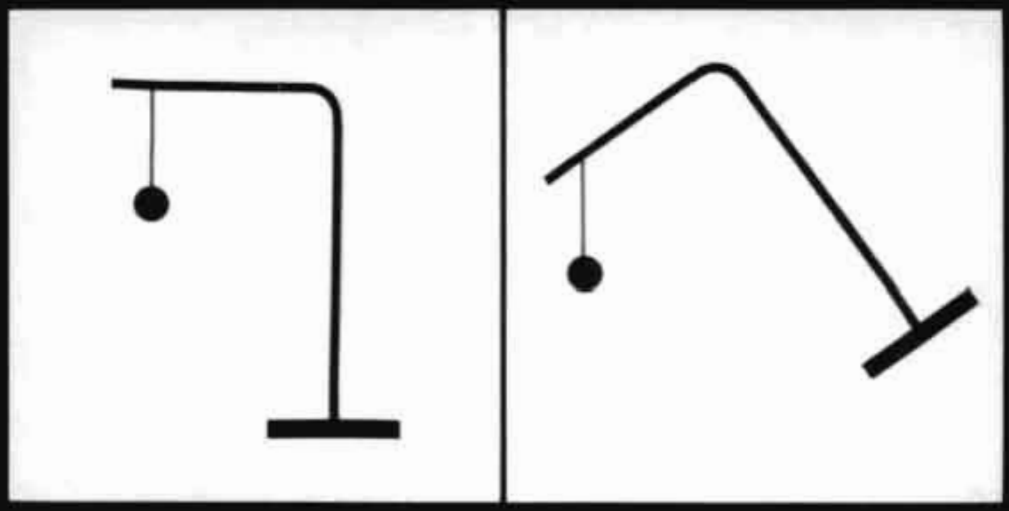
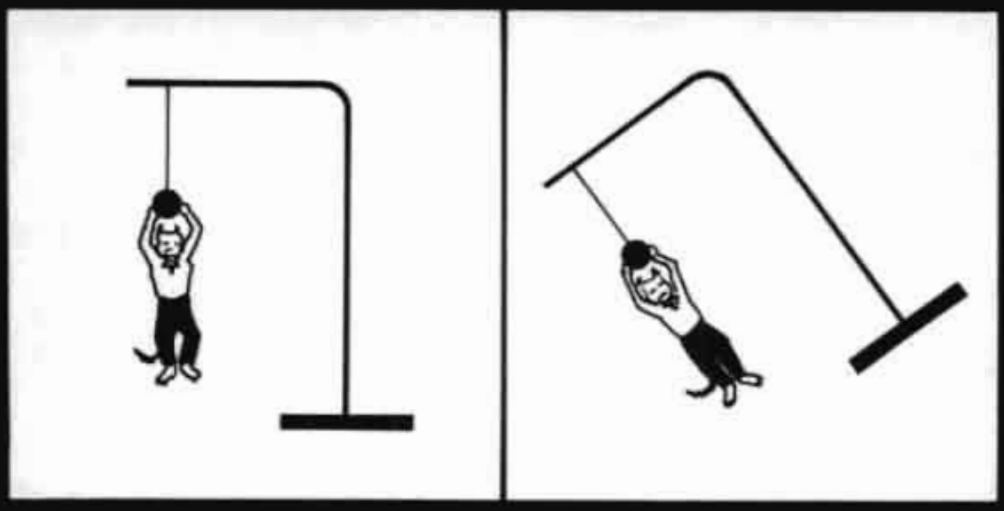
ORIGIN of SYMMETRIES





Colin D Froggatt and Holger B Nielsen

World Scientific

Published by

World Scientific Publishing Co. Pte. Ltd.

P O Box 128, Farrer Road, Singapore 9128

USA office: Suite 1B, 1060 Main Street, River Edge, NJ 07661

UK office: 73 Lynton Mead, Totteridge, London N20 8DH

The editors and publisher are grateful to the authors and the following publishers of the various journals and books for their assistance and permission to reproduce the selected articles found in this volume:

Academic Press (*Annals of Physics*); American Institute of Physics (*Rev. Mod. Phys.*); Elsevier Science Publishers (*Nucl. Phys.* and *Phys. Lett.*); Plenum Publishing Corp. (*Symmetries in Science*); Progress of Theoretical Physics (*Prog. Theor. Phys.*); Springer-Verlag (*Commun. Math. Phys.*); American Physical Society (*Phys. Rev.* and *Phys. Rev. Lett.*).

ORIGIN OF SYMMETRIES

ISBN 9971-96-630-1
ISBN 9971-96-631-X (pbk)

Printed in Singapore by JBW Printers & Binders Pte. Ltd.

CONTENTS

REPRINTED PAPERS

ORIGIN of SYMMETRIES

Chapter I

INTRODUCTION

Symmetry is a vast subject with great significance in art and nature.[1] In this book we shall confine our attention to the symmetries of the laws of nature. These symmetries play a very important role, both in the derivation of the consequences of the laws and in the development of physics [Papers 1 and 2]. A notable example is the close connection between a symmetry and a conservation law provided by Noether's theorem.[2]

It is a remarkable fact that the laws of nature have such an arsenal of symmetries as translational, rotational and Lorentz invariance, gauge symmetries, isospin invariance etc. The literature abounds with applications of and proposals for symmetries of the laws of nature. However we think it is fair to say that the question of explaining why there are such symmetries has not been so widely studied. It is the purpose of this book to review the development in our understanding of the origin of symmetries and to reprint a selection of papers giving derivations or explanations of various symmetries.

We do not intend to present explanations of an essentially philosophical nature. For example we do not discuss the anthropic principle,[3] according to which some symmetries have to be present because they are necessary for the existence of life and observers. In fact we restrict ourselves to explanations whereby a symmetry is understood as a consequence of other features of the physical laws, possibly in some restricted regime of physical phenomena. In order to illustrate what we have in mind, let us consider how one might imagine deriving parity conservation in a quantum field theory with only spinless fields.

In a toy model with only spinless fundamental fields $\phi^a(x)$, $(a = 1, 2, \ldots, N)$, the most general renormalisable Poincaré invariant action is

$$S = \int \mathrm{d}^4 x \mathcal{L}(x) \tag{1}$$

where

$$\mathcal{L}(x) = \frac{1}{2}\sum_{a=1}^{N} \partial_\mu \phi^a \partial^\mu \phi^a - V(\phi^b). \tag{2}$$

Here $V(\phi^b)$ can be any polynomial up to fourth order in the spinless fields $\phi^b(x)$. By dimensional counting, renormalisability requires that there be no coefficient in the Lagrangian $\mathcal{L}(x)$ having dimension of mass to a negative power. This requirement prevents the occurrence of terms with gradients other than the usual kinetic energy term $\frac{1}{2}\partial_\mu \phi^a \partial^\mu \phi^a$. Consequently there is no way to make use of the Levi-Civita symbol $\epsilon_{\mu\nu\rho\sigma}$ in a renormalisable action with only spinless fields. It follows that the action must be invariant under the 'parity operation'

$$\begin{gathered} \mathbf{x} \longrightarrow -\mathbf{x} \\ \phi^a(t,\mathbf{x}) \longrightarrow \phi^a(t,-\mathbf{x}). \end{gathered} \tag{3}$$

Under this operation we are treating all the spinless fields as scalar (not pseudoscalar) fields. We can consider the above argument as an explanation or derivation of parity symmetry in our toy world. It is a prediction of this "derivation" that all the fundamental particles in the toy world are scalars under the derived parity. For special forms of the polynomial $V(\phi^b)$ there could, in addition, exist an alternative conserved parity operation, under which some of the fields ϕ^b are pseudoscalar. Our derived symmetry (3) might not then be called parity, but rather be considered as a combination of parity with a symmetry under sign shift of the pseudoscalar fields.

The requirement of the existence of a renormalisable field theory for the spinless fields played a crucial role in the above derivation of parity symmetry. If the spinless particles were not really fundamental but only bound states, then our Lagrangian (2) would become an effective Lagrangian not constrained by renormalisability. An effective term such as

$$\frac{\phi^5}{M^4}\epsilon^{\mu\nu\rho\sigma}\partial_\mu\phi^1\partial_\nu\phi^2\partial_\rho\phi^3\partial_\sigma\phi^4 \tag{4}$$

would then be allowed, provided there are at least five types of spinless fields. Here M is a constant having the dimension of mass, and the parity operation (3) would be an approximate symmetry at energies low compared to M.

The above type of argument based on renormalisability, but applied to more realistic Yang-Mills theories, provides an explanation of many of the symmetries of the standard SU(3)×SU(2)×U(1) model of particle physics. Thus parity, charge conjugation, time reversal and strangeness (or, more generally, flavour) conservation are no longer considered as fundamental symmetries, but as natural consequences of a gauge field theory of strong and electromagnetic interactions. The development in the interpretation of these symmetries, from being *a priori* principles to being automatic consequences of gauge theories, is discussed in [Paper 3]. This paper

also discusses the derivation of the approximate symmetries of the standard model, such as chiral symmetry and Gell-Mann's SU(3) flavour symmetry in the limit of negligible quark masses. The symmetries of the standard model are the subject of Chapter III.

Encouraged by the derivation of the standard model based symmetries, it is tempting to speculate that even the supposedly fundamental symmetries of Poincaré invariance and gauge invariance may be derivable , in some physical limit. The extreme point of view, that all the symmetries of the laws of nature should be derivable in this way, is the subject of the last chapter. It forms the main content of the ambitious "random dynamics" program[4] [Paper 30].

The more conventional response to the successes of the standard model and its symmetries is to assume that the most fundamental laws of nature must have a large degree of symmetry. Many of these symmetries would then be broken, one way or another, as one descends in energy to the present experimentally studied energy regime. This is the point of view manifested by Grand Unified Theories and by supersymmetry or supergravity models which we discuss in Chapter IV. In effect this philosophy amounts to postulating the observed gauge symmetry group, since the symmetry is only explained by the existence of an even bigger gauge symmetry group, which is itself not explained. However, this is to take too negative an attitude. For instance, the $N = 8$ extended supergravity model can be 'explained' or rather specified, by requiring the maximum amount of supersymmetry consistent with the absence of elementary fields having spin greater than two. Currents with spin greater than two would give such strong restrictions on scattering processes that they would enforce an essentially free and therefore uninteresting theory.[5] Similarly it appeared for some time that superstring theory was only consistent for an SO(32) or $E_8 \times E_8$ gauge group [Paper 21], but now other possibilities are known (see chapter 4.6).

A distinction is often made between geometrical and non-geometrical symmetries. However the content of such a distinction can be somewhat arbitrary. Space translational and rotational invariances are clearly geometrical. But it is only since the publication of Einstein's theory of relativity that Lorentz invariance (or Galilean invariance) and time translational invariance have been considered geometrical. Further it is a matter of taste whether a supersymmetry (should it exist) is called geometrical, since it is intertwined in a non-trivial way with the geometrical symmetries of space and time translations. Also it is questionable whether the general co- ordinate invariance, or diffeomorphism symmetry, of Einstein's theory of gravity should be called a geometrical symmetry, since it is such a formal symmetry. On the other hand the gauge symmetries of Yang-Mills theories are usually not called geometrical, unless one takes a rather extended definition of "geometrical" within a fibre bundle formulation. However a natural generalisation of the work of Kaluza and Klein [Paper 20] showed how the diffeomorphism symmetry, or reparameterisation invariance, of a higher than four dimensional space-time manifold could give rise to the gauge symmetry of a Yang-Mills theory. So, using the Kaluza-Klein idea,

one may "derive" gauge symmetries from the geometrical symmetries of a higher dimensional space-time. In fact part of the diffeomorphism symmetry of gravity in the higher dimensions becomes a gauge symmetry in four dimensions. In some sense, however, this just means that one assumed symmetry is interpreted as or is transformed into another one.

As emphasized in Wigner's article [Paper 1], we must separate the symmetries of the physical laws from the symmetries of the whole world as it now exists. In general the state of the world is not invariant under the symmetries of the laws of nature. One says that the "initial conditions" are not invariant.

The distinction between the laws of nature and the initial conditions becomes rather blurred in the case of spontaneous symmetry breakdown. Spontaneous symmetry breakdown occurs when the vacuum state is not invariant under the symmetry in question. In this case, should the state of the vacuum be considered as part of the laws of nature or just as part of the initial conditions? This ambiguity plays a role in some of the derivations of gauge symmetries discussed in Chapter VI.

When the vacuum is not invariant under a symmetry of the laws of nature, most physical processes do not manifest this symmetry. However there are still physical consequences which make the symmetry experimentally testable. In the case of a spontaneously broken continuous global symmetry, the Nambu-Goldstone theorem predicts the existence of a massless particle. This particle is a boson, unless a supersymmetry is broken. The prediction can be avoided by the Higgs-Kibble mechanism when the spontaneously broken symmetry is a local gauge symmetry. The predicted Nambu-Goldstone boson can then be shown to be unphysical, by an appropriate gauge choice. At the same time, the otherwise massless spin 1 gauge particle acquires a non-zero mass. The degree of freedom associated with the Nambu-Goldstone boson is transformed into the required helicity zero component of the massive spin 1 particle. Even after "Higgsing" the gauge symmetry in this way, there are still traces of the symmetry left. For instance the coupling constants of various particles to the gauge particle are related. These traces of the gauge symmetry remain due to the requirement of renormalisability. Renormalisability forbids terms in the Lagrangian density having coefficients of dimension mass to a negative power. If such irrelevant (i.e. non-renormalisable) terms were allowed, the last traces of the gauge symmetry would vanish. We make use of this possibility in Chapter VI, when an arbitrary non-gauge invariant lattice theory is written as a "Higgsed" theory.

The successes achieved by the application of symmetry principles in physics have led to the consideration of so-called dynamical symmetries.[6] Dynamical symmetries are not true symmetries in the sense of being exhibited by the laws of nature. For dynamical symmetries, the Hamiltonian H is a function of the symmetry group generators but H does not commute with all the generators. The multiplets of the dynamical symmetry group are then split by H, but not mixed. Thus the dynamical symmetry can be important for the classification of the Hamiltonian eigenstates, even though the actual system does not really exhibit the symmetry in question. A

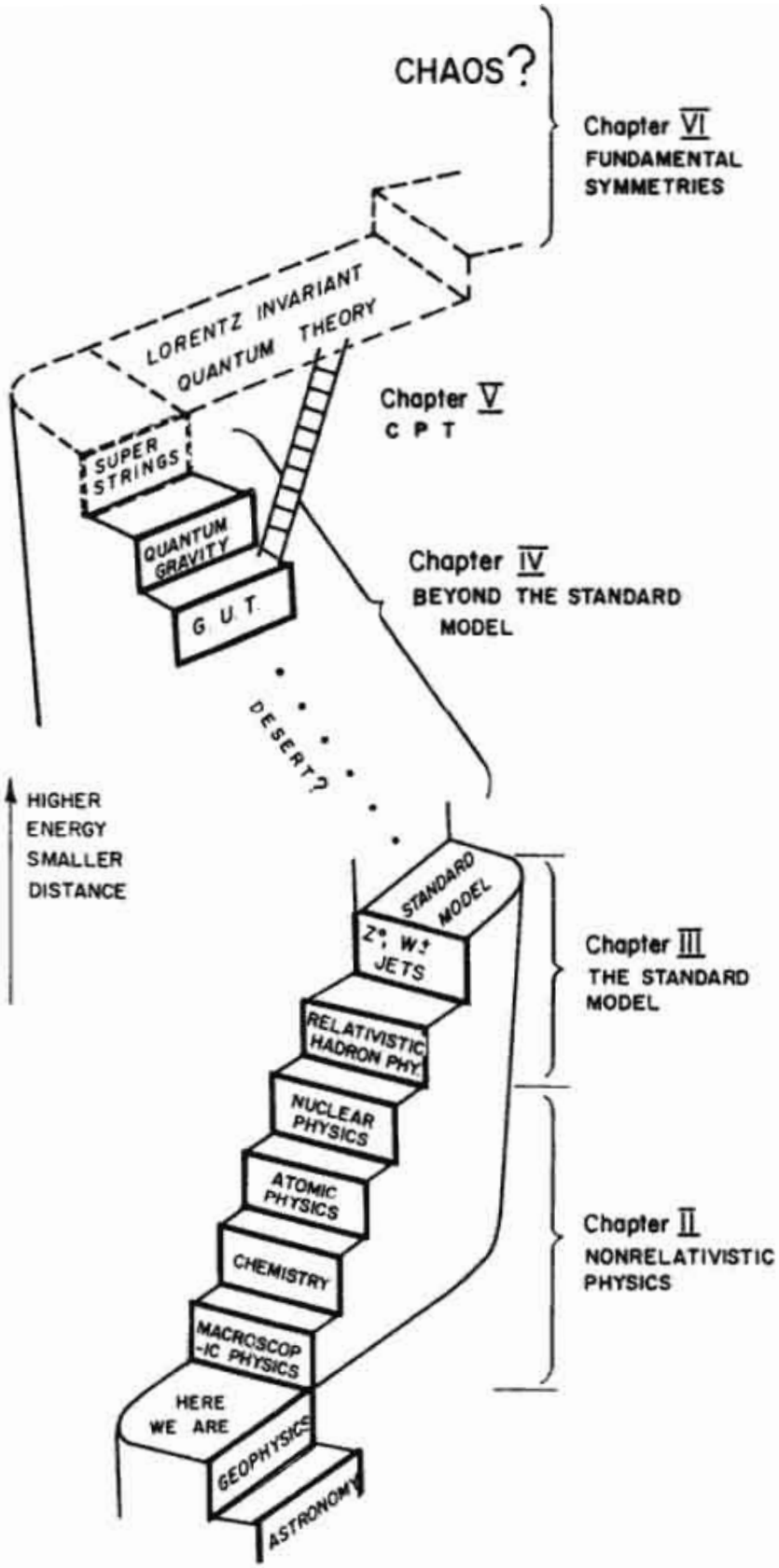


Fig. 1.1. The Quantum Staircase; the upper part is of course speculative and can be changed to accommodate the reader's favourite theory of everything.

simple example is the SU(3) symmetry of nuclear physics introduced by Elliot.[6,7] We shall not be concerned with such dynamical symmetries in this book, since they are not true symmetries of the laws of nature.

In the following chapters we give examples of symmetry derivations, which are organised to follow the quantum staircase (Fig. 1.1). That is to say we first discuss the derivation of symmetries in the lowest energy (or rather longest distance) regime. We then go, successively, to higher and higher energy (smaller and smaller distance) physics. Thus in Chapter II we consider non-relativistic examples. In Chapter III we consider the symmetries derived from, but not directly put into, the well-established standard SU(3)×SU(2)×U(1) Yang-Mills model of quarks and leptons. Chapter IV deals with symmetries derived from theories beyond the standard model. A separate Chapter V is devoted to the CPT theorem, which provides the paradigm for a symmetry derivation. In Chapter VI we present the more speculative derivations of the most fundamental symmetries of Lorentz invariance and gauge invariance. Finally, in Chapter VII, we present an overview of the progress made in explaining the origin of symmetries and a discussion of random dynamics.

It should be emphasized that, in most cases, the "symmetry derivations" are *a posteriori* understandings of symmetries, which have been known in advance on more phenomenological grounds. If one takes the random dynamics program seriously, however, one could hope to derive all the observed symmetries without having to assume any *a priori*. This would represent a new philosophy of the origin of symmetries: the observed symmetries arise naturally even if the most fundamental laws (at very high energies) do not possess them.[4,8,9] According to this philosophy, one might attempt to read the chapters of our book in reverse order (i.e. coming down the quantum staircase). Although there are many gaps to be filled, one would then take it as a logical derivation of first some very fundamental symmetries which are used in earlier chapters to derive all the experimentally observed symmetries. Needless to say, this ambitious program of random dynamics is far from completion in a convincing form. It is not presented here as the most compelling argument for believing in the symmetries.

References

1. H. Weyl, *Symmetry* (Princeton University Press, 1952).
2. E. L. Hill, *Rev. Mod. Phys.* **23**, 253 (1957).
3. J. Barrow and F. Tipler, *The Anthropic Cosmological Principle* (Oxford University Press, 1986).
4. H. B. Nielsen, *Gauge Theories of the Eighties*, "Akäslompolo, Finland 1982, eds. R. Raito and J. Lindfors, p. 288 (Springer-Verlag, 1983); H. B. Nielsen, D. L. Bennett and N. Brene, "Recent Developments in Quantum Field Theory," *Proceedings of the Niel Bohr Centennial Conference*, Copenhagen 1985, eds. J. Ambjørn, B. Durhuus and J. L. Petersen, p. 263 (North-Holland, 1985).
5. S. Weinberg, *Phys. Rev.* **B138**, 988 (1965); R. Behrends, B. de Wit, J. van Holten and P. van Nieuwenhuizen, *J. Phys.*, **A13**, 1643 (1980).
6. A. Arima and F. Iachello, *Ann. Phys.* (NY) **111**, 201 (1978).
7. J. P. Elliot, *Proc. Roy. Soc.* **A245**, 128 and 562 (1958).
8. J. Iliopolous, *Proceedings of 1979 International Conference on High Energy Physics*, Geneva, Vol. 1, p. 371 (CERN, 1980).
9. F. Englert, "Recent Developments in Quantum Field Theory", *Proceedings of the Niels Bohr Centennial Conference*, Copenhagen 1985, eds. J. Ambjørn, B. Durhuus and J. L. Petersen, p. 39 (North-Holland, 1985).

Chapter II

SYMMETRIES FROM NON-RELATIVISTIC PHYSICS

We commence with an example from the bottom of the quantum staircase and consider the derivation of the scaling symmetries of macroscopic physics. The laws of physics are not themselves invariant under a change of scale, as first pointed out 300 years ago by Galileo in his discussion of the sizes of living creatures.[1] The physical effects of scaling are very often studied using the method of dimensions. Dimensional analysis is particularly useful in the design of small scale models for testing an aeroplane or a ship.

As a simple example from hydrodynamics, suppose the motion of a body in a viscous incompressible fluid is to be modelled. The model body is geometrically similar to the prototype, differing from it only in size. Then, provided the Reynolds number

$$R = \frac{\rho l v}{\mu} \tag{1}$$

is the same for the model and the prototype, the flow patterns are dynamically similar.[2] Here ρ and μ are the density and coefficient of viscosity for the fluid, l is a characteristic length of the object and v is some representative velocity. For simplicity we shall just consider one fluid, so that ρ and μ are constants. There is then a symmetry under the operation of scaling the size of the model up by a factor ξ and, at the same time, reducing velocities by the same factor.

It is clear, from the atomic structure of matter, that the above scaling law cannot be exact. Indeed, if we scale down to a model size so small that only a fraction of a molecule of fluid is required, the scaling law becomes meaningless. Thus we can really only expect the scaling law to be approximately true, corresponding to the situation when there are many molecules present. In practice, of course, the number of molecules will be very large for any real model and the scaling law can be a very good approximation.

This scaling symmetry is usually derived by applying dimensional arguments to the hydrodynamical differential equations – Navier-Stokes equations.[2] Let us

here derive the scaling law directly from the underlying molecular dynamics, in the limit when the number of fluid molecules becomes very large. The scaling law can be stated in the following form: The time development of two flows of the same incompressible fluid around two solid bodies are dynamically similar, when the second flow is related to the first by scaling (a) all geometrical lengths by a factor ξ, (b) all (non-thermal) velocities by a factor ξ^{-1} and (c) all time intervals by a factor ξ^2.

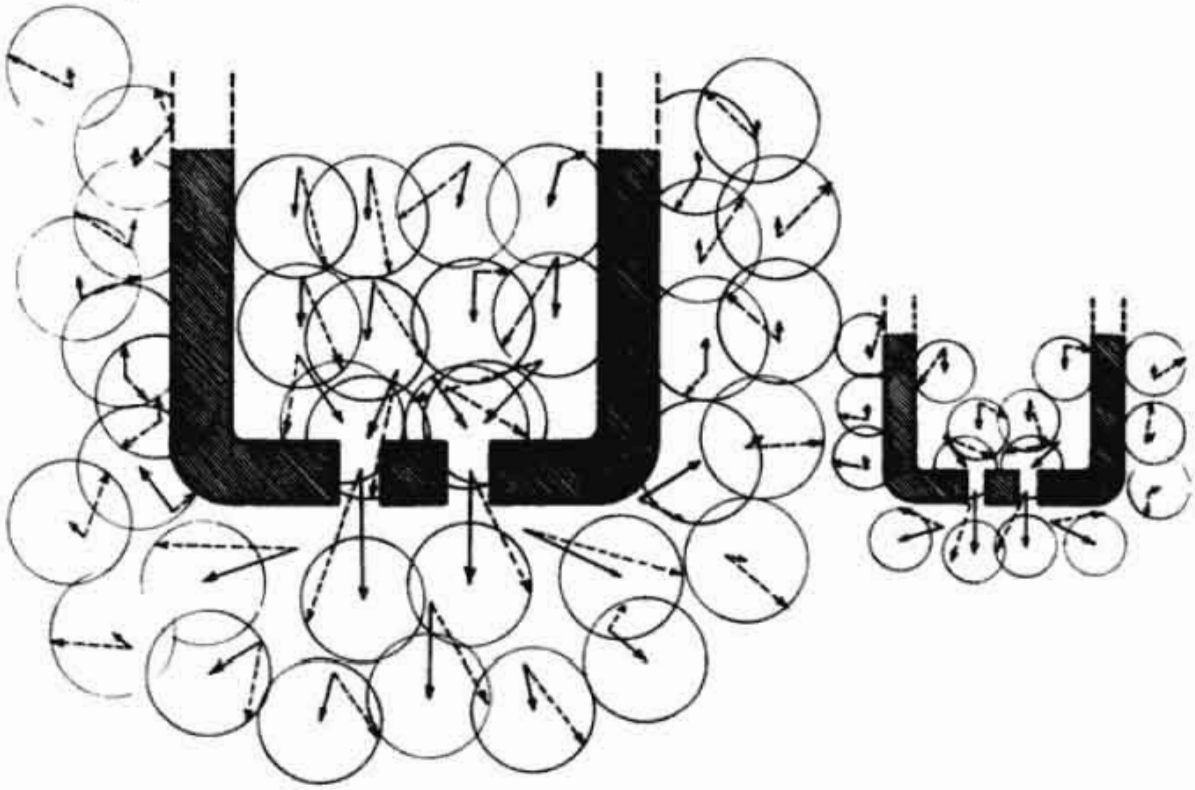


Fig. 2.1. Hydrodynamical scaling symmetry illustrated by same fluid flowing out of two similar tubes related by a scaling factor $\xi = 1/2$. The full-line arrows represent the displacements of molecules due to the average flow of the fluid during time intervals Δt and $\xi^2 \Delta t$ respectively. The dashed arrows represent the total displacements including the effects of Brownian motion.

The derivation of the scaling law from molecular physics proceeds along the following lines. The velocity of each molecule of fluid is considered to be composed of a directed motion component $\mathbf{v}_d$ and a Brownian or thermal motion component $\mathbf{v}_B$. We now consider the development of the two flow patterns, during corresponding small time intervals Δt and $\xi^2 \Delta t$ respectively. First suppose each molecule only had the directed motion component of its velocity. Then molecules from corresponding regions in the two flow patterns (see Fig. 2.1) would travel with velocities $\mathbf{v}_d$ and $\xi^{-1}\mathbf{v}_d$, for time intervals Δt and $\xi^2 \Delta t$ respectively. It follows that a molecule in the second flow travels a distance ξ times further than a corresponding molecule in the first flow, during these time intervals. Now we must consider the Brownian motion due to the thermal fluctuations of the molecules, which is superimposed onto the directed motion. The root mean square displacement of a molecule in a given fluid, due to this thermal motion, is proportional to the square root of the time interval considered. Thus the thermal displacement of the molecules in the second flow is also ξ times further than that in the first flow, during the corresponding time

intervals. This follows simply because we are considering the same fluid, with the same local properties, in both flow patterns, and the time interval for the second flow is ξ^2 times that for the first flow.

We have shown above, for both components of their motion, that the displacement of molecules from a region of the second flow is related to the displacement of molecules from the corresponding region of the first flow by the scaling factor ξ. Does it now follow, after the time intervals Δt and $\xi^2 \Delta t$ in the two flows, that the average (i.e. directed) velocities in corresponding small regions are still related by a factor ξ^{-1}, as required for dynamical similarity? Yes, indeed it does, as is easily seen by considering the molecules in small corresponding regions of the two flow patterns at some instant. Since the scaling factor ξ applies to the displacements due to both components of their motion, the same fraction of these molecules must arrive in two other small corresponding regions, after time intervals Δt and $\xi^2 \Delta t$ respectively. Thus the average molecular velocities, in corresponding regions of the two flow patterns, are still related by the scaling factor ξ^{-1}.

We have therefore derived the hydrodynamical scaling law for an incompressible fluid from molecular dynamics, in the limit of a large number of molecules where the methods of statistical mechanics apply. It should, of course, be stressed that the above argument does not imply a general scaling symmetry of macroscopic physics. In fact, under the scaling law derived above, the extra pressure p (relative to some representative value in the fluid) in corresponding regions of the two flows scales according to $p \rightarrow \xi^{-2} p$. It follows that the force exerted by the fluid on corresponding areas of the two solid bodies is the same, and the stresses on the body scale like ξ^{-2}. Hence the body would break for a sufficiently small ξ, i.e. the strength of the solid material violates the scaling law. Also, for small ξ, the velocity of the fluid and the pressure become high. In practice the velocity of a real fluid would approach the velocity of sound and the fluid could no longer be considered incompressible. The scaling law in the above form then breaks down, since the dimensionless Mach number becomes relevant to the flow pattern as well as the Reynold number.

As a second example from macroscopic physics, we consider the mechanical properties of an organic compound, such as sugar, which has parity non-invariant molecules. It has been known, since Pasteur's work in 1848, that all sugar molecules produced naturally from living things have the same kind of thread or handedness.[3] The mirror image molecule is not produced naturally. Sugar is therefore optically active and when polarised light is passed through a solution of naturally produced sugar its plane of polarisation is turned to the right. However when we manufacture sugar artificially, from substances which are not themselves asymmetrical, both kinds of molecules are produced in equal numbers. There is thus a kind of spontaneous breakdown of parity symmetry in naturally produced sugar, arising from the common origin of life on earth right back to the completely molecular level. Despite this parity violation in the structure of sugar molecules, the purely mechanical properties of a sugar crystal or of amorphous sugar are parity invariant.

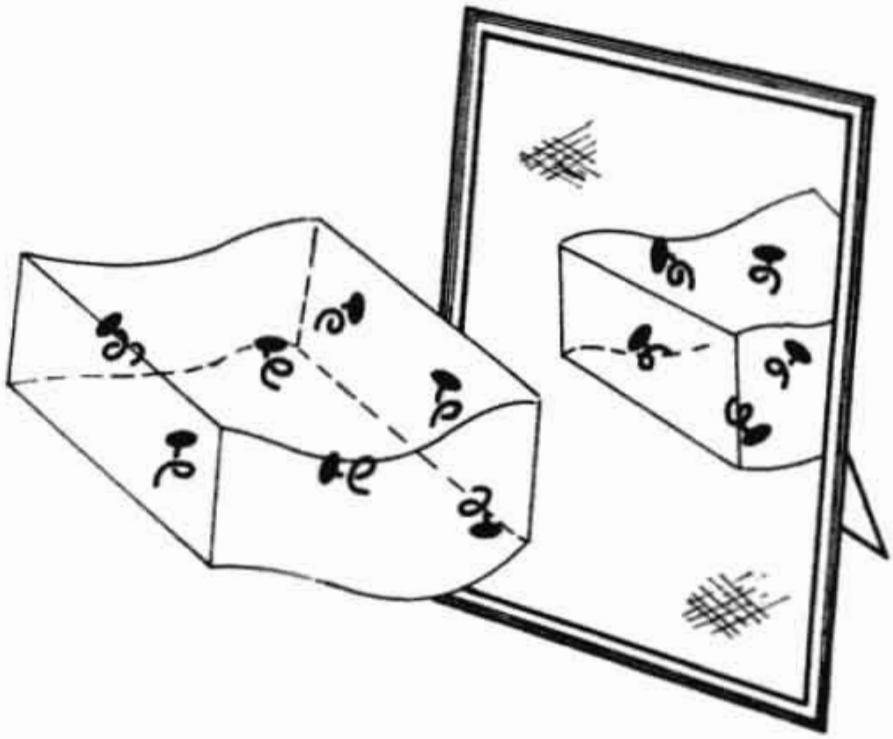


Fig. 2.2. Sugar and its mirror image have identical elasticity properties, despite the parity non-invariance of its molecules.

What is the origin of this parity symmetry in the elastic properties of sugar?

The important point is that the parity violation in sugar is microscopic (i.e. the handedness of single molecules) and thus can only affect the local form of the elastic deformation (Fig. 2.2). The local deformation of a material is described by a second rank symmetric tensor — the strain tensor.[4] The free energy density of a material therefore only depends on the local strain tensor, which is invariant under parity simply because the strain tensor is second rank and not a pseudotensor. There is simply no way the elasticity coefficients can be made complicated enough to violate parity symmetry. Even if the microscopic parity violation were due to truly parity violating forces, as in weak interactions, the elastic properties of sugar could not violate parity symmetry.

We now proceed up the quantum staircase and consider symmetries in non-relativistic quantum mechanics. We first consider the apparently accidental degeneracies of energy levels, which occur for some dynamical systems. In particular the origins of the underlying O(4) and SU(n) symmetries, responsible for the degeneracies in the hydrogen atom and the n-dimensional harmonic oscillator, are presented [Papers 4 and 5].

The best known example is that of the Schrödinger equation for the electron in a hydrogen atom, which has the symmetry of the four dimensional rotation group O(4). The same symmetry of course applies to any particle in a Coulomb-like $\frac{1}{r}$ potential, such as for planetary motion. The great progress made in science during this millennium is due in large part to the tractability of the $\frac{1}{r}$ potential

problem, which is a consequence of this very O(4) symmetry making the behaviour of the system much simpler. The O(4) symmetry has been very helpful in making the Keplerian law simple, as well as in the development of Bohr's atomic model, although Kepler and Bohr were unaware of the symmetry at the time.

The O(4) symmetry of the hydrogen atom is a rather hidden symmetry, which Fock [Paper 4] first made manifest by writing the Schrödinger equation in momentum representation. By stereographic projection onto the surface S_3 of a unit four dimensional Euclidean sphere (Fig. 2.3), the points in momentum space are associated with the points on S_3. After a simple transformation of the wave function, Fock then obtains a new Schrödinger equation for the hydrogen atom, in which the Hamiltonian operator is expressed as a convolution with the function

$$\frac{\text{constant}}{\sin^2 \frac{w}{2}} \tag{2}$$

where w is the distance along a great circle on the S_3-sphere. In this way, it is seen that the Hamiltonian for the hydrogen atom is invariant under rotations, in 4-space, of this S_3-sphere around its centre.

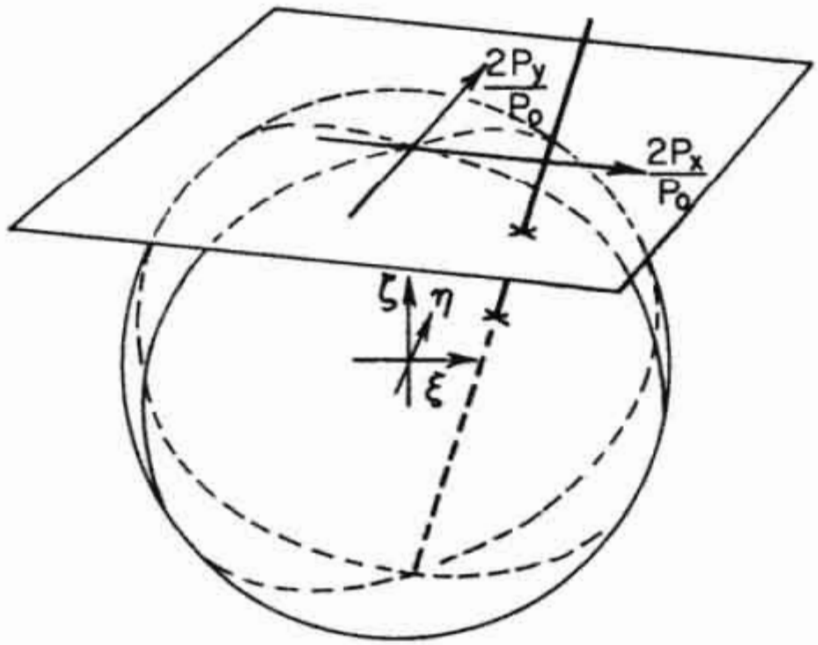


Fig. 2.3. Illustration of the stereographic projection given by Eq. 3 of Paper 4. One dimension has been suppressed.

The above derivation of the O(4) symmetry for the hydrogen atom is exact, as long as the finite size of the nucleus and the effects of spin and relativity theory can be ignored. It is simply a somewhat hidden symmetry which had to be discovered. In a similar way Jauch and Hill [Paper 5] showed that the Hamiltonian for the n-dimensional harmonic oscillator has a symmetry under the SU(n) group. Thus the Elliot SU(3) dynamical symmetry of nuclear physics, mentioned in the introduction, would be a true symmetry of the nuclear shell model Hamiltonian, if the nucleons in the outer shell of the nucleus moved in a three-dimensional harmonic oscillator

potential without any residual interactions. It has also been pointed out[5] that, in the presence of a weak magnetic field, the symmetry groups for the hydrogen atom and the three-dimensional harmonic oscillator both reduce to SU(2).

An important symmetry in non-relativistic quantum mechanics arises from the decoupling of spin in atomic physics (for low Z atoms). In atoms with low atomic number, $Z \sim 1$, it is well known that the order of magnitude of the electron velocity is given by the fine structure constant α. Here, and in the rest of the book, we use natural units with $\hbar = c = 1$. It is thus the smallness of the fine structure constant α, which makes the non-relativistic approximation applicable in atomic physics. To leading non-vanishing order in α it can be shown, from the Dirac equation for an atomic electron in the Coulomb field of the nucleus and the other atomic electrons, that the spin degree of freedom for the electron completely decouples. The spin then just acts as a label for two different types of electron, and its presence is only felt through the effects of Fermi statistics for the electron.

