. To avoid diluting the baryon asymmetry by an equally large amount it is necessary for the entropy to be generated prior to the generation of a non-zero baryon number.

6.5. Kaluza–Klein monopoles

A possible problem for Kaluza–Klein cosmology (as for the cosmology of grand unified theories (Preskill 1979, Zel' dovich and Khlopov 1978, Kibble 1979)) is the existence of magnetic monopoles which may be produced in the early universe and contribute excessively to the present energy density. (These monopoles, which appear in ordinary three-dimensional space are, of course, not to be confused with the monopole solutions on the compact manifold of § 3.5.) In the case of a grand unified theory, magnetic monopoles arise when the gauge group G_{YM} is spontaneously broken by a scalar field expectation value leaving a residual symmetry H. Then, for a single connected G_{YM}, magnetic monopoles are associated with non-trivial topological quantum numbers of the homotopy group $\pi_1(H)$. For the Kaluza–Klein case (Gross and Perry 1983, Sorkin

1983, Pollard 1983, Perry 1984) the group of $(4+D)$-dimensional coordinate transformations is spontaneously broken upon compactification to the product of four-dimensional coordinate transformations and the isometry group G, monopole solutions being classified (Perry 1984) by $\pi_1(\mathrm{G})$, e.g. for $\mathrm{G}=\mathrm{U}(1)$, $\pi_1(\mathrm{G})=Z$, for $\mathrm{G}=\mathrm{SO}(n)$, $n\geqslant 3$, $\pi_1(\mathrm{G})=Z_2$, and for $\mathrm{G}=\mathrm{SU}(n)$, $n\geqslant 3$, $\pi_1(\mathrm{G})=0$, where Z are the integers and Z_2 are the integers modulo 2.

In particular, for the five-dimensional case of § 1, where $\mathrm{G}=\mathrm{U}(1)$, the explicit (static) monopole solution is given (Gross and Perry 1983) by

$$\bar{g}_{AB}\,\mathrm{d}\bar{x}^A\,\mathrm{d}\bar{x}^B=\mathrm{d}t^2-\Phi[\tilde{R}\,\mathrm{d}\bar{x}^5-a(1-\cos\theta)\,\mathrm{d}\phi]^2-\Phi^{-1}(\mathrm{d}r^2+r^2\,\mathrm{d}\theta^2+r^2\sin^2\theta\,\mathrm{d}\phi^2) \tag{6.79}$$

where r, θ, ϕ are three-dimensional polar coordinates and $\bar{x}^5$ is the angular coordinate for S^1 (which we avoid denoting by θ here to prevent confusion). There is *no* coordinate singularity at $r=0$ provided we choose (Gross and Perry 1983, Sorkin 1983, Misner 1963)

$$4\pi a=2\pi\tilde{R}. \tag{6.80}$$

Comparing with (1.6) and (1.8) it is seen that the corresponding four-dimensional gauge field solution is given by

$$A_\phi=(a/\xi\tilde{R})(1-\cos\theta) \tag{6.81}$$

so that from a four-dimensional standpoint this solution is a Dirac monopole of magnetic charge

$$q_{\mathrm{M}}=a/\xi\tilde{R}. \tag{6.82}$$

Using (6.86) and (1.26), which shows that the unit of electric charge is

$$e=\kappa/\tilde{R} \tag{6.83}$$

it follows that

$$q_{\mathrm{M}}=1/2e. \tag{6.84}$$

Thus the monopole has unit magnetic charge. The (inertial) mass m_{M} of this five-dimensional soliton may be calculated from the energy-momentum tensor and is found (Gross and Perry 1983, Sorkin 1983) to be given by

$$m_{\mathrm{M}}^2=m_{\mathrm{p}}^2/16\alpha \tag{6.85}$$

in line with the usual expectation that the mass of a monopole is e^{-1} times the natural mass scale. Similar solutions have been studied by Perry (1984) in more than five dimensions.

Detailed monopole cosmology depends on the very early history of the universe, e.g. with $(4+D)$-dimensional isotropic behaviour until dynamical compactification the monopole number to entropy density ratio is of order $(\tilde{R}^{-1}/m_{\mathrm{p}})^3$ after compactification (Harvey *et al* 1984), assuming that one monopole is produced per causal volume, where $\tilde{R}$ is the 'radius' of the compact manifold and m_{p} is the Planck mass. If the 'radius' of the compact manifold is not too many orders of magnitude greater than the Planck length, this is a number not too much less than 1, e.g. with the estimate of (1.28) it is 10^{-3}. Such a large monopole density would contribute an unacceptable amount to the energy density of the universe unless it can be subsequently diluted, e.g. by cosmological inflation. An encouraging feature is that inflation may be more effective in diluting monopole density in Kaluza-Klein theories than in grand unified theories, because (Gross and Perry 1983) monopole-antimonopole pairs can have non-trivial topological quantum numbers so that they are *not* thermally produced after compactification (Harvey *et al* 1984).

7. Kaluza-Klein supergravity

7.1. Eleven-dimensional supergravity

Eleven dimensions has a special significance for supergravity theories (for more extensive reviews from a somewhat different perspective see Englert and Nicolai (1983) and Duff *et al* (1986)). It is the highest dimensionality for which a consistent theory exists because spins greater than 2 are inevitable for higher dimensions (Nahm 1978). Moreover, consistency is only possible for a *single* supersymmetry generator ($N=1$). The generalised Dirac matrices for eleven dimensions may be taken to be

$$\begin{aligned} \Gamma^\alpha &= \gamma^\alpha \otimes 1 \qquad \alpha = 0, 1, 2, 3 \\ \Gamma^\alpha &= \gamma_5 \otimes \tilde{\Gamma}^\alpha \qquad \alpha = 4, \ldots, 10 \end{aligned} \tag{7.1}$$

where γ^α are the Dirac matrices for four dimensions, and $\tilde{\Gamma}^\alpha$ are the Dirac matrices for seven dimensions.

Then,

$$\{\Gamma^\alpha, \Gamma^\beta\} = 2\eta^{\alpha\beta} I \tag{7.2}$$

where α, $\beta = 0, 1, \ldots, 10$ and

$$\eta_{\alpha\beta} \equiv \mathrm{diag}\{1, -1, \ldots, -1\}. \tag{7.3}$$

The supersymmetry generator $\tilde{Q}_\rho$, $\rho = 1, \ldots, 32$, is an eleven-dimensional Majorana spinor with the anticommutation relations

$$\{\tilde{Q}_\rho, \bar{\tilde{Q}}_\sigma\} = 2\Gamma^\alpha_{\rho\sigma} P_\alpha \tag{7.4}$$

where P_α is the eleven-dimensional momentum operator.

The supergravity multiplet for $N=1$ supergravity in eleven dimensions contains the following fields (Cremmer *et al* 1978). The vielbein $\tilde{e}^\alpha_A$, where $\alpha = 0, 1, \ldots, 10$ is a flat index associated with the tangent space, and $A = 0, 1, \ldots, 10$ is a curved index, a Majorana vector spinor ψ_A and an antisymmetric tensor field A_{BCD}, with corresponding field strength

$$F_{ABCD} \equiv \partial_A A_{BCD} - \partial_B A_{ACD} + \partial_C A_{ABD} - \partial_D A_{ABC} \tag{7.5}$$

as in (3.10), invariant under the Abelian gauge transformation,

$$\delta A_{ABC} = \partial_A \Lambda_{BC} - \partial_B \Lambda_{AC} + \partial_C \Lambda_{AB} \tag{7.6}$$

where

$$\Lambda_{AB} = -\Lambda_{BA}. \tag{7.7}$$

Under an infinitesimal local supersymmetry transformation,

$$\delta \tilde{e}^\alpha_A = -\tfrac{1}{2} \mathrm{i} \bar{\varepsilon} \Gamma^\alpha \psi_A \tag{7.8}$$

$$\delta A_{ABC} = \tfrac{1}{8}\sqrt{2}\,\bar{\varepsilon}\Gamma_{[AB}\psi_{C]} \tag{7.9}$$

$$\delta\psi_A = D_A\varepsilon + \frac{2\sqrt{2}\mathrm{i}}{(4!)^2}(\Gamma^{BCDE}{}_A - 8\delta_A{}^B\Gamma^{CDE})F_{BCDE}\varepsilon + (\text{fermion bilinears}) \tag{7.10}$$