In a light atom, $Z \sim 1$, we readily find the following orders of magnitude for the electron variables (in the rest frame of the atom), in terms of its mass m and the fine structure constant $\alpha = e^2/4\pi$.

$$\text{Velocity } v \sim \alpha, \quad \text{radius of orbit } a \sim \frac{1}{\alpha m}, \tag{3}$$

$$\text{momentum } p \sim \alpha m, \quad \text{energy } E = \frac{p^2}{2m} \sim \alpha^2 m. \tag{4}$$

Similarly the electromagnetic potential A_μ, in Coulomb gauge, satisfies

$$eA_0 \sim \alpha^2 m, \quad |e\mathbf{A}| \sim \alpha^3 m. \tag{5}$$

The characteristic time scale for an appreciable variation of a dynamical variable is given by

$$\Delta t \sim \frac{1}{E} \sim \frac{1}{\alpha^2 m}. \tag{6}$$

Using the above orders of magnitude, we can make a Foldy-Wouthuysen transformation[6,7]

$$\psi = e^{-iS}\psi' \tag{7}$$

which decouples the large and small components of the spinor ψ' for the atomic electron, to leading order in α. The Dirac Hamiltonian

$$H = \beta m + \boldsymbol{\alpha} \cdot (\mathbf{P} - e\mathbf{A}) + eA_0 \tag{8}$$

is then transformed into

$$H' = e^{iS}(H - i\partial_t)e^{-iS} \tag{9}$$

$$= H + i[S, H] - \frac{1}{2}[S, [S, H]] - \dot{S} + \ldots \tag{10}$$

The three contributions to H in Eq. (8) are of order $m, \alpha m$ and $\alpha^2 m$ respectively.

To leading order we take

$$S = -i\beta\boldsymbol{\alpha} \cdot \mathbf{P}/2m \ , \tag{11}$$

which is a small quantity of the order of the fine structure constant α, and obtain

$$H' = \beta m + \frac{\mathbf{P}^2}{2m} + eA_0 + O(\alpha^3 m). \tag{12}$$

The second and third terms in Eq. (12) are of order $\alpha^2 m$ and are obtained by including terms from H, in Eq. (10), up to order αm in $[S, H]$ and just to order m in $[S, [S, H]]$. It follows from Eq. (6) and Eq. (11) that $\dot{S}$ is of order $\alpha^3 m$ and therefore does not contribute.

After this Foldy-Wouthuysen transformation and neglecting terms of order $\alpha^3 m$ and higher, there are no spin dependent terms in the Hamiltonian H' at all. This leads to the separate conservation of the spin of each electron in this approximation and, *a fortiori*, the total spin **S** of all the atomic electrons together is conserved. Since the total angular momentum **J** is conserved, it follows that the total orbital angular momentum **L** of all the electrons together is also conserved. We have therefore derived, to leading order in α, the symmetry of H' which underlies the Russel-Saunders LS coupling scheme in atomic physics. It is an excellent approximation for light and medium atoms.

The neglected terms of order $\alpha^3 m$ include a spin-orbit term, which violates separate **L** and **S** conservation and acts as a perturbation, splitting the otherwise degenerate energy levels. As Z increases the spin-orbit term grows and becomes of the same order of magnitude as the angle dependent part of the electrostatic interaction (i.e. the residual Coulomb interaction between the electrons left after subtracting the central Hartree-Fock potential). The LS coupling scheme then becomes unrealistic and other coupling schemes have been considered.[8] However the alternative jj coupling and jl coupling[9] schemes have a much more limited use, being restricted to certain types of atomic configurations.[8]

The Wigner SU(4) symmetry of nuclear physics [Paper 6] is, at first sight, of a similar nature to the spin decoupling symmetry of atomic physics. In an SU(4) symmetric theory all the charge-spin states of a nucleon are completely equivalent. This would be the case if, for instance, both spin and isospin variables were absent from the nuclear interaction.

The Wigner SU(4) symmetry is defined as a symmetry of the nuclear Hamiltonian, under the simultaneous transformation of all the nucleons by the same group element of SU(4). The transformation of each nucleon is defined by the operation of the appropriate SU(4) matrix on the four component column vector consisting of the four spin-isospin states of the nucleon $p\uparrow, p\downarrow, n\uparrow$ and $n\downarrow$.

A Majorana nucleon-nucleon force is proportional to the factor $(1+\boldsymbol{\sigma}_1 \cdot \boldsymbol{\sigma}_2) \cdot (1 + \boldsymbol{\tau}_1 \cdot \boldsymbol{\tau}_2)$ times a space dependent function. Here $\boldsymbol{\sigma}_i$ and $\boldsymbol{\tau}_i (i = 1, 2)$ denote the Pauli

matrices for the spin and isospin of the two nucleons. Thus the Majorana force is an exchange force which interchanges both the spin and isospin of the nucleons. It is clearly SU(4) invariant, since the SU(4) operation simply rotates every nucleon in SU(4) space. Majorana plus Wigner (spin-isospin independent) forces alone would therefore give an SU(4) symmetric model of nucleonic interactions.

However, it is not possible to derive the Wigner SU(4) symmetry as a non-relativistic limit analogous to the spin decoupling symmetry limit in atomic physics. The spin one vector mesons exchanged in nucleon-nucleon interactions have finite masses and large (strong interaction) coupling constants. Also, at larger range, spin zero scalar (σ meson) and pseudoscalar (π meson) exchange are important.[10] In fact the nucleon-nucleon interaction is rather complicated and the Wigner SU(4) symmetry is only a very approximate symmetry. Nevertheless there is a cancellation between the tensor potentials from ρ vector meson and π meson exchanges, and sensible results have been obtained[11] applying the SU(4) symmetry to the ground state energies of medium- heavy nuclei, $30 \leq A \leq 110$.

We have no convincing explanation for the origin of the approximate Wigner SU(4) symmetry and perhaps it should not therefore have been included in this book. However it served as an inspiration for the introduction of another approximate symmetry, the SU(6) spin and unitary spin symmetry of strong interactions [Paper 7]. This SU(6) symmetry is interpreted[12] as the generalisation of the Wigner SU(4) symmetry obtained by using quarks instead of nucleons. The transformation of each quark is obtained by multiplying the appropriate 6×6 SU(6) matrix into a column vector, with elements denoted by all six combinations of the three lowest quark flavours u, d, s and the two spin states $\uparrow$ and $\downarrow$. The SU(6) transformation matrix is the same for every quark. The approximate SU(6) symmetry of hadrons is then considered to be a consequence of the additive quark model.

In the simple additive quark model, the quarks are treated as effectively free particles bound inside hadrons. In such an approximation, when the effective quark masses are taken to be equal, there is no spin dependence nor flavour dependence in the hadron spectrum. Consequently there is a symmetry under SU(6) in the simple additive quark model. It is, of course, hoped to derive the simple additive quark model as a non-relativistic limit of quantum chromodynamics (QCD), the relativistic quantum field theory of quarks. In the next chapter we consider the symmetries derived from this gauge theory of strong interactions and the Glashow-Weinberg-Salam gauge theory of the electroweak interactions.

References

1. D'Arcy Wentworth Thomson, *On Growth and Form* (Cambridge University Press, 1917).
2. G. K. Batchelor, *Fluid Dynamics* (Cambridge University Press, 1967).
3. H. Weyl, *Symmetry* (Princeton University Press, 1952).
4. L. Landau and E. M. Lifshitz, *Theory of Elasticity* (Pergamon Press, 1959).
5. M. Moshinsky, N. Mendez, E. Murow and J. W.B. Hughes, *Ann. Phys.* **155**, 231 (1984).
6. L. L. Foldy and S. A. Wouthuysen, *Phys. Rev.* **78**, 29 (1950).
7. C. Itzykson and J. B. Zuber, *Quantum Field Theory* (McGraw-Hill, 1980); J. D. Bjorken and S. D. Drell, *Relativistic Quantum Mechanics* (McGraw-Hill, 1964).

8. H. G. Kuhn, *Atomic Spectra* (Longman, 1969).
9. G. Racah, *Phys. Rev.* **61**, 537 (1942).
10. G. E. Brown, "Nuclear Spectroscopy", *Lecture Notes of the Workshop held at Gull Lake*, Michigan 1979, eds. G. F. Bertsch and D. Kurath, p. 1 (Spinger-Verlag, 1980).
11. P. Franzini and L. A. Radicati, *Phys. Lett.* **6**, 322 (1963).
12. B. Sakita, *Phys. Rev.* **136**, B1756 (1964); O. W. Greenberg, *Phys. Rev. Lett.* **13**, 598 (1964).

Chapter III

SYMMETRIES FROM THE STANDARD MODEL

3.1. The Standard Model

At present (1991) it seems that all the experimentally accessible interactions of matter are described by the standard model of elementary particle physics developed during the last two decades.[1,2,3] The standard model combines the Glashow-Weinberg-Salam theory of electroweak interactions with quantum chromodynamics [Paper 3].

The standard model is a Yang-Mills theory of quark and lepton interactions based on the gauge group S(U(2)×U(3)), defined by the following set of 5 × 5 unitary matrices **U**

$$\left\{ \mathbf{U}\ \varepsilon \begin{pmatrix} & U(2) & & 0 & 0 & 0 \\ & & & 0 & 0 & 0 \\ 0 & & 0 & & & \\ 0 & & 0 & & U(3) & \\ 0 & & 0 & & & \end{pmatrix} \;\middle|\; \det \mathbf{U} = 1 \right\} . \tag{1}$$

This S(U(2)×U(3)) group is a subgroup of SU(5) and has the same Lie algebra as U(1)×SU(2)×SU(3). Following Michel,[4] we here give a physical meaning to the Lie *group* as well as to the Lie algebra, by requiring that all the particle multiplets of the theory must form genuine (i.e. single-valued) representations of the Lie *group*. Of course the covering group associated with a given Lie algebra can always be represented in this way, but we also require the true Lie group of a theory to be the 'smallest' one that can be represented on all the particle multiplets. The standard model group S(U(2)×U(3)) turns out[5] to be representable on precisely those particle multiplets which obey the electric charge quantisation rule

$$Q = t_3 + \frac{1}{2}y \equiv -\frac{1}{3}\text{“triality”} \pmod 1 . \tag{2}$$

Here "triality" is 1 for quarks, -1 for antiquarks and zero for colourless states, while y is weak hypercharge (conventionally normalised to be integer for leptons) and t_3 the third component of weak isospin.

Let us now write down the Lagrangian density for the standard model as a sum of five terms.

$$\mathcal{L} = \mathcal{L}_1 + \mathcal{L}_2 + \mathcal{L}_3 + \mathcal{L}_4 + \mathcal{L}_5. \tag{3}$$

The first term

$$\begin{aligned}\mathcal{L}_1 = &- \frac{1}{4g_1^2} F_{\mu\nu} F^{\mu\nu} - \frac{1}{2g_2^2} F_{\mu\nu i}{}^{j} F^{\mu\nu}{}_{j}{}^{i} \\ &- \frac{1}{2g_3^2} F_{\mu\nu\alpha}{}^{\beta} F^{\mu\nu}{}_{\beta}{}^{\alpha}\end{aligned} \tag{4}$$

contains the gauge invariant kinetic energy contributions from the U(1), SU(2) and SU(3) gauge fields A_μ, $A_{\mu i}{}^{j}$ and $A_{\mu\alpha}{}^{\beta}$, expressed in terms of the field tensors

$$F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu \tag{5}$$

$$\begin{aligned}F_{\mu\nu i}{}^{j} = &\partial_\mu A_{\nu i}{}^{j} - \partial_\nu A_{\mu i}{}^{j} \\ &- (A_{\mu i}{}^{k} A_{\nu k}{}^{j} - A_{\nu i}{}^{k} A_{\mu k}{}^{j})\end{aligned} \tag{6}$$

and

$$\begin{aligned}F_{\mu\nu\alpha}{}^{\beta} = &\partial_\mu A_{\nu\alpha}{}^{\beta} - \partial_\nu A_{\mu\alpha}{}^{\beta} \\ &- (A_{\mu\alpha}{}^{\gamma} A_{\nu\gamma}{}^{\beta} - A_{\nu\alpha}{}^{\gamma} A_{\mu\gamma}{}^{\beta}) \,.\end{aligned} \tag{7}$$

Here $A_{\mu i}{}^{j}$ and $A_{\mu\alpha}{}^{\beta}$ are respectively 2×2 and 3×3 Hermitean traceless matrices with vector fields as elements. The indices $i, j = 1,2$ run through a basis for the defining representation of SU(2) and $\alpha, \beta = 1,2,3$ run through a basis for the defining representation of SU(3). As usual a repeated index implies summation. The gauge coupling constants g_1, g_2 and g_3 have been absorbed into the definition of the corresponding U(1), SU(2) and SU(3) gauge fields and therefore do not appear explicitly in the definitions, Eqs. (5)–(7), of the gauge field tensors.

Next we introduce the matter fields given in Table 3.1, making explicit all group and generation indices as in Paper 12. Here we have listed the fermion (lepton and quark) field combinations $l_{iaL}, C\overline{e_{aR}}^{T}$ etc., which are 'left-handed' in the sense that the particles which they can annihilate have negative helicities (when fermion masses are neglected). Left-handed and right-handed Weyl spinors satisfy

$$\psi_L = \frac{1}{2}(1 - \gamma_5)\psi_L \quad \text{and} \quad \psi_R = \frac{1}{2}(1 + \gamma_5)\psi_R \tag{8}$$

Table 3.1. The irreducible representations for the left-handed Weyl fermion fields and the Higgs scalar field in the standard model. t = weak isospin, t_3 = third component of weak isospin, y = weak hypercharge and Q = electric charge.

Matter fields	U(1) representation $\frac{1}{2}y = Q - t_3$	SU(2) representation t	SU(3) representation
l_{iaL}	$-\frac{1}{2}$	$\frac{1}{2}$	1
$C\overline{e_{aR}}^T$	1	0	1
$q_{i\alpha aL}$	$\frac{1}{6}$	$\frac{1}{2}$	3
$C\overline{u_{\alpha aR}}^T$	$-\frac{2}{3}$	0	3^*
$C\overline{d_{\alpha aR}}^T$	$\frac{1}{3}$	0	3^*
ϕ_i	$\frac{1}{2}$	$\frac{1}{2}$	1

respectively and C is the antisymmetric charge conjugation matrix satisfying

$$C\gamma^\mu C^{-1} = -\gamma^{\mu T} \ . \tag{9}$$

Corresponding to a right-handed field ψ_R, one can construct the linear combination $\psi^C = C\bar{\psi}_R{}^T$ of the components of the hermitean conjugate field ψ_R^+ which, formally, is the result of a charge conjugation operation. The field ψ^C is left-handed and represents the CP conjugate particles.

The fields l_{iaL} describe both neutrinos (for i=1) and left-handed charged, $Q = -1$, leptons (for i=2). Similarly $q_{i\alpha aL}$ describes left-handed quarks with two different values of the electric charge, $Q = \frac{2}{3}$ and $-\frac{1}{3}$. The right-handed charged leptons are described by the field e_{aR}, the $Q = \frac{2}{3}$ right-handed quarks by $u_{\alpha aR}$ and the $Q = -\frac{1}{3}$ right-handed quarks by $d_{\alpha aR}$. Right-handed neutrinos are not present in the standard model. Any non-zero (Dirac or Majorana) mass for a neutrino would have to be attributed to effects outside the standard model.

The generation indices $a, b = 1, 2, \ldots, N$ enumerate sets of isomorphic irreducible representations of the standard group on the fermion fields, and will be used to distinguish the $N \geq 3$ generations of physical lepton and quark states $e, \mu, \tau \ldots$; $u, c, t \ldots$; $d, s, b \ldots$. The 'bare' or 'mathematical' basis of Weyl fermion fields entering the Lagrangian density $\mathcal{L}$ do not correspond to the physical mass eigenstates, and we introduce a prime on the lepton and quark fields l'_{iaL}, $C\overline{e'_{aR}}^T$ etc. to denote this fact. The interaction of the fermions with the gauge fields is then given by the second term in Eq. (3)

$$\begin{aligned}
\mathcal{L}_2 = {} & i\bar{l'}^i{}_{aL}\gamma^\mu D_{\mu i}{}^j l'_{jaL} \\
& + i\bar{e'}_{aR}\gamma^\mu D_\mu e'_{aR} \\
& + i\bar{q'}^{i\alpha}{}_{aL}\gamma^\mu D_{\mu i}{}^j{}_\alpha{}^\beta q'_{j\beta aL} \\
& + i\bar{u'}^\alpha{}_{aR}\gamma^\mu D_{\mu\alpha}{}^\beta u'_{\beta aR} \\
& + i\bar{d'}^\alpha{}_{aR}\gamma^\mu D_{\mu\alpha}{}^\beta d'_{\beta aR} \ .
\end{aligned} \tag{10}$$

Here the covariant derivatives $D_\mu, D_{\mu i}{}^j, D_{\mu\alpha}{}^\beta$ and $D_{\mu i}{}^j{}_\alpha{}^\beta$ are those appropriate to the different irreducible representations in Table 3.1. For example, the covariant derivative for the left-handed quark doublet field $q'_{j\beta aL}$ becomes

$$D_{\mu i}{}^j{}_\alpha{}^\beta = \partial_\mu \delta_i{}^j \delta_\alpha{}^\beta - iA_{\mu i}{}^j \delta_\alpha{}^\beta - iA_{\mu\alpha}{}^\beta \delta_i{}^j - i\frac{1}{6} A_\mu \delta_i{}^j \delta_\alpha{}^\beta \ . \tag{11}$$

The factor of $\frac{1}{6}$ multiplying the U(1) gauge field is just the value of the U(1) quantum number, $\frac{1}{2}y$, for the field $q'_{j\beta aL}$ given in Table 3.1.

The final field in Table 3.1 is a complex scalar Higgs doublet ϕ_i, which generates masses for the weak gauge bosons and fermions, by spontaneous breakdown of the electroweak gauge symmetry down to the electromagnetic $U(1)_{em}$ gauge symmetry. The Higgs particle is the least reliable element of the standard model and its existence has still to be established. The gauge and self interactions of the Higgs field are given by

$$\mathcal{L}_3 = D_{\mu i}{}^j \phi_j D^{+\mu i}{}_j \phi^{+j} - V(\phi_j \phi^{+j}) \tag{12}$$

where the scalar potential is

$$V(\phi_j \phi^{+j}) = -\mu^2 \phi_j \phi^{+j} + \lambda (\phi_j \phi^{+j})^2 \ . \tag{13}$$

The Yukawa coupling between the scalars and fermions takes the form

$$\begin{aligned} \mathcal{L}_4 = & h^e_{ab} \bar{l}'^i{}_{aL} \phi_i e'_{bR} \\ & + h^u_{ab} \bar{q}'^{i\alpha}{}_{aL} \phi^{+j} u'_{\alpha bR} \varepsilon_{ij} \\ & + h^d_{ab} \bar{q}'^{i\alpha}{}_{aL} \phi_i d'_{\alpha bR} \\ & + \text{Hermitian Conjugate} \ . \end{aligned} \tag{14}$$

The coefficients h^e_{ab}, h^u_{ab} and h^d_{ab} are *a priori* arbitrary complex Yukawa coupling constants and ε_{ij} is the antisymmetric symbol with $\varepsilon_{12} = 1$.

Finally there are the SU(3) and SU(2) topological terms

$$\begin{aligned} \mathcal{L}_5 = & \theta \frac{1}{16\pi^2} F_{\mu\nu\alpha}{}^\beta \tilde{F}^{\mu\nu}{}_\beta{}^\alpha \\ & + \theta' \frac{1}{16\pi^2} F_{\mu\nu i}{}^j \tilde{F}^{\mu\nu}{}_j{}^i \end{aligned} \tag{15}$$

The dual field strength tensors are defined by

$$\tilde{F}^{\mu\nu}{}_\beta{}^\alpha = \frac{1}{2} \varepsilon^{\mu\nu\rho\sigma} F_{\rho\sigma\beta}{}^\alpha \tag{16}$$

$$\tilde{F}^{\mu\nu}{}_j{}^i = \frac{1}{2} \varepsilon^{\mu\nu\rho\sigma} F_{\rho\sigma j}{}^i \tag{17}$$

where $\varepsilon^{\mu\nu\rho\sigma}$ is the totally antisymmetric symbol with $\varepsilon^{0123} = 1$. The angles θ and θ' multiplying the topological terms are arbitrary real parameters of the theory.

By the standard Higgs-Kibble m...
λ in the potential ...

3) Gauge invariance under the gauge group S(U(2)×U(3)).
4) Matter fields belong to the irreducible representations of Table 3.1.

The constraints imposed by the above assumptions are so powerful that the Lagrangian density $\mathcal{L}$ is forced to have a rather simple form. As a consequence, the standard model Lagrangian density $\mathcal{L}$ exhibits several other symmetries which are not put into the theory as *a priori* principles. These symmetries arise because $\mathcal{L}$ is simply not allowed to be complicated enough to violate them. We will discuss these 'derived' symmetries following our general principle of going up the quantum staircase. Thus we shall first consider the low energy part of the standard model, in which the weak interactions are so weak — due to the relatively high mass, Eq. (20), of the weak intermediate vector bosons $W^{\pm}$ and Z^0 — that they can be neglected. In other words we begin by discussing just the strong and electromagnetic interactions.

3.2. Symmetries of the Strong and Electromagnetic Interactions

In order to write down the reduced Lagrangian density $\mathcal{L}_{ST+EM}$ for the strong and electromagnetic interactions, we must identify the electromagnetic field A_μ and the vector boson field Z_μ. The massless photon and the massive Z^0 particle are the eigenstates of the neutral vector boson mass matrix, obtained by substituting the Higgs field vacuum expectation value, Eq. (18) into $\mathcal{L}_3$. It follows that

$$\frac{1}{e}A_\mu = \sin\theta_W \frac{1}{g_2} 2A_{\mu 1}{}^1 + \cos\theta_W \frac{1}{g_1} A_\mu \tag{33}$$

and

$$\cos\theta_W \frac{1}{g_2} Z_\mu = \cos\theta_W \frac{1}{g_2} 2A_{\mu 1}{}^1 - \sin\theta_W \frac{1}{g_1} A_\mu \tag{34}$$

where the rotation angle θ_W is known as the Weinberg angle. The electromagnetic coupling constant e (the magnitude of the electron charge) is related to the SU(2) and U(1) gauge coupling constants by the equations

$$e = g_2 \sin\theta_W = g_1 \cos\theta_W .$$

The mass of the Z^0 particle is given by

$$m_Z = \frac{m_W}{\cos\theta_W} \simeq 90 \text{ GeV} \tag{37}$$

for the experimental value of $\sin^2\theta_W = 0.23$. The electromagnetic field tensor $\mathcal{F}_{\mu\nu}$ which enters $\mathcal{L}_{ST+EM}$ is given by

$$\mathcal{F}_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu \tag{38}$$

$$u_{\alpha a} = u_{1\alpha aL} + u_{\alpha aR} \tag{39}$$

$$d_{\alpha a} = d_{2\alpha aL} + d_{\alpha aR}. \tag{40}$$

These are all the fermion fields which enter $\mathcal{L}_{ST+EM}$. We do not include the neutrino in $\mathcal{L}_{ST+EM}$, because it does not interact. The graviton is omitted from the standard model Lagrangian density $\mathcal{L}$ on the same grounds. It follows that $\mathcal{L}_{ST+EM}$ describes a vectorlike theory, in which the left-handed and right-handed components of each fermion field have the same gauge quantum numbers. The fermion mass terms are therefore gauge invariant.

The Lagrangian density $\mathcal{L}_{ST+EM}$ becomes

$$\begin{aligned}
\mathcal{L}_{ST+EM} = &- \frac{1}{4e^2}\mathcal{F}_{\mu\nu}\mathcal{F}^{\mu\nu} - \frac{1}{2g_3^2}F_{\mu\nu\alpha}{}^{\beta}F^{\mu\nu}{}_{\beta}{}^{\alpha} \\
&+ i\overline{u}^{\alpha}{}_{a}\gamma^{\mu}(\partial_{\mu}\delta_{\alpha}{}^{\beta} - iA_{\mu\alpha}{}^{\beta} - i\frac{2}{3}\mathcal{A}_{\mu}\delta_{\alpha}{}^{\beta})u_{\beta a} \\
&+ i\overline{d}^{\alpha}{}_{a}\gamma^{\mu}(\partial_{\mu}\delta_{\alpha}{}^{\beta} - iA_{\mu\alpha}{}^{\beta} + i\frac{1}{3}\mathcal{A}_{\mu}\delta_{\alpha}{}^{\beta})d_{\beta a} \\
&+ i\overline{e}_{a}\gamma^{\mu}(\partial_{\mu} + i\mathcal{A}_{\mu})e_{a} \\
&+ m_{a}^{u}\overline{u}^{\alpha a}u_{\alpha a} + m_{a}^{d}\overline{d}^{\alpha a}d_{\alpha a} + m_{a}^{e}\overline{e}_{a}e_{a} \\
&+ \theta\frac{1}{16\pi^2}F_{\mu\nu\alpha}{}^{\beta}\tilde{F}^{\mu\nu}{}_{\beta}{}^{\alpha}.
\end{aligned} \tag{41}$$

There is of course an understood summation over the thrice repeated generation index a on the mass terms. We note that, due to the electric charge quantisation rule of Eq. (2), the strong plus electromagnetic gauge group is U(3). The Lagrangian density $\mathcal{L}_{ST+EM}$ is equivalent to the most general renormalisable U(3) gauge theory of quarks and charged leptons [Paper 8].

One immediately sees that $\mathcal{L}_{ST+EM}$ is invariant under independent global phase transformations for each fermion flavour. For example, $\mathcal{L}_{ST+EM}$ is invariant under the transformation

$$u_{\alpha a} \rightarrow \exp[i\phi_a]u_{\alpha a} \tag{42}$$

for all values of the SU(3) colour index α and fixed generation number index a, all other fields being unchanged. It follows that all the fermion flavour charges

$$Q_a^u = \int d^3x\ \overline{u}^{\alpha}{}_{a}\gamma_o u_{\alpha a} \tag{43}$$

$$Q_a^d = \int d^3x\ \overline{d}^{\alpha}{}_{a}\gamma_o d_{\alpha a} \tag{44}$$

and

$$Q_a^e = \int d^3x\ \overline{e}_a\gamma_o e_a \tag{45}$$

are conserved, where there is a sum over the repeated colour index α but not over the generation index a. These 'flavour charges' are simply the number of quarks

or leptons of a given flavour minus the number of corresponding antiparticles. For example

$$S = -Q_2^d \tag{46}$$

$$L_\mu = Q_2^e \tag{47}$$

$$B = \sum_a \frac{1}{3}(Q_a^u + Q_a^d) \tag{48}$$

$$Q = \sum_a (\frac{2}{3}Q_a^u - \frac{1}{3}Q_a^d - Q_a^e) \tag{49}$$

where S = strangeness, L_μ = muon lepton number, B = baryon number and Q = electric charge. Since we are dealing with a vectorlike theory, these flavour conservation laws cannot be spoiled by anomalies.[7]

The basic reason for the flavour conservation laws is the absence of fundamental scalar fields in the theory of strong and electromagnetic interactions. It follows that there are only two types of allowed terms in $\mathcal{L}_{ST+EM}$: kinetic (i.e. gauge covariant derivative terms) and mass terms. The two fermion fields in either term combine to form the trivial representation of the gauge group. One can then perform a simultaneous diagonalisation of the two matrices in flavour space associated with the fermion kinetic and mass terms [Paper 8]. Had there been a scalar neutral colourless field in the theory, there would have been a third matrix in flavour space corresponding to its Yukawa couplings to the fermions. It would not, in general, be possible to diagonalise the three matrices simultaneously and all the flavour conservation laws would not be obtained. If the scalar field instead had non-trivial gauge quantum numbers, other combinations of fermion field representations would be allowed in the Lagrangian. For this reason, in the full standard model, the Kobayashi-Maskawa matrix K_{ab} of Eq. (27) is not equal to the unit matrix and the W^+ gauge boson causes transitions between particles with different generation indices.

It is also easily shown that $\mathcal{L}_{ST+EM}$ is invariant under charge conjugation:

$$\mathcal{L}_{ST+EM}(\psi^c(\mathbf{x},t)) = \mathcal{L}_{ST+EM}(\psi(\mathbf{x},t)). \tag{50}$$

Here $\psi(\mathbf{x},t)$ is a generic symbol for all the fields and $\psi^c(\mathbf{x},t)$ denote the charge conjugation transformed fields. The fermion fields are taken to be Grassman (anti-commuting) variables, as in the path integral formalism. We list below the charge conjugate gauge fields and $Q = \frac{2}{3}$ quark fields (as a typical fermion field)

$$A^c{}_{\mu\alpha}{}^\beta = -A_{\mu\beta}{}^\alpha, \quad \mathcal{A}^c_\mu = -\mathcal{A}_\mu \tag{51}$$

$$F^c{}_{\mu\nu\alpha}{}^\beta = -F_{\mu\nu\beta}{}^\alpha, \quad \mathcal{F}^c{}_{\mu\nu} = -\mathcal{F}_{\mu\nu} \tag{52}$$

$$\tilde{F}^c{}_{\mu\nu\alpha}{}^\beta = -\tilde{F}_{\mu\nu\beta}{}^\alpha, \quad \tilde{\mathcal{F}}^c{}_{\mu\nu} = -\tilde{\mathcal{F}}_{\mu\nu} \tag{53}$$

$$u^c{}_{\alpha a} = C\overline{u}^T{}^\alpha{}_a, \quad \overline{u}^{c\alpha}{}_a = u^T{}_{\alpha a}C. \tag{54}$$

Here T denotes the transpose and C is a 4×4 Dirac matrix satisfying

$$C\gamma^{\mu}C^{-1} = -\gamma^{\mu T}, \quad C^{-1} = -C. \tag{55}$$

The kinetic and topological terms for the gauge fields in $\mathcal{L}_{ST+EM}$ are clearly invariant under the above charge conjugation transformation. Let us now explicitly verify that the fermion kinetic and mass terms are also invariant. We consider first the kinetic term for the $Q = \frac{2}{3}$ quarks.

$$\overline{u}^{\alpha}{}_{a}\gamma^{\mu}(\partial_{\mu}\delta_{\alpha}{}^{\beta} - iA_{\mu\alpha}{}^{\beta} - i\frac{2}{3}A_{\mu}\delta_{\alpha}{}^{\beta})u_{\beta a} \tag{56}$$

$$\xrightarrow[C]{} \quad u^{T}{}_{\alpha a}C\gamma^{\mu}(\partial_{\mu}\delta^{\alpha}_{\beta} + iA_{\mu\beta}{}^{\alpha} + i\frac{2}{3}A_{\mu}\delta^{\alpha}_{\beta})C\overline{u}^{T\beta}{}_{a} \tag{57}$$

$$= u^{T}{}_{\alpha a}\gamma^{\mu T}(\partial_{\mu}\delta^{\alpha}_{\beta} + iA_{\mu\beta}{}^{\alpha} + i\frac{2}{3}A_{\mu}\delta^{\alpha}_{\beta})\overline{u}^{T\beta}{}_{a} \tag{58}$$

$$= -\overline{u}^{\beta}{}_{a}\gamma^{\mu}(\overleftarrow{\partial}_{\mu}\delta^{\alpha}_{\beta} + iA_{\mu\beta}{}^{\alpha} + i\frac{2}{3}A_{\mu}\delta^{\alpha}_{\beta})u_{\alpha a} \tag{59}$$

$$= \overline{u}^{\beta}_{a}\gamma^{\mu}(\overrightarrow{\partial}_{\mu}\delta^{\alpha}_{\beta} - iA_{\mu\beta}{}^{\alpha} - i\frac{2}{3}A_{\mu}\delta^{\alpha}_{\beta})u_{\alpha a}\ . \tag{60}$$

The minus sign in Eq. (59) arises from commuting the order of the Grassman variables $\overline{u}^{\beta}_{a}$ and $u_{\alpha a}$. The arrow on the derivative in Eq. (59) indicates that it is acting to the left. The derivative acts to the right as usual in Eq. (60), which follows from Eq. (59) modulo an irrelevant pure derivative term. Hence the kinetic term for the $Q = \frac{2}{3}$ quarks is charge conjugation invariant, as are those for the $Q = -\frac{1}{3}$ quarks and charged leptons.

The mass term for the $Q = \frac{2}{3}$ quarks transforms as follows

$$m^{u}_{a}\overline{u}^{\alpha}{}_{a}\delta^{\beta}_{\alpha}u_{\beta a} \xrightarrow[C]{} m^{u}_{a}u^{T}{}_{\alpha a}C\delta^{\beta}_{\alpha}C\overline{u}^{T\beta}{}_{a} \tag{61}$$

$$= -m^{u}_{a}u^{T}{}_{\alpha a}\delta_{\alpha}{}^{\beta}\overline{u}^{T\beta}{}_{a} \tag{62}$$

$$= m^{u}_{a}\overline{u}^{\beta}{}_{a}\delta_{\alpha}{}^{\beta}u_{\alpha a} \tag{63}$$

where we have used $C^2 = -1$ in Eq. (62) and the Grassman variable nature of $u_{\alpha a}$ in Eq. (63). The fermion mass terms are thus seen to be invariant under charge conjugation. So charge conjugation is a symmetry of the strong and electromagnetic Lagrangian density $\mathcal{L}_{ST+EM}$.

In the absence of the topological term, the vectorlike nature of the theory of strong and electromagnetic interactions would automatically ensure that $\mathcal{L}_{ST+EM}$ is invariant under parity

$$\mathcal{L}_{ST+EM}(\psi^{P}(\mathbf{x},t)) = \mathcal{L}_{ST+EM}(\psi(-\mathbf{x},t)). \tag{64}$$

The parity transformed fields $\psi^{P}(\mathbf{x},t)$ are defined as follows:

$$A^{P}_{\mu\alpha}{}^{\beta} = -(-1)^{\delta^{0}_{\mu}}A_{\mu\alpha}{}^{\beta}(-\mathbf{x},t) \quad A^{P}_{\mu} = -(-1)^{\delta^{0}_{\mu}}A_{\mu}(-\mathbf{x},t) \tag{65}$$

$$F^P_{\mu\nu\alpha}{}^{\beta} = (-1)^{\delta^0_\mu+\delta^0_\nu} F_{\mu\nu\alpha}{}^{\beta}(-\mathbf{x},t) \quad \mathcal{F}^P_{\mu\nu} = (-1)^{\delta^0_\mu+\delta^0_\nu}\mathcal{F}_{\mu\nu}(-\mathbf{x},t) \tag{66}$$

$$\tilde{F}^P_{\mu\nu\alpha}{}^{\beta} = -(-1)^{\delta^0_\mu+\delta^0_\nu}\tilde{F}_{\mu\nu\alpha}{}^{\beta}(-\mathbf{x},t) \tag{67}$$

$$u^P{}_{\alpha a} = \gamma^0 u_{\alpha a}(-\mathbf{x},t) \quad \overline{u}^{P\alpha}{}_a = \overline{u}^{\alpha}{}_a(-\mathbf{x},t)\gamma^0 \, . \tag{68}$$

The topological term changes sign under parity

$$F_{\mu\nu\alpha}{}^{\beta}(\mathbf{x},t)\tilde{F}^{\mu\nu}{}_{\beta}{}^{\alpha}(\mathbf{x},t)$$
$$\xrightarrow[P]{} -F_{\mu\nu\alpha}{}^{\beta}(-\mathbf{x},t)\tilde{F}^{\mu\nu}{}_{\beta}{}^{\alpha}(-\mathbf{x},t) \tag{69}$$

and violates parity symmetry, Eq. (64).

Phenomenologically, the coefficient θ of the topological term must be very small or zero. The experimental limit on the possible value of the neutron electric dipole moment implies[8] that $|\theta| \lesssim 10^{-9}$. Here we shall therefore neglect the topological term by setting $\theta = 0$. However a true derivation of parity P (and time reversal T) invariance for strong and electromagnetic interactions must explain the smallness of θ. This requires us to go beyond the standard model and we return to the problem later in this section, when we discuss Paper 9.

Neglecting the topological term, it is easy to verify the validity of Eq. (64). We illustrate this result, by explicitly giving the parity transformation of the $Q = \frac{2}{3}$ quark kinetic energy term:

$$\overline{u}^{\alpha}{}_a(\mathbf{x},t)\gamma^{\mu}[\partial_\mu\delta^{\beta}_{\alpha} - iA_{\mu\alpha}{}^{\beta}(\mathbf{x},t) - i\frac{2}{3}\mathcal{A}_\mu(\mathbf{x},t)\delta^{\beta}_{\alpha}]u_{\beta a}(\mathbf{x},t) \tag{70}$$

$$\xrightarrow[P]{} -(-1)^{\delta^0_\mu}\overline{u}^{\alpha}{}_a(-\mathbf{x},t)\gamma^0\gamma^{\mu}$$
$$\times [\partial_\mu\delta^{\beta}_{\alpha} - iA_{\mu\alpha}{}^{\beta}(-\mathbf{x},t) - i\frac{2}{3}\mathcal{A}_\mu(-\mathbf{x},t)\delta^{\beta}_{\alpha}]\gamma^0 u_{\beta a}(-\mathbf{x},t) \tag{71}$$

$$= \overline{u}^{\alpha}{}_a(-\mathbf{x},t)\gamma^{\mu}[\partial_\mu\delta^{\beta}_{\alpha} - iA_{\mu\alpha}{}^{\beta}(-\mathbf{x},t) - i\frac{2}{3}\mathcal{A}_\mu(-\mathbf{x},t)\delta^{\beta}_{\alpha}]u_{\beta a}(-\mathbf{x},t). \tag{72}$$

The anticommutation relations of the Dirac γ matrices are used in passing from Eq. (71) to Eq. (72).

The vectorlike or non-chiral nature of the fermion representations under the strong and electromagnetic gauge group is, of course, crucial in the above derivation of C and P invariance [Paper 8]. So we may ask if and why this feature has been put into the standard model. Since the U(3) gauge group is unbroken, the assumption that all fermions have a non-zero mass requires the fermions to belong to vector-like representations and have parity invariant gauge interactions.[9] However there could exist massless fermions with parity non-invariant gauge interactions. This theoretical possibility can be ruled out, by requiring sufficiently small representations of the U(3) gauge group and no gauge anomaly for the photon and gluons.[10] In fact it is sufficient to assume that all fermions have electric charge $|Q| \leq 1$ and belong

to the trivial $\underline{1}$, triplet $\underline{3}$ or anti-triplet $\underline{3}^*$ representation of colour SU(3). Then just seven different irreducible representations of the gauge group U(3) are allowed. These representations are $(0,\underline{1})$ and three pairs of mutually conjugate U(1)×SU(3) representations $(1,\underline{1})$ and $(-1,\underline{1})$, $(\frac{2}{3},\underline{3})$ and $(-\frac{2}{3},\underline{3}^*)$ and finally the pair $(-\frac{1}{3},\underline{3})$ and $(\frac{1}{3},\underline{3}^*)$. The $(0,\underline{1})$ representation corresponds to a particle which has no strong and electromagnetic interaction and is therefore omitted from $\mathcal{L}_{ST+EM}$.

There are three combinations of gauge particles which can act as external lines for potentially anomalous triangle diagrams, Fig. 3.1. Each of these three combinations gives a gauge anomaly proportional to a linear combination of the three differences

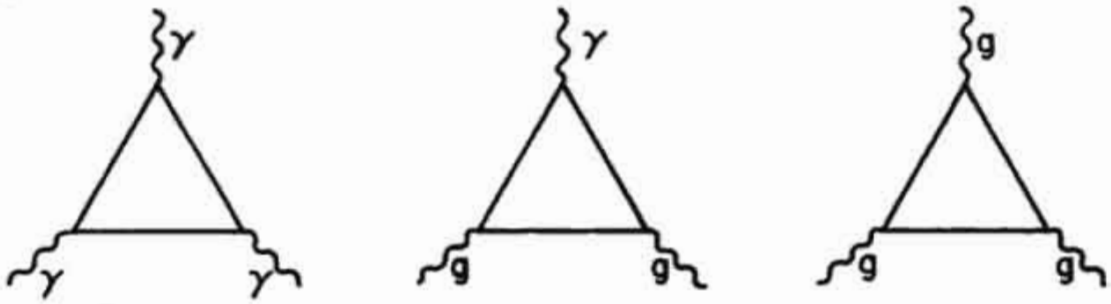


Fig. 3.1. Three triangle diagrams which can contribute to anomalies in the U(3) gauge theory.

$$N_{(1,\underline{1})} - N_{(-1,\underline{1})} \tag{73}$$

$$N_{(\frac{2}{3},\underline{3})} - N_{(-\frac{2}{3},\underline{3}^*)} \tag{74}$$

and

$$N_{(-\frac{1}{3},\underline{3})} - N_{(\frac{1}{3},\underline{3}^*)} \,. \tag{75}$$

Here N_D denotes the number of left-handed fermion species in the representation D of the gauge group. It is easily seen that the three equations, resulting from the conditions of anomaly cancellation, enforce all three of the above differences to be zero. This means that the model is then vectorlike and automatically parity conserving.

Similar considerations can be applied to the full standard model gauge group S(U(2)×U(3)). The assumptions of small representations, the absence of gauge and mixed anomalies and mass protection principles again lead to the experimentally observed representation pattern of quarks and leptons. Mass protection means that fermion representations which can combine to form S(U(2)×U(3)) gauge invariant mass terms will, in fact, form very heavy fermion states. The mass scale associated with these states is expected to be given by the characteristic energy scale of some new interaction, beyond the standard model. These fermions, which are allowed to have masses without appealing to the standard model Higgs mechanism, should thus be heavy compared to the $W^{\pm}$ and Z^0 boson masses and unobservable at presently available energies.

Since the theory of strong and electromagnetic interactions involves only gauge bosons and fermions, it has CP symmetry (still neglecting any topological term).

The above 'explanation' of the vectorlike nature of the fermion representations in $\mathcal{L}_{ST+EM}$, therefore, implies C invariance as well as P invariance.

Time reversal, T, symmetry of $\mathcal{L}_{ST+EM}$ (with $\theta = 0$) now follows as a consequence of the CPT theorem [Paper 23] to be discussed in Chapter V. It is however instructive to verify T invariance explicitly. Time reversal is an anti-linear operator and the condition for invariance is

$$\mathcal{L}_{ST+EM}(\psi^T(\mathbf{x},t)) = \mathcal{L}^*_{ST+EM}(\psi(\mathbf{x},-t)) \tag{76}$$

where $\mathcal{L}^*$ denotes the complex conjugate of the functional form $\mathcal{L}$. In other words, the complex conjugation applies to the coefficients of the field combinations occurring in $\mathcal{L}$ but not to the fields themselves. The time reversal transformed fields $\psi^T(\mathbf{x},t)$ are defined as follows:

$$A^T{}_{\mu a}{}^{\beta} = -(-1)^{\delta^0_\mu} A_{\mu a}{}^{\beta}(\mathbf{x},-t) \qquad \mathcal{A}^T_\mu = -(-1)^{\delta^0_\mu}\mathcal{A}_\mu(\mathbf{x},-t) \tag{77}$$

$$\begin{aligned} F^T_{\mu\nu a}{}^{\beta} &= -(-1)^{\delta^0_\mu+\delta^0_\nu} F_{\mu\nu a}{}^{\beta}(\mathbf{x},-t) \\ \mathcal{F}^T_{\mu\nu} &= -(-1)^{\delta^0_\mu+\delta^0_\nu}\mathcal{F}_{\mu\nu}(\mathbf{x},-t) \end{aligned} \tag{78}$$

$$\tilde{F}^T_{\mu\nu a}{}^{\beta} = (-1)^{\delta^0_\mu+\delta^0_\nu}\tilde{F}_{\mu\nu a}{}^{\beta}(\mathbf{x},-t) \tag{79}$$

$$u^T{}_{\alpha a} = T u_{\alpha a}(\mathbf{x},-t) \qquad \bar{u}^{T\alpha}{}_a = \bar{u}^{\alpha}{}_a(\mathbf{x},-t)T. \tag{80}$$

In the discussion of time reversal invariance the superscript T is used to denote the time reversal transformed fields and not the transpose. In Eq. (80), T denotes a 4×4 Dirac matrix satisfying

$$T\gamma^\mu T^{-1} = -(-1)^{\delta^0_\mu}\gamma^{\mu *} \qquad T = T^+ = T^{-1}\ . \tag{81}$$

It is now straightforward to see that $\mathcal{L}_{ST+EM}$ fulfills Eq. (76), provided we neglect the topological term, which changes sign under time reversal

$$\begin{aligned} &F_{\mu\nu a}{}^{\beta}(\mathbf{x},t)\tilde{F}^{\mu\nu}{}_{\beta}{}^{\alpha}(\mathbf{x},t) \\ &\quad\xrightarrow[T]{} -F_{\mu\nu a}{}^{\beta}(\mathbf{x},-t)\tilde{F}^{\mu\nu}{}_{\beta}{}^{\alpha}(\mathbf{x},-t). \end{aligned} \tag{82}$$

Again we explicitly verify the invariance of the $Q = \frac{2}{3}$ quark kinetic term

$$i\bar{u}^{\alpha}{}_a(\mathbf{x},t)\gamma^\mu[\partial_\mu\delta^\beta_\alpha - iA_{\mu a}{}^{\beta}(\mathbf{x},t) - i\frac{2}{3}\mathcal{A}_\mu(\mathbf{x},t)\delta^\beta_\alpha]u_{\beta a}(\mathbf{x},t) \tag{83}$$

$$\begin{aligned} \xrightarrow[T]{} &- i\bar{u}^{\alpha}{}_a(\mathbf{x},-t)T\gamma^{\mu *}(-1)^{\delta^0_\mu} \\ &\times[\partial_\mu\delta^\beta_\alpha - iA_{\mu a}{}^{\beta}(\mathbf{x},-t) - i\frac{2}{3}\mathcal{A}_\mu(\mathbf{x},-t)\delta^\beta_\alpha]Tu_{\beta a}(\mathbf{x},-t) \end{aligned} \tag{84}$$

$$= i\overline{u}^{\alpha}{}_{a}(\mathbf{x},-t)\gamma^{\mu}[\partial_{\mu}\delta^{\beta}_{\alpha} - iA_{\mu\alpha}{}^{\beta}(\mathbf{x},-t) - i\frac{2}{3}\mathcal{A}_{\mu}(\mathbf{x},-t)\delta^{\beta}_{\alpha}]u_{\beta a}(\mathbf{x},-t). \qquad (85)$$

Thus we have derived the three discrete symmetries C, P and T for the strong and electromagnetic sector of the standard model. Strictly speaking, however, the 'derivations' of P and T invariance are not valid, in the absence of an explanation for the vanishing of the coefficient θ of the topological term. Such an explanation requires us to somewhat extend the standard model [Paper 9].

As mentioned in our discussion of the diagonalisation of the quark mass matrices M^u and M^d, chiral transformations of the quark fields Eq. (24)-(25), induce a change in the value of the coefficient θ. In particular, an axial phase rotation of all the quark fields

$$\psi \rightarrow \psi' = e^{i\alpha\gamma_5}\psi, \qquad (86)$$

under which right handed quark fields are multipled by $e^{i\alpha}$ and left handed ones by $e^{-i\alpha}$, leads to the transformations

$$M^u \rightarrow e^{-2i\alpha}M^u \qquad (87)$$

$$M^d \rightarrow e^{-2i\alpha}M^d \qquad (88)$$

$$\theta \rightarrow \theta + 2N_F\alpha. \qquad (89)$$

Here N_F denotes the number of quark flavours under consideration. The need for the change in the value of θ arises from the axial anomaly[7] in QCD. Due to the anomaly, the functional integral measure for the quark fields is not invariant under Eq. (86), but undergoes the transformation[6]

$$\mathcal{D}\psi\mathcal{D}\psi^{+} \rightarrow e^{-i\Delta}\mathcal{D}\psi\mathcal{D}\psi^{+} \qquad (90)$$

where

$$\Delta = 2\alpha N_F \int d^4x \frac{1}{16\pi^2} F_{\mu\nu\alpha}{}^{\beta}\tilde{F}^{\mu\nu}{}_{\beta}{}^{\alpha}\ . \qquad (91)$$

The combination

$$\overline{\theta} = \theta + \arg[\det M^u \cdot \det M^d] \qquad (92)$$

is invariant under all chiral transformations. It is therefore possible to rotate away the original topological term in $\mathcal{L}_{ST+EM}$, obtaining instead an overall phase $e^{i\theta/N_F}$ for the quark mass matrices. So all the P and T (or CP) symmetry breaking can be arranged to come in via the quark mass matrices. Consequently, if the masses of the relevant quarks are small, the effect of the original θ-term is suppressed. In the case of a massless quark, say the up quark u, the effect of the topological term can be entirely transformed away, by an axial phase transformation

$$u \rightarrow e^{i\alpha\gamma_5}u \qquad (93)$$

on the massless quark alone. This transformation just changes θ to $\theta + 2\alpha$, without any change in the mass matrix if $m_u = 0$. The zero mass implies chiral invariance under the transformation Eq. (93), except for the anomaly. Thus θ can be

transformed into any value, without changing the physics of the field theory. It follows that P, T and CP are symmetries of $\mathcal{L}_{ST+EM}$. The problem of explaining the vanishing of θ is then replaced by explaining why $m_u = 0$. Phenomenologically it seems unlikely that any quark mass is zero [Paper 10], although $m_u = 0$ cannot be entirely ruled out.[11]

More quantitatively the effect of the θ-term can be calculated, treating the mass term for the $N_F = 3$ lightest quarks u, d and s as a perturbation. The θ-term is replaced, to lowest order in θ, by the CP violating mass term

$$\mathcal{L}_{CP} = -i\frac{\theta m_u m_d m_s}{m_u m_d + m_u m_s + m_d m_s}(\bar{u}\gamma_5 u + \bar{d}\gamma_5 d + \bar{s}\gamma_5 s). \tag{94}$$

The contribution of $\mathcal{L}_{CP}$ to the electric dipole moment of the neutron can then be calculated,[8] using current algebra techniques and standard quark masses [Paper 10] to be $D_n \simeq 5 \times 10^{-16}\theta$ cm. Experimentally it is known that $|D_n| < 10^{-24}$ cm and hence $|\theta| < 10^{-9}$.

We have seen that chiral phase invariance is broken in the standard model, explicitly by the quark masses, as well as by the axial anomaly. Consequently the topological term gives P and CP violations. Peccei and Quinn [Paper 9] therefore suggested embedding the standard model in a larger theory possessing a chiral U(1) invariance, which is broken only by the anomaly. This global chiral U(1) invariance is introduced into the Higgs sector of the theory, by the addition of an extra weak isodoublet scalar field. The U(1) symmetry is spontaneously broken when the two Higgs fields develop vacuum expectation values. The relative phase of the vacuum expectation values is determined dynamically, by the anomaly and strong interaction effects (instantons). It turns out that instanton activity is very likely to align the vacuum, in just such a direction that this relative phase precisely cancels out the effect of the θ term. In other words, the effective parameter $\bar{\theta}$ of Eq. (92) becomes zero, in agreement with the P and CP symmetry of strong interactions.

In fact it can be readily shown that the vacuum energy is invariant under the sign change $\bar{\theta} \to -\bar{\theta}$ (i.e. under parity). The effective potential $V_\theta(\phi)$ therefore has an extremum at $\bar{\theta} = 0$ (and also at $\bar{\theta} = \pi$). Dynamical calculations suggest that $\bar{\theta} = 0$ corresponds to a true minimum of the vacuum energy density. So the vacuum adjusts itself to take precisely that value of the $\bar{\theta}$ parameter, which ensures P and CP invariance for the strong and electromagnetic interactions. These symmetries are thereby explained.

The spontaneous breakdown of the Peccei-Quinn chiral U(1) invariance leads to a Nambu-Goldstone boson. Explicit symmetry breaking by instantons generates a small mass for this particle, called the axion.[12] Experimentally a particle with the properties of this axion does not seem to exist.[13] However it is possible to make the axion in practice invisible, by further extending the model. It is necessary to introduce a third Higgs field, with a non-zero Peccei-Quinn U(1) charge and trivial under the standard group, which develops a very large vacuum expectation value.[14]

Of course all these models, involving an extra chiral symmetry, really take us outside the standard model.

We now turn our attention to some approximate symmetries of the strong interaction sector of the standard model [Papers 8 and 10]. These approximate symmetries arise due to the existence, for some reason, of a set of N_F quark flavours with exceptionally small masses,

$$m_q \ll \Lambda_{QCD}. \tag{95}$$

Here Λ_{QCD} refers to the characteristic strong interaction mass scale of a few hundred MeV. The quark mass m_q refers to the mass parameter appearing in the Lagrangian density and not to the constituent quark mass (used in the 'naive' quark model), which cannot be small compared to Λ_{QCD} [Paper 10]. See Fig. 3.2.

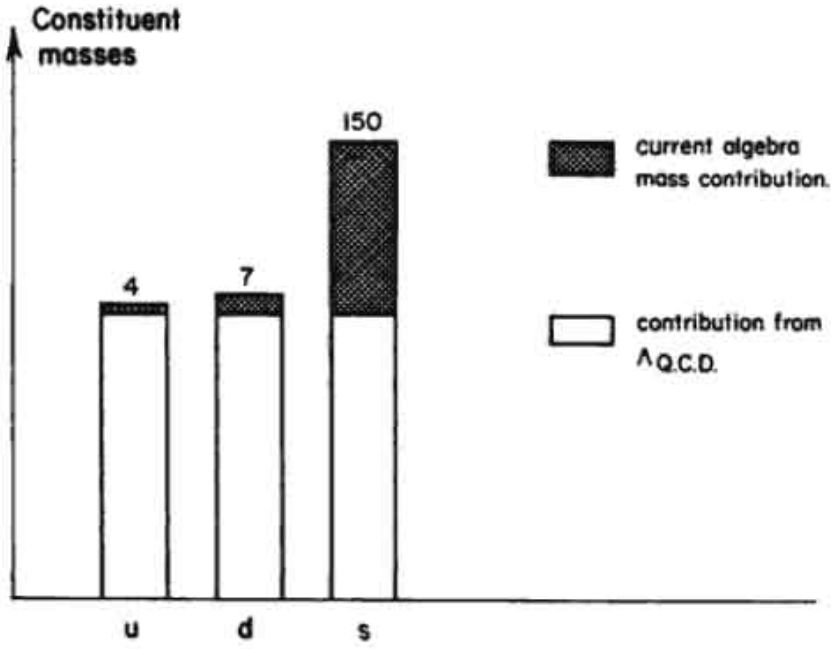


Fig. 3.2. Relationship between the constituent quark masses of non-relativistic quark models and the current algebra masses occurring in the QCD Lagrangian. The difference between them is due to the momentum fluctuations arising from confinement and is proportional to Λ_{QCD}. The numbers indicate the current algebra masses in MeV.

In a first approximation, we can neglect the masses of the N_F quarks satisfying Eq. (95). There is then a symmetry under global unitary transformations on the massless left-handed quark Weyl fields, belonging to the same representation of the colour gauge group (i.e. the $\underline{3}$ and $\underline{3}^*$ states respectively of Table 3.1, restricted to the massless subspace). Alternatively, we can consider the N_F left- and right-handed quark colour triplet, $\underline{3}$, fields. For $N_F = 3$ massless flavours u, d and s, we have chiral invariance under 3×3 unitary transformations U_L and U_R

$$\begin{pmatrix} u'_L \\ d'_L \\ s'_L \end{pmatrix} = U_L \begin{pmatrix} u_L \\ d_L \\ s_L \end{pmatrix}, \quad \begin{pmatrix} u'_R \\ d'_R \\ s'_R \end{pmatrix} = U_R \begin{pmatrix} u_R \\ d_R \\ s_R \end{pmatrix} . \tag{96}$$

In general, for N_F massless quark flavours, the chiral invariance group becomes *a priori*, $U(N_F)_L \times U(N_F)_R$. Using the Lie algebra decomposition $U(N) \approx U(1) \times SU(N)$, we immediately recognise that the diagonal (vector) $U(1)_V$ subgroup corresponds to baryon number conservation, Eq. (48). However, the antisymmetric (axial) combination of the two U(1) subgroups has an anomaly, as discussed above, Eq. (90). The divergence of the axial baryon number current is thus non-vanishing

$$\partial^\mu J^5_\mu = 2N_F \frac{1}{16\pi^2} F_{\mu\nu\alpha}{}^\beta \tilde{F}^{\mu\nu}{}_\beta{}^\alpha \tag{97}$$

and the axial $U(1)_A$ subgroup is not a true symmetry group.

The remaining chiral symmetry group is $SU(N_F)_L \times SU(N_F)_R$. Again we can readily recognise the diagonal (vector) subgroup $SU(N_F)_V$, consisting of the elements $(g, g) \in SU(N_F)_L \times SU(N_F)_R$ where $g \in SU(N_F)$. It is the Gell-Mann flavour $SU(N_F)$ symmetry[15], under which hadrons are classified into multiplets. The Cartan subalgebra corresponds to conserved quark flavour charges, Eq. (43) and (44). The antisymmetric combination or axial vector subgroup $SU(N_F)_A$ is known, phenomenologically, to be realised by spontaneous symmetry breakdown. There are good dynamical reasons[16] for believing that the QCD vacuum does, indeed, break the global chiral symmetry $SU(N_F)_L \times SU(N_F)_R$ down to its diagonal vector subgroup $SU(N_F)_V$ in this way.

The successes of soft pion theorems and current algebra[7,17] show that, in nature, there is an almost exact $N_F = 2$ (u and d quark) flavour chiral symmetry. The Nambu-Goldstone bosons, corresponding to the spontaneously broken $SU(2)_A$ symmetry, are the pions π^+, π^0 and π^-. The non-zero pion masses arise from the explicit chiral symmetry breaking quark mass terms, $m_u \bar{u}u$ and $m_d \bar{d}d$, in the Lagrangian density. The low value of the pion mass, compared to that of a typical hadron, is evidence that $m_u, m_d \ll \Lambda_{QCD}$. Isospin or $SU(2)_V$ symmetry is obtained in the same approximation.

We should stress here that isospin symmetry is explained by the u and d quarks both being light, without requiring equality of the masses m_u and m_d. In fact it is estimated, from meson masses [Paper 10], that

$$\frac{m_d}{m_u} = 1.8. \tag{98}$$

Therefore it is clear that the accuracy of isospin symmetry is not due to the degeneracy of m_u and m_d. It is, rather, the mass difference $m_d - m_u$ which matters, and it is small because both m_d and m_u are small (approximately 7 MeV and 4 MeV respectively). So isospin symmetry should not be considered as a fundamental symmetry of the strong interactions, broken only by electromagnetism. It is an approximate symmetry, which follows dynamically from the existence of two quarks satisfying Eq. (95). Similarly the 'eightfold way'[18] of flavour $SU(3)_V$ is a symmetry of the strong interactions, accurate to 10 or 20%, in spite of the s quark being about twenty times heavier than the d quark.

A complete derivation of chiral symmetry would, of course, explain why at least two quark flavours are much lighter than Λ_{QCD}. In the standard model quark masses arise, via the Higgs mechanisms, from the Yukawa coupling constants h^u_{ab} and h^d_{ab}, as in Eq. (22). These coupling constants are pure numbers and, *a priori*, one would have expected them to be of order unity, in the absence of any reason for them to be otherwise.

In principle the QCD mass scale parameter Λ_{QCD} cannot be appreciably larger than the $W^{\pm}$ or Z^0 boson masses, since QCD would then act as a technicolour interaction[19], generating $W^{\pm}$ and Z^0 masses of order Λ_{QCD}. In practice Λ_{QCD} is much smaller than the W-Z mass scale. So it is not possible, with the order of unity coupling constant philosophy, to obtain a good isospin symmetry. In fact all the quark and lepton masses would be of the same order of magnitude as the $W^{\pm}$ and Z^0 masses. This is just one facet of the fermion mass problem. It follows that the order of unity coupling philosophy must be totally wrong or further explanations, beyond the standard model, must be postulated.

A natural way to explain the observed fermion mass hierarchy is to postulate the existence of a new set of approximately conserved chiral quantum numbers [Paper 11]. These chiral charges are supposed to be different for the various left- and right-handed quark and lepton Weyl fields. The quantum number differences between the various left- and right-handed Weyl states lead to approximate selection rules for the corresponding Yukawa coupling constants. The fermion Yukawa coupling constants are thereby suppressed, by factors determined by the required amounts of chiral charge symmetry breaking. The corresponding fermion mass matrix elements are therefore suppressed, relative to the W-Z mass scale, by different orders of magnitude.

We have considered [Paper 11] a very general model of this type, postulating a random system of approximate selection rules. The existence of light quarks and leptons and the observed hierarchical structure of the fermion mass spectrum are naturally reproduced. Furthermore the qualitative features of the quark mixing matrix elements K_{ab} of Eq. (26) are also reproduced[20], if the standard model is further unified by, for example, introducing right-handed currents and leptoquark currents.

Another approach to the mass problem is to replace the Higgs field by a new Yang-Mills interaction, technicolour[19], analogous to QCD. The spontaneous breakdown of the global chiral technisymmetry gives rise to the $W^{\pm}$ and Z^0 masses. At first sight, in such technicolour schemes, we seem to have too good a solution to the problem of the existence of light fermions. All the quarks and leptons have zero mass! It is possible to avoid this result, by introducing a further 'extended technicolour' gauge interaction. However it is clear that a realistic model of this type must be rather complicated. Nonetheless it might be relatively easy to obtain some quarks much lighter than other mass scales, such as the W-Z mass scale and Λ_{QCD}.

Again we see that a full understanding of some strong interaction symmetries, in

this case isospin and chiral symmetry, requires us to go beyond the standard model.

Let us now look at which symmetries and conservation laws are valid for the whole standard model, and not just its low energy limit.

3.3. Symmetries of the Full Standard Model

In the previous section we neglected weak interaction processes since, at low energy, their amplitudes are of order α/m_W^2, where $\alpha = e^2/4\pi$ is the fine structure constant. Similarly weak radiative corrections, which violate symmetries of $\mathcal{L}_{ST+EM}$ such as strangeness and parity, are automatically of order α/m_W^2 rather than of order α [Paper 8]. In this section we discuss the symmetries which remain, when weak interaction effects are included.

The flavour charges of Eqs. (43)-(45) are no longer generally conserved. However, baryon number B is conserved and we can define a conserved lepton number L_a for each generation.

It is easy to see that the total number of quarks minus antiquarks (i.e. $3B$) is still conserved, due to exact colour invariance. The covariant derivative $D_{u_i}{}^{j}_{\alpha}{}^{\beta}$ and the colourless Higgs field ϕ_i must therefore connect quark fields of compensating colour representations. It follows that the quark fields can only occur in combinations of the type $\bar{\psi}\ldots\psi$, where ψ denotes a generic quark colour triplet field. Thus all terms in the standard model Lagrangian density $\mathcal{L}$ create and annihilate equally many quarks. *A priori* this gives baryon number conservation, but the baryon number current has an anomaly,[7] due to its interaction with the $W^{\pm}$ and Z^0 bosons. However the amplitude for the occurrence of instantons in the SU(2) gauge field, which could produce baryon number violation, is exponentially suppressed by the factor[21]

$$\exp(-8\pi^2/g_2^2) \simeq \exp(-180). \tag{99}$$

The predicted baryon number violation, at normal temperatures and energies, is therefore unobservable.

Similarly it is easily seen that there is a lepton conservation law for each fermion generation. In the standard model, neutrinos are all massless and we are free to define the electron neutrino ν_e, say, as that linear combination of neutrino species which couples, via the $W^{\pm}$ boson, to the electron. The $W^{\pm}$ coupling then, by definition, conserves electron number — the number of electrons and electron neutrinos minus the number of corresponding antiparticles. As we shall discuss below, the couplings of the Z^0 boson to fermions do not cause flavour changing transitions, due to the GIM mechanism [Paper 13]. It follows that there are separate conservation laws for electron number L_e, muon number L_μ and tau lepton number L_τ (for $N = 3$ generations). Again, there is an anomaly in each of the lepton number currents. Due to the enormous suppression factor, Eq. (99), for SU(2) gauge field instantons, however, there is in practice no violation of any of the lepton number conservation laws.

It is even possible to construct anomaly free conservation laws, by taking differences in which the anomalies[21] cancel out. In fact the lepton number differ-

ences $L_\mu - L_e$ and $L_\tau - L_e$ are exactly conserved. Furthermore $B - L$, where $L = L_e + L_\mu + L_\tau$ is the total lepton number, is a conserved quantity, even with anomalies taken into account.

It is interesting to note that the quantity $B - L$ is still conserved, when the standard model is extended to include arbitrary SU(3)×SU(2)×U(1) invariant, baryon number violating, four-fermion interactions [Paper 12]. These four-fermion interactions, such as

$$O^{(1)}_{abcd} = (\bar{d}^c{}_{\alpha a R} u_{\beta b R})(\bar{q}^c{}_{i\gamma c L} l_{j d L}) \epsilon_{\alpha\beta\gamma} \epsilon_{ij}, \tag{100}$$

are, of course, supposed to be an effective, non-renormalisable, low energy parameterisation of new physics, which occurs at a large mass scale $m_X \gg m_W$. In general, the coupling for an effective interaction of mass dimension d will be of order $(m_X)^{4-d}$. The four-fermion operators have d=6. Other baryon number violating effective interactions have higher dimensions and are suppressed by higher powers of $\frac{1}{m_X}$.

In order to classify the four-fermion operators, it is useful to introduce the concept of F parity.[22] The F parity of a field is defined to be

$$F = (-1)^{2t+2a}, \tag{101}$$

where t is its weak isospin and $2a$ is the number of left-handed spinor indices in the Lorentz group representation (a, b) of the field. Due to Lorentz invariance and SU(2) gauge invariance, F parity is multiplicatively conserved. F parity takes the values: F=+1 for quarks and leptons; $F = -1$ for antiquarks, antileptons, gauge fields, the Higgs field and the differential operator ∂_μ. Baryon number violating four-fermion operators must contain three quark (or antiquark) fields, coupled together to from a colour singlet, and a lepton (or antilepton) field. It then follows, from F parity conservation, that the allowed terms either annihilate three quarks and one lepton or three antiquarks and one antilepton. Thus the 'charge' $B - L$ is conserved by the allowed d=6 operators.

It must be admitted, however, that there are allowed d=5 terms, involving the Higgs boson field ϕ_i, which violate lepton number L without changing the baryon number B:

$$\bar{l}^c{}_{iaL} l_{jbL} \phi_k \phi_l \epsilon_{ik} \epsilon_{jl}, \tag{102}$$

$$\bar{l}^c{}_{iaL} l_{jbL} \phi_k \phi_l \epsilon_{ij} \epsilon_{kl}. \tag{103}$$

These terms clearly violate $B - L$, although they do not contribute to proton decay. If no elementary Higgs field is available, as in technicolour models, then both ϕ's in Eqs. (102) and (103) should be replaced by fermion bilinear expressions. The dimension of these terms is thereby increased from $d = 5$ to $d = 9$. In such a technicolour-like model $B - L$ would in fact be conserved including all effective interactions up to and including dimension $d = 6$. However in the pure standard model extended to include all effective interactions up to and including dimension $d = 6$, $B - L$ is only conserved for certain processes, such as proton decay.

We now consider another conservation law, which is only valid for a certain special type of process: flavour conservation for neutral currents [Paper 13]. As we have already seen, the fermion flavour charges, Eqs. (43)-(45), are conserved by the strong and electromagnetic interactions. However strangeness and other flavours are not conserved in charged current weak interactions, Eq. (28), mediated by the $W^{\pm}$ boson. The absence, experimentally, of strangeness-changing neutral current weak interactions was, for several years, considered to be evidence against the existence of the neutral weak intermediate vector boson Z^0. It was then pointed out by Glashow, Iliopolous and Maiani (GIM)[23] that the suppression of strangeness-changing neutral currents could be explained by the introduction of a fourth quark flavour: charm. More generally, flavour conservation for neutral currents is a consequence of the fermion generations all having the same representation, Table 3.1, under the standard group and the existence of precisely one neutral Higgs boson in the standard model [Paper 13].

The Z^0 couplings cause transitions between fermions of the same electric charge, belonging to the same weak isodoublet. By using the doublets of Eq. (24), $(u, d_c)_L$ etc., as a basis for the $Q = \frac{2}{3}$ quarks (and the doublets of Eq. 25, (u_c, d) etc., as a basis for the $Q = -\frac{1}{3}$ quarks), it is easily seen that the Z^0 boson couples each quark to itself only. Similarly the Z^0 couplings are diagonal in lepton flavour. The Yukawa couplings of the Higgs particle are proportional to the fermion mass matrix and, thus, also diagonal in the mass eigenstate basis. As noted earlier, the $W^{\pm}$ couplings conserve the lepton number for each generation. Hence all the interactions respect the separate lepton number conservation laws.

However the $W^{\pm}$ couplings to quarks are not diagonal in generation number space due to the non-vanishing off-diagonal elements K_{ab} of the Kobayashi-Maskawa matrix. Quark generation number is conserved by the neutral currents, but not by the charged currents. In addition to quark flavour conservation in second order (single boson exchange) neutral weak interactions, the GIM mechanism suppresses higher order induced neutral current effects, in which a quark undergoes a flavour-changing but electric charge conserving transition. All higher order flavour-changing neutral current interactions vanish, in the limit that all the $Q = \frac{2}{3}$ quark masses are equal and all the $Q = -\frac{1}{3}$ quark masses are equal. In practice this limit is used to mean the approximation, in which all relevant quark masses are negligible compared to m_W.

The fact that all flavour-changing neutral current interactions vanish, in the limit of totally degenerate quark masses, is easily seen, by analogy to the separate lepton number conservation laws. In this degenerate situation, the 'Cabibbo-rotated' quarks, d_c etc. and u_c etc. are all mass eigenstates. Let us now consider a single quark line, corresponding to a $Q = \frac{2}{3}$ quark in the initial state, passing through a Feynman diagram, Fig. 3.3. We choose the doublets of Eq. (24), $(u, d_c)_L$ etc., to define a quark generation number basis in this mass degenerate situation. The $W^{\pm}$ couplings to the quark line then conserve generation number, as do all the other standard model interactions. A similar result applies to a $Q = -\frac{1}{3}$ quark in

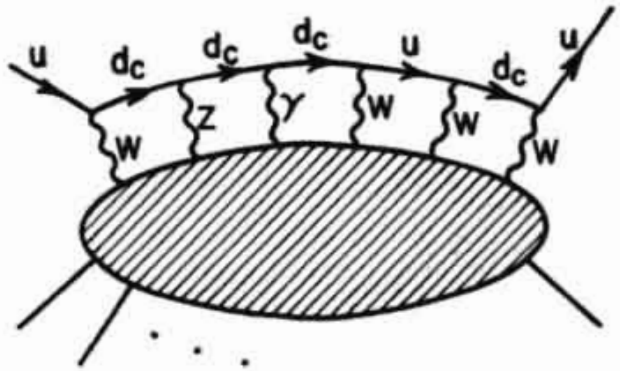


Fig. 3.3. The interactions of a massless Q=2/3 quark in an arbitrary neutral current process.

the initial state, in a generation number basis defined by the doublets of Eq. (25), $(u_c, d)_L$ etc. Since quark generation numbers are unchanged, it follows that quark flavour is conserved in neutral current processes to all orders of perturbation theory.

When real non-degenerate quark masses are used, the actual degree of suppression of any higher order flavour-changing neutral current process depends on the convergence properties of the Feynman diagrams under consideration. A truly fourth-order weak process, in which two heavy bosons are exchanged, suffers a power law suppression,[24] so that the amplitude is of order

$$\frac{g_2^4 \Delta m^2}{m_W^4} \sim G_F^2 \Delta m^2. \tag{104}$$

Here Δm^2 is an appropriate combination of quark mass squared differences. However if the process is second-order weak and second-order electromagnetic, there is a very mild GIM suppression factor and the amplitude is of order [24]

$$\frac{g_2^2 \alpha}{m_W^2} \sim G_F \alpha. \tag{105}$$

The amplitude can be further suppressed by small Kobayashi-Maskawa mixing matrix elements.

The full standard model, including weak interactions, is not invariant under any of the discrete operations C, P or T. However T and CP would be symmetries, if there had only been $N = 1$ or 2 generations [Paper 14] (again setting the coefficient θ of the topological term equal to zero).

After diagonalisation of the fermion mass matrices, the CP violating Yukawa couplings to the Higgs particle disappear. Then, apart from the θ-term, the only CP violating terms occur in the couplings of the charged current to the $W^{\pm}$ bosons:

$$K_{ab} A_{\mu 1}{}^{2} \bar{u}^{1\alpha}{}_{aL} \gamma^{\mu} d_{2\alpha bL} + K^{*}_{ab} A_{\mu 2}{}^{1} \bar{d}^{2\alpha}{}_{bL} \gamma^{\mu} u_{1\alpha aL}. \tag{106}$$

The physical quark fields $u_{1\alpha aL}$ and $d_{2\alpha bL}$ are defined in Eqs. (24) and (25). In discussing the CP transformation properties of the charged current, one must allow for relative phases between these quark fields. All other terms are diagonal in the quark flavours.

The fields transform as follows under CP:

$$A_{\mu i}{}^{j}(t,\mathbf{x}) \rightarrow (-1)^{\delta^0_\mu} A_{\mu j}{}^{i}(t,-\mathbf{x}) \tag{107}$$

$$u_{1\alpha aL}(t,\mathbf{x}) \rightarrow \eta_{1a}\gamma^0 C\bar{u}^{T1\alpha}{}_{aL}(t,-\mathbf{x}) \tag{108}$$

$$\bar{u}^{1\alpha}{}_{aL}(t,\mathbf{x}) \rightarrow \eta^*_{1a} u^T{}_{1\alpha aL}(t,-\mathbf{x})C\gamma^0 \tag{109}$$

$$d_{2\alpha bL}(t,\mathbf{x}) \rightarrow \eta_{2b}\gamma^0 C\bar{d}^{T2\alpha}{}_{bL}(t,-\mathbf{x}) \tag{110}$$

$$\bar{d}^{2\alpha}{}_{bL}(t,\mathbf{x}) \rightarrow \eta^*_{2b} d^T{}_{2\alpha bL}(t,-\mathbf{x})C\gamma^0. \tag{111}$$

Here η_{1a} and η_{2b} are phase factors and T denotes the transpose. It then follows, using Eq. (55) and the anti-commutation properties of the Grassman fields $u_{1\alpha aL}$ and $d_{2\alpha bL}$, that the first term in the Eq. (106) transforms as follows under CP:

$$\mathrm{K}_{ab} A_{\mu 1}{}^{2}\bar{u}^{1\alpha}{}_{aL}\gamma^\mu d_{2\alpha bL} \rightarrow \mathrm{K}_{ab}\eta^*_{1a}\eta_{2b}\bar{d}^{2\alpha}{}_{bL}\gamma^\mu u_{1\alpha aL}. \tag{112}$$

For CP invariance of the charged current couplings, we require that the transformed expression of Eq. (112) should equal the second term of Eq. (106). The condition for CP invariance thus becomes

$$\eta^*_{1a}\eta_{2b}\mathrm{K}_{ab} = \mathrm{K}^*_{ab} \tag{113}$$

or, in other words, the matrix

$$\mathrm{K}^1_{ab} = \eta^{*\frac{1}{2}}_{1a}\eta^{\frac{1}{2}}_{2b}\mathrm{K}_{ab} \tag{114}$$

must be real.