where ε is an infinitesimal 32-component Majorana spinor parameter, $\Gamma^{A_1A_2\ldots A_n}$

denotes the completely antisymmetrised product of n generalised Dirac matrices, divided by the number of permutations, $\Gamma_{[AB}\psi_{C]}$ is the antisymmetrised sum over all permutations divided by their number of $\Gamma_{AB}\psi_C$ and D_A is the covariant derivative. The locally supersymmetric action is

$$\bar{I}=\int d^{11}\bar{x}|\det \bar{e}|\Big(-\tfrac{1}{2}\bar{\mathbb{R}}-\tfrac{1}{48}F_{ABCD}F^{ABCD}-\tfrac{1}{2}i\bar{\psi}_A\Gamma^{ABC}D_B\psi_C$$
$$+\frac{4\sqrt{2}}{(4!)^3}\frac{\varepsilon^{A_1\ldots A_{11}}}{|\det \bar{e}|}F_{A_1\ldots A_4}F_{A_5\ldots A_8}A_{A_9A_{10}A_{11}}$$
$$+\frac{3\sqrt{2}}{(4!)^2}(\bar{\psi}_A\Gamma^{ABCDEF}\psi_B+12\bar{\psi}^C\Gamma^{DE}\psi^F)F_{CDEF}$$
$$+(\text{quartic terms in fermion fields})\Big) \tag{7.11}$$

where units with

$$8\pi\bar{G}=1 \tag{7.12}$$

have been used, and $\varepsilon^{A_1\ldots A_{11}}$ is the eleven-dimensional Levi-Civita symbol, such that

$$\varepsilon^{01\ldots 11}=1. \tag{7.13}$$

7.2. Compactification of eleven-dimensional supergravity

The field equations arising from the above Lagrangian (assuming vanishing expectation values for the spinor fields) are

$$\bar{\mathbb{R}}_{AB}-\tfrac{1}{2}\bar{\mathbb{R}}\bar{g}_{AB}=\bar{T}_{AB}=-\tfrac{1}{6}(F_{CDEA}F_B{}^{CDE}-\tfrac{1}{8}F_{CDEF}F^{CDEF}\bar{g}_{AB}) \tag{7.14}$$

and

$$D_AF^{ABCD}=|\det \bar{e}|^{-1}\bar{\partial}_A(|\det \bar{e}|F^{ABCD})$$
$$=\frac{-\sqrt{2}}{2(4!)^2}\frac{\varepsilon^{EFGHIJKLBCD}}{|\det \bar{e}|}F_{EFGH}F_{IJKL}. \tag{7.15}$$

These field equations have a solution of the type

$$F^{\mu\nu\rho\sigma}=|\det e|^{-1}\varepsilon^{\mu\nu\rho\sigma}F \qquad \mu,\nu,\rho,\sigma=0,1,2,3 \tag{7.16}$$

and all other entries zero, where $\det e$ refers to four-dimensional space, as in (3.14). The term on the right-hand side of (7.15), which was *not* present in (3.12), does *not* contribute for a solution of this type, or indeed whenever F^{ABCD} has *no* components connecting four and seven dimensions. (It is also possible (Englert 1982) to find solutions where F^{mnpq}, $m, n, p, q=4,\ldots,10$, are also non-zero. However, these solutions have the arguably undesirable property of having all supersymmetries broken in the four-dimensional theory resulting from compactification (Biran *et al* 1983, D'Auria *et al* 1983, Englert *et al* 1983) and we shall not consider them further here.) Now, as in § 3.2, with $\bar{\Lambda}$ equal to zero as required by supersymmetry, one finds

$$\mathbb{R}=\frac{4(D-1)}{(D+2)}F^2=\tfrac{8}{3}F^2 \tag{7.17}$$

$$\tilde{\mathbb{R}}=-\frac{3D}{(D+2)}F^2=-\tfrac{7}{3}F^2 \tag{7.18}$$

$$\bar{\mathbb{R}}_{\mu\nu}=\frac{(D-1)}{(D+2)}F^2\bar{g}_{\mu\nu}=\tfrac{2}{3}F^2\bar{g}_{\mu\nu} \tag{7.19}$$

and

$$\bar{\mathbb{R}}_{mn} = -\frac{3F^2}{(D+2)}\bar{g}_{mn} = -\tfrac{1}{3}F^2\bar{g}_{mn}. \tag{7.20}$$

The compact manifold and four-dimensional space are thus inevitably Einstein spaces, and the separating off of exactly four non-compact dimensions is a consequence of the four indices on the tensor field strength F^{ABCD} which arises from the supergravity multiplet (Freund and Rubin 1980).

Whenever F is non-zero, four-dimensional space has to be anti-de Sitter rather than Minkowski, and compactifications of this type with compact manifold the 7-sphere (Duff 1982, Duff and Pope 1983), the squashed 7-spheres (Awada *et al* 1983, Duff *et al* 1983a, b), or M^{pqr} (Castellani *et al* 1984a) have been studied. (The squashed 7-spheres are topologically spheres, but as coset spaces are $G/H \equiv (SO(5)\times SO(3))/(SO(3)\times SO(3))$ for a particular embedding of H in G (Bais *et al* 1983).) When F is zero, four-dimensional space is allowed to be Minkowski, and the compact manifold is Ricci flat. Possible Ricci flat compact manifolds to have been considered are the 7-torus T^7 (Cremmer and Julia 1978, 1979) and $T^3\times K3$ (Duff *et al* 1983a, b). These last two possibilities have been less favoured than the others since the isometry group for these manifolds has only Abelian gauge symmetries.

At first sight, the manifolds M^{pqr} of § 2.6 are most promising, because they have isometry group $SU(3)\times SU(2)\times U(1)$ as required for the known strong and electroweak interactions. However, they become less compelling once it is realised that the quark and lepton representations of $SU(3)\times SU(2)\times U(1)$ do *not* arise in the harmonic expansion on M^{pqr} of elementary eleven-dimensional spinors (Randjbar-Daemi *et al* 1984a, D'auria and Fre 1984a), and that (see § 7.3) they never lead to $N=1$ supersymmetry in four dimensions.

Most attention (see the reviews of Englert and Nicolai 1983 and Duff *et al* 1986) has been given to the 7-sphere and squashed 7-spheres, for which the isometry groups are SO(8) and $SO(5)\times SU(2)$, respectively. In neither case is the isometry group large enough to contain $SU(3)\times SU(2)\times U(1)$, so that the pure Kaluza-Klein philosophy that all gauge fields come from the metric in the fashion of § 2 has to be abandoned. It is nonetheless possible to obtain an $SU(3)\times SU(2)\times U(1)$ gauge group in a fairly natural way as follows. The spin connection on the compact manifold, $(\omega_{\alpha\beta})_m$ of (2.39), provides a naturally arising *composite* gauge field (Cremmer and Julia 1979) with gauge group SO(7), because the spin connection is the 'gauge field' of the tangent space group. (Because ω_m can be calculated from the vielbein, for zero torsion, as in (2.35), it should not be regarded as elementary.) Then $SU(2)\times U(1)$ of electroweak theory may be obtained from the isometry group SO(8) or $SO(5)\times SU(2)$ of the 7-sphere or squashed 7-spheres, respectively, and the SU(3) of quantum chromodynamics from the SO(7) tangent space group associated with the spin connection on the compact manifold.

Now that the gauge group contains composite gauge fields, it is natural that the quarks and leptons should also be composite, arising from supergravity preons (Cremmer and Julia 1979, Ellis *et al* 1980, de Wit and Nicolai 1982, Duff *et al* 1984b, c). That the quarks and leptons should be composite, rather than arising from the harmonic expansion of elementary spinors in eleven dimensions, is in any case suggested by the results on the non-existence of chiral fermions in odd-dimensional Kaluza-Klein theories (see § 4). One advantage of this odd-dimensional supergravity theory over generic even-dimensional theories is that chiral gauge field-graviton anomalies do *not* arise (see § 4).

The gauge symmetry of the four-dimensional Lagrangian for 7-sphere compactification may be greater than SO(8)×SO(7). Indeed it has been suggested (Cremmer and Julia 1979, de Wit and Nicolai 1982) that the four-dimensional gauge symmetry may be enlarged to SO(8)×SU(8), and that the SO(8) gauge fields may provide a confining force for the formation of composite SU(8) gauge fields and fermions from supergravity preons. A recent development of this idea is the model of Duff *et al* (1984a) where the SO(5)×SO(3) gauge fields of the (right) squashed 7-sphere produce SU(5)×SU(3)×U(1) bound states, with the SU(5) identified with the standard SU(5) of grand unified theory, though there is the possible drawback that there is then *no* residual supersymmetry in four dimensions.

7.3. *Unbroken supersymmetries in four dimensions*

We may use the properties of the fields of the theory under a local eleven-dimensional supersymmetry transformation (7.8)-(7.10) to investigate how many unbroken supersymmetries appear in four dimensions. For four-dimensional Lorentz covariance the expectation value of all components of the (vector) spinor ψ_A must vanish. Then, from (7.8) and (7.9) it is automatic that

$$\delta \bar{e}^{\alpha}_{A} = 0 \tag{7.21}$$

and

$$\delta A_{ABC} = 0. \tag{7.22}$$

Thus, the necessary and sufficient condition for an unbroken supersymmetry in four dimensions is that

$$\delta \psi_A = 0. \tag{7.23}$$

In general, this is the condition

$$D_A \varepsilon + \frac{2\sqrt{2}\mathrm{i}}{(4!)^2}(\Gamma^{BCDE}{}_A - 8\delta_A{}^B \Gamma^{CDE})F_{BCDE}\varepsilon = 0. \tag{7.24}$$

In certain circumstances this simplifies considerably. First, for an eleven-dimensional space which is the direct product of a four-dimensional space and a seven-dimensional compact manifold we may assume that the infinitesimal supersymmetry transformation parameter $\varepsilon(x, y)$ factorises as the product of a four-component four-dimensional Majorana spinor $\xi(x)$, and an eight-component seven-dimensional Majorana spinor $\eta(y)$. Second, let us consider a solution for F^{ABCD} as in (7.16). Then, using the decomposition (7.1) of the eleven-dimensional Dirac matrices, (7.24) yields the two conditions

$$\bar{D}_\mu \xi \equiv D_\mu \xi(x) - (F/\sqrt{18})\gamma_5 \gamma_\mu \xi(x) = 0 \qquad \mu = 0, 1, 2, 3 \tag{7.25}$$

and

$$\bar{D}_m \eta \equiv D_m \eta(y) - (F/2\sqrt{18})\tilde{\Gamma}_m \eta(y) = 0 \qquad m = 4, \ldots, 10. \tag{7.26}$$

A necessary condition for the existence of a solution η of (7.26) is

$$[\bar{D}_m, \bar{D}_n]\eta(y) = 0. \tag{7.27}$$

For the spinorial representation of SO(7), (2.37)-(2.39) give the covariant derivative

$$D_m \eta = (\partial_m + \tfrac{1}{4}(\omega_{\alpha\beta})_m \tilde{\Gamma}^{\alpha\beta})\eta. \tag{7.28}$$

Evaluating the commutator in (7.27) with the aid of (2.36) puts (7.27) in the form

$$(R_{mnpq}\tilde{\Gamma}^{pq}+\tfrac{1}{9}F^2\tilde{\Gamma}_{mn})\eta=0. \tag{7.29}$$

The number N of unbroken supersymmetries appearing in four dimensions is the number of independent solutions (Killing spinors) of (7.29). This may be interpreted geometrically (Duff *et al* 1983b) as the number of spinors left invariant by the holonomy group $\mathscr{H}$ of the generalised spin connection specified by $\bar{D}_m$.

The corresponding consistency condition deriving from (7.25)

$$[\bar{D}_\mu, \bar{D}_\nu)\xi(x)=0 \tag{7.30}$$

yields

$$(R_{\mu\nu\rho\sigma}\Gamma^{\rho\sigma}-\tfrac{4}{9}F^2\Gamma_{\mu\nu})\xi(x)=0 \tag{7.31}$$

which is always satisfied when four-dimensional space is maximally symmetric with 'radius' consistent with (7.19)

$$R_{\mu\nu\rho\sigma}=\tfrac{2}{9}F^2(\bar{g}_{\mu\nu}\bar{g}_{\rho\sigma}-\bar{g}_{\mu\rho}\bar{g}_{\nu\sigma}) \tag{7.32}$$

The maximum number of unbroken supersymmetries is $N=8$ which occurs when *every* (eight-component) spinor η satisfies (7.29). (Then the holonomy group $\mathscr{H}$ is *I*.) This happens when the compact manifold is maximally symmetric with 'radius' consistent with (7.20):

$$R_{mnpq}=-\tfrac{1}{18}F^2(\tilde{g}_{mp}\tilde{g}_{nq}-\tilde{g}_{mq}\tilde{g}_{np}) \tag{7.33}$$

so that

$$R_{mnpq}\tilde{\Gamma}^{pq}+\tfrac{1}{9}F^2\tilde{\Gamma}_{mn}=0. \tag{7.34}$$

When $F\neq 0$, the compact manifold is S^7 and the 4-space is anti-de Sitter (Duff and Pope 1983, Duff 1982), and when $F=0$ the compact manifold is the 7-torus T^7 and the 4-space is Minkowski (Cremmer and Julia 1978, 1979).

Other cases (without torsion) leading to some unbroken supersymmetries in four dimensions are the (left) squashed 7-sphere (Awada *et al* 1983) with $N=1$ supersymmetry ($\mathscr{H}=G_2$) and anti-de Sitter 4-space, M^{pqr} for $p=q$ (Castellani *et al* 1984a) with $N=2$ supersymmetry ($\mathscr{H}=SU(3)$) and anti-de Sitter 4-space, and $T^3\times K3$ (Duff *et al* 1983a, b) with $N=4$ supersymmetry ($\mathscr{H}=SU(2)$) and Minkowski 4-space.

Throughout the above derivation it has been assumed that F^{ABCD} is as in (7.16), with the components F^{mnpq} on the compact manifold all zero. If a non-zero F^{mnpq} is introduced (Englert 1982) then consistency with the Einstein field equations (7.14) implies that there are *no* unbroken supersymmetries in four dimensions (Biran *et al* 1983, D'Auria *et al* 1983).

A slightly different perspective is obtained by *assuming* from the outset that 4-space is Minkowski. Then (Candelas and Raine 1984) the existence of an unbroken supersymmetry in four dimensions implies that the components F^{mnpq} are zero, and that the compact manifold is Ricci flat, *without* using the field equations. This conclusion is thus applicable to situations where quantum corrections to the classical field equations are important.

7.4. Minimal ten-dimensional supergravity

For it to be possible for chiral fermions to arise from the harmonic expansion of elementary spinor fields (in the presence of explicit gauge fields) it is necessary to use an even dimensionality (see § 4). Ten dimensions is especially important because ten-dimensional supergravity theories may be derived from superstring theories for energies very much less than the Planck mass. Although the ten-dimensional supergravity theory is non-renormalisable, the superstring theory, of which it is a limit, is probably renormalisable, the relationship between the two theories being analogous to that between renormalisable electroweak theory and non-renormalisable four-fermion theory. (For reviews of superstring theory see Schwarz 1982, 1984, Green 1983, Brink 1984.)

Two different $N=2$ (two supersymmetry generators) ten-dimensional supergravity theories arise from type IIA and IIB superstrings, respectively (Green *et al* 1982, Green and Schwarz 1982, 1983, Marcus and Schwarz 1982, Schwarz and West 1983). The first of these can be obtained by dimensional reduction of eleven-dimensional supergravity (Cremmer and Julia 1979, Scherk and Schwarz 1979), and is thus non-chiral. The second is a chiral theory but, because it does *not* allow explicit ten-dimensional gauge fields, there may be no way of transmitting this property to four dimensions, upon compactification (see § 4). We shall therefore not discuss these theories further here, but shall concentrate instead on $N=1$ ten-dimensional supergravity, which can be derived from type-I superstrings. In this section the *minimal* $N=1$ ten-dimensional supergravity theory will be described, and in the next section we shall describe the modified anomaly-free theory which can be derived from the type-I superstring.