The weak mixing matrix K_{ab} is unitary and thus has N^2 real degrees of freedom, where N is the number of generations. For CP invariance, we require K^1_{ab} to belong to the $N(N-1)/2$ dimensional manifold of orthogonal matrices. Therefore, in general, we need $N^2 - N(N-1)/2$ independent adjustable parameters, in order to ensure CP symmetry. There are $2N$ phase factors, η_{1a} and η_{2b}, which can be adjusted. However a common phase factor for all the quark fields would leave all the parameters $\eta^*_{1a}\eta_{2b}$ unchanged. Hence there are $2N-1$ independent phases $\eta^*_{1a}\eta_{2b}$ available. Thus, in order to satisfy the general condition for CP symmetry, we require

$$N^2 - \frac{1}{2}N(N-1) \leq 2N - 1. \tag{115}$$

This condition is equivalent to

$$\frac{1}{2}(N-1)(N-2) \leq 0 \tag{116}$$

which is only satisfied for N=1 or 2. For N=3 generations we have $(N-1)(N-2)/2 = 1$ and obtain just one nontrivial CP-violating phase factor in the Kobayashi-Maskawa matrix [Paper 14].

In the second order (one boson exchange) process involving only external quark flavours belonging to, say, the first two generations, there can be no CP violation. This follows, since the relevant quark couplings only involve a 2×2 sub-matrix of the full mixing matrix K_{ab}. A CP-violating process must either (i) be fourth order in the weak interactions or (ii) involve external quarks from three different generations.

The only place where CP violation has been observed is in the neutral kaon system. The coefficient of admixture ε of the CP- even state K_1, relative to the CP-odd state K_2, in the long- lived eigenstate K_L is given approximately by[25]

$$\varepsilon = \frac{1}{2\sqrt{2}} \cdot \frac{\text{Im } M_{12}}{\text{Re } M_{12}} e^{\frac{i\pi}{4}}. \tag{117}$$

Here

$$M_{12} = \langle K^0 | H_{\text{eff}} | \bar{K}^0 \rangle \tag{118}$$

is the $K^0 - \bar{K}^0$ transition matrix element of the effective fourth order weak Hamiltonian density H_{eff}, changing strangeness by two units. The CP violating amplitude Im M_{12} necessarily involves couplings to a third generation quark and is, therefore, naturally suppressed relative to the CP conserving amplitude Re M_{12}. This suppression is due to the observed smallness of the quark mixing matrix elements K_{ab}, connecting the two lowest mass generations to the third generation, compared to the 'Cabibbo' matrix element K_{12}.[26] The empirically small value of the CP violation parameter

$$|\varepsilon| \simeq 2.3 \times 10^{-3} \tag{119}$$

is thereby understood, without requiring the CP violating phase in the mixing matrix to be small (or close to π).

Finally we consider scaling symmetry in the standard model. We use natural units $\hbar = c = 1$, and measure dimensions in units of mass. By analogy with the discussion in Chapter II of scaling symmetries in macroscopic physics, we can apply dimensional analysis to the action, $S = \int d^4x \mathcal{L}$, of a Lagrangian field theory.

The action for a field theory with no dimensional parameter (i.e. masses and non-dimensionless coupling constants all set equal to zero) is invariant under a scale transformation[27]

$$x_\mu \to x'_\mu = \lambda x_\mu, \tag{120}$$

where the transformed fields are given by

$$\phi'(x) = \lambda^{-1} \phi(\lambda^{-1} x) \tag{121}$$

$$\psi'(x) = \lambda^{-\frac{3}{2}} \psi(\lambda^{-1} x). \tag{122}$$

Here $\phi(x)$ and $\psi(x)$ denote generic boson and fermion fields respectively. The scale invariance of the action is broken by particle masses. These masses should be negligible at very high energy or very small distances, where we formally obtain

asymptotic scale invariance. However the scale symmetry of a massless theory is broken by a quantum anomaly,[7,27] similar to the axial U(1) anomaly which breaks the chiral invariance of massless fermions.

The divergence of the scale current S_μ, for a renormalisable field theory, is given by the trace of the energy-momentum tensor $\theta_{\mu\nu}$:

$$\partial^\mu S_\mu = \theta^\mu_\mu. \tag{123}$$

Formally the trace θ^μ_μ is given by mass terms, but another anomalous contribution arises from the renormalisation of the theory. The renormalisation procedure introduces a momentum scale, $p^2 = -\mu^2$, at which the renormalised dimensionless coupling constant $g(\mu)$ is defined (for simplicity we consider a theory with a single coupling constant). The renormalisation scale μ is an arbitrary parameter, which breaks scale invariance and introduces an anomaly into the trace of the energy-momentum tensor. This trace is proportional to the renormalisation group β-function:[28]

$$\beta(g) = \mu \frac{dg}{d\mu}. \tag{124}$$

Scale invariance is thus restored if the physical coupling constant corresponds to a zero of the β-function. However, if $g(\mu)$ is non-zero at such a fixed point β=0, the fields will in general have anomalous scaling dimensions,[27] $d \neq 1$ for bosons and $d \neq \frac{3}{2}$ for fermions.

Non-Abelian gauge theories have the property that, for small values of $g(\mu)$, the β-function is given by

$$\beta(g) = bg^3 + 0(g^5) \tag{125}$$

with a negative value of b. Hence it is consistent to assume that $g(\mu) \to 0$ and $\beta(g) \to 0$ as $\mu \to \infty$. So we expect naive scale invariance, Eq. (121)-(122), to be restored asymptotically at large momenta. This phenomenon is known as asymptotic freedom.[1,3,27]

The only dimensional parameter in the standard model Lagrangian density $\mathcal{L}$ is the tachyonic mass associated with the Higgs field. It sets the scale of the vacuum expectation value of the Higgs field, Eq. (19), and the masses of the $W^\pm$ and Z^0 bosons. These masses are too large to be neglected at presently available energies, unlike the quark masses which are suppressed by the ill-understood small values of the Yukawa coupling constants. The electromagnetic interaction is Abelian and not asymptotically free. We therefore expect asymptotic scale invariance to be valid only for the strong interaction sector (QCD), at energies where the relevant quark masses can be neglected. For $\mu \gg \Lambda_{QCD}$, the QCD scale parameter, the strong interaction coupling constant has the form

$$\alpha_s(\mu) = \frac{g_3^2(\mu)}{4\pi} = \frac{12\pi}{(33 - 2N_F)\ln\, \mu^2/\Lambda^2_{QCD}} \tag{126}$$

provided the number of quark flavours N_F is less than $\frac{33}{2}$. Asymptotic scaling behaviour should set in for energies and momenta large compared to Λ_{QCD} and the relevant quark masses.

Hadrons are extended objects on a scale of order Λ_{QCD}^{-1} and, thus, momenta of order Λ_{QCD} are relevant even for high energy hadron beams. It is therefore difficult to test scaling symmetry in purely hadronic processes. Experimentally, scaling symmetry is studied in deep inelastic lepton-nucleon scattering processes, which involve one photon or weak vector boson exchange. The cross-sections are described in terms of dimensionless structure functions[1,3,15] $F_i(q^2, X)$. Here q_μ is the 4-momentum transfer carried by the virtual photon or intermediate vector boson exchanged, p_μ is the 4-momentum of the nucleon and

$$X = -\frac{q^2}{2p \cdot q}. \tag{127}$$

Invariance under the canonical scale transformations of Eqs. (121)-(122) would imply that the structure functions $F_i(q^2, X)$ depend only upon the dimensionless variable X and not on q^2. This is the famous Bjorken scaling behaviour (for q^2 and $p \cdot q$ large compared to the relevant quark masses). The asymptotic scaling symmetry of QCD does not give exact Bjorken scaling. However the predicted scaling violations are only logarithmic; the moments

$$M_i^n(q^2) = \int_0^1 dX \; X^n F_i(q^2, X) \tag{128}$$

of the structure functions are not constants, but definite powers of logarithms.[1,3] The predicted approximate Bjorken scaling, with logarithmic violations, appears to be in agreement with experiment.[29]

It is interesting to note that scale invariance implies conformal invariance for a renormalisable field theory.[28] Conformal transformations are a four parameter manifold of space-time transformations of the form

$$x^\mu \to x'^\mu = \frac{x^\mu + c^\mu x^2}{1 + 2c_\lambda x^\lambda + c^2 x^2}. \tag{129}$$

These transformations can be written as an inversion

$$x^\mu \to x'^\mu = \frac{x^\mu}{x^2} \tag{130}$$

followed by a translation

$$x'^\mu \to x''^\mu = x'^\mu + c^\mu \tag{131}$$

and then followed by another inversion. When combined with Lorentz transformations, translations and scale transformations, they form a 15 dimensional Lie group, called the conformal group.

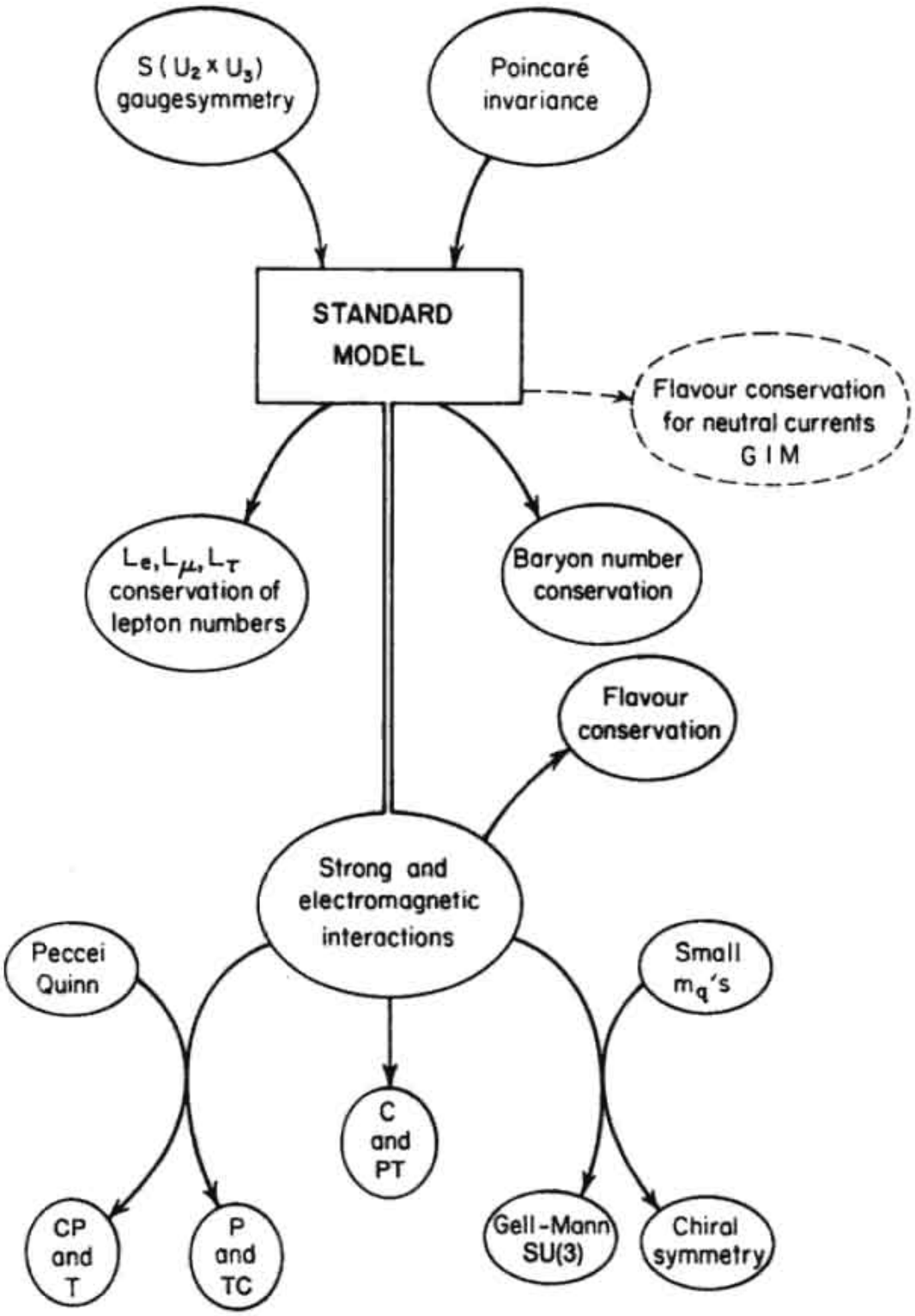


Fig. 3.4. The assumed and derived symmetries of the standard model.

For a renormalisable field theory, the divergence of the four conformal currents $K^{\lambda\mu}$, associated with the transformations of Eq. (129), can be expressed in terms of the trace of the energy-momentum tensor

$$\partial_\mu K^{\lambda\mu} = 2x^\lambda \theta^\mu_\mu. \tag{132}$$

Apart from mass terms, θ^μ_μ is given by the trace anomaly discussed above. Thus the conditions for asymptotic scale invariance and for asymptotic conformal invari-

ance are the same in a renormalisable field theory, namely the vanishing of the β-function.[28]

It has also been pointed out[30] that, for an effective field theory action which is invariant under general co-ordinate transformations, local scale invariance (Weyl invariance) implies conformal invariance in flat space.

We have seen that a rather large number of symmetries arise naturally out of the standard model. Poincaré invariance and gauge invariance are the only fundamental input symmetries of the model. The symmetries derived from the model, together with the auxiliary assumptions, are summarised in Fig. 3.4.

References

1. T.-P. Cheng and L.-F. Li, *Gauge Theory of Elementary Particle Physics* (Oxford University Press, 1984).
2. K. Huang, *Quarks, Leptons and Gauge Fields* (World Scientific, 1982).
3. C. Quigg, *Gauge Theories of Strong, Weak and Electromagnetic Interactions* (Benjamin/ Cummings, 1983).
4. L. Michel, *Group Theoretical Concepts and Methods in Elementary Particle Physics, Lectures of the Istanbul Summer School of Theoretical Physics 1962*, ed. F. Gursey, p. 135 (Gordon and Breach, 1964).
5. L. O' Raifeartaigh, *Group Structure of Gauge Theories* (Cambridge University Press, 1986).
6. K. Fujikawa, *Phys. Rev. Lett.* **42**, 1195 (1979).
7. S. B. Treiman, R. Jackiw, B. Zumino and E. Witten, *Current Algebra and Anomalies* (World Scientific, 1985).
8. V. Baluni, *Phys. Rev.* **D19**, 2227 (1979); R. Crewther, P. di Vecchia, G. Veneziao and E. Witten, *Phys. Lett.* **89B**, 123 (1979).
9. Ken-ichi Hikasa, *Physica Scripta* **34**, 204 (1986).
10. S Chadha and H. B. Nielsen (unpublished).
11. D. B. Kaplan and A. V. Manohar, *Phys. Rev. Lett.* **56**, 2004 (1986).
12. S. Weinberg, *Phys. Rev. Lett.* **40**, 223 (1978); F. Wilczek, *Phys. Rev. Lett.* **40**, 279 (1978).
13. S. Yamada, *Proceedings of 1983 International Symposium on Lepton and Photon Interactions at High Energies*, Cornell, p. 525 (Cornell University, 1983); W. A. Bardeen, R. D. Peccei and T. Yanagida, *Nucl. Phys.* **B279**, 401 (1987).
14. M. Dine, W. Fischler and M. Srednicki, *Phys. Lett.* **104B**, 199 (1981).
15. F. E. Close, *An Introduction to Quarks and Partons* (Academic Press, 1979).
16. C. Vafa and E. Witten, *Nucl. Phys.* **B234**, 173 (1984).
17. S. L. Adler and R. F. Dashen, *Current Algebras and Application to Particle Physics* (Benjamin, 1968); V. de Alfaro, S. Fubini, G. Furlan and C. Rossetti, *Currents in Hadron Physics* (North-Holland, 1973).
18. M. Gell-Mann and Y. Ne'eman, *The Eightfold Way* (Benjamin, 1964)
19. E. Farhi and R. Jackiw, *Dynamical Gauge Symmetry Breaking* (World Scientific, 1982).
20. A. Conkie, C. D. Froggatt and H. B. Nielsen, *Phys. Lett.* **161B**, 347 (1985).
21. G. 't Hooft, *Phys. Rev. Lett.* **37**, 8 (1976); V. Kuzmin, V. Rubakov and M. Shaposhnikov, Phys. Lett. **155B**, 36 (1985).
22. S. Weinberg, *Phys. Rev.* **D22**, 1694 (1980).
23. S. Glashow, J. Iliopolous and L. Maiani, *Phys. Rev.* **D2**, 1285 (1970).
24. M. K. Gaillard and B. W. Lee, *Phys. Rev.* **D10**, 897 (1974); M. K. Gaillard, B. W. Lee and R. E. Schrock, *Phys. Rev.* **D13**, 2674 (1976).
25. J. Ellis, M. K. Gaillard and D. V. Nanopolous, *Nucl. Phys.* **B109**, 213 (1976); F. J. Gilman and M. B. Wise, *Phys. Rev.* **D20**, 2392 (1979).
26. F. J. Gilman, K. Kleinknecht and B. Renk, Particle Data Group, *Phys. Lett.* **204B**, 107 (1988).
27. S. Coleman, *Aspects of Symmetry* (Cambridge University Press, 1985).
28. B. Schroer, *Nuovo Cimento Lett.* **2**, 867 (1971); N. K. Nielsen, *Nucl. Phys.* **B97**, 527 (1975); *Nucl. Phys.* **B120**, 212 (1977).

29. P. Söding and G. Wolf, *Ann. Rev. Nucl. Part. Sci.* **31**, 231 (1981); F. Sciulli, *Proceedings of 1985 International Symposium on Lepton and Photon Interactions at High Energies*, Kyoto, p. 8, 1986
30. B. Zumino, *Lectures in Elementary Particle Physics and Quantum Field Theory, Proceedings of 1970 Brandeis Summer Institute*, Vol. 2, p. 441 (MIT Press, 1970).

Chapter IV

BEYOND THE STANDARD MODEL

4.1. Grand Unification

The standard model, together with Einstein's theory of gravity, seems to explain all the fundamental interactions of physics known today. However there are several reasons[1] to look for theories that go beyond the standard model, in the sense of having the latter as a low-energy approximation. The standard three generation S(U(2)×U(3)) model involves 20 arbitrary parameters; three independent gauge coupling constants, six quark masses, four quark mixing parameters, three charged lepton masses, the W-mass, the Higgs particle mass and the coefficients θ and θ' of the SU(3) and SU(2) topological terms. Many of these parameters are unnaturally[2] small. The smallness of the quark masses are responsible for the chiral symmetry and the Gell-Mann SU(3) flavour symmetry of the strong interactions. Similarly parity and *CP* invariance require the θ-parameter to be effectively zero. We therefore need a model for the parameters, if we are to understand these symmetries. We should also like to understand the origin of the gauge symmetry group S(U(2)×U(3)) itself. These symmetry and other[1] considerations strongly suggest the existence of a more complete theory containing the standard model.

An attractive method of extending the standard model is to postulate the unification[3,4] of the strong, weak and electromagnetic interactions in a simple gauge group G. The simplest such grand unified theory (GUT) is the Georgi-Glashow G = SU(5) model [Paper 15]. Twelve new gauge bosons are introduced by the SU(5) group, which mediate quark-lepton transitions and must be made superheavy. The global baryon and lepton number conservation laws of the standard model are thereby violated. In fact, proton decay and the existence of superheavy magnetic monopoles are generic features of GUTs.

The grand unified gauge group G is arranged to be spontaneously broken, at a superheavy mass scale m_X, by the Higgs mechanism, down to the standard model group. In fact precisely the standard model group S(U(2)×U(3)), and not the group

$SU(3)\times SU(2)\times U(1)$, is a subgroup of SU(5). The charge quantisation rule

$$Q = -\frac{1}{3}\text{"triality"} \pmod 1 \tag{1}$$

is thus automatically obeyed in the SU(5) model. Each generation of left-handed fermions fits exactly into the anomaly-free reducible $\mathbf{5}^* + \mathbf{10}$ representation of SU(5). However it is an exaggeration to claim that the standard model group has been derived within such a GUT model. In reality the standard model symmetry is built into the gauge group G *ab initio*. So, strictly speaking, GUT models do not qualify for inclusion in this book as examples of symmetry derivations.

Nevertheless, some GUT models do have symmetries which are not *a priori* obvious and are not imposed on the model from the outset. The most important example of this type is the conservation of $B - L$ (baryon number minus lepton number) in the simplest version of the SU(5) theory, despite the separate violation of B and L. In its simplest version, SU(5) is broken down to the standard $S(U(2)\times U(3))$ group by an adjoint representation of Higgs, $\mathbf{24}_H$, developing a supermassive vacuum expectation value. The standard group is then broken by a fundamental representation of Higgs, $\mathbf{5}_H$, which is just the SU(5) extension of the familiar $SU(2)\times U(1)$ Higgs doublet of the standard model. The Lagrangian for this minimal SU(5) model possesses a global U(1) symmetry[5], with the following U(1) quantum number (called X) assignments.

$$X(\mathbf{5}_H) = -\frac{2}{5}, \quad X(\mathbf{5}^*_F) = -\frac{3}{5}, \quad X(\mathbf{10})_F) = \frac{1}{5}\ . \tag{2}$$

Here the subscript F refers to left-handed Weyl fermion fields. The remaining fields namely the $\mathbf{24}_H$ Higgs field and the gauge fields, are assigned zero X-charge. The $SU(2)\times U(1)$ breaking vacuum expectation value of the $\mathbf{5}_H$ Higgs field also breaks the X-charge spontaneously. However the linear combination

$$Z = X + \frac{4}{5}\cdot\frac{Y}{2}\ , \tag{3}$$

where Y denotes the weak hypercharge leaves the $\mathbf{5}_H$ vacuum expectation value invariant. The Z-charge is therefore conserved globally and the appearance of a Goldstone boson is avoided.[6] It turns out that for fermions

$$Z = B - L\ . \tag{4}$$

We must treat Z as a generalised $B-L$ charge, in the sense that some bosons carry this quantum number although neither B nor L quantum numbers are assigned to them.

Exact $B-L$ conservation appears more or less accidentally in the minimal SU(5) model and fails if additional Higgs fields, such as a $\mathbf{15}_H$ or $\mathbf{10}^*_H$, are introduced. For instance, a $\mathbf{15}_H$ Higgs field would generate Majorana mass terms for the neutrinos,

obviously violating L while B is unchanged. Extending SU(5) to the SO(10) GUT, Z becomes a generator of the gauge group.[5] In order to avoid a massless gauge boson coupled to the Z charge, it must be spontaneously broken and $B - L$ conservation is no longer exact in the SO(10) model.

Apart from the perfect fit of quarks and leptons into SU(5) representations, the prediction of the Weinberg angle[3]

$$\sin^2 \theta_W = 0.228 \pm 0.0044 \tag{5}$$

is the main experimental evidence in favour of grand unification. In theories like SU(5), where the complete spectrum of fermions is already contained in the standard model, the Weinberg angle is given by

$$\sin^2 \theta_W = \frac{g_1^2}{g_1^2 + g_2^2} = \frac{3}{8} \text{ (at the unification scale)} \tag{6}$$

at super-high energies, where the breaking of the grand unified group G can be neglected [Paper 16]. Also theories based on larger groups, such as SO(10) or E_6,[3] in which the standard model group is embedded in an SU(5) subgroup, give the same prediction, Eq. (6), for $\sin^2 \theta_W$ at the unification scale. Renormalisation group techniques can be used to calculate the dependence of the gauge coupling constants $g_i(\mu)$, and hence $\sin^2 \theta_W(\mu)$, on the momentum scale μ at which they are defined [Paper 16]. Below the unification mass scale m_X, the couplings $g_i(\mu)$ vary in a G non-invariant way. For the minimal SU(5) model, or any simple GUT satisfying Eq. (6) with no new physics in the 'great desert' between $\mu = m_W$ and $\mu = m_X$, the corrected value of the Weinberg angle at $\mu = m_W$ agrees with the experimental value, Eq. (5), within errors. (*Note added in proof*: Recent LEP precision measurements of $\sin^2 \theta_W$ are inconsistent with the minimal SU(5) value.)

Assuming the great desert hypothesis of no new physics between $\mu = m_W$ and $\mu = m_X$, the unification mass scale is predicted to be[3]

$$m_X \simeq 1.3 \Lambda_{\rm QCD} \times 10^{15} \ . \tag{7}$$

Here $\Lambda_{\rm QCD}$ denotes the effective four flavour QCD scale parameter (in the modified minimal subtraction scheme of Ref. 7). Superheavy gauge bosons of mass m_X mediate proton decay at a rate proportional to m_X^{-4}, leading to a predicted partial lifetime for the decay $p \to e^+ \pi^0$ of[3]

$$\tau(p \to e^+ \pi^0) = 6.6 \times 10^{28 \pm 0.9} \left[\frac{\Lambda_{\rm QCD}}{100 \text{ MeV}} \right]^4 \text{ yr} \tag{8}$$

This decay rate is calculated assuming that all the fermion mixing matrices, which appear in baryon number violating processes, are equal to the usual Kobayashi-Maskawa matrix, as predicted in minimal SU(5). The factor $10^{\pm 0.9}$ is an indication

of the theoretical uncertainties in the calculation. We therefore obtain the following theoretical bounds[3] on proton decay, using present experimental limits on Λ_{QCD}

$$\tau(p \to e^+\pi^0) < 8.4 \times 10^{30} \text{ yr for } \Lambda_{QCD} = 100^{+100}_{-50} \text{ MeV} \tag{9}$$

and

$$\tau(p \to e^+\pi^0) < 1.4 \times 10^{32} \text{ yr for } \Lambda_{QCD} < 400 \text{ MeV} . \tag{10}$$

These results are in contradiction with the experimental bound[8]

$$\tau(p \to e^+\pi^0) > 3.3 \times 10^{32} \text{ yr} \quad (\text{at } 90\% \text{ C.L.}) . \tag{11}$$

The simple SU(5) model is thus ruled out by the experimental bound on proton decay. It is necessary to introduce new physics at a mass scale below m_X, in the form of scalars, fermions and/or gauge bosons, in order to obtain a GUT with

$$m_X \geq 5 \times 10^{14} \text{ GeV} , \tag{12}$$

compatible with the limits on proton decay. This suggests the existence of a larger gauge group, such as the superstring motivated $E_8 \times E_8$ symmetry [Papers 21, 22], which is broken down in steps as one goes down the quantum staircase. The standard model group then consists of the symmetries which survive near the bottom of the staircase, and can be considered to arise from a sort of "survival of the fittest" mechanism.[9]

An appealing modification of the standard model, with new physics below the unification mass scale m_X, involves the introduction of simple $N = 1$ supersymmetry.[4,10] Supersymmetry relates bosons and fermions and their interactions. In the supersymmetric version of the standard model, the spin 1 gauge bosons are accompanied by new spin $\frac{1}{2}$ gluinos, photinos etc., spin $\frac{1}{2}$ fermions are accompanied by new spin zero squarks and sleptons, and spin zero Higgs particles are accompanied by new spin $\frac{1}{2}$ Higgsinos. Supersymmetric theories have remarkable renormalisation properties and in some cases are completely finite in perturbation theory. These properties can be exploited to make the mass hierarchy, $m_X >> m_W$, technically natural in supersymmetric grand unification, if global $N = 1$ supersymmetry is broken at an energy around 1 TeV. In addition the grand unification mass scale m_X is increased, as required for consistency between experiment and proton decay mediated by superheavy boson exchange.

The commutator of two successive supersymmetry transformations is equivalent to a space-time translation. Thus, if the supersymmetry transformations are made local and gauged, the spin 2 graviton gauge field is automatically included, as a partner to the spin $\frac{3}{2}$ gravitino gauge field associated with local supersymmetry transformations. Local supersymmetry therefore includes Einsteinian gravity and is usually called supergravity. Gravity can be unified with other particle interactions in N-extended supergravity, which contains N spinorial supersymmetry generators and N gravitinos. The largest of these models is $N = 8$ supergravity [Paper 18]. For

$N > 8$, the model necessarily contains massless particles with spin greater than two, and the corresponding conserved currents are so restrictive on scattering processes that the S-matrix must be trivial.[11]

The $N = 8$ supergravity theory is truly unified, in the sense that all fundamental fields of the theory appear in a single supermultiplet. However, despite the improved convergence due to supersymmetry, the theory is not renormalisable and divergences very likely appear at three or perhaps eight loops.[12] It has a manifest SO(8) global symmetry. The supermultiplet contains 28 vector bosons and, in fact, it is possible to gauge the SO(8) group. This alternative $N = 8$ gauged supergravity[13] has a new independent gauge coupling constant g, in addition to the gravitational constant $\kappa(= \sqrt{4\pi G_N}$ in four dimensions, where G_N is Newton's constant).

The SO(8) group does not contain the minimal grand unified group SU(5) as a subgroup. Consequently SO(8) does not contain candidate states for all the observed elementary particles of the standard model. However, when properly interpreted, $N = 8$ supergravity is found to have a hidden local SU(8) symmetry [Paper 18]. This raises the possibility that the SU(8) somehow gauges itself dynamically and acts as a realistic grand unified group. The $N = 8$ supergravity model must then be treated as a preon theory, in which all the spin 1, spin $\frac{1}{2}$ and spin 0 particles of the standard model are composite. Unfortunately attempts to obtain a successful phenomenology based on this idea have been beset with difficulties.[14] Here we are interested in explaining the origin of the local SU(8) group together with a global exceptional E_7 group, as hidden symmetries of $N = 8$ supergravity.

4.2. Hidden Local Symmetry and Dynamical Gauge Bosons in Non-Linear Sigma Models

As we shall see later, the scalar and pseudoscalar particles of $N = 8$ supergravity form a non-linear sigma model field theory, in which the field takes values on the homogeneous space corresponding to the set of cosets $E_{7(+7)}/\mathrm{SU}(8)$. In order to get acquainted with such coset-valued field theories, we shall first consider how such models are revealed to be gauge theories.

Let us consider the general sigma model consisting of a set of spin zero fields, which take values on a coset space G/H rather than in number sets as for usual fields. By the coset space G/H, we understand the set

$$G/H = \{gH | g \in G\} \tag{13}$$

of cosets $gH = \{gh | h \in H\}$, where G is a group and H is a subgroup of G. We are interested in connected Lie groups. In the case relevant for $N = 8$ supergravity, the group $G = E_{7(+7)}$ is in fact a non-compact Lie group. It is the non-compact version of the exceptional group E_7, in which $+7 = 70 - 63$ is the signature of G equal to the number of non-compact generators minus the number of compact generators. The corresponding group $H \approx \mathrm{SU}(8)$ is the maximal compact subgroup of $E_{7(+7)}$, having 63 generators.

As an explicit example of our general discussion, we shall refer to the simpler $\mathbb{C}P^{n-1}$ model, which is renormalisable in two dimensions and has been studied in detail using the $1/n$ expansion [Paper 17]. The coset space for the $\mathbb{C}P^{n-1}$ model is obtained by taking

$$G = \mathrm{SU}(n) \tag{14}$$

and the subgroup

$$H = \mathrm{S}(\mathrm{U}(1) \times \mathrm{U}(n-1)) \approx \mathrm{U(n - 1)} \; . \tag{15}$$

Here $\mathrm{S}(\mathrm{U}(1)\times\mathrm{U}(n-1))$ denotes the group of $n \times n$ unitary matrices with unit determinant of the block form

$$\left(\begin{array}{c|ccc} \mathrm{U}(1) & 0 & \dots & 0 \\ \hline 0 & & & \\ \vdots & & \mathrm{U}(n-1) & \\ 0 & & & \end{array}\right) \tag{16}$$

having a $\mathrm{U}(1)$ and a $\mathrm{U}(n-1)$ matrix along the diagonal.

The coset space G/H for the $\mathbb{C}P^{n-1}$ model is, in fact, the set of directions, i.e. rays, in the space $\mathbb{C}^n$ of n complex dimensions. It is easily seen that the group $G = \mathrm{SU}(n)$ acts transitively on the set of rays

$$[z] = \{\lambda z | \lambda \in \mathbb{C}\} \tag{17}$$

obtained by letting z be any ordered set $z = (z_1, z_2, \dots, z_n)$ of n complex numbers not all being zero, $z \neq 0$. This means that, for any pair of rays $[z]$ and $[y]$, there exists an element g of G such that

$$g[z] = [y], \quad g \in \mathrm{SU}(n) \; . \tag{18}$$

Here the operation of g on $[z]$ is defined to be matrix multiplication, treating z as an n dimensional column vector. The subgroup of G leaving a ray $[z]$ invariant is isomorphic to the group H. For example, the ray corresponding to the vector $z = (1, 0, \dots, 0)$ is clearly left invariant by the $n \times n$ matrix of Eq. (16) and we have

$$h[(1, 0, \dots, 0)] = [(1, 0, \dots, 0)], \quad h \in H \; . \tag{19}$$

In fact, there is a one-to-one correspondence between the rays $[z]$ and the cosets gH given by

$$[z] \leftrightarrow \{g \in G | g[1, 0, \dots, 0] = [z]\} \; . \tag{20}$$

It follows from Eq. (19) that if $g \in \{g \in G | g[1,0,\dots,0] = [z]\}$ then $gh \in \{g \in G | g[1, 0, \dots, 0] = [z]\}$ and, thus, the set $\{g \in G | g[1,0,\dots,0] = [z]\}$ is of the form of a coset $gH = \{gh | h \in H\}$.

It is a remarkable fact that, by a rather trivial rewriting of the theory, a sigma model taking values in G/H is found to have a gauge symmetry for the group H.

We may simply express the model in terms of a field theory taking values in the group G itself, rather than in the coset space. However we must then declare the replacement of a field $g(x) \in G$ by $g(x)h(x)$ to be a gauge transformation, where $h(x)$ is an arbitrary 'gauge function' taking values in H.

The Lagrangian density for such a sigma model will, in general, contain derivatives of the coset-valued field $\phi(x) = g(x)H$. How should we translate this derivative $\partial_\mu \phi(x)$ into the G-valued formulation? Mathematically $\partial_\mu \phi$ is a tangent vector to the coset space. This tangent vector space is in natural correspondence to a subspace of the tangent vector space of G, which is orthogonal to the tangent vectors along the coset (viewed as a sub-manifold of the group manifold G). Orthogonality is defined here with respect to the Killing form derived metric for the Lie group G. In the G-valued formulation, we therefore want to replace $\partial_\mu \phi$ by a derivative $D_\mu g(x)$ which is guaranteed to be a tangent vector to G orthogonal to the tangent vectors along $g(x)H$. Formally this is achieved by introducing the covariant derivative

$$D_\mu g(x) = \partial_\mu g(x) + i g(x) A_\mu(x) \tag{21}$$

where $A_\mu(x)$ is constructed from $g(x)$ and $\partial_\mu g(x)$, using the projection operator P_H from the Lie algebra $\mathfrak{G}$ of G onto the Lie algebra $\mathfrak{H}$ of H:

$$iA_\mu(x) = -P_H(g(x)^{-1}\partial_\mu g(x)) \; . \tag{22}$$

These equations are most easily understood by thinking of them as expressed in some representation; abstractly Eq. (21) may be rewritten as a relationship between Lie algebra elements:

$$g(x)^{-1} D_\mu g(x) = g(x)^{-1}\partial_\mu g(x) + iA_\mu(x) \; . \tag{23}$$

We should now like to include $A_\mu(x)$ as an auxiliary field, in such a way that varying the action with respect to $A_\mu(x)$ gives us Eq. (22). It is easily seen, Fig. 4.1, that we require $A_\mu(x)$ to lie in the Lie algebra $\mathfrak{H}$ and to minimise the expression

$$\mathrm{Tr}(|g(x)^{-1}\partial_\mu g(x) + iA_\mu(x)|^2) \tag{24}$$

Here the norm of a matrix M is defined by

$$|M|^2 = M^+ M \; . \tag{25}$$

More abstractly, we minimise the Killing form for G of $g(x)^{-1}\partial_\mu g(x) + iA_\mu(x)$ with itself.

We can immediately recognise the expression, Eq. (24), after minimisation, as the kinetic energy term

$$\partial^\mu \phi^\alpha \partial_\mu \phi^\beta g_{\alpha\beta}(\phi) \tag{26}$$

for the G/H valued sigma model. Here $\phi^\alpha (\alpha = 1, 2, \ldots, \dim(G/H))$ is a set of co-ordinates enumerating the points on the sigma model target space G/H, and

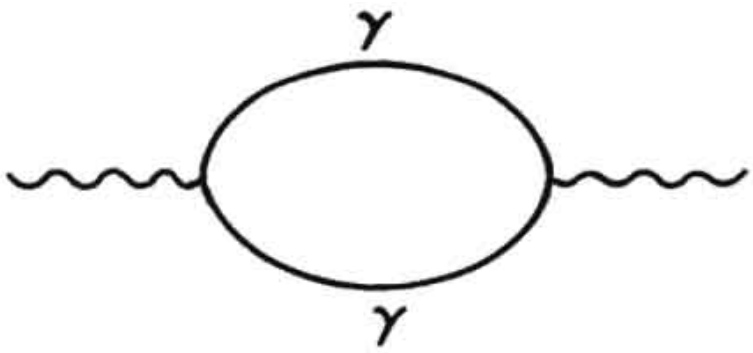


Fig. 4.2. Self energy diagram for the $A_\mu(x)$ field. The solid lines indicate propagators for the Lie algebra valued scalar field $\gamma(x)$.

In order to perform ordinary perturbation theory for this interaction, we linearise the group-valued field $g(x)$, by writing

$$g(x) = \exp(i\gamma(x)) \tag{33}$$

where $\gamma(x)$ is a Lie algebra valued field. A perturbation expansion of the action, Eq. (28), then gives

$$\frac{1}{2f}\mathrm{Tr}(|g(x)^{-1}\partial_\mu g(x) + iA_\mu(x)|^2)$$

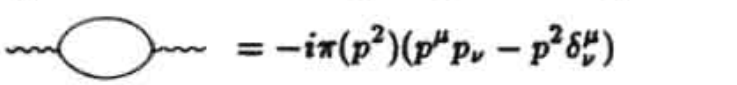
$$= \frac{1}{2f}\mathrm{Tr}((\partial_\mu\gamma(x) + A_\mu(x))^2)$$

$$- \frac{i}{2f}\mathrm{Tr}(\partial_\mu\gamma(x)[\gamma(x), \partial^\mu\gamma(x)])$$

$$- \frac{i}{2f}\mathrm{Tr}(A_\mu(x)[\gamma(x), \partial^\mu\gamma(x)]) + \dots . \tag{34}$$

The existence of a term of the form $\frac{1}{f}\mathrm{Tr}(A_\mu(x)\partial^\mu\gamma(x))$ in the action means that, in general, the $A_\mu(x)$ and $\gamma(x)$ fields mix. However we choose a gauge in which this mixing vanishes. In the Lorentz gauge, $\partial^\mu A_\mu(x) = 0$, the integral $\int d^dx\mathrm{Tr}(A_\mu(x)\partial^\mu\gamma(x))$ becomes a pure boundary term. In the "unitary gauge", $P_H(\gamma(x)) = 0$, the term $\mathrm{Tr}(A_\mu(x)\partial^\mu\gamma(x))$ simply vanishes. The propagators for $A_\mu(x)$ and $\gamma(x)$ do not then mix. The bare propagator for $\gamma(x)$ becomes that of a massless scalar particle, while the bare propagator for the $A_\mu(x)$ field is $-i2f\delta^\mu_\nu$ and does not have a particle pole.