In its minimal form, $N=1$ ten-dimensional supergravity has been constructed by Gliozzi *et al* (1977), Scherk (1978), Chamseddine (1981a), Bergshoeff *et al* (1982) and Chapline and Manton (1983). The supergravity multiplet $\{\bar{e}^{\alpha}_{A}, \psi_A, B_{AB}, \lambda, \phi\}$ contains the vielbein $\bar{e}^{\alpha}_{A}$, where $\alpha=0, 1, \ldots, 9$ is a flat index, and $A=0, 1, \ldots, 9$ is a curved index, a Rarita–Schwinger Majorana–Weyl spinor ψ_A, an antisymmetric tensor field B_{AB}, a right-handed Majorana–Weyl spinor field λ and a scalar field ϕ. (An alternative version of ten-dimensional supergravity, which we do not discuss here, contains instead a sixth-rank antisymmetric tensor field (Chamseddine 1981b, Fré and Zizzi 1984).) The theory is coupled to explicit Yang–Mills gauge fields A_A, where a runs over the adjoint representation of the gauge group G_{YM}, and the corresponding left-handed Majorana–Weyl gauginos χ^a. The field strength corresponding to A_A is given (as usual) by

$$F^a_{AB}\tilde{t}_a \equiv F_{AB} = \partial_A A_B - \partial_B A_A + g[A_A, A_B] \tag{7.35}$$

where g is the gauge coupling constant and

$$A_A \equiv A^a_A \tilde{t}_a \tag{7.36}$$

and $\tilde{t}_a$ is the defining representation of G_{YM} by anti-Hermitian matrices with

$$[\tilde{t}_a, \tilde{t}_b) = f_{abc}\tilde{t}_c. \tag{7.37}$$

The field strength H_{ABC} corresponding to B_{AB} is given by

$$H_{ABC} = \partial_{[A}B_{BC]} - (\bar{\kappa}/\sqrt{2})\widetilde{\mathrm{Tr}}(A_{[A}F_{BC]} - \tfrac{2}{3} g A_{[A}A_B A_{C]}) \tag{7.38}$$

where $\partial_{[A}B_{BC]}$ is the antisymmetrised sum over all permutations divided by the number of permutations, etc, $\widetilde{\mathrm{Tr}}$ is the trace in the defining representation and

$$\bar{\kappa} = 8\pi\bar{G}. \tag{7.39}$$

(To agree with Chapline and Manton (1983) we have adopted here a notation different from (1.17).) This field strength is invariant under the generalised infinitesimal gauge transformation

$$\delta A_A = D_A \Lambda \qquad \delta B_{AB} = \sqrt{2}\bar{\kappa}\ \tilde{\mathrm{Tr}}(\Lambda \partial_{[A} A_{B]}). \tag{7.40}$$

Under an infinitesimal local supersymmetry transformation with infinitesimal spinor parameter ε,

$$\delta \bar{e}^{\alpha}_{A} = \tfrac{1}{2}\bar{\varepsilon}\Gamma^{\alpha}\psi_A \tag{7.41}$$

$$\delta\psi_A = D_A\varepsilon + \tfrac{1}{32}\sqrt{2}\phi^{-3/4}(\Gamma_A^{\ BCD} - 9\delta_A^B\Gamma^{CD})H_{BCD}\varepsilon$$
$$+ (\text{fermion bilinears}) \tag{7.42}$$

$$\delta B_{AB} = \tfrac{1}{4}\sqrt{2}\phi^{3/4}(2\bar{\varepsilon}\Gamma_{[A}\psi_{B]} - (1/\sqrt{2})\bar{\varepsilon}\Gamma_{AB}\lambda) + (\bar{\kappa}/\sqrt{2})\phi^{3/8}\bar{\varepsilon}\Gamma_{[A}\,\mathrm{Tr}(\chi A_{B]}) \tag{7.43}$$

$$\delta\lambda = -\tfrac{3}{8}\sqrt{2}(\not{D}\phi/\phi)\varepsilon + \tfrac{1}{8}\phi^{-3/4}\Gamma^{ABC}\varepsilon F_{ABC} + (\text{fermion bilinears}) \tag{7.44}$$

$$\delta A_A = \tfrac{1}{2}\phi^{3/8}\bar{\varepsilon}\Gamma_A\chi \tag{7.45}$$

$$\delta\chi = -\tfrac{1}{4}\phi^{-3/8}\Gamma^{AB}F_{AB}\varepsilon + (\text{fermion bilinears}). \tag{7.46}$$

(The complete transformations are given in Chapline and Manton (1983).) In these equations $\Gamma^{A_1A_2\cdots A_n}$ denotes the completely antisymmetrised product of n ten-dimensional Dirac matrices Γ^A divided by the number of permutations and D_A is the covariant derivative.

The locally supersymmetric action is

$$\bar{I} = \int \mathrm{d}^{10}\bar{x}(\mathscr{L}_{\mathrm{B}} + \mathscr{L}_{\mathrm{F}} + \mathscr{L}_{\mathrm{FB}})(+\text{four-Fermi interactions}) \tag{7.47}$$

with the bosonic term

$$|\det \bar{e}|^{-1}\mathscr{L}_{\mathrm{B}} = -\tfrac{1}{2}\bar{\kappa}^{-2}\bar{\mathbb{R}} - \tfrac{3}{4}\phi^{-3/2}H_{ABC}H^{ABC} - \tfrac{9}{16}\bar{\kappa}^{-2}\phi^{-2}D_A\phi D^A\phi$$
$$-\tfrac{1}{4}\phi^{-3/4}F^{a}_{AB}F^{AB}_{a} \tag{7.48}$$

the fermionic kinetic term

$$|\det \bar{e}|^{-1}\mathscr{L}_{\mathrm{F}} = -\tfrac{1}{2}\bar{\psi}_A\Gamma^{ABC}D_B\psi_C - \tfrac{1}{2}\lambda\not{D}\lambda - \tfrac{1}{2}\tilde{\mathrm{Tr}}(\bar{\chi}\not{D}\chi)$$
$$-(3/4\sqrt{2})\bar{\psi}_A(\not{\partial}\phi/\phi)\Gamma^A\lambda \tag{7.49}$$

and fermion–boson interaction

$$|\det \bar{e}|^{-1}\mathscr{L}_{\mathrm{FB}} = -\tfrac{1}{4}\bar{\kappa}\phi^{-3/8}\,\tilde{\mathrm{Tr}}(\bar{\chi}\Gamma^A\Gamma^{BC}F_{BC})(\psi_A + \tfrac{1}{12}\sqrt{2}\Gamma_A\lambda)$$
$$+\tfrac{1}{16}\sqrt{2}\bar{\kappa}\phi^{-3/4}H_{ABC}(\tilde{\mathrm{Tr}}(\bar{\chi}\Gamma^{ABC}\chi) + \bar{\psi}_D\Gamma^{DABCE}\psi_E + 6\bar{\psi}^A\Gamma^B\psi^C$$
$$-\sqrt{2}\bar{\psi}_D\Gamma^{ABC}\Gamma^D\lambda). \tag{7.50}$$

It has been observed by Freedman *et al* (1983) and Randjbar-Daemi *et al* (1983d) that there is a basic difficulty with the theory as it stands, resulting from the field equation for the field ϕ (in the absence of fermion condensates) which may be written as

$$D_AD^A\sigma = -(3/2\sqrt{2})\bar{\kappa}\exp(-\sqrt{2}\bar{\kappa}\sigma)H_{ABC}H^{ABC}$$
$$-(1/4\sqrt{2})\bar{\kappa}\exp(-(1\sqrt{2})\bar{\kappa}\sigma)\,\tilde{\mathrm{Tr}}(F_{AB}F^{AB}) \tag{7.51}$$

where σ is defined by

$$\phi = \exp(\tfrac{2}{3}\sqrt{2}\bar{\kappa}\sigma). \tag{7.52}$$

For maximal symmetry of four-dimensional space, $H_{\mu\nu\rho}$, $F_{\mu\nu}$ and $\partial_\nu\phi$ must vanish for μ, ν, $\rho=0,1,2,3$. It may then be deduced, by integrating (7.51) over the compact manifold, that

$$H_{ABC}=0 \qquad F_{AB}=0 \qquad A, B, C=0,1,\ldots,9. \tag{7.53}$$

By multiplying (7.51) by σ and integrating over the compact manifold, it may also be shown that

$$\sigma=\text{constant}. \tag{7.54}$$

The Einstein field equation then shows that

$$\bar{R}_{AB}=0 \qquad A, B=0,1,\ldots,9. \tag{7.55}$$

Thus, 4-space is Minkowski and the compact manifold is Ricci flat, so that it is impossible to obtain any non-Abelian symmetries from the isometry group of the compact manifold. Moreover, because F_{AB} is zero in (7.53), chiral fermions *cannot* be obtained.

If we wish to obtain compactification with chiral fermions it is necessary to evade the above result, e.g. by adding an *ad hoc* supersymmetry breaking term $-V(\phi)$ to the Lagrangian, perhaps arising dynamically. Proceeding in this way Randjbar-Daemi *et al* (1983d) have looked for G-invariant compactifications on coset spaces G/H, with the gauge field expectation value forming a generalised monopole solution (Randjbar-Daemi and Percacci 1982) as discussed in § 3.7. Compactification is possible in this way when the compact manifold is the coset space $S^6\equiv SO(7)SO(6)$ or $G_2/SU(3)$ but *not* if it is CP^3 or $CP^2\times S^2$. A key element in this discussion is the observation that the expressions (7.35) and (7.38) for the field strengths require H_{ABC} to satisfy the Bianchi identity

$$[\partial_m H_{npq}]_{\text{antisymmetrised}}=-(\bar{\kappa}/2\sqrt{2})[\widetilde{\mathrm{Tr}}(F_{mn}F_{pq})]_{\text{antisymmetrised}}. \tag{7.56}$$

This is a strong constraint on possible compactifications.

The cosmology of these theories has been considered by Chapline and Gibbons (1984). A stable generalised Friedmann-Robertson-Walker solution may be found for which the radius of the compact manifold tends to a constant at late times, as required for the constancy of the gauge coupling constant associated with the isometry group of the compact manifold (see § 6.2). However, the introduction of the *ad hoc* term $-V(\phi)$ in the Lagrangian, discussed above, is required. Specifically, Chapline and Gibbons (1984) take $V(\phi)$ to be a mass term for the field σ of (7.52).

7.5. Anomaly-free ten-dimensional supergravity

Two $N=1$ ten-dimensional supergravities free of all gauge, graviton and mixed gauge-graviton anomalies, have been derived from superstring theory, one (Green and Schwarz 1984) with explicit Yang-Mills group $G_{YM}=SO(32)$, and the other (Gross *et al* 1985) with $G_{YM}=E_8\times E_8$. In the minimal theory described in § 7.4, *not* all anomalies cancel even when the gauge group is SO(32) or $E_8\times E_8$. However, the supergravity theory derived as a low-energy limit of the superstring contains non-minimal higher derivative interactions, some of which are non-invariant under Yang-Mills and local Lorentz gauge transformations in just such a way as to cancel the anomalous non-invariant contributions of fermion loops, for $G_{YM}=SO(32)$ or $E_8\times E_8$. (Such terms have earlier been introduced by Wess and Zumino (1971) in another context.)

It has been observed by Frampton *et al* (1984a) and Frampton and Yamamoto (1984) that earlier attempts to construct anomaly-free Einstein-Yang-Mills theories (see § 5) should be re-examined in the light of the anomaly cancellation mechanism in ten-dimensional supergravity. The lesson is that it may not be necessary for all fermion anomalies to cancel directly if some can be cancelled by adding gauge-dependent contact terms arising from the exchange of superheavy fields.

The non-minimal theory has the same supersymmetry transformation laws as the minimal theory (equations (7.42)-(7.46).) However, the Lorentz transformation property of B_{AB} is modified so that H_{ABC} defined by (7.38) is no longer Lorentz invariant. To obtain a Lorentz-invariant field strength, it is necessary to modify the definition to

$$H = \mathrm{d}B - \omega_{3Y} + \omega_{3L} \tag{7.57}$$

where we are using differential form notation (see § 2.5) and

$$\omega_{3Y} \equiv \tfrac{1}{30}\,\mathrm{Tr}((A \wedge F) - \tfrac{1}{3}(A \wedge A \wedge A)) \tag{7.58}$$

and

$$\omega_{3L} \equiv \mathrm{Tr}((\omega \wedge R) - \tfrac{1}{3}(\omega \wedge \omega \wedge \omega)) \tag{7.59}$$

where Tr denotes a trace in the adjoint representation of $\mathrm{G_{YM}}$, and tr denotes a trace in the defining representation of the tangent space group SO(1, 9) (the local Lorentz group). In (7.58), the gauge field 1-form is

$$A \equiv A_A\, \mathrm{d}\bar{x}^A \tag{7.60}$$

with A_A as in (7.36) but with the matrices now in the adjoint representation, and the Yang-Mills field strength 2-form is

$$\tfrac{1}{2} F_{AB}(\mathrm{d}\bar{x}^A \wedge \mathrm{d}\bar{x}^B) \equiv F = \mathrm{d}A + A \wedge A \tag{7.61}$$

with F_{AB} as in (7.35). The gauge coupling constant g has been absorbed in the definition of A. In (7.59), the 2-form for the second-rank tensor field is

$$B \equiv B_{AB}(\mathrm{d}\bar{x}^A \wedge \mathrm{d}\bar{x}^B) \tag{7.62}$$

the 3-form for the field strength is

$$H \equiv H_{ABC}(\mathrm{d}\bar{x}^A \wedge \mathrm{d}\bar{x}^B \wedge \mathrm{d}\bar{x}^C) \tag{7.63}$$

and the spin connection 1-form is

$$\omega \equiv \omega_A\, \mathrm{d}\bar{x}^A \tag{7.64}$$

with ω_A defined analogously to (2.39), but with $M^{\alpha\beta}$ being the generators of SO(1, 9), and indices being lowered by $\eta_{\alpha\beta}$, the ten-dimensional Minkowski metric, rather than $\delta_{\alpha\beta}$. Also, R is the curvature 2-form defined by

$$R \equiv R_{\alpha\beta} M^{\alpha\beta} = \mathrm{d}\omega + \omega \wedge \omega \tag{7.65}$$

(analogously to (2.36)). Units are being used where $8\pi\bar{G}$ is 1. The term ω_{3L} in (7.57) does *not* occur in the minimal theory, and constitutes a basic difference between the two theories.

It is not difficult to deduce from (7.57) that

$$\mathrm{d}H = \mathrm{tr}(R \wedge R) - \tfrac{1}{30}\,\mathrm{Tr}(F \wedge F) \tag{7.66}$$

which is (7.56) with an extra $\mathrm{tr}(R \wedge R)$ term added. This is again a strong constraint on possible compactifications. Adding a term $-V(\phi)$ to the Lagrangian, as in the last section, it is possible using Randjbar-Daemi-Percacci (1982) gauge field configurations (see § 3.7) on symmetric coset spaces K to obtain compactifications consistent with (7.66) which give a satisfactory spectrum of fermion zero modes (Witten 1984, Frampton *et al* 1984b). However, at least for $K = S^6$, $S^4 \times S^2$, $S^2 \times S^2 \times S^2$ or CP^3, it does not seem possible to obtain simultaneously consistency with (7.66), stability under classical perturbations in the gauge fields (see § 3.10), and a realistic fermion spectrum (Bailin and Love 1985b).

If, in order to solve the hierarchy problem of grand unified theory, it is demanded that $N = 1$ supersymmetry is unbroken at the compactification scale (Candelas *et al* 1985), and additionally that four-dimensional space is maximally symmetric, then at least for

$$H_{ABC} = 0 \qquad A, B, C = 0, 1, \ldots, 9 \tag{7.67}$$

four-dimensional space is Minkowski, and six-dimensional space is a Ricci flat Kahler manifold (see Eguchi *et al* (1980) for a discussion of Kahler manifolds)

$$\bar{\mathbb{R}}_{mn} = 0 \tag{7.68}$$

with the SU(3) holonomy group, and there is the additional constraint on the Yang-Mills field strength

$$F^a_{mn}\Gamma^{mn}\eta = 0 \tag{7.69}$$

where η is a covariently constant spinor on the compact manifold. (It is not clear whether solutions with $H_{mnp} \neq 0$, and unbroken $N = 1$ supersymmetry, are possible.) Six-(real)-dimensional Ricci flat Kahler manifolds of SU(3) holonomy are known to exist (Calabi 1957, Yau 1977). These manifolds have *no* continuous isometries. Thus there are *no* gauge fields in four dimensions *from the metric*, and these theories are at the opposite extreme to pure Kaluza-Klein theories. For $H = 0$, the Bianchi identity (7.66) may be satisfied (Candelas *et al* 1985) for the Yang-Mills group SO(32) or $E_8 \times E_8$ by identifying the expectation value of the gauge field with the spin connection for the Calabi-Yau manifold

$$A_m{}^{\alpha\beta} = \omega_m{}^{\alpha\beta} \tag{7.70}$$

where ω_m is as defined by (2.38), and α, β are flat indices for the tangent space of the compact manifold. Because the Calabi-Yau manifold has holonomy group SU(3), (7.70) requires an embedding of an SU(3) gauge field in G_{YM} to be defined. The SO(32) case leads (Candelas *et al* 1985) to unrealistic models, with only real vector-like representations of SO(32). The case $G_{YM} = E_8 \times E_8$ leads to a realistic supergravity GUT. With the above SU(3) embedded in the maximal SU(3) $\times$ E_6 subgroup of a single E_8, the expectation value (7.70) breaks the $E_8 \times E_8$ gauge group to $E_6 \times E_8$. If known particles are singlets under the E_8 factor, the E_6 is a promising GUT. The adjoint representation **248** of E_8 (to which the fermions belong in the superstring theory) has the decomposition under SU(3) $\times$ E_6,

$$\mathbf{248} = (\mathbf{3}, \mathbf{27}) + (\bar{\mathbf{3}}, \overline{\mathbf{27}}) + (\mathbf{8}, \mathbf{1}) + (\mathbf{1}, \mathbf{78}). \tag{7.71}$$