There is a one loop self energy diagram, Fig. 4.2, for the A_μ field, where the vertex corresponds to the term $-\frac{i}{2f}\mathrm{Tr}(A_\mu[\gamma, \partial^\mu\gamma])$ in Eq. (34). The computation of this Feynman diagram in momentum space gives an expression of the form

 $= -i\pi(p^2)(p^\mu p_\nu - p^2\delta^\mu_\nu)$ (35)

where $\pi(p^2)$ turns out to be divergent in general. Now we recognise $(p^\mu p_\nu - p^2\delta^\mu_\nu)$ as the usual gauge field kinetic energy term written in momentum space. So, the one loop correction has generated a kinetic energy for A_μ and, to this order, the A_μ bilinear part of the effective Lagrangian density becomes

$$-\frac{\pi(p^2)}{2}\mathrm{Tr}(F_{\mu\nu}F^{\mu\nu}) + \frac{1}{2f}\mathrm{Tr}(A_\mu^2) \tag{36}$$

where

$$F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu \ . \tag{37}$$

Thus, in the $P_H(\gamma(x)) = 0$ gauge, quantum effects have generated a term corresponding to a vector A_μ particle of mass m_A given by

$$m_A^2 = \frac{1}{2f\pi(m_A^2)} \ . \tag{38}$$

In this sense, the dynamically generated A_μ gauge field is *a priori* Higgsed.

We note that an n-loop Feynman diagram corresponds to a term of order $\hbar^n$, in a power series expansion in Planck's constant $\hbar$. Hence $\pi(p^2)$ is of order $\hbar$ and thus m_A is proportional to $\hbar^{-1/2}$. So, in the classical limit $\hbar \to 0$, the mass m_A diverges and there is really no A_μ field particle. This is reasonable, since the A_μ field is a purely formal object initially.

We now consider whether the $A_\mu(x)$ gauge particles can obtain zero mass by quantum fluctuations of the $g(x)$ field. Surprisingly enough, it is indeed possible to generate an unbroken local gauge symmetry dynamically in this way. This is illustrated by our example of the $\mathbb{C}\mathrm{P}^{n-1}$ model.

In the $\mathbb{C}\mathrm{P}^{n-1}$ model, the gauge particle associated with the Abelian U(1) factor group of H is shown to be massless in $d = 2$ dimensions, using the $1/n$ expansion [Paper 17]. A massless gauge field in two dimensions has no real physical particle associated with it, since it has $d - 2 = 0$ transverse components. However it gives rise to a linear confining Coulomb potential, $V(r) \approx r$. On the other hand, a massive vector field in two dimensions gives rise to a Yukawa potential of the form $V(r) \approx r\exp(-m_A r)$. Thus, even in $d = 2$ dimensions, there is some physical content in a gauge field being massless. It follows that the $\mathbb{C}\mathrm{P}^{n-1}$ model provides an example of the reverse of the Higgs effect, dynamically generating a gauge symmetry [Paper 19].

It is a detailed dynamical question whether or not quantum fluctuations can generate a physical gauge symmetry, corresponding to the formal gauge group H (or to some subgroup of H). When an H gauge group symmetry is generated, it appears that the quantum fluctuations cause the probability distribution of the field $g(x)$ over G to be invariant, under the symmetry operation of right multiplication by elements of H, at any point in the vacuum. For instance, in the case when the fundamental model has a global G invariance, the distribution of $g(x)$ could be completely evenly distributed over G, like the Haar measure (see Fig. 4.3). When

the probability distribution of $g(x)$ over G in the vacuum is invariant under H, there is no longer any spontaneous breakdown of the global part of the H gauge symmetry. Thus the formal gauge bosons of the gauge group H are not Higgsed. We should then expect the presence of massless gauge particles, due to the dynamical generation of a kinetic energy term like Eq. (35).

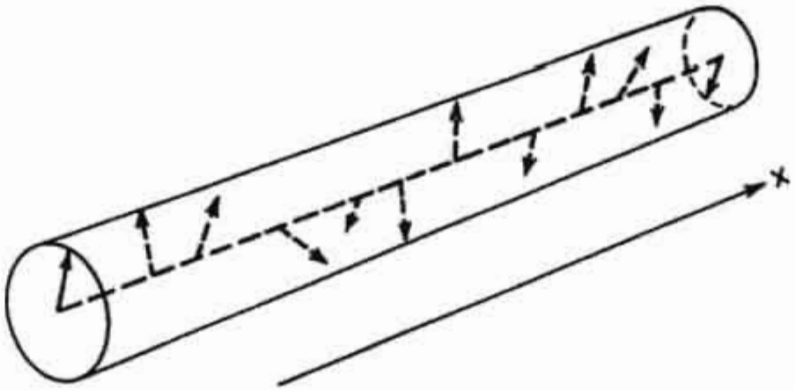


Fig. 4.3. Restoration of a vacuum symmetry in a non-linear sigma model. The figure illustrates symbolically how quantum fluctuations in the field $g(x)$ can be invariant under right multiplication by elements of H, using the oversimplified example of an S_1 target space.

The other possibility is that the probability distribution of $g(x)$ over G in the vacuum is not invariant under H transformations, but rather is peaked somewhere in G. Then we should expect the theory to remain in a Higgs-like phase with massive, or perhaps non-existent, gauge particles.

If G (or G/H) is a non-compact space, it seems unreasonable to expect the quantum fluctuations to spread the field $g(x)$ over the whole non-compact space. In fact, a G-invariant distribution cannot be normalised in a non-compact space and, so, cannot be used as a probability distribution. Therefore we do not expect the inverse Higgs effect to work in a non-compact group space. Indeed, detailed analysis[15] seems to confirm that the phenomenon, of dynamical generation of massless gauge bosons, does not occur in non-compact sigma models.

The idea, behind the analysis of Ref. 15, is to consider the effect of a renormalisation group block-spinning on the vacuum values of the G/H valued sigma model field $\phi(x)$. This block-spinning corresponds to integrating out the momenta in an interval between, say, μ and μ/n, where μ is the cut-off energy scale. There is, thereby, a transition to a more coarse-grained field formulation, in which n^d values of the field are blocked together. However, in order to take averages over the non-linear sigma model field $\phi(x)$, we must allow it to take values in a larger linear space, corresponding to the convex closure of the G/H manifold. The model must therefore be expressed in terms of a field $\Phi(x)$, which takes values on the target space of the corresponding linear sigma model (i.e. the convex closure of G/H). At the regularisation scale μ of the model, $\Phi(x)$ is constrained to lie in G/H. This is symbolised in Fig. 4.4 by (a) a full-line circle for a compact target space, and (b) a full-line hyperbola for a non-compact target space. A typical block of field values is shown and the mean value is denoted by a star. The target space of the

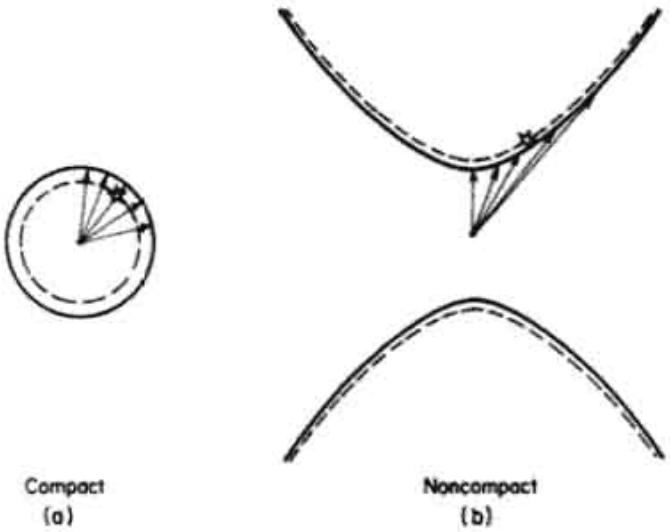


Fig. 4.4. Illustration of the effect of block-spinning the vacuum values of the sigma model field $\Phi(x)$. (a) For a compact sigma model, the target space becomes smaller and smaller in the infra-red. (b) For a non-compact sigma model, the target space becomes larger in the infra-red.

block-spinned field values is symbolised by a dashed curve.

As suggested by Fig. 4.4 we find, on the average, that the mean value of a block of n^d field values lies closer to the origin, in the linearised space, for a compact G/H manifold. On the other hand, we see that the mean value is very likely to lie further away from the origin for a non-compact G/H manifold. We therefore have the possibility, in the compact group case, of obtaining a phase in the infra-red corresponding to zero vacuum expectation value of $\Phi(x)$. For a non-compact group G, however, the vacuum expectation value of $\Phi(x)$ will not become zero. In this latter non-compact case, we still have the Higgs mechanism at work and the gauge bosons remain massive. It is only in the compact group G case that the gauge particles for the group H can become massless.

The 70 spin zero fields of the $N = 8$ supergravity model are described by a non-compact G/H sigma model, with $G = E_{7(+7)}$ and $H \approx \mathrm{SU}(8)$. Since H is the maximal compact subgroup of G, the theory has no ghosts and the Hamiltonian is positive definite [Paper 18]. However, the discussion of the previous paragraphs suggests that quantum fluctuations will not dynamically generate an unbroken SU(8) Yang-Mills interaction in the model. For this and other reasons,[14] it seems unlikely that $N = 8$ supergravity can provide a realistic theory of particle interactions.

The occurrence of hidden global $E_{7(+7)}$ and local SU(8) symmetries in $N = 8$ supergravity remains a suprising result. We now turn to a discussion of the origin of these symmetries, which is the aspect of supergravity theory relevant to this book.

4.3. Hidden Symmetries in N = 8 Supergravity

The $N = 8$ supergravity model [Paper 18] has an SU(8) gauge symmetry, which arises formally in much the same way as for the G/H valued non-linear sigma model described in the previous section. A more truly hidden symmetry is the invariance of the equations of motion under the global non-compact exceptional group $E_{7(+7)}$, which has $\mathrm{SU}(8)/Z_2$ as a maximal compact subgroup. In fact the 70 scalar and

pseudoscalar fields of $N = 8$ supergravity can be shown [Paper 18] to parameterise the coset space $E_{7(+7)}/(\mathrm{SU}(8)/Z_2)$. The main features of the above symmetries are illustrated in [Paper 19], by a qualitative discussion of the local SO(7) and global SL(7, IR) subgroups.

The 70 spin zero fields of $N = 8$ supergravity can be packed into a 56×56 matrix, called V, in such a way that these physical fields are unchanged, if V is multiplied on the left by a matrix belonging to the 56 dimensional representation of SU(8). Multiplication on the right, by a matrix belonging to the fundamental 56 dimensional representation of $E_{7(+7)}$, transforms V into another matrix corresponding, in general, to another set of spin zero fields. This transformation of the spin-zero fields is non-linear.

The simplest way of understanding the construction of the matrix V is to use a classification of the spin zero fields, resulting from the dimensional reduction of the $N = 1$ supergravity model in 11 dimensions [Paper 18]. Dimensional reduction involves compactifying seven of the dimensions to a point and assuming that the fields do not depend on the seven extra co-ordinates. The 11 dimensional $N = 1$ supergravity Lagrangian depends on these gauge fields: an elfbein $e_M{}^A$, a Rarita-Schwinger field ψ_M and a rank 3 antisymmetric tensor field A_{MNP}. The index M on the elfbein denotes a 'curved' vector index, which transforms according to the rules of general relativity under a reparameterisation of the 11 dimensional co-ordinates. On the other hand, the index A is a 'flat' tangent-space index, which transforms as a vector under local SO(1,10) Lorentz transformations. The flat metric has the 11 dimensional Minkowski form $\eta_{AB} = (+1, -1, -1, \ldots, -1)$. In general, letters at the beginning of the alphabets are used to denote flat indices, and the later characters are used for curved indices. Since we intend to rewrite $d = 11$ supergravity under a $4 + 7$ split of co-ordinates and indices, we let A, α, a (respectively M, μ, m) run from 0 to 10, 0 to 3 and 4 to 10 respectively.

The elfbein $e_M{}^A$ describes a graviton with $9 \times 10/2 - 1 = 44$ massless spin states. The Rarita-Schwinger field ψ_M satisfies a Majorana condition and describes a gravitino with $(11 - 3) \times 16 = 128$ spin states. Finally, the antisymmetric tensor (Kalb-Ramond) field A_{MNP} provides the extra $9 \times 8 \times 7/(3 \times 2) = 84$ boson spin states, required for supersymmetry. By dimensional reduction, the above 256 states become the 256 states of the CPT self-conjugate field representation of $N = 8$ supergravity in $d = 4$ dimensions: 2 forming a spin 2 graviton, 16 forming 8 spin $\frac{3}{2}$ gravitinos, 56 forming 28 spin 1 vector bosons, 112 forming 56 Majorana spin $\frac{1}{2}$ fermions and the rest forming 70 spin zero particles.

The spin zero particles are made up of 35 scalars and 35 pseudoscalars. The scalar particles arise in two ways. There are 28 scalar states described by the 7×7 internal metric tensor,

$$g_{mn} = e_m{}^a e_n{}^b \eta_{ab}, \quad m, n = 4, 5 \ldots, 10 \tag{39}$$

associated with the seven compactified dimensions. Here the internal flat metric $\eta_{ab} = -\delta_{ab}$. The remaining seven scalar fields $\phi_m (m = 4, 5, \ldots 10)$ are obtained from

the components $A_{\mu\nu m}(\mu,\nu = 0,1,2,3)$ of the antisymmetric tensor field A_{MNP}. These components $A_{\mu\nu m}$ form 7 second order antisymmetric tensor gauge fields in the four dimensional space. A simple analysis given below shows that, by a duality transformation, any second order antisymmetric gauge field $B_{\mu\nu}$ in four dimensions is equivalent to a spin zero field. This field is pseudoscalar if $B_{\mu\nu}$ is an ordinary tensor, and scalar if $B_{\mu\nu}$ is a pseudotensor as $A_{\mu\nu m}$ is taken to be.

The gauge invariant field strength

$$F_{\mu\nu\sigma m}(x) = \partial_{[\mu} A_{\nu\sigma]m} \tag{40}$$

obeys the equation of motion

$$\partial^{\mu} F_{\mu\nu\sigma m}(x) = 0 \tag{41}$$

in the free field approximation. Here and subsequently, we use square brackets [ ] to denote the antisymmetrized sum over all permutations of bracketed indices, divided by the number of these permutations. It follows that the dual field strength

$$\widetilde{F}^{\rho}{}_{m}(x) = \frac{1}{2}\varepsilon^{\rho\mu\nu\sigma} F_{\mu\nu\sigma m}(x) \tag{42}$$

obeys the equation

$$\partial^{\nu} \widetilde{F}^{\rho}{}_{m}(x) - \partial^{\rho} \widetilde{F}^{\nu}{}_{m}(x) = 0 \ . \tag{43}$$

The Levi-Civita tensor $\varepsilon^{\cdots}$, in any dimension, with upper indices is defined to be equal to +1 for even permutations of the natural order. From Eq. (43), neglecting all topological effects, we see that the dual field strength must be of the form

$$\widetilde{F}^{\rho}{}_{m}(x) = \partial^{\rho} \phi_{m}(x) \ . \tag{44}$$

It then follows from the Bianchi identity

$$\partial_{\rho} \widetilde{F}^{\rho}{}_{m}(x) = 0$$

that ϕ_m behaves like a scalar field, obeying

$$\Box \phi_{m}(x) = 0 \tag{45}$$

in the free approximation.

The pseudotensor nature of A_{MNP} in four dimensions implies that the components A_{mnp} describe $7 \times 6 \times 5/(3 \times 2) = 35$ pseudoscalar fields. We will return to the pseudoscalar fields, after first showing that the 35 scalar fields span the coset space SL(8, IR)/SO(8).

It is relatively easy to extend the 7×7 internal 'curved' metric tensor g_{ij} into an 8×8 matrix $S_{i'j'}(i',j' = 4,5,\ldots,11)$, which describes all the $28 + 7 = 35$ scalar

fields. This matrix $S_{i'j'}$, with primed indices running over 8 values each, is constructed from a field $v^{a'}{}_{i'}$, which is an extension of the siebenbein $e^a{}_i = e_i{}^a$. As indicated here, the order of the upper and lower indices on the siebenbein is irrelevant. The upper index is 'flat' on the siebenbein $e^a{}_i$ and 'curved' on its inverse $e_a{}^i$.

The extended siebenbein, or achtbein, $v^{a'}{}_{i'}$ has unit determinant and is constructed from the siebenbein $e^a{}_i$ and the 7 scalar fields ϕ_m as follows:

$$v^{a'}{}_{i'} = \Delta^{-1/8} \begin{pmatrix} e^a{}_i & \phi^j e^a{}_j \\ 0 \ldots 0 & \sqrt{\Delta} \end{pmatrix} \tag{46}$$

where

$$\Delta = (\det e^a{}_i)^2 = -\det g_{ij} \ . \tag{47}$$

The extended metric tensor $S_{i'j'}$ is then constructed from $v^{a'}{}_{i'}$,

$$\begin{aligned} S_{i'j'} &= v^{a'}{}_{i'} v^{b'}{}_{j'} \eta_{a'b'} \\ &= \Delta^{-1/4} \begin{pmatrix} g_{ij} & \phi_i \\ \phi_j & \phi^l \phi_l - \Delta \end{pmatrix} \end{aligned} \tag{48}$$

where $\phi_j = g_{jk}\phi^k$ and $\eta_{a'b'} = -\delta_{a'b'}$. It follows that the real symmetric 8×8 matrix $S_{i'j'}$ satisfies $\det S_{i'j'} = 1$, and has $8\times 9/2 - 1 = 35$ degrees of freedom to represent the 35 scalar fields.

It is not difficult to see that the values of the $S_{i'j'}$ tensor are in one-to-one correspondence with the coset space SL(8, IR)/SO(8). The achtbein $v^{a'}{}_{i'}$ is an 8×8 matrix with $\det v^{a'}{}_{i'} = 1$ and can, therefore, be considered to be an element of the group SL(8, IR) of special linear maps of an 8 dimensional real vector space. However $v^{a'}{}_{i'}$ and the transformed achtbein $R^{a'}{}_{b'} v^{b'}{}_{i'}$ (with $\det R^{a'}{}_{b'} = 1$, so that the transformed achtbein has unit determinant) correspond to the same tensor $S_{i'j'}$, when $R^{a'}{}_{b'}$ belongs to the group SO(8). Thus a scalar field configuration, or equivalently a particular tensor $S_{i'j'}$, corresponds to the set of SL(8, IR) matrices

$$\begin{aligned} &\{v^{a'}{}_{i'} \in \mathrm{SL}(8,\ \mathbb{R}) | v^{a'}{}_{i'} v^{b'}{}_{j'} \eta_{a'b'} = S_{i'j'}\} \\ &\quad = \{R^{a'}{}_{b'} v^{(0)b'}{}_{i'} | R^{a'}{}_{b'} \in \mathrm{SO}(8)\} \end{aligned} \tag{49}$$

where $v^{(0)b'}{}_{i'}$ is some representative achtbein, obeying $v^{(0)b'}{}_{i'} v^{(0)c'}{}_{j'} \eta_{b'c'} = S_{i'j'}$. We immediately see that this set is a coset gH, where $g \in \mathrm{SL}(8, \mathbb{R})$ and $H = \mathrm{SO}(8)$.

It follows, from the general considerations of the previous section, that the scalar field sector formally has an SO(8) gauge symmetry. In fact, it also has a hidden SL(8, IR) global symmetry. However we will postpone a discussion of the global symmetry, until we have included the pseudoscalar sector.

The 35 pseudoscalar fields are described by the components A_{mnp} $(m, n, p = 4, 5, \ldots, 10)$ of the antisymmetric pseudotensor field. By contraction with the siebenbein, we introduce pseudoscalar fields

$$A_{abc} = e_a{}^m e_b{}^n e^p{}_c \ A_{mnp} \tag{50}$$

with 'flat' indices. Then we formally add an extra index, set equal to 11, by defining the totally antisymmetric field

$$A_{a'b'c'd'} = 4A_{[a'b'c'}\delta_{d']}{}^{11} \tag{51}$$

with primed indices taking 8 values (4,5, ..., 11). Using the Levi-Civita symbol with 8 indices, we can also define the dual field

$${}^*A^{a'b'c'd'} = \frac{1}{24}\varepsilon^{a'b'c'd'e'f'g'h'}A_{e'f'g'h'} \tag{52}$$

which is zero if one of its indices is equal to 11. It is now possible to pack the pseudoscalar fields into a 56×56 matrix

$$V_- = \exp\begin{pmatrix} 0 & {}^*A^{a'b'c'd'} \\ A_{a'b'c'd'} & 0 \end{pmatrix} \tag{53}$$

where we consider the fields $A_{a'b'c'd'}$ and ${}^*A^{a'b'c'd'}$ as 28×28 matrices with rows and columns labelled by non-ordered pairs, $(a'b')$ and $(c'd')$ respectively, of unequal indices. The matrix V_- can be recognised as an element of the E_7 group in its fundamental 56 dimensional representation (Appendix B of Paper 18).

We must now introduce a 56×56 E_7 matrix V_+ to describe the scalar fields, so that we can construct the 56×56 matrix $V = V_-V_+$ referred to earlier, which describes all 70 spin zero fields. The matrix V_+ is formed by 'squaring' the achtbein $v^{a'}{}_{i'}$

$$V_+ = \begin{pmatrix} v^{[a'}{}_{[i'}v^{b']}{}_{j']} & 0 \\ 0 & v_{[a'}{}^{[i'}v_{b']}{}^{j']} \end{pmatrix} \tag{54}$$

where

$$v_{a'}{}^{i'} = \eta_{a'b'}v^{b'}{}_{j'}S^{i'j'} \ . \tag{55}$$

The rows of V_+ are labelled by a non-ordered pair of unequal flat indices $(a'b')$ — upper indices for the first 28 and lower indices for the last 28 rows. The columns are labelled, in an analogous way, by curved indices (i', j').

Under an SO(8) gauge transformation on the achtbein, we have

$$v^{a'}{}_{i'} \longrightarrow R^{a'}{}_{b'}v^{b'}{}_{i'} \tag{56}$$

$$v_{a'}{}^{i'} \longrightarrow R_{a'}{}^{b'}v_{b'}{}^{i'} \tag{57}$$

where

$$R_{a'}{}^{b'} = (R^{-1})^{b'}{}_{a'} = R^{a'}{}_{b'} \ . \tag{58}$$

It follows that the corresponding local SO(8) transformation on V_+ is given by

$$V_+ \longrightarrow \tilde{R}V_+$$

where

$$\tilde{R} = \begin{pmatrix} R^{[a'}{}_{[c'} R^{b']}{}_{d']} & 0 \\ 0 & R_{[a'}{}^{[c'} R_{b']}{}^{d']} \end{pmatrix} . \tag{59}$$

Under this SO(8) transformation on the flat indices, V_- transforms as follows

$$V_- \longrightarrow \tilde{R} V_- \tilde{R}^{-1} . \tag{60}$$

Thus the matrix of spin zero fields

$$V = V_- V_+ \tag{61}$$

transforms in the same way as V_+,

$$V \longrightarrow \tilde{R} V \tag{62}$$

The SO(8) transformation matrix $\tilde{R}$ can actually be extended to an SU(8) (or rather SU(8)/Z_2) transformation matrix U. Since V_+ and V are transformed by left multiplication, we are really only interested here in the transformation of a 56-column

$$\begin{pmatrix} x^{a'b'} \\ y_{a'b'} \end{pmatrix} \tag{63}$$

obtained by selecting specific values for the curved indices (i', j'). In order to consider the general form of the SU(8) transformation matrix U, it is convenient to pack the above 56-column into a complex 28-column

$$\frac{x^{a'b'} + i y_{a'b'}}{\sqrt{2}} \tag{64}$$

The SO(8) group corresponding to the transformation matrix $\tilde{R}$ is not embedded in SU(8) in the standard trivial way, but is related to the trivial embedding by one of the outer automorphisms of SO(8). In other words, the components of the 28-column in Eq. (64) do not transform simply as the antisymmetric product Z_{AB} of two fundamental 8 dimensional representations of SU(8). However, as discussed in the Appendices of Paper 18, they can be carried into such a representation by the linear transformation

$$Z_{AB} = \frac{1}{4} \left(\frac{x^{a'b'} + i y_{a'b'}}{\sqrt{2}} \right) (\Gamma^{a'b'})_{AB} . \tag{65}$$

Here

$$\Gamma^{a'b'} = \Gamma^{[a'} \Gamma^{b']} \tag{66}$$

is the antisymmetric product of two 8×8 γ^5-projected γ-matrices, appropriate to a Weyl spinor in 8 dimensions.

Calculation then shows that an infinitesimal SU(8) transformation on Z_{AB} translates into the following infinitesimal transformation on V

$$V \longrightarrow V + \begin{pmatrix} 2\Lambda^{[a'}{}_{[c'}\delta^{b']}{}_{d']} & {}^*\Sigma^{a'b'c'd'} \\ \Sigma_{a'b'c'd'} & 2\Lambda_{[a'}{}^{[c'}\delta_{b']}{}^{d']} \end{pmatrix} V \ . \tag{67}$$

The 28 antisymmetric infinitesimal parameters $\Lambda^{a'}{}_{c'} = -\Lambda^{c'}{}_{a'}$ describe the SO(8) transformations. They are related to the parameters $\Lambda'_A{}^B = -\Lambda'_B{}^A$, associated with the SO(8) generators in the defining representation of SU(8), by the automorphism

$$\Lambda'{}_A{}^B = \frac{1}{4}\Lambda_{a'b'}(\Gamma^{a'b'})_A{}^B = -\Lambda'_B{}^A \ . \tag{68}$$

We note that the metric $\eta_{a'b'} = -\delta_{a'b'}$ is used to lower indices and that the position of the indices A, B on $\Gamma^{a'b'}$ is of no significance. The traceless set of symmetric infinitesimal parameters $\Lambda''{}_A{}^B = \Lambda''{}_B{}^A$, corresponding to the other 35 generators of SU(8) in the defining 8 dimensional representation, is given by

$$\Lambda''{}_A{}^B = -\frac{1}{48}\Sigma_{a'b'c'd'}(\Gamma^{a'b'c'd'})_A{}^B = \Lambda''{}_B{}^A \tag{69}$$

Here $\Gamma^{a'b'c'd'}$ is the antisymmetric product

$$\Gamma^{a'b'c'd'} = \Gamma^{[a'}\Gamma^{b'}\Gamma^{c'}\Gamma^{d']} \tag{70}$$

and the symmetry property $\Lambda''{}_A{}^B = \Lambda''{}_B{}^A$ translates into the anti-self-duality condition

$$\Sigma_{a'b'c'd'} = -{}^*\Sigma^{a'b'c'd'} \ . \tag{71}$$

We now wish to establish the advertised result that the scalar and pseudoscalar fields parameterise the homogeneous space $E_{7(+7)}/(\mathrm{SU}(8)/Z_2)$, by showing:

I) The SU(8) transformations U acting on the set of matrices V obtained from the spin zero fields,

$$UV = UV_-V_+ \ , \tag{72}$$

generate all the 56×56 representation matrices of $E_{7(+7)}$.

and

II) For no matrix $V = V_-V_+$ does the coset $\{UV|U \in \mathrm{SU}(8)\}$ contain another V, representing a different spin zero field configuration.

We shall also see below that the action for the spin zero field sector is invariant under global $E_{7(+7)}$ transformations, but this will not be relevant to our discussion of local SU(8) invariance.

The group $E_{7(+7)}$ may be defined, in the fundamental 56 dimensional representation, as the set of matrices

$$\exp\begin{pmatrix} 2\Lambda^{[a'}{}_{[c'}\delta^{b']}{}_{d']} & {}^*\Sigma^{a'b'c'd'} \\ \Sigma_{a'b'c'd'} & 2\Lambda_{[a'}{}^{[c'}\delta_{b']}{}^{d']} \end{pmatrix} \ . \tag{73}$$

Here, $\Sigma_{a'b'c'd'}$ is any real totally antisymmetric fourth rank tensor (70 parameters) and $\Lambda^{a'}{}_{b'} = -\Lambda_{b'}{}^{a'}$ is any real traceless second order tensor (63 parameters). Its SU(8) subgroup is obtained by restricting $\Lambda^{a'}{}_{b'}$ to be antisymmetric and $\Sigma_{a'b'c'd'}$ to be anti-selfdual. The symmetric part of $\Lambda^{a'}{}_{b'}$ and the selfdual part of $\Sigma_{a'b'c'd'}$ parameterise E_7 transformations, corresponding to the generators orthogonal to SU(8). Property I above requires that these transformations can be constructed by multiplying an SU(8) matrix, U, by a matrix of spin zero fields $V = V_- V_+$.

We have already seen that the scalar field configurations parameterise the homogeneous space SL(8, IR)/SO(8), with the general achtbein field $v^{a'}{}_{i'}$ being any element of SL(8, IR). Thus V_+ parameterises the SL(8, IR) subgroup of $E_{7(+7)}$. Following Cremmer and Julia, we shall be satisfied to show that UV_-V_+ parameterises the $E_{7(+7)}$ group in the neighbourhood of the identity I.

An infinitesimal scalar field configuration gives an achtbein of the form

$$v^{a'}{}_{i'} = \delta^{a'}{}_{i'} + \Delta v^{a'}{}_{i'} \tag{74}$$

where $\Delta v^{a'}{}_{i'}$ is an arbitrary infinitesimal traceless second order tensor. An infinitesimal pseudoscalar field configuration is described by an infinitesimal fourth order antisymmetric tensor $A_{a'b'c'd'}$, having 35 independent components with $d' = 11$. The corresponding matrices V_+ and V_- take the forms

$$V_+ = I + \begin{pmatrix} 2\Delta v^{[a'}{}_{[i'}\delta^{b']}{}_{j']} & 0 \\ 0 & 2\Delta v_{[a'}{}^{[i'}\delta_{b']}{}^{j']} \end{pmatrix} \tag{75}$$

and

$$V_- = I + \begin{pmatrix} 0 & {}^*A^{a'b'c'd'} \\ A_{a'b'c'd'} & 0 \end{pmatrix} . \tag{76}$$

An infinitesimal SU(8) transformation takes the form

$$U = I + \begin{pmatrix} 2\tilde{\Lambda}^{[a'}{}_{[c'}\delta^{b']}{}_{d']} & {}^*\tilde{\Sigma}^{a'b'c'd'} \\ \tilde{\Sigma}_{a'b'c'd'} & 2\tilde{\Lambda}_{[a'}{}^{[c'}\delta_{b']}{}^{d']} \end{pmatrix} \tag{77}$$

where

$$\tilde{\Lambda}^{a'}{}_{b'} = -\tilde{\Lambda}^{b'}{}_{a'} , \quad \Sigma_{a'b'c'd'} = -{}^*\Sigma^{a'b'c'd'} . \tag{78}$$

It follows that, to leading order, UV_-V_+ is an $E_{7(+7)}$ matrix of the form of Eq. (73) (multiplied by a unit matrix converting flat to curved indices) with

$$\Lambda^{a'}{}_{b'} = \tilde{\Lambda}^{a'}{}_{b'} + \Delta v^{a'}{}_{b'} \tag{79}$$

and

$$\Sigma_{a'b'c'd'} = \tilde{\Sigma}_{a'b'c'd'} + A_{a'b'c'd'} . \tag{80}$$

In order to derive property I, we must now show that any set of E_7 parameters $(\Lambda^{a'}{}_{b'}, \Sigma_{a'b'c'd'})$ can be obtained in this way. From Eq. (74) and (79), $\Delta v^{a'}{}_{b'}$

and $\Lambda^{a'}{}_{b'}$ are arbitrary infinitesimal traceless second order tensors. The infinitesimal anti-selfdual tensor $\widetilde{\Sigma}_{a'b'c'd'}$ has 35 independent components, which can be chosen to be the components with no index equal to 11 (for which Eq. (51) and (80) give $\Sigma_{a'b'c'd'} = \widetilde{\Sigma}_{a'b'c'd'}$). The infinitesimal pseudoscalar fields $A_{a'b'c'd'}$ can, then, be chosen to give any required infinitesimal value to the other 35 independent components of $\Sigma_{a'b'c'd'} = \widetilde{\Sigma}_{a'b'c'd'} + A_{a'b'c'd'}$ with $d' = 11$. Thus we generate arbitrary infinitesimal values for the $E_{7(+7)}$ parameters $(\Lambda^{a'}{}_{b'}, \Sigma_{a'b'c'd'})$ and establish statement I in the neighbourhood of the unit element.

Let us now verify statement II above, but again to lowest order in the pseudoscalar fields. We first notice that the matrix V^+V is invariant under the SU(8) transformation $V \longrightarrow UV$. In order to establish property II, we must show that the invariant matrix $R = V^+V (= V^TV$ when V is simply $V = V_-V_+$ and not transformed with $U)$, which has only curved indices, specifies the values of the physical scalar $S_{i'j'}$ and pseudoscalar A_{mnp} fields.

We find, to linear order in the pseudoscalar fields, that

$$R = \begin{pmatrix} \frac{1}{2}(S_{i'k'}S_{j'l'} - S_{i'l'}S_{j'k'}) & S_{i'm'}S_{j'n'}{}^*A^{m'n'k'l'} + A_{i'j'm'n'}S^{m'k'}S^{n'l'} \\ S^{i'm'}S^{j'n'}A_{m'n'k'l'} + {}^*A^{i'j'm'n'}S_{m'k'}S_{n'l'} & \frac{1}{2}(S^{i'k'}S^{j'l'} - S^{i'l'}S^{j'k'}) \end{pmatrix} \tag{81}$$

where the rows and columns are labelled by the non-ordered pairs (i', j') and (k', l'). The achtbein $v^{a'}{}_{i'}$ has been used to change from flat to curved indices, on the four index antisymmetric pseudoscalar field $A_{i'j'k'l'}$. It follows from Eq. (46) and (51) that $A_{i'j'k'l'} = 0$ if all indices are different from 11. The inverse metric tensor $S^{i'j'}$ is given by

$$S^{i'j'} = \Delta^{-3/4} \begin{pmatrix} \Delta g^{ij} - \phi^i\phi^j & \phi^i \\ \phi^j & -1 \end{pmatrix} . \tag{82}$$

It is not difficult to see that the diagonal 28×28 blocks of R completely specify the metric tensor $S_{i'j'}$, because the blocks have a superfluous number of components, compared to the 35 independent scalar field components of $S_{i'j'}$. Similarly, knowing $S_{i'j'}$, the off-diagonal blocks of R determine the 35 independent pseudoscalar field components of $A_{i'j'k'l'}$ (or equivalently of its dual ${}^*A^{i'j'k'l'}$). We have therefore verified statement II, for all scalar fields $S_{i'j'}$ and infinitesimal pseudoscalar fields A_{mnp}.

We have thus established that the 70 spin zero fields of $N = 8$ supergravity parameterise the coset space $E_{7(+7)}/(\mathrm{SU}(8)/Z_2)$, in the neighbourhood of the identity. We now simply assume, as in [Paper 18], that the manifold of the spin zero fields is the full coset space. According to the general discussion of the previous section, the theory should then have a formal SU(8) gauge invariance, with a set of auxiliary gauge fields $Q_{\mu A}{}^B$. Here it is convenient to use the basis of Eq. (65), with capital letters, $A, B = 4, 5, \ldots 11$, which features the SU(8) subgroup of E_7. In this basis,

the 56×56 matrix describing the spin zero fields takes the block form

$$V' = \begin{pmatrix} u_{AB}{}^{MN} & v_{ABMN} \\ \overline{v}^{ABMN} & \overline{u}^{AB}{}_{MN} \end{pmatrix} \tag{83}$$

where the bar denotes complex conjugation

$$\overline{u}^{AB}{}_{MN} = (u_{AB}{}^{MN})^* \quad \overline{v}^{ABMN} = (v_{ABMN})^* \ . \tag{84}$$

The local SU(8) symmetry acts on the left of V' (flat indices A, B) and the global E_7 symmetry, considered below, acts on the right (curved indices M, N).

Under an infinitesimal SU(8) transformation, we have

$$(\delta_{SU(8)} V')V'^{-1} = \begin{pmatrix} 2\Lambda_{[A}{}^{[C}\delta_{B]}{}^{D]} & 0 \\ 0 & 2\overline{\Lambda}^{[A}{}_{[C}\delta^{B]}{}_{D]} \end{pmatrix} \tag{85}$$

where the traceless antihermitian 8×8 matrix $\Lambda_A{}^C(\overline{\Lambda}^A{}_C)$ acts on an $\mathbf{8}(\mathbf{8}^*)$ representation. The set of (traceless antihermitian) gauge fields $Q_{\mu A}{}^B$ is defined, by projecting the E_7 Lie algebra elements $(\partial_\mu V')V'^{-1}$ onto the SU(8) subalgebra, which we have just seen, Eq. (85), corresponds to the block diagonal elements:

$$(\partial_\mu V')V'^{-1} = \begin{pmatrix} 2Q_{\mu[A}{}^{[C}\delta_{B]}{}^{D]} & P_{\mu ABCD} \\ \overline{P}_\mu{}^{ABCD} & 2\overline{Q}_\mu{}^{[A}{}_{[C}\delta^{B]}{}_{D]} \end{pmatrix} \ . \tag{86}$$

Using $Q_{\mu A}{}^B$ as SU(8) gauge fields, the covariant derivative of the spin zero fields is given by

$$(D_\mu V')V'^{-1} = \begin{pmatrix} 0 & P_{\mu ABCD} \\ \overline{P}_\mu{}^{ABCD} & 0 \end{pmatrix} \tag{87}$$

which is clearly orthogonal to the SU(8) Lie algebra. The complex totally antisymmetric rank-4 tensor $P_{\mu ABCD}$ satisfies a self-duality condition and, therefore, has 70 degrees of freedom. We note that $Q_{\mu A}{}^B$ and $P_{\mu ABCD}$ are invariant under the global E_7 transformations, which act on the right of V'.