Thus, since chiral fermions must be non-singlet under both E_6 and SU(3) (since coupling to a gauge field expectation value is needed to produce zero modes, as in § 5), the quarks and leptons must belong to **27** or $\overline{\mathbf{27}}$ of E_6. With the gauge field expectation value given by (7.70), the number N_g of generations of fermions may be related using (5.148) to the Euler characteristic $\chi(K)$ of the compact manifold K, by (Candelas *et al* 1985)

$$N_g = \tfrac{1}{2}\chi(K). \tag{7.72}$$

To obtain a reasonably small number of generations it turns out to be necessary to use a *non*-simply connected Calabi–Yau manifold. This has the desirable spin-off of providing a mechanism for the first-stage breaking of the GUT symmetry at the compactification scale. Non-trivial Wilson loops $\exp(\int A_m \, dy^m)$ of gauge fields on the compact manifold then act as effective Higgs scalars in the adjoint representation **78** of E_6, by a mechanism discussed previously by Hosotani (1983a, b). Intense effort is being devoted at the time of writing to working out patterns of symmetry breaking for the E_6 GUT (Cecotti *et al* 1985, Breit *et al* 1985, Dine *et al* 1985a, Witten 1985a), to exploring mechanisms for breaking the $N = 1$ supersymmetry (Derendinger *et al* 1985, Dine *et al* 1985b), to investigating possible compactification mechanisms (Nepomechie *et al* 1985), and to deriving the details of the softly broken four-dimensional supergravity theory (Witten 1985b, Cohen *et al* 1985).

The difficulty of obtaining a consistent solution for $F_{AB} \neq 0$ of the field equation for the scalar field ϕ, which arose in the minimal ten-dimensional supergravity (Freedman *et al* 1983, Randjbar-Daemi *et al* 1983d) may be resolved in the present non-minimal theory. There arises (Candelas *et al* 1985) from the $E_8 \times E_8$ superstring theory (Gross *et al* 1985) a non-minimal curvature term which results in $\frac{1}{30}\mathrm{Tr}(F_{AB}F^{AB})$ in the Lagrangian being replaced by $\frac{1}{30}\mathrm{Tr}(F_{AB}F^{AB}) + \bar{\mathbb{R}}_{ABCD}\bar{\mathbb{R}}^{ABCD}$. With the gauge field expectation value given by (7.70) this causes the troublesome $\mathrm{Tr}(F_{AB}F^{AB})$ term in (7.51) to cancel, so that a consistent solution may be obtained with ϕ constant and $F_{AB} \neq 0$. (We have already assumed $H_{ABC} = 0$.) The complete set of field equations is then satisfied with 4-space Minkowski, and 6-space Ricci flat. It has been suggested (Zwiebach 1985) that the correct truncation of the superstring theory is obtained by using $\frac{1}{30}\mathrm{Tr}(F_{AB}F^{AB}) + \bar{\mathbb{R}}_{ABCD}\bar{\mathbb{R}}^{ABCD} - 4\bar{\mathbb{R}}_{AB}\bar{\mathbb{R}}^{AB} + \bar{\mathbb{R}}^2$ instead. For a Ricci flat compact manifold, this achieves the same effect so far as the ϕ field equation is concerned, but it also removes ghosts from the supergravity theory (Zwiebach 1985, Zumino 1985), and moreover removes a catastrophic instability of Friedmann–Robertson–Walker cosmology due to higher-order derivatives (Bailin and Love 1985c, Bailin *et al* 1985b).

8. Conclusions

In the purest form of Kaluza–Klein theory all gauge fields in four dimensions would arise from components of the metric in higher dimensions and there would be no explicit gauge fields present in the higher dimensionality. However, because observed light fermions are chiral (i.e. have left-handed components transforming differently from right-handed components under the observed $SU(3) \times SU(2) \times U(1)$ gauge group) it seems difficult to sustain this extreme position. Instead one probably has to assume that some gauge fields are already present in the higher-dimensional theory. Once such 'elementary' gauge fields are present, consistent theories are very tightly constrained by the requirement that all gauge, gravitational and mixed gauge-gravitational anomalies should cancel, and theories in which this cancellation is able to occur tend to have rather large higher-dimensional gauge groups. Particularly promising examples (with gauge group $E_8 \times E_8$ or $SO(32)$) are provided by the supergravity limit of superstring theory, and, at the time of writing, these models have progressed some distance towards a phenomenologically satisfactory theory.

One of the most direct ways of 'observing' the extra spatial dimensions may be in the cosmological context. Here, Kaluza-Klein theory can give a satisfactory late-time cosmology (without rapid variation of the gauge coupling constants or gravitational constant), and also suggests various mechanisms for achieving the desired cosmological inflation at early times, though none of these has so far proved entirely compelling when detailed calculations have been performed.

In summary, it seems quite likely that even if the original *pure* Kaluza-Klein theory cannot be sustained, extra spatial dimensions will play an important role in the eventual unified theory of interactions, and in understanding early cosmology.

References

Abbott R B, Barr S M and Ellis S D 1984 *Phys. Rev.* D **30** 720
—— 1985 *Phys. Rev.* D **31** 673
Adler S L 1969 *Phys. Rev.* **177** 2426
Alvarez-Gaumé L and Witten E 1984 *Nucl. Phys.* B **234** 269
Appelquist T and Chodos A 1983a *Phys. Rev. Lett.* **50** 141
—— 1983b *Phys. Rev.* D **28** 772
Appelquist T, Chodos A and Myers E 1983 *Phys. Lett.* **127B** 51
Atiyah M and Hirzebruch F 1970 *Essays on Topology and Related Topics* ed A Haeflinger and R Narasiman (Berlin: Springer) p 18
Awada M A, Duff M J and Pope C N 1983 *Phys. Rev. Lett.* **50** 294
Bailin D and Love A 1984a *Phys. Lett.* **137B** 348
—— 1984b *Phys. Lett.* **144B** 359
—— 1985a *Nucl. Phys.* B **254** 543
—— 1985b *Phys. Lett.* **157B** 375
—— 1985c *Phys. Lett.* **163B** 135
—— 1985d *Z. Phys.* C **29** 477
Bailin D, Love A and Stein-Schabes J 1985a *Nucl. Phys.* B **253** 387
Bailin D, Love A and Vayonakis C 1984 *Phys. Lett.* **142B** 344
Bailin D, Love A and Wong D 1985b *Phys. Lett.* **165B** 270
Bais F A, Nicolai H and van Nieuwenhuizen P 1983 *Nucl. Phys.* B **228** 333
Bardeen W A 1969 *Phys. Rev.* **184** 1848
Barrow J D and Stein-Schabes J 1985 *Phys. Rev.* D **32** 1595
Bell J and Jackiw R 1969 *Nuovo Cimento* A **60** 47
Bergshoeff M, de Roo B, de Wit B and van Nieuwenhuizen P 1982 *Nucl. Phys.* B **195** 97
Biran B, Casler A, Englert F, Rooman M and Spindel P 1984 *Phys. Lett.* **134B** 179
Biran B, Englert F, de Wit B and Nicolai H 1983 *Phys. Lett.* **124B** 45
Brans C and Dicke R H 1961 *Phys. Rev.* **124** 925
Breit J D, Ovrut B A and Segre G C 1985 *Phys. Lett.* **158B** 33
Brink L 1984 *CERN Preprint TH 4006/84*
Brink L, Schwarz J H and Scherk J 1977 *Nucl. Phys.* B **121** 77
Calabi E 1957 *Algebraic Geometry and Topology: A symposium in honour of S Lefschetz* (Princeton, NJ: Princeton University Press)
Candelas P, Horowitz G T, Strominger A and Witten E 1985 *Nucl. Phys.* B **258** 46
Candelas P and Raine D J 1984 *Nucl. Phys.* B **248** 415
Candelas P and Weinberg S 1984 *Nucl. Phys.* B **237** 397
Cartan E and Schouten J A 1926 *Proc. Kon. Ned. Adad., Amsterdam* **29** 803
Castellani L, D'Auria R and Fré P 1984a *Nucl. Phys.* B **239** 610
Castellani L, D'Auria R, Fré P, Pilch K and van Nieuwenhuizen P 1984b *Class. Quantum Grav.* **1** 339
Castellani L, Romans L J and Warner N P 1984c *Ann. Phys., NY* **157** 394
Cecotti S, Derendinger J P, Ferrara S, Girardello L and Rocandelli M 1985 *Phys. Lett.* **156B** 318
Chadha S and Daniel M 1985a *Rutherford preprint RAL 85-040*

Chadha S and Daniel M 1985b *Rutherford preprint RAL 85-039*
Chamseddine A H 1981a *Nucl. Phys.* B **185** 403
—— 1981b *Phys. Rev.* D **24** 3065
Chapline G F and Gibbons G W 1984 *Phys. Lett.* **135B** 43
Chapline G and Manton N S 1981 *Nucl. Phys.* B **184** 391
—— 1983 *Phys. Lett.* **120B** 105
Chapline G and Slansky R 1982 *Nucl. Phys.* B **209** 461
Charap J M and Duff M J 1977 *Phys. Lett.* **69B** 445
Cho Y M 1975 *J. Math. Phys.* **16** 2029
Cho Y M and Freund P G O 1975 *Phys. Rev.* D **12** 1711
Cohen E, Ellis J, Enquist K and Nanopoulos D V 1985 *CERN preprint TH 4195*
Coleman S 1975 *Phys. Rev.* D **11** 2088
Coleman S and Weinberg E 1973 *Phys. Rev.* D **7** 1888
Cremmer E and Julia B 1978 *Phys. Lett.* **80B** 48
—— 1979 *Nucl. Phys.* B **159** 141
Cremmer E, Julia B and Scherk J 1978 *Phys. Lett.* **76B** 409
Cremmer E and Scherk J 1976 *Nucl. Phys.* B **108** 409
—— 1977 *Nucl. Phys.* B **118** 61
D'Auria R and Fré P 1984a *Ann. Phys., NY* **157** 1
—— 1984b *Class. Quantum Grav.* **1** 447
D'Auria R, Fré P and van Nieuwenhuizen P 1983 *Phys. Lett.* **122B** 225
de Alwis S P and Koh I G 1984 *Phys. Lett.* **148B** 428
de Alwis S P, Koh I G and Yi W S 1985 *Phys. Rev.* D **31** 3086
Delbourgo R and Salam A 1972 *Phys. Lett.* **40B** 381
Derendinger J P, Ibanez L E and Nilles H P 1985 *Phys. Lett.* **155B** 65
Destri C, Orzalesi C A and Rossi P 1983 *Ann. Phys., NY* **147** 321
de Wit B and Nicolai H 1982 *Nucl. Phys.* B **208** 323
DeWitt B 1964 *Relativity, Groups and Topology Lectures at 1963 Les Houches School* ed B DeWitt and C DeWitt (New York: Gordon and Breach)
Dine M, Kaplunovski V, Mangano M, Nappi C and Seiberg N 1985a *Nucl. Phys.* B **259** 549
Dine M, Rohm R, Seiberg N and Witten E 1985b *Phys. Lett.* **156B** 55
Dolan L 1983 *Rockefeller preprint RU83/B/90*
—— 1984a *Rockefeller preprint RU84/B/89* (unpublished)
—— 1984b *Proc. Conf. on Vertex Operators in Mathematics and Physics, Berkeley 1983* (Berlin: Springer)
—— 1984c *Phys. Rep.* **109** 1
Dolan L and Duff M J 1984 *Phys. Rev. Lett.* **52** 14
Duff M J 1982 *Supergravity '81* ed S Ferrara and J G Taylor (Cambridge: Cambridge University Press)
—— 1984 *Imperial College preprint TP/83-84/45*
Duff M J, Koh I G and Nilsson B E W 1984a *Phys. Lett.* **148B** 431
Duff M J, Nilsson B E W and Pope C N 1983a *Phys. Lett.* **129B** 39
—— 1983b *Phys. Rev. Lett.* **50** 2043
—— 1984b *Phys. Lett.* **139B** 154
—— 1984c *Nucl. Phys.* B **233** 433
—— 1986 *Phys. Rep.* **130** 1
Duff M J, Nilsson B E W, Pope C N and Warner N P 1984d *Phys. Lett.* **149B** 90
Duff M J and Pope C N 1983 *Supersymmetry and Supergravity '82* ed S Ferrara, J G Taylor and P van Nieuwenhuizen (Singapore: World Scientific)
—— 1984 *NSF-ITP-84-166*
Duff M J, Pope C N and Warner N P 1983c *Phys. Lett.* **130B** 254
Eguchi T and Freund P G O 1976 *Phys. Rev. Lett.* **37** 1251
Eguchi T, Gilkey Pl B and Hanson A J 1980 *Phys. Rep.* **66** 213
Eichten E, Kang K and Koh I G 1982 *J. Math. Phys.* **23** 2529
Ellis J, Gaillard M, Maiani L and Zumino B 1980 *Unification of the fundamental interactions* ed J Ellis, S Ferrara and P van Nieuwenhuizen (New York: Plenum)
Englert F 1982 *Phys. Lett.* **119B** 339
Englert F and Nicolai H 1983 *CERN preprint TH 3711*
Englert F, Rooman M and Spindel P 1983 *Phys. Lett.* **127B** 47
Ezawa Z F and Koh I G 1984a *Phys. Lett.* **142B** 153
—— 1984b *Phys. Lett.* **142B** 157