We conclude that, at least formally, $N = 8$ supergravity contains a composite SU(8) Yang Mills field $Q_{\mu A}{}^B$. However, whether or not this gauge group appears with zero mass gauge bosons is a dynamical question. We now turn to the global $E_{7(+7)}$ symmetry, which is not purely formal: it is a true, but hidden, symmetry of $N = 8$ supergravity.

As we have already seen, detailed calculations establish that the 70 scalar and pseudoscalar fields of the $N = 8$ supergravity multiplet take values in the coset space $E_{7(+7)}/\mathrm{SU}(8)$. This means that there is a way in which the group $E_{7(+7)}$ acts on the spin zero fields, but does not in itself require the action to be invariant under global $E_{7(+7)}$ transformations. In fact, direct calculations [Paper 18] show that the Lagrangian density for the spin-zero fields can be written in the form

$$L_S = \frac{e}{48\kappa^2}\mathrm{Tr}[(D_\mu V' V'^{-1})^2] \tag{88}$$

$$= \frac{e}{24\kappa^2} P_{\mu ABCD}\overline{P}_\nu{}^{ABCD} g^{\mu\nu} \tag{89}$$

where $e = \det e_\mu{}^\alpha = \sqrt{-\det g_{\mu\nu}}$ is the determinant of the vierbein, and $\kappa^2 = 4\pi G_N$ is the gravitational coupling constant. The Lagrangian density L_S is manifestly invariant under global $E_{7(+7)}$ transformations, as well as under local SU(8) transformations. The local SU(8) and global $E_{7(+7)}$ groups act non-linearly on the spin zero fields. However there is a global SU(8) symmetry acting linearly on the fields, which corresponds to combining a constant SU(8) transformation of local SU(8) with the same SU(8) transformation of $E_{7(+7)}$. When restricted to SO(8), this is the usual global SO(8) invariance of $N = 8$ supergravity.

We have restricted our above discussion of the local SU(8) and global $E_{7(+7)}$ symmetries to the spin zero sector of the theory. These symmetries are also established for the fermion and vector fields in Paper 18. The fermion fields (8 gravitinos and 56 Majorana spin-$\frac{1}{2}$ fields) are invariant under $E_{7(+7)}$, but transform convariantly under local SU(8). The 28 vector fields of the supergravity multiplet are Abelian and invariant under SU(8). The only term in the $N = 8$ supergravity Lagrangian density, which is not invariant under $E_{7(+7)}$, is the kinetic term for these spin-1 vector fields. However, the global symmetry $E_{7(+7)}$ is realised on the field equations — exchanging the equations of motion of the vector fields with their Bianchi identities. When both field equations and Bianchi identities are satisfied (on shell), there is invariance under $E_{7(+7)}$ acting linearly on the 28 vector field strengths and their 28 generalised dual tensors, which together form a fundamental 56 dimensional representation of $E_{7(+7)}$. Thus the global $E_{7(+7)}$ group is a symmetry of the equations of motion, although not of the complete action. The complete action can be made manifestly invariant under the local SU(8) symmetry.

Finally, we mention the symmetries of the alternative gauged $N = 8$ supergravity model,[13] in which the SO(8) subgroup of $E_{7(+7)}$ is gauged. This SO(8) subgroup is realised on the 28 vector fields of the supergravity multiplet, which become the SO(8) Yang-Mills fields. The global $E_{7(+7)}$ invariance of the field equations is thereby broken. Nevertheless, the $E_7/\mathrm{SU}(8)$ coset structure of the spin-zero fields remains intact, and the local SU(8) invariance is not affected by the gauging of SO(8). Consequently, the gauged $N = 8$ supergravity model has a local SO(8)×SU(8) symmetry.

Unfortunately, it is difficult to reconcile either version of $N = 8$ supergravity with the well-established standard model phenomenology. At least some of the SU(8) Yang-Mills fields $Q_{\mu A}{}^B$ must dynamically develop poles corresponding to physical gauge particles, such as the $\mathrm{W}^\pm$ bosons. Quarks and leptons must also appear as bound states of the fundamental supergravity multiplet. Attempts to obtain realistic particle spectra in this way are beset with unresolved dynamical problems.[14,15]

4.4. Kaluza-Klein Theories

In the previous section, we saw how $N = 8$ supergravity is constructed, by the dimensional reduction of a simpler eleven dimensional theory to four dimensional space-time. The extra seven dimensions were introduced there as a purely mathe-

matical device. However it is tempting to attach physical significance to the simpler theory and to the higher dimensional space. The idea of introducing extra physical dimensions at the fundamental level goes back to the 1920's and the work of Kaluza[16] and Klein.[17] They showed how electrodynamics could be obtained, by starting with Einstein's general theory of relativity in five dimensions and compactifying the extra spatial dimension into a very small circle. The extra dimension is then not directly observable, but the associated components of the metric tensor field can be interpreted as the Maxwell electromagnetic field together with a Brans-Dicke scalar field [Paper 20].

Einstein's equations, in the five dimensional Kaluza-Klein theory, contain not only the four dimensional gravity equations, but also Maxwell's equations for the electromagnetic field and the Klein-Gordon equation for the scalar field. Electromagnetic gauge invariance is a remnant of reparameterisation invariance in the extra co-ordinate. A few years ago, there was a renewed interest in generalised Kaluza-Klein theories, motivated by the hope of explaining the origin of the non-Abelian gauge invariances of the standard model in a similar way [Paper 20].

In higher dimensional Kaluza-Klein theories,[18] we start with a general relativistic theory in $4+n$ dimensions, having a metric tensor g_{AB} together with additional matter fields. The $4+n$ dimensional manifold is assumed to have the product form $M^4 \times B$, where M^4 is the usual Minkowski space and B is some compact Riemannian manifold of dimension n. More precisely, we assume the vacuum state to be $M^4 \times B$, described by the ground state metric,

$$\langle g_{AB}(x^\alpha, \phi^\kappa) \rangle = \begin{pmatrix} g_{\mu\nu}(x^\alpha) & 0 \\ 0 & \gamma_{ij}(\phi^k) \end{pmatrix} . \tag{90}$$

Here $g_{\mu\nu}(x^\alpha)$ is the ordinary metric tensor of the four-dimensional world, and $\gamma_{ij}(\phi^k)$ is the metric tensor of the internal space B having co-ordinates $\phi^k, k = 1, 2, \ldots, n$.

Symmetries of the compact internal space B, which generically has dimensions of the order of the Planck length ${G_N}^{1/2} \sim 10^{-33}$ cm, appear as gauge symmetries in the effective four-dimensional world. If the space B has a continuous group G of isometries (transformations leaving the metric tensor $\gamma_{ij}(\phi^k)$ form-invariant), we obtain a four-dimensional Yang Mills theory with gauge group G. The condition that the infinitesimal co-ordinate transformation

$$\phi_i \rightarrow \phi_i + \varepsilon K_i(\phi) \tag{91}$$

should leave $\gamma_{ij}(\phi^k)$ form-invariant leads to the Killing vector equation

$$D_i K_j + D_j K_i = 0 , \tag{92}$$

where D_i denotes the covariant derivative for the space B. For an N dimensional symmetry group G, there are N Killing vector fields $K_i^a(\phi), a = 1, 2, \ldots, N$, which obey the Lie algebra

$$\left[K_i^a \frac{\partial}{\partial \phi_i}, K_j^b \frac{\partial}{\partial \phi_j} \right] = f^{abc} K_k^c \frac{\partial}{\partial \phi^k} \tag{93}$$

where f^{abc} are the structure constants for the group G.

We now consider small fluctuations corresponding to massless excitations in the off-diagonal components of the metric tensor

$$g_{\mu j}(x^\alpha, \phi^k) = A_\mu{}^a(x^\alpha) K_j^a(\phi^k) \ . \tag{94}$$

The four dimensional fields $A_\mu{}^a(x^\alpha)$ are massless gauge fields of the gauge group G. This may be seen by noting that there is a symmetry under the infinitesimal gauge transformation

$$A_\mu^a(x) \to A_\mu^a(x) + \partial_\mu \varepsilon^a(x) + f^{abc} \varepsilon^b(x) A_\mu^c(x) \tag{95}$$

which results from the special co-ordinate transformation (see Fig. 4.5):

$$(x^\alpha, \phi_i) \to (x^\alpha, \ \phi_i + \varepsilon^{(a)}(x) K_i^a(\phi)) \ . \tag{96}$$

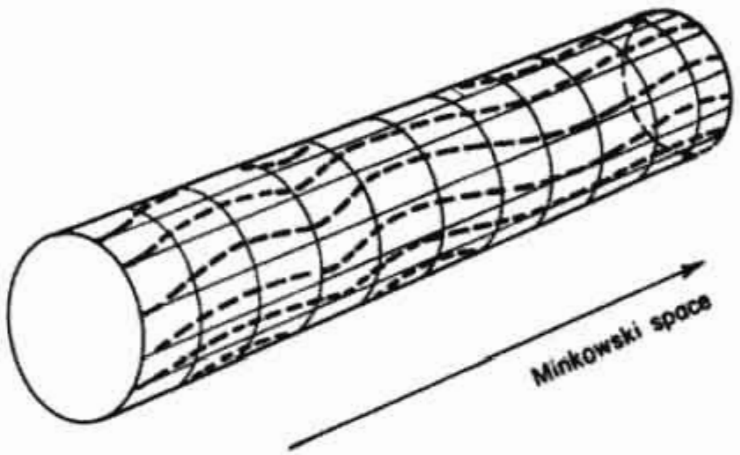


Fig. 4.5. The deformation of the co-ordinate system, Eq. (96), corresponding to a gauge transformation, Eq. (95), in a Kaluza-Klein model.

This diffeomorphism transformation leaves $\gamma_{ij}(\phi)$ and, in the linearised approximation, $g_{\mu\nu}(x)$ form invariant. In order to obtain exact form-invariance of $g_{\mu\nu}(x)$, the linear fluctuation ansatz of Eq. (94) is often generalised to include quadratic terms in $A_\mu^a(x)$ as follows:

$$g_{AB}(x, \phi) = \begin{pmatrix} g_{\mu\nu}(x) + A_\mu^a(x) A_\nu^b(x) K_i^a(\phi) K_j^b(\phi) \gamma^{ij}(\phi) & A_\mu^a(x) K_j^a(\phi) \\ A_\nu^a(x) K_i^a(\phi) & \gamma_{ij}(\phi) \end{pmatrix} \ . \tag{97}$$

Substitution of this Kaluza-Klein ansatz into the gravitational action,

$$S = -\frac{1}{4\pi\kappa^2} \int d^4x d^n\phi \sqrt{|\det g_{AB}|} R \tag{98}$$

or into some other diffeomorphism invariant action, must clearly give an action for the $A_\mu^a(x)$ fields which is invariant under the gauge symmetry of Eq. (95). The effective action, obtained after integration over the compactified co-ordinates ϕ^k, therefore contains the usual Yang Mills kinetic term

$$S_A = -\frac{1}{4g^2} \int d^4x \sqrt{|\det g_{\mu\nu}(x)|} F_{\mu\nu}{}^a F^{\mu\nu a} \tag{99}$$

for the case of a simple group; if the symmetry group of the internal space B is not simple a different coefficient $1/(4g_i^2)$ is needed for each basic invariant subgroup.

In the above discussion we neglected the effects of any matter fields. In particular, we implicitly assumed that matter fields with non-zero vacuum expectation values transform as singlets under G. The gauge symmetry group G is then not spontaneously broken by the matter fields. However the value of the gauge coupling constant g^2 and the structure of the non-linear terms in the massless excitation ansatz for g_{AB} depend on the detailed form of the matter-field interactions, as do the spin-zero massless modes.

At first sight we seem to have an example of a derived symmetry, namely Yang Mills gauge symmetry for the gauge group G of isometries for the space B. However this symmetry is simply a re-expression of the diffeomorphism symmetry under the special transformation of Eq. (96). In reality we do not have a derivation of a symmetry, but rather a metamorphosis of one already present. Even to obtain such a metamorphosis of the diffeomorphism symmetry into a gauge symmetry, it is crucial to assume that the compactified Riemannian space B has an isometry group G. Otherwise the produced gauge fields get Higgsed.

The symmetries of the space B should arise as a result of minimising the vacuum energy. The expression for the vacuum energy of a space B has a diffeomorphism symmetry, and its minimisation can easily lead to the space B being symmetric under some isometry group G. It is really a question of whether or not there is a spontaneous breakdown of the diffeomorphism symmetry. In practice the form of the compactified space B and the consequent continuous isometry group G are usually just assumed, which is tantamount to putting in the gauge symmetry by hand.

As illustrations of the Kaluza-Klein mechanism for obtaining gauge symmetries, we briefly reconsider the two versions of $N = 8$ supergravity discussed in the previous section as different compactifications of eleven dimensional supergravity.[18] The Cremmer-Julia dimensional reduction corresponds to taking the seven dimensional compact space B to be the seven-torus T^7, with isometry group $U(1)^7$. There are thus seven Kaluza-Klein U(1) gauge bosons in the Cremmer-Julia $N = 8$ supergravity. In addition 21 Abelian spin-1 gauge bosons arise from the components $A_{\mu mn}$ of the third rank antisymmetric tensor present in supergravity. The deWit-Nicolai gauged SO(8) version of $N = 8$ supergravity corresponds to taking B to be the seven sphere S^7, which has SO(8) as its isometry group. In this case all 28 of the SO(8) gauge bosons arise as Kaluza-Klein gauge fields, and the four index gauge invariant curl of the antisymmetric tensor A_{NPQ}

$$F_{MNPQ} = \partial_{[M} A_{NPQ]} \tag{100}$$

develops a non-zero vacuum expectation value, in a necessarily four dimensional space. The existence of the four index field strength F_{MNPQ} here provides a dynamical reason for the compactification to four space-time dimensions.[19] The potential spin-1 gauge bosons of the hidden local SU(8) symmetry of $N = 8$ supergravity

would be dynamical bound states of the spin-0 fields and not Kaluza-Klein gauge fields.

There are many other possible compactifications of eleven dimensional supergravity, with some or all of the $N = 8$ supersymmetries broken in four dimensions. In Paper 20, compactifications to a space B having SU(3)×SU(2)×U(1) as its isometry group are considered. However the fermion spectra obtained from these models are vectorlike, in contradiction to the complex representations of the standard model. This is a general problem for Kaluza-Klein theories, in which all the gauge fields arise as components of the metric tensor in higher dimensions.[20] A theorem due to Atiyah and Hirzebruch shows that, in the absence of elementary gauge fields, the Dirac equation in $4 + n$ dimensions always leads to non-chiral fermion quantum numbers in 4 dimensions. There is a similar chirality problem for the Rarita-Schwinger field.

In order to circumvent the chiral fermion problem in Kaluza-Klein theory, both (i) chiral fermions and (ii) explicit gauge fields must be present already in $4 + n$ dimensions. The extra number of dimensions n is required to be even, since it is not possible to define a generalised γ_5 Dirac matrix, and thus chirality, in an odd number of dimensions. Also to ensure the survival of chirality on compactification to 4-dimensions, the gauge fields must assume a topologically non-trivial structure on the compactified space B.

As a simple example[21] with $n = 2$, the space B could be an S_2 sphere having a constant magnetic field F_{56} on it, corresponding to a Dirac monopole in the embedding space for the S_2 sphere. The integral of the magnetic field over the S_2 sphere is, apart from a normalisation factor, the Pontryagin index for the two-dimensional space. The Atiyah-Singer index theorem then ensures that a fermion field, chiral in 6 dimensions, will have one or more zero modes of the Dirac operator for the two-dimensional space B. These zero modes are observed as massless chiral fermions in four dimensions. Here the natural mass scale is given by the inverse of the radius of the S_2 sphere, which is generically of the order of the Planck mass ${G_N}^{-1/2} \sim 10^{19}$GeV. The 'massless fermions' can obtain small masses from low energy gauge symmetry breaking, as in the standard model.

Inclusion of explicit gauge fields in the higher dimensions is fundamentally different in spirit to the Kaluza-Klein idea. The original idea of getting gauge fields from higher dimensional gravity is given up, and the gauge fields are put in by hand.

The presence of chiral fermions coupled to gauge fields in $4+n$ dimensions means that the problem of anomalies must be faced.

4.5. Anomaly Cancellation

In higher dimensional theories with an even number $D = 2m$ dimensions, it is possible to have a parity violating gauge theory, in which the left-handed fermions transform differently under the gauge group from the way the right-handed fermions transform. Fermion loop diagrams with $m + 1$ external gauge particles are potentially anomalous in such a chiral theory.[22] Anomalies thus arise from triangle dia-

grams, such as Fig. 3.1, in $D = 4$ dimensions and from hexagon diagrams, Fig. 4.6, in $D = 10$ dimensions. These quantum anomalies violate gauge invariance and unitarity of the S-matrix. Consistency then requires the cancellation of all anomalies.

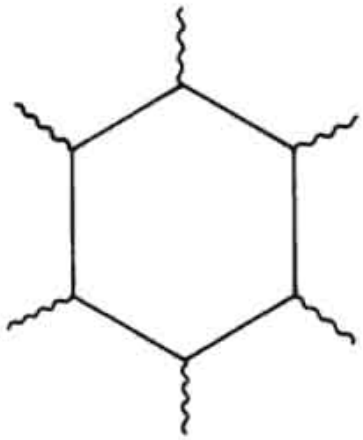


Fig. 4.6. Hexagon diagram with external gauge bosons or gravitons, which can give gauge, gravitational or mixed anomalies in ten dimensions.

Furthermore, in $D = 4k+2$ dimensions, the particle and antiparticle associated with a Weyl field have the same helicity or chirality. This is in contrast to the situation in $D = 4k$ dimensions, where the helicities of particle and antiparticle are opposite. The gravitational coupling of a Weyl fermion in $4k+2$ dimensions is therefore chiral, and we have the possibility of gravitational anomalies.[23] Gravitational anomalies violate diffeomorphism symmetry and lead to non-conservation of the energy-momentum tensor. In addition to purely gauge and purely gravitational anomalies, there is the possibility of mixed anomalies, arising from fermion loop diagrams with both gauge particles and gravitons as external lines.

Higher dimensional chiral theories in $4k+2$ dimensions are severely constrained, by the requirement of gravitational and gauge anomaly cancellation.[23,24] This is illustrated by $N = 1$ supergravity in $D = 10$ dimensions, which is the highest dimensionality consistent with a chiral supergravity theory having no particles with spin greater than 2. The pure $D = 10$ supergravity multiplet consists of the massless particles listed in Table 4.1. The numbers of physical degrees of freedom associated with each of the massless fields are indicated in brackets. The gravitational anomalies, due to the chiral gravitino ψ_μ and spinor λ fields in the supergravity multiplet, do not cancel. Thus pure $N = 1$ supergravity, without matter fields, is not consistent quantum mechanically in 10 dimensions. Local Lorentz invariance and hence energy momentum conservation are violated. However the gravitational anomalies can be cancelled [Paper 21] by including as matter fields, a supersymmetric Yang-Mills multiplet, Table 4.2, *provided* the gauge group G is SO(32) or $E_8 \times E_8$. Moreover the gauge and mixed anomalies can also be cancelled for these two groups.

Table 4.1. The $N = 1$ supergravity multiplet in 10 dimensions.

Field (Degrees of Freedom)	Description
$g_{\mu\nu}(35)$	graviton
$\psi_\mu(56)$	left-handed Majorana-Weyl gravitino
$B_{\mu\nu}(28)$	antisymmetric tensor
$\lambda(8)$	right-handed Majorana-Weyl spinor
$\phi(1)$	scalar

Table 4.2. Supersymmetric Yang-Mills multiplet in 10 dimensions for a gauge group G of dimension n.

Field (Degrees of freedom)	Descriptions
$A_\mu{}^a(8n)$	gauge particle
$\chi^a(8n)$	left-handed Majorana-Weyl spinors in adjoint representation of G

The requirement of a non-trivial cancellation of anomalies, between the supergravity and super Yang-Mills multiplets in 10 dimensions, thus determines the gauge group to be SO(32) or $E_8 \times E_8$ (apart from the relatively trivial cases of $\mathrm{U}(1)^{496}$ and $E_8 \times \mathrm{U}(1)^{248}$). In a sense, therefore, the gauge group G can be said to be derived and it is for this reason that Paper 21 is included in this book. However it is more correct to say that the gauge group is selected, rather than derived, by the anomaly cancellation requirement.

Let us now examine how the anomaly cancellation requirement specifies the super Yang Mills group [Paper 21]. The anomaly is defined as the variation $\delta_\lambda \Gamma$ of the one loop fermion effective action

$$\Gamma = \mathrm{tr}(\log \not{D}) \tag{101}$$

under an infinitesimal gauge or general co-ordinate transformation $\lambda(x)$. Here $\not{D}$ is the inverse propagator for a Weyl fermion, in a background of both Yang Mills and gravitational fields. The variation of a formally gauge invariant action Γ in D dimensions must be an integral, local in $\lambda(x)$, of the form

$$\delta_\lambda \Gamma = \int \lambda(x) a_D \ . \tag{102}$$

Here a_D is a D form, as is required to make the integral sensible.

It is very convenient to use the language of differential forms to construct the anomaly D form λa_D, in terms of a formal gauge invariant $D+2$ form I_{D+2}. The construction is given by the descent equations

$$I_{D+2} = dI_{D+1} \tag{103}$$

and

$$\delta_\lambda I_{D+1} = d(\lambda a_D) \ . \tag{104}$$

This construction is ambiguous up to a closed form, but the integral of Eq. (102) specifies a unique anomaly corresponding to I_{D+2}.

The gravitational, gauge and mixed anomalies in $D = 10$ super Yang Mills plus supergravity theory can be constructed from a 12 form, which is proportional to

$$\begin{aligned} I_{12} = & -\frac{1}{15}\mathrm{Tr}\, F^6 + \frac{1}{24}\mathrm{Tr}\, F^4 \mathrm{tr}\, R^2 \\ & - \frac{1}{960}\mathrm{Tr}\, F^2 [4\mathrm{tr}\, R^4 + 5(\mathrm{tr}\, R^2)^2] \\ & + (n - 496)\left[\frac{\mathrm{tr}\, R^6}{7560} + \frac{\mathrm{tr}\, R^2 \mathrm{tr}\, R^4}{5760} + \frac{(\mathrm{tr}\, R^2)^3}{13824}\right] \\ & + \frac{1}{8}\mathrm{tr}\, R^2\ \mathrm{tr}\, R^4 + \frac{1}{32}(\mathrm{tr}\, R^2)^3 . \end{aligned} \tag{105}$$

Here F is the gauge field strength two form

$$F = \frac{1}{2} F_{\mu\nu} dx^\mu dx^\nu \tag{106}$$

in the adjoint representation of the gauge algebra, and R is the curvature two form

$$R = \frac{1}{2} R_{\mu\nu} dx^\mu dx^\nu \ . \tag{107}$$

The Riemann curvature tensor $R_{\mu\nu}$ is expressed in the zehnbein formalism, with two suppressed flat indices. Thus the two form R is a 10×10 matrix, belonging to the fundamental representation of the Lorentz algebra SO(9,1). In the above formula we suppress the wedge product symbol so that, for example, F^2 denotes $F \wedge F$. The symbols Tr and tr are used to denote traces in adjoint and fundamental representations respectively. The number n of matter Weyl fermion species is equal to the dimension of the Yang Mills gauge group G.

The anomalies associated with I_{12} are clearly non-vanishing for any gauge group G. However Green and Schwarz pointed out the existence of a non gauge invariant local counterterm that cancels these anomalies, whenever I_{12} factorises into an expression of the form

$$I_{12} = (\mathrm{tr}\, R^2 + k\mathrm{Tr}\, F^2) X_8 \tag{108}$$

where k is a constant and X_8 is a gauge invariant eight form made out of F and R. The two necessary and sufficient conditions for this factorisation to be possible are

$$n = \dim G = 496 \tag{109}$$

and

$$\mathrm{Tr}\, F^6 = \frac{1}{48}\mathrm{Tr}\, F^4 \mathrm{Tr}\, F^2 - \frac{1}{14400}(\mathrm{Tr}\, F^2)^3 \tag{110}$$

with k uniquely determined to be

$$k = -\frac{1}{30} \ . \tag{111}$$

It is these two conditions which characterise the allowed groups SO(32) and $E_8 \times E_8$. In addition there are just two other groups, $U(1)^{496}$ and $E_8 \times U(1)^{248}$, which satisfy the conditions rather trivially.

When the I_{12} factorisation condition is satisfied, the key to the anomaly cancellation lies in the addition to the action of a higher derivative counterterm of the form

$$S_c \sim \int B X_8 \ . \tag{112}$$

Here B is the two form

$$B = B_{\mu\nu} dx^\mu dx^\nu \tag{113}$$

and $B_{\mu\nu}$ is the antisymmetric tensor field belonging to the supergravity multiplet, which is required to transform non trivially under both local Lorentz and Yang Mills gauge transformations

$$\delta B = \omega_{2L}^1 - \frac{1}{30}\omega_{2Y}^1 \ . \tag{114}$$

The two forms ω_{2L}^1 and ω_{2Y}^1 are obtained by descent from $\mathrm{tr}\, R^2$ and $\mathrm{Tr}\, F^2$ respectively. The superscript 1 indicates that they are linear in the infinitesimal gauge parameters $\lambda(x)$. They are constructed by applying infinitesimal transformations to the Lorentz and Yang-Mills Chern-Simons three forms ω_{3L} and ω_{3Y}

$$\delta\omega_{3L} = d\omega_{2L}^1 \tag{115}$$

$$\delta\omega_{3Y} = d\omega_{2Y}^1 \ . \tag{116}$$

The Chern-Simons three forms themselves satisfy

$$d\omega_{3L} = \mathrm{tr}\, R^2 \tag{117}$$

$$d\omega_{3Y} = \mathrm{Tr}\, F^2 \ . \tag{118}$$

For example, the Yang-Mills Chern-Simons form is given explicitly by

$$\omega_{3Y} = \mathrm{Tr}\left(AF - \frac{1}{3}A^3\right) = \mathrm{Tr}\left(AdA + \frac{2}{3}A^3\right) \ . \tag{119}$$

The field strength $H_{\mu\nu\rho}$ associated with $B_{\mu\nu}$ is made invariant under local Lorentz and Yang-Mills gauge transformations, by the definition

$$H = dB + \omega_{3L} - \frac{1}{30}\omega_{3Y} \tag{120}$$

of the three form field strength H.

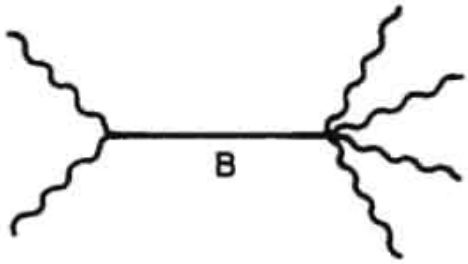


Fig. 4.7. Anomalous B exchange diagram, which cancels the hexagon anomaly.

The basic point in the Green-Schwarz mechanism is that the new terms introduced into the action generate anomalous B exchange diagrams, Fig. 4.7, which just cancel the anomaly from hexagon diagrams, Fig. 4.6.

It is thus the Green-Schwarz factorisation condition, Eq. (108), for anomaly cancellation which selects SO(32) and $E_8 \times E_8$ as the allowed gravity-coupled super Yang-Mills gauge groups in 10 dimensions. However what is the motivation for considering $N = 1$ supergravity in 10 dimensions? It is possible to have non-trivial anomaly cancellations between particles of different spin in higher than 10 dimensions,[26,27] but these theories suffer from inconsistencies. They either contain a massless spin $\frac{3}{2}$ particle (a gravitino) without being supersymmetric, or massless particles with spin greater than 2. Chiral supergravity theories in (even) dimensions higher than 10 have more than $N = 8$ supersymmetry generators when reduced to 4 dimensions, as is easily verified by counting the number of spinor components for the gravitino. The graviton multiplet for chiral $N = 1$ supergravity therefore contains massless particles with spin greater than 2, unless the number of dimensions $D \leq 10$.

So the maximum number of dimensions in which it is possible to construct a consistent $N = 1$ supergravity model with chiral fermions is $D = 10$. However in itself this is not a convincing reason for selecting a ten dimensional model. In fact, at present, the main interest in $N = 1, D = 10$ supergravity coupled to super Yang-Mills theory lies in its identification as the low energy effective field theory limit of a consistent superstring theory. Indeed the anomalous B exchange diagram, Fig. 4.6, and the associated higher derivative counterterms, discussed in Paper 21, were suggested by the low energy expansion of type I superstring theory. The critical dimension, $D = 10$, of superstring theory is determined by the requirement of invariance under infinitesimal reparameterisations of the two-dimensional world sheet of the string (conformal invariance).[25]

In string language the above anomaly analysis means that the unoriented closed superstring,[(25)] which has the $N = 1, D = 10$ supergravity multiplet of Table 4.1 as its spectrum of massless states, suffers from anomalies. For consistency it is necessary to introduce an open superstring, with Chan Paton factors[(25)] corresponding to the unique gauge group SO(32). The massless states of the open string correspond to the Yang-Mills supermultiplet of Table 4.2. Thus we have the remarkable result that anomaly cancellation, which really means quantum mechanical consistency, implies that once we consider non-orientable closed strings then open strings are forced to exist also.

The realisation of $E_8 \times E_8$, as a string theory gauge group, required the con-

struction of the oriented closed heterotic string [Paper 22]. It is just the Yang-Mills gauge groups SO(32) or $E_8 \times E_8$, which uniquely emerge as required, but rather hidden, symmetries of the $D = 10$ heterotic string with $N = 1$ spacetime supersymmetry. The origin of these specific groups is discussed in the next section. The crucial ingredient, which guarantees the anomaly factorisation and cancellation by the Green Schwarz mechanisms, is the invariance of the two-dimensional string world sheet under co-ordinate transformations not continuously connected to the identity (modular invariance).[27]

4.6. Strings

Superstring theory[25] has recently aroused great interest, as a candidate theory for unifying all fundamental forces in a 'theory of everything'. In particular quantum string theories require gravity in most cases.[28] It also appears, under very restrictive conditions though, that they could be consistent and finite. This is in contrast to quantum point particle theories of gravity, such as supergravity models, which all have non-renormalisable infinities.

In the critical space-time dimension $D = 10$, there is a near uniqueness of the superstring theory. There are just three models with $N = 1$ space-time supersymmetry and a Yang Mills gauge group: the SO(32) type I superstring and the $E_8 \times E_8$ and SO(32) heterotic strings. The gauge groups, SO(32) and $E_8 \times E_8$, are the same as those suggested by the anomaly cancellation requirement from ten dimensional $N = 1$ supergravity. Of these models, the $E_8 \times E_8$ heterotic string seems the most promising phenomenologically.

On reduction to four space-time dimensions, however, the uniqueness of the heterotic string essentially disappears.[29] Generalised 'compactifications' to four dimensions can be made, which treat the co-ordinates for left-moving and right-moving modes on the closed string differently. It is then difficult to interpret these co-ordinates conventionally, as describing six extra space dimensions compactified on a scale of the order of the Planck length. It is more appropriate to formulate such models directly in 4 space-time dimensions, and to interpret the remaining string world-sheet degrees of freedom (which are required to avoid an anomaly in conformal invariance) as internal degrees of freedom. There are very many four dimensional superstring models, having gauge groups of rank 22 or less. These may be constructed, in a fermionic formulation[29], by modifying the periodicity conditions (boundary conditions in the case of open strings). By taking combinations of different sets of boundary (periodicity) conditions, in the two string co-ordinates σ and τ, it is possible to construct many modular invariant and, thereby, consistent models.

In this section we consider (1) the Frenkel-Kac-Segal mechanism responsible for the origin of a non-Abelian symmetry in the heterotic string, and (2) the general appearance of gauge symmetry in string theories.

4.6.1. The Heterotic String and Kac-Moody Algebras

Here we shall discuss the by now classical example of the 10-dimensional heterotic string [Paper 22]. Its characteristic property is that it has 16 bosonic fields $X^I(\sigma,\tau), I = 1,2,\ldots 16$ defined over a closed orientable string and taking values on a certain somewhat skew torus T^{16}. The co-ordinates σ and τ parameterise the $1+1$ dimensional world sheet of the string and the 16 fields $X^I(\sigma,\tau)$ are thus described by a 2-dimensional field theory. These fields X^I are peculiar in having, when quantised, a symmetry under either $E_8 \times E_8$ or $\mathrm{Spin}(32)/Z_2$ (which has the same Lie algebra as SO(32)). The classical Kaluza-Klein symmetry, obtained from such a toroidal compactification, is just the Abelian subgroup $\mathrm{U}(1)^{16}$; the isometry group of the torus T^{16}.

An important ingredient in obtaining the symmetries $E_8 \times E_8$ or Spin $(32)/Z_2$ is the assumption that the fields $X^I(\sigma,\tau)$ are purely left-moving. That is to say that the fields X^I are assumed to be only functions of $\tau+\sigma$, where τ is the time co-ordinate and σ the space co-ordinate, in the $1+1$ dimensional field theory. Thus any disturbance in these fields moves with the velocity of light to the left and is massless, in a 2-dimensional world sheet sense, when viewed as a quasiparticle on the string.

Now a torus-valued field theory has the possibility of having the fields wound up around the torus. These 'topological' winding modes of the fields X^I are known as solitons. They exist in addition to the usual oscillator modes or phonons on the string (Fig. 4.8).

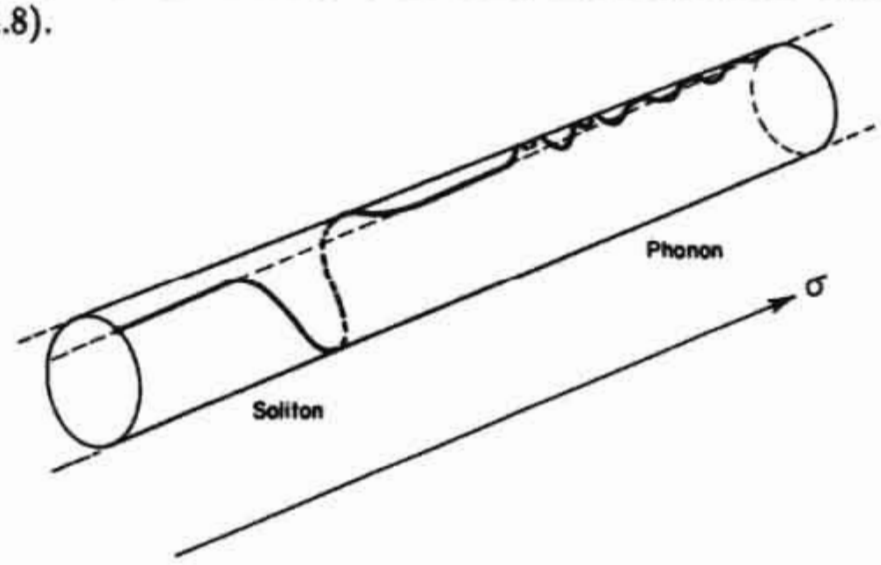


Fig. 4.8. Illustration of the difference between a soliton and a phonon in an S_1 sigma model.

The torus-valued fields X^I can take as values all real numbers, where we then identify sets of X^I values congruent modulo a lattice Λ in the R^{16} space.[30] This means we regard the torus as a quotient space $T^{16} = R^{16}/\Lambda$. It is convenient to write $\Lambda = \pi l\Gamma$, where the fundamental string length scale $l = (\pi T_0)^{-1/2}$ will henceforth be set equal to unity. We thus choose units in which the string tension $T_0 = 1/2\pi\alpha' = 1/\pi$.

From the requirement that the disturbances in the X^I variables corresponding to soliton states be purely left-moving, it follows that the soliton momenta P^I should

lie on the lattice Γ. Quantum mechanically, the soliton momenta P^I must also lie on the dual lattice $\tilde{\Gamma}$, since the string has a single-valued wave function on the torus. Physical states are constructed by combining right- and left-moving oscillator and soliton modes, which satisfy the condition

$$N = \tilde{N} - 1 + \frac{1}{2}\sum_I (P^I)^2 \ , \tag{121}$$

and the ten-dimensional mass operator is

$$(\text{mass})^2 = 8N \ . \tag{122}$$

Here N and $\tilde{N}$ are the normal ordered number operators for the right- and left-moving oscillator modes. Since N and $\tilde{N}$ are integer-valued, it follows that the squares of the allowed soliton momenta $\sum_I (P^I)^2$ must be even integers. Finally the consistency of string theory requires that the lattice be self-dual $\Gamma = \tilde{\Gamma}$, otherwise modular invariance, and thus unitarity, breaks down. This means that one-loop amplitudes develop anomalies, which destroy global reparameterisation invariance of the string world sheet, unless $\Gamma = \tilde{\Gamma}$.