Forgacs P and Manton N 1980 *Commun. Math. Phys.* **72** 15
Frampton P H 1983 *Phys. Lett.* **122B** 351
Frampton P H and Kephart T W 1983a *Phys. Rev. Lett.* **50** 1343, 1347
—— 1983b *Phys. Rev.* D **28** 1010
—— 1983c *Phys. Lett.* **131B** 80
Frampton P H, Moxhay P J and Yamamoto K 1984a *Phys. Lett.* **144B** 350
Frampton P H, Van Dam H and Yamamoto K 1984b *IFP-242-UNC*
Frampton P H and Yamamoto K 1984 *University of North Carolina preprint*
—— 1985 *Phys. Lett.* **156B** 345
Fré P and Zizzi P A 1984 *CERN Preprint TH4016/84*
Freedman D Z, Gibbons G W and West P C 1983 *Phys. Lett.* **124B** 491
Freund P G O 1982 *Nucl. Phys.* B **209** 146
Freund P G O and Rubin M A 1978 *Phys. Lett.* **76B** 401
—— 1980 *Phys. Lett.* **97B** 233
Friedman J A and Kolb E W 1985 *Phys. Rev. Lett.* **55** 1435
Fujikawa K 1979 *Phys. Rev. Lett.* **42** 1195
—— 1980a *Phys. Rev.* D **21** 2848
—— 1980b *Phys. Rev.* D **22** 1499 (Erratum)
Furusawa T and Hosoya A 1985 *Prog. Theor. Phys.* **73** 467
Geroch R P 1967a *J. Math. Phys.* **8** 782
—— 1967b *PhD thesis* Princeton University
—— 1970 *J. Math. Phys.* **11** 437
Gilbert G and McClain B 1984 *Texas preprint UTTG-7-84*
Gilbert G, McClain B and Rubin M A 1984 *Phys. Lett.* **142B** 28
Gleiser M, Rajpoot S and Taylor J G 1984 *Phys. Lett.* **138B** 377
Gliozzi F, Scherk J and Olive D 1977 *Nucl. Phys.* B **122** 253
Gourdin M 1982 *Basics of Lie Groups* (Gif-sur-Yvette: Editions Frontières)
Green M B 1983 *Surv. High Energy Phys.* **3** 127
Green M B and Schwarz J H 1982 *Phys. Lett.* **109B** 444
—— 1983 *Phys. Lett.* **122B** 143
—— 1984 *Phys. Lett.* **149B** 117
Green M B, Schwarz J H and Brink L 1982 *Nucl. Phys.* B **198** 474
Gross D J, Harvey J A, Martinec E and Rohm R 1985 *Phys. Rev. Lett.* **54** 502
Gross D J and Perry M J 1983 *Nucl. Phys.* B **226** 29
Guth A 1981 *Phys. Rev.* D **23** 347
Harvey J A, Kolb E W and Perry M J 1984 *Phys. Lett.* **149B** 465
Horvath Z and Palla L 1978 *Nucl. Phys.* B **142** 327
Horvath Z, Palla L, Cremmer E and Scherk J 1977 *Nucl. Phys.* B **127** 57
Hosotani Y 1983a *Phys. Lett.* **126B** 309
—— 1983b *Phys. Lett.* **129B** 193
Iliopoulos J, Itzykson C and Martin A 1975 *Rev. Mod. Phys.* **47** 165
Inami T and Yasudo O 1983 *Phys. Lett.* **133B** 180
Irvine J M and Humphreys R 1984 *Manchester preprint M/C TH8406*
Jordan P 1959 *Z. Phys.* **157** 112
Kac V G 1968 *Izv. Akad. Nauk. SSR Ser. Mat.* **32** 1323 (*Math. USSR Izu.* **2**)
Kaluza Th 1921 *Sitzungsber. Preuss Akad. Wiss. Berlin Math. Phys.* **K1** 966
Kephart T W 1984 *University of North Carolina preprint*
Kerner R 1968 *Ann. Inst. H Poincaré* **9** 143
Kibble T W B 1979 *J. Phys. A: Math. Gen.* **9** 1387
Klein O 1926 *Z. Phys.* **37** 895
Kogan Y I, Voloshin M B and Voronov N A 1983 *Phys. Lett.* **130B** 174
Koikawa T and Yoshimura M 1985 *Phys. Lett.* **155B** 137
Kolb E W, Lindley D and Seckel D 1984 *Phys. Rev.* D **30** 1205
Kolb E W and Slansky R 1984 *Phys. Lett.* **135B** 378
Lichnerowitz A 1963 *C.R. Acad. Sci., Paris* A–B **257** 7
Linde A D 1982 *Phys. Lett.* **108B** 389
Luciani J F 1978 *Nucl. Phys.* B **135** 111
Maeda K 1986 *Class. Quantum Grav.* **3** 233
Mandelstam S 1975 *Phys. Rev.* D **11** 3026

Manton N S 1979 *Nucl. Phys.* B **158** 141
—— 1981 *Nucl. Phys.* B **193** 391
Marcus N and Schwarz J H 1982 *Phys. Lett.* **115B** 111
Matsuki Y 1983 *Phys. Rev.* D **28** 2107
Matsuki Y and Hill A 1983 *OSU preprint*
Misner C W 1963 *J. Math. Phys.* **4** 924
Moody R 1968 *J. Algebra* **10** 211
Nahm W 1978 *Nucl. Phys.* B **135** 149
Nepomechie R I, Wu Y S and Zee A 1985 *University of Washington preprint*
Nilsson B E W and Pope C N 1984 *Class. Quantum Grav.* **1** 499
Okada Y 1985 *Tokyo preprint UT-448*
Page D N 1984 *Texas preprint*
Perry M J 1984 *Phys. Lett.* **137B** 171
Preskill J P 1979 *Phys. Rev. Lett.* **43** 1365
Pollard D 1983 *J. Phys. A: Math. Gen.* **16** 565
Rajpoot S 1980 *Phys. Rev.* D **22** 2244
Randjbar-Daemi S 1983 *Proc. 12th Int. Conf. on Differential Geometric Methods in Theoretical Physics, Clausthal*
Randjbar-Daemi S and Percacci R 1982 *Phys. Lett.* **117B** 41
Randjbar-Daemi S, Salam A, Sezgin E and Strathdee J 1984a *Preprint IC/84/218*
Randjbar-Daemi S, Salam A and Strathdee J 1983a *Nucl. Phys.* B **214** 491
—— 1983b *Phys. Lett.* **124B** 345
—— 1983c *Phys. Lett.* **132B** 56
—— 1983d *Phys. Lett.* **124B** 349
—— 1984b *Preprint IC/84/15*
—— 1984c *Preprint IC/84/62*
—— 1984d *Phys. Lett.* **135B** 388
Rivers R J 1984 *Z. Phys.* C **22** 137
Rubin M A and Roth B D 1983 *Nucl. Phys.* B **226** 444
Sahdev D 1984 *Phys. Lett.* **137B** 155
Salam A and Strathdee J 1982 *Ann. Phys., NY* **141** 316
Schellekens A N 1984 *Phys. Lett.* **143B** 121
—— 1985a *Nucl. Phys.* B **250** 252
—— 1985b *Nucl. Phys.* B **262** 661
Scherk J 1978 *Recent Developments in Gravitation, Cargese (1978)* ed M Levy and S Deser (New York: Plenum)
Scherk J and Schwarz J H 1975 *Phys. Lett.* **57B** 463
—— 1979 *Nucl. Phys.* B **153** 61
Schrödinger E 1932 *Phys. Math.* **K1** **25** 102
Schwarz J H 1982 *Phys. Rep.* **89** 223
—— 1984 *Comment. Nucl. Part. Phys.* **13** 103
Schwarz J H and West P C 1983 *Phys. Lett.* **126B** 301
Shafi Q and Wetterich C 1983 *Phys. Lett.* **129B** 387
Sorkin R D 1983 *Phys. Rev. Lett.* **51** 87
Steinhardt P J 1983 *Pennsylvania preprint 0242T*
Symanzik K 1964 *Analysis in Function Space* ed Martin and Segal (Cambridge, MA: MIT Press)
—— 1970 *Commun. Math. Phys.* **16** 48
Tipler F J 1977 *Ann. Phys., NY* **108** 1
—— 1985 *Phys. Lett.* **165B** 67
Townsend P K and Sierra G 1983 *Nucl. Phys.* B **222** 493
Trautman A 1970 *Rep. Math. Phys.* 1 29
Watamura S 1983 *Phys. Lett.* **129B** 188
—— 1984 *Phys. Lett.* **136B** 245
Weinberg S 1972 *Gravitation and Cosmology* (New York: Wiley)
—— 1983 *Phys. Lett.* **125B** 265
Wess J and Zumino B 1971 *Phys. Lett.* **37B** 15
Wetterich C 1981 *Nucl. Phys.* B **187** 343
—— 1982a *Phys. Lett.* **110B** 379

Wetterich C 1982b *Phys. Lett.* **113B** 377
—— 1982c *CERN Preprint TH3310*
—— 1982d *CERN Preprint TH3418*
—— 1983 *CERN Preprint TH3517*
Wheeler J A 1968 *Battelle Rencontres* ed C DeWitt and J A Wheeler (New York: Benjamin)
Wilczek F 1977 *Quark Confinement and Field Theory* ed Stump and E Weingarten (New York: Wiley Interscience)
Witten E 1981 *Nucl. Phys.* B **186** 412
—— 1982 *Nucl. Phys.* B **195** 481
—— 1983 *Princeton Preprint*
—— 1984 *Phys. Lett.* **149B** 351
—— 1985a *Phys. Lett.* **155B** 151
—— 1985b *Nucl. Phys.* B **258** 75
Wu Y S and Zee A 1984 *Washington preprint 40048-25* p 3
Yau S T 1977 *Proc. Natl Acad. Sci. USA* **74** 1798
Yoshimura M 1984 *Phys. Rev.* D **30** 344
—— 1985 *KEK preprint KEK TH107*
Young R E 1984 *Phys. Lett.* **142B** 149
Zee A 1981 *Proc. 4th Summer Inst. on Grand Unified Theories and Related Topics* ed M Konuma and Thashawa (Singapore: World Scientific)
Zel'dovich Ya B and Khlopov M Y 1978 *Phys. Lett.* **79B** 239
Zumino B 1985 *Berkeley preprint*
Zumino B, Yong-Shi W and Zee A 1983 *Washington preprint*
Zwiebach B 1985 *Phys. Lett.* **156B** 315