The above requirements can only be satisfied if the lattice Γ is even and self-dual. A lattice is said to be even if the length squared of any lattice vector is an even integer. There are just two even, self-dual, 16-dimensional lattices, $\Gamma_8 \times \Gamma_8$ and Γ_{16}. The first possibility is the lattice of roots for the group $E_8 \times E_8$; the second is a certain subset of the possible weights of the Lie group Spin(32), in fact the weight lattice of the factor group $\text{Spin}(32)/Z_2$. So remarkably the same groups, $E_8 \times E_8$ and $\text{Spin}(32)/Z_2$, are selected as those consistent with anomaly cancellation in 10 dimensions. The crucial requirement in selecting these two groups, rather than any other simply-laced Lie group of rank 16, is modular invariance. As mentioned in section 4.5, it is also modular invariance which guarantees anomaly cancellation, by the Green-Schwarz mechanism, for the heterotic string.[27]

However, *a priori*, the string running on the torus T^{16}, defined by the lattice $\Gamma_8 \times \Gamma_8$ or Γ_{16}, has no symmetry under the groups $E_8 \times E_8$ or $\text{Spin}(32)/Z_2$. The X^I degrees of freedom correspond to the Cartan sub-algebra of the Lie group, but there are no fields corresponding to the other directions in the Lie algebra. Nonetheless it turns out — and here we have an example of a hidden symmetry — that the quantised system of the X^I variables, running on the torus T^{16}, indeed does have the full symmetry under the group $E_8 \times E_8$ or $\text{Spin}(32)/Z_2$. The groups are represented, in this $1+1$ dimensional field theory, by transforming X^I-phonons and X^I-solitons into each other (Fig. 4.8). It is, of course, a prerequisite for such a symmetry to occur that the two-dimensional world sheet energy and momentum should be the same for the phonons and solitons. Classically this is not possible, since then the soliton energy is quantised but the phonon energy is not. Quantum mechanically, the equality is arranged by a fine-tuning of the radii and angles of the torus. The sixteen X^I-phonons with $\tilde{N} = 1$ can then lie in the same multiplet as the

480 solitons with $\sum_I (P^I)^2 = 2$. Nevertheless it still seems somewhat miraculous that such fine-tunings are sufficient to make the quantum system invariant under the group $E_8 \times E_8$ or $\mathrm{Spin}(32)/Z_2$.

We will now demonstrate the presence of the symmetry by writing down the group generators as quantum operators,[30] using the Frenkel-Kac-Segal construction.[31-34] These generators commute with the mass squared operator and, thus, the physical string states provide a representation of the symmetry group at each mass level.

The commuting generators H^I, of the Cartan sub-algebra $\mathrm{U}(1)^{16}$, are simply represented by the momentum operators along the torus directions

$$H^I = P^I \ . \tag{123}$$

The basis generators E^K, not belonging to the Cartan sub-algebra, are each represented by an integral over a normal-ordered vertex operator

$$V(K,z) =: \exp[2i \sum_I K^I X^I(z)] : \tag{124}$$

where K^I is a root vector, of length squared two, and

$$z = \exp[2i(\sigma + \tau)] \ . \tag{125}$$

The operator E^K is required to create (or annihilate) a soliton, characterised by a string lying from the identity element to the lattice point K^I and, hence, also by momentum K^I.

The generators E^K are actually taken to be

$$E^K = A_K C_K(P) \tag{126}$$

where

$$A_K = \frac{1}{\pi} \int_0^\pi V(K,z) d\tau \tag{127}$$

$$= \oint \frac{dz}{2\pi i z} V(K,z) \tag{128}$$

and $C_K(P)$ is a sign factor. The factors $C_K(P)$ are functions of the 16-momentum operators, which must be constructed to obey the equation

$$C_K(P - K') C_{K'}(P) = (-1)^{K \cdot K'} C_{K'}(P - K) C_K(P) \tag{129}$$

in order that the generators E^K have simple commutation rules. The scalar product $K \cdot K'$ is of course defined to be

$$K \cdot K' = \sum_{I=1}^{16} K^I K'^I \ . \tag{130}$$

In order to close the algebra and obtain the correct commutation rules, we also require

$$C_K(P-K)C_{K'}(P) = \varepsilon(K,K')C_{K+K'}(P) \tag{131}$$

where $\varepsilon(K,K') = \pm 1$.

There are many solutions to Eq. (129)-(131). A particular explicit solution[25] is expressed in terms of an ordered * product, which is defined by choosing some definite ordering of the 16 basis vectors $e_i^I, i = 1,2,\ldots 16, e_i \cdot e_i = 2$, of the lattice $\Gamma_8 \times \Gamma_8$ or Γ_{16}. Any two lattice vectors K and K' can be expanded in terms of e_i with integer coefficients

$$K^I = \sum_{i=1}^{16} m_i e_i^I \tag{132}$$

$$K'^I = \sum_{i=1}^{16} m_i' e_i^I \tag{133}$$

and the * product is defined by

$$K * K' = \sum_{i>j} m_i m_j e_i \cdot e_j \ . \tag{134}$$

Then the sign factors

$$C_K(P) = (-1)^{P*K} \tag{135}$$

satisfy Eq. (129)-(131) for

$$\varepsilon(K,K') = (-1)^{K*K'} \ . \tag{136}$$

With the definitions Eq. (123) and (126) of the generators, the Lie algebra commutation relations are readily verified. In fact Frenkel and Kac[31] and Segal[32] constructed not just the Lie algebra but a larger infinite-dimensional Kac-Moody algebra. The Kac-Moody algebra, in string theory, acts as a partial spectrum generating algebra, linking states of different mass. The extra generators E_n^K and H_n^I, of the affine $E_8 \times E_8$ or affine SO(32) algebras, are constructed by introducing a factor z^n, for integer n, under the integral signs in the expressions, Eq. (126) and (128), defining E^K, and in a similar expression for H^I as an integral over $\frac{dX^I}{dz}$.

By the above construction of the generators, we have demonstrated the global $E_8 \times E_8$ or $\mathrm{Spin}(32)/Z_2$ symmetry of the ten-dimensional heterotic string. Among the physical zero mass, $N = 0$, string states are vector particles belonging to the adjoint representation of $E_8 \times E_8$ or $\mathrm{Spin}(32)/Z_2$. These string states are gauge particles, in the $9+1$ dimensional space-time, and would thus signify the presence of a gauge symmetry, if the string theory was written as an infinite component field theory.

The gauge particle string states are of the form

$$|i\rangle_R \times \tilde{\alpha}^I_{-1}|0\rangle_L \tag{137}$$

for the neutral vector bosons, arising from the Cartan sub-algebra, and

$$|i\rangle_R \times |P^I;\ \sum_I (P^I)^2 = 2\rangle \tag{138}$$

for the other 480 charged vector bosons. Here $|i\rangle_R$ are the ten-dimensional spin 1 ground states of the right-handed sector, which are singlets under the symmetry group $E_8 \times E_8$ or $\mathrm{Spin}(32)/Z_2$. The left-moving sector X^I-phonon modes $\tilde{\alpha}^I_{-1}|0\rangle_L$, having $\tilde{N} = 1$, and the X^I-soliton modes $|P^I; \Sigma_I (P^I)^2 = 2\rangle$ together make up the adjoint representation of the group $E_8 \times E_8$ or $\mathrm{Spin}(32)/Z_2$. These left-moving modes are scalars in 10-space.

This leads us into a general discussion of gauge symmetries resulting from string theory.

4.6.2. Gauge Symmetry from Strings

An interesting example of a 'derived' symmetry, or rather the metamorphosis of one symmetry into another one, is the appearance of gauge particles and thus of gauge symmetry in string theory. For simplicity, we shall restrict our explicit discussion to Abelian gauge symmetry, although it is true that Yang-Mills gauge symmetry appears in open string models with Chan-Paton factors and in the closed heterotic string, where the charges are distributed along the length of the string.

The above gauge symmetry was in fact a problem for the original dual models of hadronic physics. The dual model was shown to be consistent with unitarity only in $D = 26$ dimensions and when it contained massless particles having spin 1.

Since gauge symmetry is really a formal symmetry, it will naturally depend upon the formalism whether or not a second quantised string theory will exhibit gauge invariance. For instance, writing string theory in the light-cone formalism, with only physical degrees of freedom,[35] leads to a formalism without gauge symmetry. This, of course, may just mean that, in some way, the gauge has thereby been fixed. In fact, even in the light-cone formalism, there are massless vector bosons with only $D-2$ components, rather than the $D-1$ components of a massive vector boson. In this sense, there is the physical imprint of a gauge symmetry.

When the low energy or zero Regge slope $\alpha' \to 0$ limit is taken, string theories can be described by an effective quantum field theory, which possesses Yang-Mills gauge invariance. In this sense, we can say that gauge symmetry is contained in string theory. What now is the origin of this gauge symmetry; is it put directly into the theory or does it appear dynamically?

It has been known for a long time that the gauge symmetry in string theory is connected with 2-dimensional diffeomorphism symmetry; more precisely with the subset of conformal transformations generated by the Virasoro algebra. In order

to see this connection, it is convenient to choose a parameterisation of the string world sheet, in which the world sheet metric is the flat 2-dimensional Minkowski metric. The string action is still invariant under conformal transformations in this parameterisation, which is characterised classically by the conditions

$$\left(\frac{\partial X^\mu}{\partial \tau} \pm \frac{\partial X^\mu}{\partial \sigma}\right)^2 = 0 \,. \tag{139}$$

The quantum creation operators $\alpha_n^\mu, n < 0$, for the oscillator modes, i.e. the phonons, on the open bosonic 26-dimensional string, are then defined by the Fourier expansion of the position variables in the 26-dimensional Minkowski space-time manifold

$$X^\mu(\sigma,\tau) = q^\mu + \alpha_0^\mu \tau + i\sum_{n\neq 0} \frac{1}{n}\alpha_n^\mu e^{-in\tau}\cos n\sigma \,. \tag{140}$$

Here q^μ and $\alpha_0^\mu = p^\mu$ are the centre of mass position and momentum of the string.

Physical states $|\psi\rangle$ in the string theory formalism must obey the conditions

$$L_n|\psi\rangle = 0 \quad n = 1,2,\ldots \tag{141}$$

and

$$(L_0 - 1)|\psi\rangle = 0 \,. \tag{142}$$

The Virasoro operators L_n are defined as the Fourier components of the two-dimensional energy-momentum tensor $T_{\alpha\beta}$ and are given by the equations

$$L_n = \frac{1}{2}\sum_{m=-\infty}^{\infty} \alpha_{m\mu}\alpha_{n-m}^\mu \,. \tag{143}$$

The massless gauge boson, which we shall call the photon, is described by the string states

$$|\varepsilon\rangle = \varepsilon_\mu \alpha_1^{\mu+}|0\rangle \tag{144}$$

$$= \varepsilon_\mu \alpha_{-1}^\mu |0\rangle \,. \tag{145}$$

Here ε_μ is an arbitrary polarisation vector, and we consider string states having a definite centre of mass momentum k_μ. The photon states must satisfy the non-trivial physical condition

$$L_1|\varepsilon\rangle = 0 \tag{146}$$

where

$$L_1 = \alpha_1^\mu p_\mu + \sum_{n=1}^{\infty} \alpha_{n+1}^\mu \alpha_{-n\mu} \,. \tag{147}$$

This condition implies that the polarisation vector satisfies the equation

$$k^\mu \varepsilon_\mu = 0 \ . \tag{148}$$

What then, in the string theory, corresponds to the gauge transformation

$$A_\mu(x) \longrightarrow A_\mu(x) + \partial_\mu \Lambda(x) \tag{149}$$

of the photon field? For a photon wave function

$$A_\mu(x) = \varepsilon_\mu e^{ik\cdot x} \tag{150}$$

the gauge transformation of Eq. (149) becomes

$$\varepsilon_\mu \rightarrow \varepsilon_\mu + i\lambda k_\mu \tag{151}$$

where $\lambda = \lambda(k)$ is the appropriate Fourier component of $\Lambda(x)$. In string language, this means the addition of the state

$$i\lambda k_\mu \alpha^\mu_{-1}|0\rangle = i\lambda L_{-1}|0\rangle \tag{152}$$

to the photon state $|\varepsilon\rangle$ of Eq. (145). The state $i\lambda L_{-1}|0\rangle$ represents the change in the state $|0\rangle$, under a reparameterisation of the co-ordinates (σ, τ) of the string world-sheet. It is an example of a null physical state, being orthogonal to all physical string states including itself. In fact for any physical state $|\psi\rangle$, satisfying Eq. (141), we have

$$(i\lambda L_{-1}|0\rangle)^+|\psi\rangle = \langle 0| - i\lambda^* L_1|\psi\rangle = 0 \ . \tag{153}$$

Thus the string states $|\varepsilon\rangle$ and $|\varepsilon + i\lambda k\rangle = |\varepsilon\rangle + i\lambda L_{-1}|0\rangle$ have the same physical significance. The unphysical nature of a contribution to ε_μ of the form $i\lambda k_\mu$ and Eq. (148) together ensure the disappearance, as physical states, of both time-like and longitudinally polarised vector particles, just as in quantum electrodynamics. It follows that the gauge symmetry of the photon, in string theory, is just part of the reparameterisation symmetry of the string formalism.

We should also emphasize here that the diffeomorphism symmetry of general relativity arises, in string theory, in a similar way to gauge symmetry. There is a massless spin 2 state in the closed string sector. The existence of zero norm states, which decouple from the S-matrix, ensures general covariance and the identification of this string state with the graviton.

Finally we mention an *a priori* different way of obtaining gauge particles in string theory, proposed by Virasoro.[36] He considered a model having a condensate of closed strings in the vacuum.

References

1. H. Harari, Fundamental Forces, *Proceedings of the 27th Scottish Universities Summer School in Physics*, St. Andrews, p. 357 (SUSSP Publications, Edinburgh University Press, 1985), M. Duff, *Proceedings of the International Europhysics Conference on High Energy Physics*, Bari, 1985, p. 679 (European Physical Society, 1986).
2. G. 't Hooft, *Recent Developments in Gauge Theories*, p. 135 (Plenum Press, 1980).
3. P. Langacker, *Phys. Rep.* **72C**, 185 (1981); P. Langacker, *Proceedings of 1985 International Symposium on Lepton and Photon Interactions at High Energies*, Kyoto, p. 186, 1986; P. Langacker, *Ninth Workshop on Grand Unification*, Aix-Les-Bains, 1988, p. 3 (World Scientific, 1988).
4. G. G. Ross, *Grand Unified Theories* (Benjamin/Cummings, 1985).
5. F. Wilczek and A. Zee, *Phys. Lett.* **88B**, 311 (1979).
6. G. 't Hooft, *Nucl. Phys.* **B35**, 49 (1971).
7. W. A. Bardeen, A. J. Buras, D. W. Duke and T. Muta, *Phys. Rev.* **D18**, 3998 (1978).
8. Y. Totsuka, *Proceedings of 1985 International Symposium on Lepton and Photon Interactions at High Energies*, Kyoto, p. 120, 1986.
9. H. B. Nielsen and N. Brene, *Nucl. Phys.* **B224**, 396 (1983).
10. J. Wess and J. Bagger, *Supersymmetry and Supergravity* (Princeton University Press, 1983), P. G. O. Freund, *Introduction to Supersymmetry* (Cambridge University Press, 1986), P. West, *Introduction to Supersymmetry and Supergravity* (World Scientific, 1986).
11. F. A. Berends, J. W. van Holten, B. De Wit and P. van Nieuwenhuizen, *J. Phys.* **A13**, 1643 (1980).
12. P. Howe and U. Lindström, *Nucl. Phys.* **B181**, 487 (1981); N. Marcus and A. Sagnotti, *Nucl. Phys.* **B256** 77 (1985).
13. B. De Wit and H. Nicolai, *Nucl. Phys.* **B208**, 323 (1982).
14. J. Ellis, M. K. Gaillard and B. Zumino, *Phys. Lett.* **94B**, 343 (1980); J. Ellis, M. K. Gaillard and B. Zumino, *Acta Physica Polonica* **B13**, 253 (1982).
15. A. C. Davis, M. D. Freeman and A. J. Macfarlane, *Nucl. Phys.* **B258**, 393 (1985).
16. Th. Kaluza, Sitzungsber. Preuss. Akad. Wiss. Berlin, *Math. Phys.* **K1**, 966 (1921).
17. O. Klein, *Z. Phys.* **37**, 895 (1926); O. Klein, *Nature* **118**, 516 (1926).
18. M. J. Duff, B. E. W. Nilsson and C. N. Pope, *Phys. Reports* **130**, 1 (1986); T. Appelquist, A. Chodos and P. G. O. Freund, *Modern Kaluza-Klein Theory and Applications* (Benjamin-Cummings, 1987).
19. P. G. O. Freund and M. A. Rubin, *Phys. Lett.* **97B**, 233 (1980).
20. E. Witten, *Shelter Island II: Proceedings of the 1983 Shelter Island Conference on Quantum Field Theory and the Fundamental Problems of Physics*, eds. R. Jackiw *et al.* p. 227 (MIT Press, 1985).
21. S. Randjbat-Daemi, A. Salam and J. Strathdee, *Nucl. Phys.* **B214**, 491 (1983).
22. P. H. Frampton and T. W. Kephart, *Phys. Rev. Lett.* **50**, 1343, 1347 (1983).
23. L. Alvarez-Gaumé and E. Witten, *Nucl. Phys.* **B234**, 269 (1983).
24. L. Alvarez-Gaumé and P. Ginsparg, *Ann. Phys.* **161**, 423 (1985).
25. M. B. Green, J. H. Schwarz and E. Witten, *Superstring Theories*, Volumes 1 and 2 (Cambridge University Press, 1987).
26. J. Thierry-Mieg, *Phys. Lett.* **171B**, 163 (1986).
27. A. N. Schellekens and N. P. Warner, *Nucl. Phys.* **B287**, 317 (1987).
28. A. Bilal and J. L. Gervais, *Phys. Lett.* **187B**, 39 (1987).
29. J. H. Schwarz, *Int. J. Mod. Phys.* **A2**, 593 (1987).
30. D. J. Gross, J. A. Harvey, E. Martinec and R. Rohm, *Nucl. Phys.* **B256**, 253 (1985); *Nucl. Phys.* **B267**, 75 (1986).
31. I. B. Frenkel and V. G. Kac, *Invent. Math.* **62**, 23 (1980).
32. G. Segal, *Comm. Math. Phys.* **80**, 301 (1981).
33. P. Goddard and D. Olive, *Vertex Operators in Mathematics and Physics*, Proceedings of a Conference, November 10-17, 1983, eds. J. Lepowsky, S. Mandelstam and I. M. Singer (Springer-Verlag, 1985), p. 51.
34. P. Goddard and D. Olive, *Int. J. Mod. Phys.* **A1**, 303 (1986).

35. P. Goddard, J. Goldstone, C. Rebbi and C. B. Thorn, *Nucl. Phys.* **B56**, 109 (1973).
36. M. A. Virasoro, *Phys. Lett.* **82B**, 436 (1979).

Chapter V

THE CPT THEOREM

The clearest example of a derived symmetry is provided by the famous CPT theorem [Paper 23]. This theorem states that any quantum field theory is invariant under the operation CPT, corresponding to the combined action of the three discrete operators: charge conjugation C, parity P and time reversal T.

The CPT theorem has been proved in great generality[1], and under very mild assumptions. These assumptions, from axiomatic field theory, are:

(1) Poincaré invariance
(2) Microscopic causality as expressed by local commutativity
(3) Continuity of quantum field operators — their matrix elements are taken to be tempered distributions.

In fact only a weak form of the second assumption is required. This weak local commutativity condition is the statement that the vacuum expectation value of any product of field operators $\psi(x)$ satisfies the equation

$$\langle 0|\psi(x_1)\dots\psi(x_n)|0\rangle = i^F \langle 0|\psi(x_n)\dots\psi(x_1)|0\rangle \qquad (1)$$

when all the co-ordinate differences $x_i - x_j$ are spacelike. Here F is the total number of half-odd integer spin fields in the product.

The CPT theorem may also be proved directly, in the more restrictive framework of Lagrangian field theory. In this approach [Paper 23], the most general form of Lorentz invariant local hermitian Lagrangian density is constructed from the field operators. After appropriate normal ordering and quantization with the normal spin-statistics relation, CPT invariance is readily verified.

We will now outline the ideas underlying the general argument for CPT invariance. Then, as an example, we will verify it, by explicit construction of the CPT operation, for the standard model Lagrangian density discussed in Chapter III.

One way of understanding the CPT theorem is via the related concepts of crossing symmetry and the S-matrix[2], which involves introducing the asymptotic condition necessary for the existence of in-states and out-states. Any S-matrix element,

or scattering amplitude, is a function of the external momenta with analyticity properties following from assumptions (2) and (3). We imagine continuing a scattering amplitude analytically in one or more of the external momenta P_μ, but keeping the external particles on mass-shell, i.e. $P_\mu^2 = m^2$. Initially we consider spinless particles. By going into the complex plane, it is possible to continue from values of P_μ having positive energy, $P_0 > 0$, to negative energy, $P_0 < 0$. The resulting new on-shell amplitude can be related to the residue of poles in a vacuum expectation value of a product of fields. It is then, in general, possible to argue that these poles must correspond to physical particles, whose scattering is described, kinematics permitting, by the new analytically continued amplitude. The particles with $P_0 < 0$ have to be interpreted as incoming instead of outgoing (or vice-versa), and consequently must carry opposite additive-conserved quantum numbers such as electric charge. Hence we are led to the existence of antiparticles.

We have thus crudely argued for crossing symmetry in the following form. By analytically continuing a scattering amplitude in the external momenta, so that some of the particles have their energies P_0 turned negative, a scattering amplitude is obtained for a new corresponding set of particles. In this corresponding set, those incoming (outgoing) particles having negative P_0 are replaced by outgoing (incoming) antiparticles with opposite values of the four-momentum.

Once we accept the above crossing relation, we may obtain the CPT-transform of a given amplitude by taking a special analytic continuation, which corresponds to a continuation in the parameters of a proper Lorentz transformation. By going through a sequence of complex Lorentz transformations connected to the identity, it is possible to generate the transformation $x_\mu \to -x_\mu$ or $P_\mu \to -P_\mu$ for all the particles. Any Lorentz invariant amplitude is of course unaltered by such an analytic continuation. It has been proved[1] that the analytic continuation, $P_\mu \to -P_\mu$, is possible for any scattering amplitude, if weak local commutativity holds.

We are thereby led to the CPT theorem for spinless particles: The scattering amplitude for a set of particles is equal to the scattering amplitude for the corresponding set of antiparticles, but with all incoming particles replaced by outgoing particles and vice versa.

Let us now consider the generalisation to particles with spin. We must take into account that, under a Lorentz transformation, the various spin components of a particle are transformed into linear combinations of each other. The analytically continued proper Lorentz transformation Λ, leading to $P_\mu \to -P_\mu$, can be considered as composed of a boost in, say, the x_1-direction, by an imaginary rapidity $i\pi$, followed by a rotation about the x_1-axis by an angle π. The Lorentz boost is illustrated in Fig. 5.1, where we show the effect on a unit time-like 4-vector (e.g. the 4-velocity of a particle), $V^\mu = (1,0,0,0)$, of a boost by a rapidity ϕ in the x_1-direction. For any real value of ϕ, the boosted 4-vector V^μ lies on the upper branch of the hyperbola drawn in the figure. However if ϕ is allowed to take on complex values, then the components V^0 and V^1 are continued into the complex plane and can end up on the lower branch of the hyperbola. In particular for purely

imaginary values of ϕ, the boost actually corresponds to a rotation in the (Re V^0, Im V^1)-plane by an angle ϕ/i. The 4-vector V^μ then runs through the dashed circle drawn in perspective in Fig. 5.1. It follows that a boost by a rapidity $\phi = i\pi$ in the x_1-direction changes the signs of P_0 and P_1, leaving P_2 and P_3 unchanged. If this boost is followed by a rotation of π about the x_1-axis, P_2 and P_3 also change sign and the full transformation Λ leads to $P_\mu \to -P_\mu$.

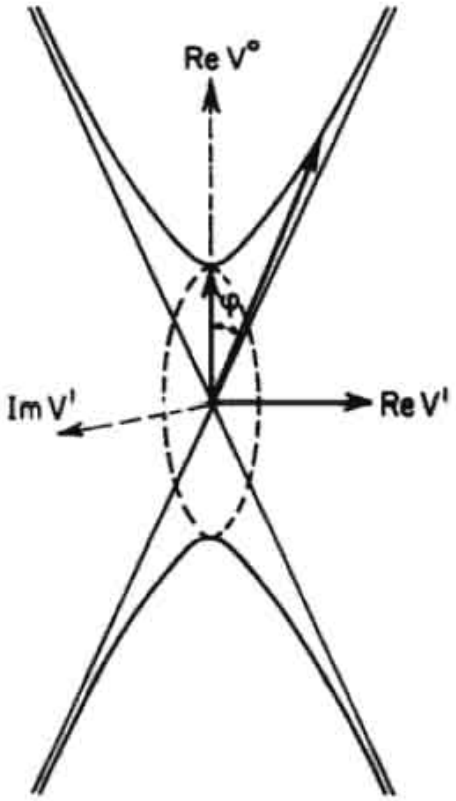


Fig. 5.1. Illustration of the sign-reversal of the V^0-component of a unit time-like 4-vector V^μ, by analytic continuation of a Lorentz boost to complex values of the rapidity ϕ.

The above complex Lorentz transformation Λ is easily shown to leave the spin angular momentum of a particle invariant. For instance, the result immediately follows if we choose the x_1-axis along the direction of the spin component under consideration. So, under the complex transformation Λ, the spin of a particle is unchanged while its 4-momentum P_μ changes sign. Hence the helicity of a particle also changes sign, $\lambda \to -\lambda$, under the transformation Λ. By crossing, we now replace incoming (outgoing) particles by outgoing (incoming) antiparticles with the opposite sign of 4-momentum and spin angular momentum. So each antiparticle carries the same 4-momentum P_μ, but the opposite helicity $-\lambda$, to that which the corresponding particle had before applying the transformation Λ. Thus, ignoring the phase factors which arise in a precise definition of particle and antiparticle helicity states, we are led to the equality of the scattering amplitudes for two processes related by crossing all the particles.

For example, the scattering amplitude for the process

$$(A, \mathbf{P}_A, \lambda_A) + (B, \mathbf{P}_B, \lambda_B) \to (C, \mathbf{P}_C, \lambda_C) + (D, \mathbf{P}_D, \lambda_D) + (E, \mathbf{P}_E, \lambda_E) \qquad (2)$$

is the same as the scattering amplitude for the process

$$(\overline{C}, \mathbf{P}_C, -\lambda_C) + (\overline{D}, \mathbf{P}_D, -\lambda_D) + (\overline{E}, \mathbf{P}_E, -\lambda_E) \\ \rightarrow (\overline{A}, \mathbf{P}_A, -\lambda_A) + (\overline{B}, \mathbf{P}_B, -\lambda_B) \ . \tag{3}$$

Here we have used the notation $(X, \mathbf{P}_X, \lambda_X)$ for a particle X having 3-momentum $\mathbf{P}_X$ and helicity λ_X and $\overline{X}$ denotes the antiparticle of X. This result can be written in terms of elements of the S-matrix

$$S_{fi} = I_{fi} + i(2\pi)^4 \delta^4(P_f^\mu - P_i^\mu) T_{fi} \tag{4}$$

up to a helicity dependent phase factor, as follows:

$$\langle C, \mathbf{P}_C, \lambda_C;\ D, \mathbf{P}_D, \lambda_D;\ E, \mathbf{P}_E, \lambda_E | T | A, \mathbf{P}_A, \lambda_A;\ B, \mathbf{P}_B, \lambda_B \rangle \\ = \langle \overline{A}, \mathbf{P}_A, -\lambda_A;\ \overline{B}, \mathbf{P}_B, -\lambda_B | T | \overline{C}, \mathbf{P}_C, -\lambda_C;\ \overline{D}, \mathbf{P}_D, -\lambda_D;\ \overline{E}, \mathbf{P}_E, -\lambda_E \rangle \ . \tag{5}$$

Simply writing the right-hand side of Eq. (5) in terms of its complex conjugate, we can express the result as a relationship between amplitudes, for which the antiparticles corresponding to incoming (outgoing) particles are also incoming (outgoing):

$$\langle C, \mathbf{P}_C, \lambda_C;\ D, \mathbf{P}_D, \lambda_D;\ E, \mathbf{P}_E, \lambda_E | T | A, \mathbf{P}_A, \lambda_A;\ B, \mathbf{P}_B, \lambda_B \rangle \\ = \langle \overline{C}, \mathbf{P}_C, -\lambda_C;\ \overline{D}, \mathbf{P}_D, -\lambda_D;\ \overline{E}, \mathbf{P}_E, -\lambda_E | T^+ | \\ \overline{A}, \mathbf{P}_A, -\lambda_A;\ \overline{B}, \mathbf{P}_B, -\lambda_B \rangle^* \ . \tag{6}$$

This is the expression of the CPT theorem for the processes (2) and (3).

By definition, the CPT operator

$$\Theta = \mathrm{CPT} \tag{7}$$

has the following action, up to a phase factor, on a single particle helicity state

$$\Theta | A, \mathbf{P}_A, \lambda_A \rangle = | \overline{A}, \mathbf{P}_A, -\lambda_A \rangle \ . \tag{8}$$

The momentum $\mathbf{P}_A$ is left invariant, because it is inverted by both P and T. The spin is inverted by T and unchanged by C and P. Hence the helicity λ_A is inverted by Θ. It is, of course, C that transforms the particle A into its antiparticle $\overline{A}$. The combined operation, $\Theta = \mathrm{CPT}$, may exist even when the individual operators C, P and T are not defined.

The symmetry derived by applying the complex Lorentz transformation Λ and crossing to a scattering amplitude, for general multiparticle states $|a\rangle$ and $|b\rangle$, can therefore be written in the form

$$\langle b | T | a \rangle = \langle \Theta b | T^+ | \Theta a \rangle^* \\ = \langle \Theta a | T | \Theta b \rangle \ . \tag{9}$$

This invariance property

$$\Theta T \Theta^{-1} = T^{+} \tag{10}$$

of the S-matrix is the general expression of the CPT theorem. The hermitian conjugation, which appears on the right-hand side of Eq. (10), is due to the antiunitary nature of Θ.

In place of the above general argument, based on the analyticity properties of axiomatic field theory, the CPT theorem may also be derived by considering the most general form, a Lorentz invariant local hermitian Lagrangian density $\mathcal{L}(\psi(\mathbf{x},t))$ can take. It is then verified [Paper 23] that $\mathcal{L}(\psi(\mathbf{x},t))$ satisfies

$$\mathcal{L}(\psi(\mathbf{x},t)) = \mathcal{L}^{*}(\psi^{\Theta}(-\mathbf{x},-t)) \tag{11}$$

and the quantised field theory described by $\mathcal{L}(\psi(\mathbf{x},t))$ is therefore CPT invariant. Here $\psi(\mathbf{x},t)$ is a generic symbol for all the fields and $\psi^{\Theta}(-\mathbf{x},-t)$ denotes the CPT transformed fields. The $*$ on $\mathcal{L}^{*}$ denotes complex conjugation of all the constants in the Lagrangian density, and its presence is a consequence of the antiunitary nature of the CPT transformation. The half-odd integer spin fields are taken to be anti-commuting Grassman variables, as in the path integral formalism.

As an illustration of the second approach, we now briefly consider the standard model Lagrangian density, which consists of five terms.

$$\mathcal{L} = \mathcal{L}_1 + \mathcal{L}_2 + \mathcal{L}_3 + \mathcal{L}_4 + \mathcal{L}_5 \tag{12}$$

as discussed in Chapter III. Using the notation of Chapter III, we list below the CPT transformations of typical Bose and Fermi fields selected from Table 3.1. The CPT transforms of the SU(2) gauge field $A_{\mu i}{}^{j}(\mathbf{x},t)$, its field strength tensor $F_{\mu\nu i}{}^{j}(\mathbf{x},t)$ and dual field strength tensor $\tilde{F}_{\mu\nu i}{}^{j}(\mathbf{x},t)$ are given by

$$A^{\Theta}_{\mu i}{}^{j}(\mathbf{x},t) = -A_{\mu j}{}^{i}(-\mathbf{x},-t) \tag{13}$$

$$F^{\Theta}_{\mu\nu i}{}^{j}(\mathbf{x},t) = F_{\mu\nu j}{}^{i}(-\mathbf{x},-t) \tag{14}$$

$$\tilde{F}^{\Theta}_{\mu\nu i}{}^{j}(\mathbf{x},t) = \tilde{F}_{\mu\nu j}{}^{i}(-\mathbf{x},-t)\ . \tag{15}$$

The left-handed quark field $q'_{i\alpha aL}(\mathbf{x},t)$ and the right-handed $Q = 2/3$ quark field $u'_{\alpha aR}$ have the following CPT transforms:

$$q'^{\Theta}{}_{i\alpha aL}(\mathbf{x},t) = i\gamma^{0}\gamma^{5}\bar{q}'^{Ti\alpha}{}_{aL}(-\mathbf{x},-t) \tag{16}$$

$$\bar{q}'^{\Theta i\alpha}{}_{aL}(\mathbf{x},t) = iq'^{T}{}_{i\alpha aL}(-\mathbf{x},-t)\gamma^{5}\gamma^{0} \tag{17}$$

$$u'^{\Theta}{}_{\alpha aR}(\mathbf{x},t) = i\gamma^{0}\gamma^{5}\bar{u}'^{T\alpha}{}_{aR}(-\mathbf{x},-t) \tag{18}$$

$$\bar{u}'^{\Theta\alpha}{}_{aR}(\mathbf{x},t) = iu'^{T}{}_{\alpha aR}(-\mathbf{x},-t)\gamma^{5}\gamma^{0}\ . \tag{19}$$

Here the superscript T on a spinor denotes its transpose. We note that ${q'}^{\Theta}{}_{i\alpha aL}(\mathbf{x},t)$ is a right-handed antiquark field and ${u'}^{\Theta}{}_{\alpha aR}(\mathbf{x},t)$ a left-handed antiquark field, as required by Eq. (8). As a final example, the CPT transformed Higgs field $\phi_i(\mathbf{x},t)$ is

$$\phi_i^{\Theta}(\mathbf{x},t) = \phi^{+i}(-\mathbf{x},-t) \tag{20}$$

$$\phi^{+\Theta i}(\mathbf{x},t) = \phi_i(-\mathbf{x},-t)\ . \tag{21}$$

The CPT transformations of the other gauge, quark and lepton fields in Table 3.1 follow, from these examples, by analogy.

It is now readily seen that, under the above CPT transformations of the fields and the reversal of the space-time co-ordinates,

$$x^{\mu} \to -x^{\mu}, \quad \partial_{\mu} \to -\partial_{\mu} \tag{22}$$

the standard model Lagrangian density, Eq. (12), satisfies the condition, Eq. (11), for CPT invariance. The CPT invariance of a general Lorentz invariant local hermitian field theory is similarly proved, by explicit construction, in Paper 23.

References

1. R.F. Streater and A.S. Wightman, *PCT, Spin and Statistics, and All That* (Benjamin/ Cummings, 1964; second printing, with additions and corrections, 1978).
2. R.J. Eden, P.V. Landshoff, D.I. Olive and J.C. Polkinghorne, *The Analytic S-Matrix* (Cambridge University Press, 1966).

Chapter VI

THE FUNDAMENTAL SYMMETRIES

6.1. Introduction

Poincaré invariance and gauge invariance are the basic input symmetries of the standard model. These symmetries are usually considered to be truly fundamental and not derivable, even in theories beyond the standard model. For example, they are both input assumptions in Grand Unified Theories and Poincaré invariance is needed in the proof of the CPT theorem. Therefore, from the conventional standpoint, Yang-Mills gauge invariance, Lorentz and translation invariance are all expected to be valid at the top of the quantum staircase (Fig. 1.1). In this chapter we consider the alternative point of view, according to which even these "fundamental" symmetries can be derived, albeit at a rather high point on the quantum staircase.

We have already touched on the possibility of the dynamical generation of a gauge symmetry in our discussion of Papers 17-19 in Chapter IV. We have also discussed, in Chapter IV, the origin of gauge invariance in string theory. A third approach to deriving gauge invariance is the renormalisation group method, in which the symmetry becomes a better and better approximation as one goes down the energy scale (or quantum staircase). These three approaches have also been applied to the derivation of Poincaré invariance.

We classify the attempts to derive Poincaré invariance and gauge invariance in Table 6.1, according to the following three methods:

1. Formal Appearance. Derivations of this type reveal the symmetry to be a purely formal one, which nevertheless attains physical significance, in some phase of the vacuum, due to quantum fluctuations.
2. Renormalisation Group. In this method a symmetry, which is not assumed valid at short distances, becomes more and more accurate at larger and larger distances. In other words, the renormalisation group β-functions imply the suppression of symmetry breaking terms towards the infra-red.
3. String Theory. This method applies to symmetries which arise, without being

Table 6.1. Classification of methods for deriving Poincaré invariance and gauge invariance.

Symmetry / Method	Poincaré Invariance		Local Gauge Invariance
	Lorentz Invariance	Translational Invariance	
Formal Appearance	Special case of reparameterisation invariance	Special case of reparameterisation invariance	Papers 17,18,19and26
Renormalisation Group	Papers 24, 25 and 29	Chapter VI Section 2.3	Papers 28 and 29
String Theory	Chapter VI Reference 8		Chapter IV Section 6.2

explicitly put in, as a consequence of other symmetry properties of string theory.

We now consider, in turn, these derivations of Poincaré invariance and gauge invariance. This chapter concludes with a brief discussion of attempts to derive supersymmetry by the renormalisation group method.

6.2. Poincaré Invariance

In this section we discuss the various methods proposed to derive Poincaré invariance.

6.2.1. Formal appearance of Poincaré invariance

Poincaré invariance can be considered to be a consequence of general co-ordinate invariance, and the absence of "prior geometry",[1] in gravitation theory. By "prior geometry" one means[1] any aspect of the geometry of space-time that is fixed immutably, i.e. that cannot be changed by changing the distribution of gravitating sources. A "prior geometry" is an example of an "absolute object", which is defined[2] to be a function, with one or more components, of space-time that is not dependent on the state of matter.

An example of an absolute object is the concept of absolute time, as introduced in Newtonian physics. In this example, the physical theory contains a real number function $I^0(x^\rho)$ of space-time, which specifies the time in an absolute sense. The value of the absolute time $I^0(x^\rho)$ is not affected by the presence of matter, but matter can be influenced by absolute time through the physical laws. For instance, one could imagine constructing a clock, which displayed the absolute time independent of its state of motion. Relativistic physics, of course, does not allow the introduction of absolute time. Any absolute object in relativity theory must be invariant under Poincaré transformations.

The action for any physical theory can be written in a reparameterisation invariant way. This formal diffeomorphism symmetry is, in general, obtained at the expense of introducing absolute objects into the theory. The special diffeomorphisms,

corresponding to Poincaré transformations, only generate a true physical Poincaré symmetry if the absolute objects are invariant under these transformations. The basic idea, for deriving Poincaré invariance from diffeomorphism symmetry, is then to find dynamical, quantum mechanical arguments for potential absolute objects being Poincaré invariant or non-existent.

Historically, of course, general relativity was constructed with Poincaré symmetry in mind. However, in the spirit of the present book, we can take the point of view that general relativity arises from some pregeometry model,[1,3] without assuming special relativity from the outset. For the purposes of our initial discussion, we temporarily assume the absence of non-Poincaré invariant absolute objects. We can then take either of the following two attitudes towards Poincaré symmetry:

I. Poincaré invariance is just a subgroup of the formal reparameterisation invariance group of general relativity.
II. Poincaré invariance is an approximate symmetry, valid when the gravitational field is neglected.

According to attitude I, Poincaré invariance is a purely formal symmetry of general relativity under the special co-ordinate transformation

$$x^\mu \longrightarrow x'^\mu = \Lambda^\mu_\nu x^\nu + a^\mu \ . \tag{1}$$

The two co-ordinate systems x'^μ and x^μ are, by definition, related to each other by a Lorentz transformation Λ^μ_ν followed by a translation a^μ. This is illustrated, for an arbitrary co-ordinate system x^μ, in Figs. 6.1 and 6.2. The co-ordinate systems x^μ and x'^μ are in each case denoted by full and dashed curves respectively, while the co-ordinate axes are drawn in heavier print. In this attitude of Poincaré symmetry being just part of diffeomorphism symmetry, one only has Poincaré invariance provided the gravitational field is transformed as well as the matter fields.

According to attitude II, the matter fields are transformed under the Poincaré transformation, but the gravitational field is left unchanged. There is then, of course, only Poincaré symmetry to the extent that the gravitational field can be neglected, i.e. in the approximation of flat space-time. If, however, space-time is flat and we choose co-ordinates x^μ so that the metric tensor is constant as a function of x^μ, then diffeomorphism symmetry can be used to derive Poincaré symmetry. This means that, once we have a reparameterisation invariant theory, the question of deriving Poincaré symmetry, in attitude II, has been transformed into the cosmological question of why space-time is almost flat nowadays. This approximate flatness is due to the Hubble expansion, having scaled up the size of the Universe by a big factor.

Let us now consider rotational invariance, as observed in daily life. It would seem that the up-down axis is singled out, and that there is only rotational invariance around a vertical axis. For instance, a pendulum suspended from a support will direct itself downwards, after it has had time to lose any kinetic energy via friction, independent of the orientation of the support. So rotational invariance is broken,

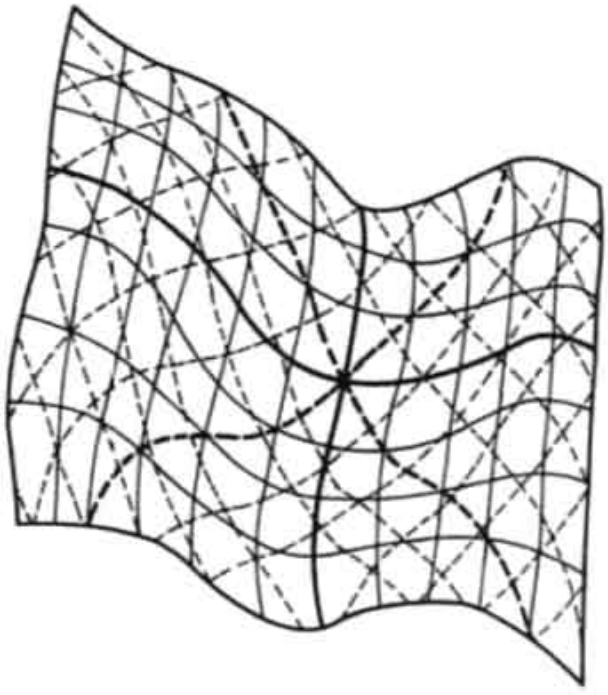


Fig. 6.1. Two co-ordinate systems related by a Lorentz transformation.

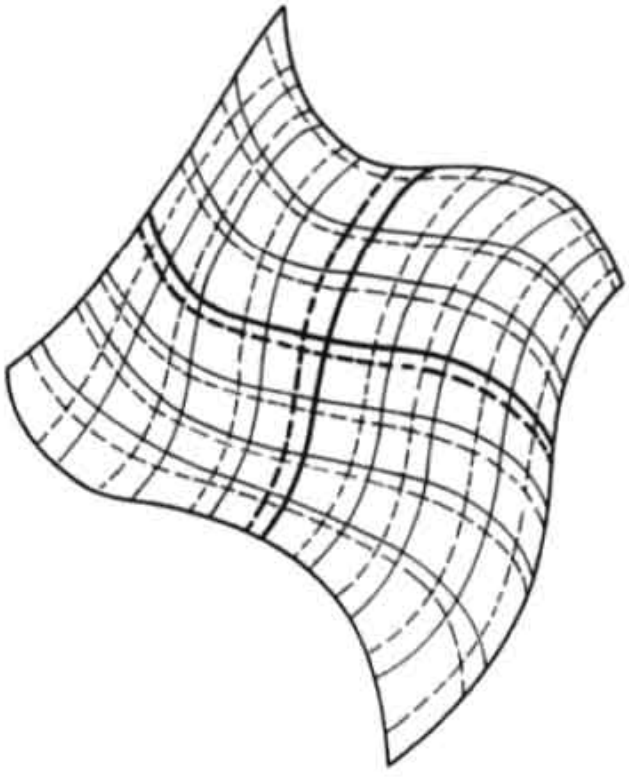


Fig. 6.2. Two co-ordinate systems related by a translation.

with the up-down direction being special, in attitude II. It is possible to formally restore exact rotational symmetry by taking attitude I. In attitude I, when rotating the apparatus, we must also remember to rotate the gravitational field and the earth creating it. This is achieved by a diffeomorphism of the form $x'^{\mu} = \Lambda^{\mu}_{\nu} x^{\nu}$ which rotates everything physical relative to the co-ordinate system.

The above remarks are illustrated in Fig. 6.3 by diagrams of a simple pendu-

lum experiment. When the supporting stand is rotated about a horizontal axis, rotational invariance appears to be broken as shown in the top two diagrams. It is easy to introduce a formal rotational symmetry; one invents the existence of a gravitational field — symbolised by a little demon in the bottom two diagrams — and rotates it as well as the supporting stand. Rotational invariance of the laws of nature is thereby formally restored.

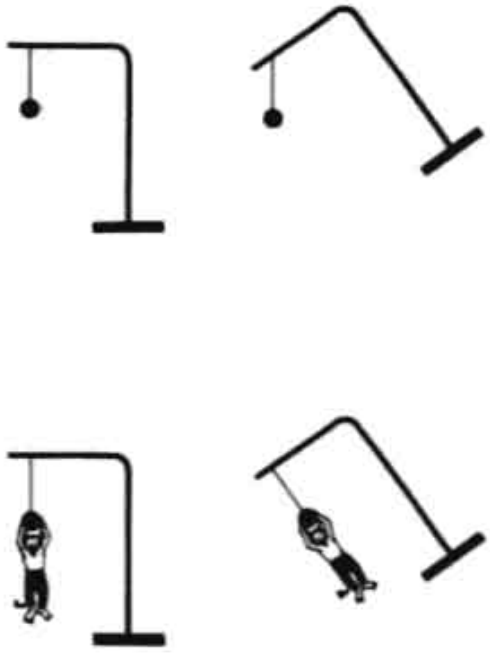


Fig. 6.3. The top two diagrams illustrate the lack of rotational invariance in daily life, by means of a pendulum experiment. The bottom two diagrams show how rotational invariance is ensured when the gravitational field, represented by a little demon, is rotated as well as the pendulum.

As long as we only consider the usual daily life situation, we can ignore the idea of associating the gravitational field with the earth. The gravitational field of everyday experience can be treated as an immutable property of our region of space-time, giving a constant downward acceleration. The introduction of a gravitational field of this type, uninfluenced by the presence of matter, means that we now have a non-rotational invariant absolute object in the theory. Rotational invariance would then indeed only be a formal symmetry, introduced basically by definition. One simply postulates a field, or little demon, giving the rotational symmetry breaking effect. The existence of rotational symmetry becomes very much a question of semantics. By identifying a symmetry breaking effect with a field in this way, one can formally introduce symmetries wherever one wishes.

However if we view the situation from a larger distance scale, we immediately recognise that the gravitational field is due to the earth. The gravitational field would be modified by moving the earth. So the demon, i.e. the gravitational field, is not really an absolute object but a dynamical object, determined by the distribution of matter. We can therefore conclude that the formal rotational invariance, in fact, corresponds to a true physical symmetry.

Analogously it is quite easy to introduce Poincaré invariance formally; it simply

corresponds to the invariance of the action, under the special diffeomorphism of Eq. 1. However if the theory was not truly Poincaré invariant to begin with, there would now exist a non-Poincaré invariant absolute object in the formalism. In this case some regions of space-time would have specific properties, which could be described by introducing a set of *a priori* physically significant co-ordinates, I^a say, as a set of four scalar fields. These scalar fields $I^a(x^\rho), a = 0, 1, 2, 3$, would depend on the co-ordinates x^ρ in an arbitrary co-ordinate system. Then any dependence of the Lagrangian density upon the fundamental co-ordinates I^a, which in the fundamental language breaks diffeomorphism symmetry, just becomes a dependence on the fields $I^a(x^\rho)$. We thereby formally construct a diffeomorphism symmetric Lagrangian density.

The presence, or absence, of physical Poincaré invariance now becomes the question of whether, or not, the absolute object $I^a(x^\rho)$ is in a Poincaré invariant state in the vacuum. Here we assume the vanishing of the cosmological constant and take the gravitational vacuum to be flat Riemannian space-time. If $I^a(x^\rho)$ is not Poincaré invariant in the vacuum, it will act as a source of spontaneous symmetry breakdown. So the question is really whether or not the formal Poincaré symmetry is spontaneously broken. A natural way for such a spontaneous breakdown to occur would be for the scalar fields $I^a(x^\rho)$ to take on vacuum values

$$I^a(x^\rho) \approx x^a \tag{2}$$

in some co-ordinate system, where the metric tensor $g_{\mu\nu}(x^\rho)$ is approximately constant. This does not occur in standard general relativity theory, where any scalar fields present take on constant values in the vacuum.

A true derivation of Poincaré invariance must therefore explain the absence of fields, like $I^a(x^\rho)$, taking space-time dependent vacuum values, such as in Eq. 2. There appear, at first, to be two possible ways in which this could happen:

(a) The fields $I^a(x^\rho)$ may take on Poincaré invariant (i.e. constant) values in the vacuum.
(b) The fields $I^a(x^\rho)$ may quantum fluctuate with the same distribution for all x^ρ.

The degrees of freedom corresponding to the fields $I^a(x^\rho)$ have to be interpreted as some matter degrees of freedom, e.g. four, as yet to be discovered, scalar particles.

Possibility (a) would seem attractive, since scalar fields can easily take on constant values in the vacuum. However, if the fields $I^a(x^\rho)$ are to be interpreted as "fundamental co-ordinates" at some level, the ordered set $I^a(x^\rho)$ should not take on the same value more than once:

$$\hat{x}^\rho \neq x^\rho \Longrightarrow I^a(\hat{x}^\rho) \neq I^a(x^\rho)\ . \tag{3}$$

One could imagine that $I^a(x^\rho)$ fluctuates quantum mechanically over a broad-peaked distribution, without taking the same value twice. Then the average could

both exist and be constant. This would require the existence of weak correlations between the values of $I^a(x^\rho)$ at different points. There could be a problem with topological obstructions, in constructing such a probability distribution on a set of continuous functions $I^a(x^\rho)$. This problem with topology can be avoided, by introducing a fundamental space-time lattice instead of fundamental space-time co-ordinates.[3]

The above picture for possibility (a) relies so much on quantum fluctuations that it differs very little, physically, from possibility (b). However the quantum fluctuations, in possibility (b), may be so large that it is not possible to assign well-defined average values to the fields $I^a(x^\rho)$ in the vacuum. This essentially occurs in the latticised quantum gravity model of Ref. 3, in which the supposedly fundamental lattice sites (and links) are dynamical variables. The sites fluctuate quantum mechanically over a Riemannian space-time manifold, in the functional integration, like the molecules of a gas. The fields $I^a(x^\rho)$ must be interpreted, in this model, as numbers specifying the name of the site, which happens to be present near the point x^ρ on the manifold. One of the main conclusions of Ref. 3 is the existence of a phase, in which the sites fluctuate, with a flat distribution, all over space-time. This means that the $I^a(x^\rho)$ take on extremely uncertain values in the vacuum, but the probability distribution for the $I^a(x^\rho)$ is the same for all x^ρ. Thus Poincaré invariance is derived in this pregeometry model.

It is now natural to ask why only Poincaré invariance is derived in this way, and not any other part of diffeomorphism symmetry. The vacuum values of the fields $I^a(x^\rho)$ may indeed be fully reparameterisation invariant, but the vacuum state itself is not. The metric $g_{\mu\nu}$ of the flat space-time vacuum of general relativity is invariant under precisely the diffeomorphisms corresponding to Poincaré transformations. The full diffeomorphism symmetry of the general relativity action is thereby spontaneously broken down to Poincaré symmetry. Indeed the graviton may be considered to be the Goldstone boson associated with the broken GL(4, IR) group of transformations.[4]

Recently an attempt was made to give a more explicit derivation of time translational invariance as a formal symmetry.[5] In this approach the original time variable of the time non-translational invariant model is made into a dynamical variable, which can essentially be identified with the variable $I^o(x^\rho)$ discussed above. However, a severe problem arises with this treatment of time as a dynamical variable since, as pointed out by Pauli, there is then no lower bound to the Hamiltonian. This problem is not manifest, and therefore easily forgotten, in a Euclidean formulation.

6.2.2. Lorentz invariance from the renormalisation group

In this subsection we consider the derivation of Lorentz invariance as an approximate symmetry which becomes more and more accurate towards the infrared, in contrast to the exact formal Lorentz invariance of general relativity. Lorentz invariance can be shown to arise as such a low energy symmetry for non-covariant quan-

tum electrodynamics [Paper 24] and non-covariant Yang-Mills theory [Paper 25]. The main assumptions made in these non-Lorentz invariant field theory models are translational invariance, gauge invariance, chiral (massless) fermions and renormalisability. Scalar fields are ignored on the grounds that they generically become very massive and therefore irrelevant to low energy physics, unless protected by some additional symmetry. A similar mass protection principle, for light fermions, was used in Chapter 3.2 to exclude observable fermion states which are vectorlike under the standard model gauge group S(U(2)×U(3)).

Under the above assumptions, the most general allowed form of local Lagrangian density has terms of the following type.

a) The kinetic term for gauge fields in an anisotropic vacuum, having magnetic susceptibility and dielectric properties depending on direction and Lorentz frame:

$$\mathcal{L}_a = -\frac{1}{4}\eta^{\mu\nu\rho\sigma}F_{\mu\nu}(x)F_{\rho\sigma}(x)\ . \tag{4}$$

Here $F_{\mu\nu}(x)$ is the gauge field strength tensor and the 256 components of $\eta^{\mu\nu\rho\sigma}$ are dimensionless coupling constants, which do not transform under Lorentz transformation. There are just 20 independent components of $\eta^{\mu\nu\rho\sigma}$ which effectively contribute to $\mathcal{L}_a$.

b) The gauge covariant kinetic term for each (left-handed) Weyl fermion field $\psi_P = \frac{1}{2}(1-\gamma_5)\psi_P$:

$$\mathcal{L}_b = i\overline{\psi}_P\gamma^\alpha V_\alpha^{(P)\mu}D_\mu\psi_P\ . \tag{5}$$

Here P denotes the fermion flavour and, for Yang-Mills theory, ψ_P is also an irreducible representation of the gauge group. The vierbein $V_\alpha^{(P)\mu}$ for the fermion of flavour P is just a set of 16 dimensionless coupling constants. The gauge covariant derivative is

$$D_\mu = \partial_\mu - i\tilde{A}_\mu \tag{6}$$

where $\tilde{A}_\mu$ is the gauge field in the appropriate representation of the Yang-Mills group or, for the Abelian case, $\tilde{A}_\mu = qA_\mu$ and q is the U(1) charge of ψ_P. A motivation for using such Weyl fields ψ_P, without Lorentz invariance, is given in [Paper 30].

c) Momentum shift terms for the fermions:

$$\mathcal{L}_c = \overline{\psi}_P\gamma^\alpha V_\alpha^{(P)\mu}\xi_\mu^P\psi_P\ . \tag{7}$$

The four coupling constants ξ_μ^P here have the dimension of mass. There could also be flavour mixing terms of this type.

d) A topological term, originally suggested by C.H. Woo, of the Chern-Simons type:

$$\mathcal{L}_d = \eta_\mu\varepsilon^{\mu\nu\rho\sigma}\mathrm{Tr}\left(A_\nu F_{\rho\sigma} - \frac{2}{3}A_\nu A_\rho A_\sigma\right)\ . \tag{8}$$

The four coupling constants η_μ also have the dimension of mass.

On purely dimensional grounds we should expect the constants ξ_μ^P and η_μ to be proportional to some fundamental mass, presumably the Planck mass. Therefore, unless we can find arguments for setting them to zero, the terms $\mathcal{L}_c$ and $\mathcal{L}_d$ would completely dominate $\mathcal{L}_a$ and $\mathcal{L}_b$ in the low energy regime and Lorentz invariance would not be obtained. In fact, the fermion momentum shift terms $\mathcal{L}_c$ are easily removed, by an additive renormalisation of momentum achieved with simple field redefinitions of the type

$$\psi_P(x) \longrightarrow \psi'_P(x) = \exp(-i\xi_\mu^P x^\mu)\psi_P(x) \; . \tag{9}$$

The fermion kinetic energy term $\mathcal{L}_b$, when written in terms of ψ'_P, develops a term which cancels $\mathcal{L}_c$. Thus $\mathcal{L}_c$ can be simply transformed away and ignored. The term $\mathcal{L}_d$ can also be forbidden by appealing to gauge invariance, provided that we assume the existence of magnetic monopoles. This argument is explained below. Thus we should really include the existence of magnetic monopoles as one of our input assumptions; otherwise an alternative method of excluding $\mathcal{L}_d$ is called for.

We now explicitly show, for the Abelian model, that the non-covariant topological term $\mathcal{L}_d$ is not gauge invariant in the presence of a non-zero magnetic monopole current density $J^\mu_{\rm mag}$. The dual field tensor

$$\widetilde{F}^{\mu\nu} = \frac{1}{2}\varepsilon^{\mu\nu\rho\sigma}F_{\rho\sigma} \tag{10}$$

then satisfies the equation

$$\partial_\mu \widetilde{F}^{\mu\nu} = J^\nu_{\rm mag} \neq 0 \; . \tag{11}$$

Let us consider the variation of the action

$$S_d = \int d^4x \mathcal{L}_d \tag{12}$$

corresponding to the topological term

$$\mathcal{L}_d = \eta_\mu \varepsilon^{\mu\nu\rho\sigma} A_\nu F_{\rho\sigma} \tag{13}$$

under the gauge transformation

$$\delta A_\mu = \partial_\mu \Lambda \; . \tag{14}$$

We obtain

$$\delta S_d = \int \eta_\mu \varepsilon^{\mu\nu\rho\sigma} (\partial_\nu \Lambda) F_{\rho\sigma} d^4x \tag{15}$$

$$= -\int \eta_\mu \varepsilon^{\mu\nu\rho\sigma} \Lambda \partial_\nu F_{\rho\sigma} d^4x \tag{16}$$

$$= -\int \eta_\mu \Lambda \partial_\nu (2\widetilde{F}^{\mu\nu}) d^4x \tag{17}$$

$$= \int 2\eta_\mu \Lambda J^\mu_{\text{mag}} d^4x \tag{18}$$

$$\neq 0 \, . \tag{19}$$

So the action S_d is not gauge invariant and is therefore forbidden if the theory contains magnetic monopoles.

We therefore drop the terms $\mathcal{L}_c$ and $\mathcal{L}_d$, leaving us with the model non-covariant Lagrangian density

$$\mathcal{L} = -\frac{1}{4}\eta^{\mu\nu\rho\sigma} F_{\mu\nu} F_{\rho\sigma} + \sum_P i\overline{\psi}_P \gamma^\alpha V_\alpha^{(P)\mu} D_\mu \psi_P \tag{20}$$

which is formally scale invariant. However, as discussed in Chapter 3.3, quantum anomalies cause a breakdown of scale invariance and require the renormalised coupling constants to depend on the energy scale λ. Thus we introduced renormalised or effective coupling constants $\eta^{\mu\nu\rho\sigma}_{\text{eff}}(\lambda)$ and $V^{(P)\mu}_{\text{eff}\alpha}(\lambda)$, which depend on the renormalisation point λ. There is a large degree of freedom in the precise definitions of these effective coupling constants. However, to leading order in the gauge coupling constant $g^2(\lambda)$, the expressions for

$$\beta_\eta^{\mu\nu\rho\sigma} = \frac{d\eta^{\mu\nu\rho\sigma}_{\text{eff}}(\lambda)}{d\log\lambda} \tag{21}$$

and the corresponding β-functions for the scale dependent vierbeins $V^{(P)\mu}_{\text{eff}\alpha}(\lambda)$ are unambiguous. Explicit calculation of the β-functions [Papers 24 and 25] shows that the effective coupling constants approach Lorentz covariant values more and more closely, as the energy scale λ is lowered.

The physically significant content of the vierbeins is provided by the associated normalised metric tensors

$$g^{(P)\mu\nu}_{\text{eff}}(\lambda) = \frac{V^{(P)\mu}_{\text{eff}\alpha}(\lambda) V^{(P)\nu}_{\text{eff}\beta}(\lambda) \eta^{\alpha\beta}}{\sqrt{|\det V^{(P)}_{\text{eff}}(\lambda)|}} \tag{22}$$

where

$$\eta^{\alpha\beta} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{pmatrix} . \tag{23}$$

Note that each Weyl fermion field $\psi_P(x)$ has its own scale dependent metric $g_{\rm eff}^{(P)\mu\nu}(\lambda)$. It is also possible to define a metric $g_{\rm eff}^{(A)\mu\nu}(\lambda)$ for the gauge field, by requiring it to minimise the expression

$$\|\eta_{\rm eff}^{\mu\nu\rho\sigma}(\lambda) - \eta_{\rm cov}^{\mu\nu\rho\sigma}(\lambda)\| \tag{24}$$

where $\|\ldots\|$ is a norm on fourth rank tensors and

$$\eta_{\rm cov}^{\mu\nu\rho\sigma}(\lambda) = \frac{1}{2g^2(\lambda)} \left[g_{\rm eff}^{(A)\mu\rho}(\lambda) g_{\rm eff}^{(A)\nu\sigma}(\lambda) - g_{\rm eff}^{(A)\mu\sigma}(\lambda) g_{\rm eff}^{(A)\nu\rho}(\lambda) \right] . \tag{25}$$

The metric $g_{\rm eff}^{(A)\mu\nu}(\lambda)$ is normalised so that

$$\det g_{\rm eff}^{(A)\mu\nu}(\lambda) = -1 \tag{26}$$

and the gauge coupling constant $g^2(\lambda)$ is free to be adjusted in the minimisation. We then denote the deviation from covariance, with respect to this metric, of $\eta_{\rm eff}^{\mu\nu\rho\sigma}(\lambda)$ by

$$\delta\eta^{\mu\nu\rho\sigma}(\lambda) = \eta_{\rm eff}^{\mu\nu\rho\sigma}(\lambda) - \eta_{\rm cov}^{\mu\nu\rho\sigma}(\lambda) . \tag{27}$$

In the case of massless non-covariant electrodynamics there are three metrics: one each for the photon field and the left-handed electron and positron Weyl fields. Introducing the notation of [Paper 24], we have the photon metric

$$g_\gamma^{\mu\nu}(\lambda) \equiv g_{\rm eff}^{(A)\mu\nu}(\lambda) \tag{28}$$

the electron metric

$$g_-^{\mu\nu}(\lambda) \equiv g_{\rm eff}^{(e^-)\mu\nu}(\lambda) \tag{29}$$

and the positron metric

$$g_+^{\mu\nu}(\lambda) \equiv g_{\rm eff}^{(e^+)\mu\nu}(\lambda) . \tag{30}$$

For Lorentz invariance, of course, we require the three metrics to be identical and the non-covariant part of the photon coupling constants $\delta\eta^{\mu\nu\rho\sigma}(\lambda)$ to vanish. It is therefore convenient to introduce the deviation metrics

$$\delta g_\pm^{\mu\nu}(\lambda) = g_\pm^{\mu\nu}(\lambda) - g_\gamma^{\mu\nu}(\lambda) . \tag{31}$$

In this notation, the renormalisation group equations become

$$\beta_{g_\gamma}^{\mu\nu}(\lambda) = \frac{\partial g_\gamma^{\mu\nu}(\lambda)}{\partial \log\lambda} \tag{32}$$

$$= -\frac{\alpha(\lambda)}{3\pi} \left[\delta g_+^{\mu\nu}(\lambda) + \delta g_-^{\mu\nu}(\lambda) \right] \tag{33}$$

and

$$\beta_{g\pm}^{\mu\nu} = \frac{\partial g_{\pm}^{\mu\nu}(\lambda)}{\partial \log \lambda} \tag{34}$$

$$= \frac{4\alpha(\lambda)}{3\pi} \delta g_{\pm}^{\mu\nu}(\lambda) \; . \tag{35}$$

Here

$$\alpha(\lambda) = \frac{1}{4\pi} g^2(\lambda) \tag{36}$$

is the running fine structure constant in non-covariant electrodynamics.

It is easily seen, from the signs in the expressions above for the β-function, that the various metrics approach each other as the renormalisation point goes towards the infrared, i.e. as the energy scale $\lambda \longrightarrow 0$. This is illustrated in Fig. 6.4, where the plane symbolizes the space of all normalised metric tensors. The development of the photon, electron and positron metrics as λ decreases, due to renormalisation group effects, is indicated by arrows. In the infrared limit all three metrics meet at the weighted average

$$\frac{1}{6}[4g_{\gamma}^{\mu\nu}(\lambda) + g_{+}^{\mu\nu}(\lambda) + g_{-}^{\mu\nu}(\lambda)] \; . \tag{37}$$

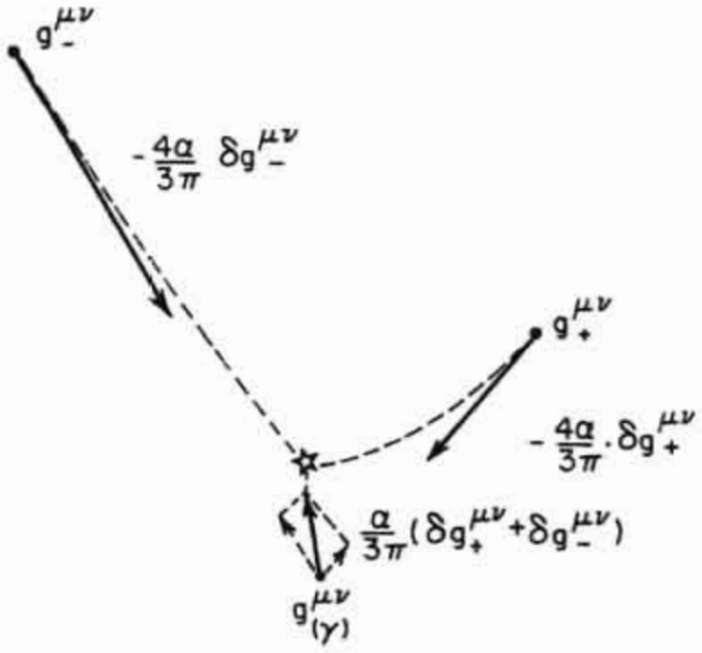


Fig. 6.4. Approach of the three metrics $g_{\gamma}^{\mu\nu}$ (λ), $g_{-}^{\mu\nu}$ (λ) and $g_{+}^{\mu\nu}$ (λ) towards the Lorentz invariant infrared fixed point, in non-covariant electrodynamics.

Also it follows, from the expression for $\beta_{\eta}^{\mu\nu\rho\sigma}$ in [Paper 24], that the non-covariant part $\delta\eta^{\mu\nu\rho\sigma}(\lambda)$ of the coupling constants tends to zero relative to the covariant part $\eta_{\text{cov}}^{\mu\nu\rho\sigma}(\lambda)$ in the infrared limit. Thus the β-functions show that Lorentz invariance is simulated more and more accurately as the energy scale is lowered.

When various non-covariant Yang-Mills fields are included in the model [Paper 25], different metrics are introduced for each simple gauge group, by minimising

expressions analogous to Eq. (24). For example, the metrics $g^{\mu\nu}_{SU(2)}(\lambda)$ and $g^{\mu\nu}_{SU(3)}(\lambda)$ would be derived from the coupling constants $\eta^{\mu\nu\rho\sigma}_{\text{eff }SU(2)}(\lambda)$ and $\eta^{\mu\nu\rho\sigma}_{\text{eff }SU(3)}(\lambda)$ for SU(2) and SU(3) gauge groups. Provided that all the various gauge groups couple to some common Weyl fermion, all the metrics approach a common limit in the infrared. The β-functions also show that the non-covariant part $\delta\eta^{\mu\nu\rho\sigma}_G(\lambda)$ of the coupling constants $\eta^{\mu\nu\rho\sigma}_{\text{eff }G}(\lambda)$, for each simple gauge group G, becomes smaller and smaller towards the infrared, relative to the covariant part $\eta^{\mu\nu\rho\sigma}_{\text{cov }G}(\lambda)$.

We have thus shown that if Lorentz invariance breaking is appreciable at some very high energy scale, the breaking will be much smaller at the low energy scale where present-day experiments are performed. We would, of course, like to assume that the coupling constants $\eta^{\mu\nu\rho\sigma}_G$ and $V^{(P)\mu}_\alpha$ of the non-covariant model are all of order unity at the Planck mass scale. The renormalisation group development of the coupling constants must then ensure Lorentz invariance, to the observed high accuracy, at low energy. This requires the β-functions discussed above to be of order unity, so that they can act effectively over a large range of energy scales. In other words we require a strongly interacting theory over many orders of magnitude of energy, so that Lorentz invariance improves at an appreciable rate as λ goes down the energy scale. The gauge group fine structure constant $\alpha_G(\lambda)$ must therefore be of order unity over a similarly large range of energy scales. This in turn requires the Yang-Mills β-function, determining the λ dependence of $\alpha_G(\lambda)$, to be small, which can be arranged by introducing an appropriate number of fermion fields. There can then be an almost exact cancellation between gauge boson and fermion contributions to the Yang-Mills β-function. However it is not possible to achieve the required degree of Lorentz invariance in this way in the standard model, for all three gauge groups U(1), SU(2) and SU(3) simultaneously, by simply introducing extra generations of quarks and leptons.[6] In particular, for the U(1) gauge particle, we would then expect Lorentz invariance breaking effects to be of order $\alpha \simeq \frac{1}{137}$, in obvious conflict with experiment. It seems that extra structure, in the form of new particles and new interactions, is required between presently available energies of 100 GeV and the Planck mass of 10^{19} GeV, if the high accuracy of Lorentz invariance is to be explained as a renormalisation group effect.

The renormalisation group method derives Lorentz invariance as an approximate low energy symmetry, in contrast to the exact formal Lorentz invariance obtained from general relativity. It is therefore natural to ask why non-covariant coupling constants, such as $\eta^{\mu\nu\rho\sigma}$, do not occur in gravitation theory. Their absence from the gravitational Lagrangian density, in fact, follows from the assumption that they should be constant over space-time, i.e. translational invariance. In general relativity, this means that they should be covariantly constant and obey

$$D_\tau \eta^{\mu\nu\rho\sigma} = 0 \,. \tag{38}$$

The Christoffel symbols $\Gamma^\rho_{\mu\nu}$, occurring in the gravitational covariant derivative D_μ, are defined to make the metric tensor $g^{\mu\nu}$ covariantly constant. It is then, in general,

impossible to satisfy Eq. (38) in a curved space-time, unless $\eta^{\mu\nu\rho\sigma}$ takes the usual covariant form proportional to $(g^{\mu\rho}g^{\nu\sigma} - g^{\mu\sigma}g^{\nu\rho})$.

6.2.3. Translational invariance from dimensional analysis

In this subsection we describe the rudiments of an argument for deriving translational invariance, as a low energy symmetry, analogous to the renormalisation group derivation of Lorentz invariance. The argument is rather formal, but we have in mind an amorphous or glass-like model for the translationally non-invariant vacuum state. A glass-like spacetime, of course, is only chaotic at short distances and has translational invariance built into its long distance structure. So our "derivation" really only shows that it is possible for macroscopic physics to be translationally invariant even though microscopic physics is not.

In a real glass, it is well known that light propagates, to a very good approximation, as if there was no breaking of translational invariance. Any lack of translational invariance manifests itself as a breakdown of momentum conservation. The essential point in our argument is that coupling constants associated with interactions which break translational invariance have, in natural units, a mass dimension lower than the corresponding momentum conserving interactions.

Let us now outline the argument, by writing the Lagrangian density $\mathcal{L}(x)$ as the sum of a translational invariant term $\mathcal{L}_{ti}(x)$ and a term breaking translational invariance $\mathcal{L}_{tb}(x)$:

$$\mathcal{L}(x) = \mathcal{L}_{ti}(x) + \mathcal{L}_{tb}(x) \ . \tag{39}$$

We imagine constructing vertices for Feynman diagrams, in momentum space, from the terms $\mathcal{L}_{ti}(x)$ and $\mathcal{L}_{tb}(x)$. A Feynman vertex constructed from $\mathcal{L}_{ti}(x)$ will contain a momentum conserving factor $\delta^4(\Sigma_i P^i_\mu)$. This δ-function will be absent from a Feynman vertex constructed from $\mathcal{L}_{tb}(x)$. The momentum conserving factor $\delta^4(\Sigma_i P^i_\mu)$ has dimension

$$\left[\delta^4\left(\sum_i P^i_\mu\right)\right] = [M^{-4}] \ . \tag{40}$$

We now consider two interaction vertices $g_{ti} \blacksquare \delta^4(\Sigma_i P^i_\mu)$ and $g_{tb} \blacksquare$ arising from $\mathcal{L}_{ti}(x)$ and $\mathcal{L}_{tb}(x)$ respectively, which have the same form $\blacksquare$ apart from the momentum conserving δ-function. The coefficients g_{ti} and g_{tb} must have different dimensions. In fact we have

$$\left[g_{ti} \blacksquare \delta^4\left(\sum_i P^i_\mu\right)\right] = [g_{tb}\blacksquare] \tag{41}$$

and hence

$$\left[\frac{g_{tb}}{g_{ti}}\right] = [M^{-4}] \ . \tag{42}$$

We now assume that the interactions derived from $\mathcal{L}_{ti}$ and $\mathcal{L}_{tb}$ are of the same order of magnitude at some fundamental mass scale m_{fund}, such as the Planck mass. We then have the simple dimensional result

$$\frac{g_{tb}}{g_{ti}} \simeq m_{\text{fund}}^{-4} \,. \tag{43}$$

For a low energy experiment, conducted at a momentum scale P, the translation symmetry breaking interactions will be suppressed, relative to the translational invariant interactions, by a factor of order P^4/m_{fund}^4 in amplitude. Taking $m_{\text{fund}} = 10^{19}$ GeV and $P = 100$ GeV, we obtain a violation of translational invariance of order

$$\frac{P^4}{m_{\text{fund}}^4} = 10^{-68} \tag{44}$$

in amplitude, and thus a suppression in the rate of a genuine momentum violating process by a factor of 10^{-136}.

The above remarks are based purely on engineering dimensions, without even considering the effects of the renormalisation group. There could be a logarithmic dependence on the energy scale coming from the renormalisation group, but this effect is completely dominated by the dependence given by Eq. (44). We therefore conclude that translational invariance will be accurately satisfied in a low energy scattering process, in which all the external particles have small momenta of order P. However once we allow energy non-conservation, typically of the order $\Delta E \simeq m_{\text{fund}}$, an initial state containing only low energy particles may lead to a final state containing some high energy particles. The amplitude for such an energy non-conserving process does not contain the suppression factor of Eq. (44), because some of the external momenta are of order m_{fund} rather than P. Thus scattering experiments with low energy particle beams would not necessarily conserve energy and momentum.

It would be natural for all final state particles to have small momenta, in experiments with low energy particle beams, if we postulate exact energy conservation. We thereby assume time translational invariance and derive just space translational invariance at low energy. In such an energy conserving but momentum non-conserving theory, the interaction vertices arising from the translational symmetry breaking interaction $\mathcal{L}_{tb}$ must contain an energy conserving δ-function. The suppression factor, Eq. (44), in the amplitude for a 3-momentum non-conserving process then becomes

$$\frac{P^3}{m_{\text{fund}}^3} = 10^{-51} \,. \tag{45}$$

Any remnant of translational invariance breaking at low energy is expected to manifest itself by a very small fraction of scattering events violating momentum conservation strongly, rather than by a tiny violation in all events. Actually such momentum non-conserving events are analogous to the Umklapp process in a crystal.[7] The amount of momentum non-conservation in an Umklapp process

must correspond to an inverse lattice vector for the crystal. Thus for a crystalline model of the vacuum, having a lattice constant of order m_{fund}^{-1}, the amount of momentum non-conservation would have to be of order m_{fund}. Since we have really failed to derive time translation invariance from dimensional analysis, let us take exact energy conservation as one of our assumptions. So we only need to latticise three-dimensional space and can take time to be continuous. For a regular lattice momentum conservation is then exact at low energy. This follows because all the particles have momenta P small compared to m_{fund}, due to energy conservation, and any violation of momentum conservation must be of order m_{fund}.

In an amorphous or glass-like model of the vacuum, the amount of momentum non-conservation in an "Umklapp reaction" could be of any size and magnitude. The phase space for momentum violation by an amount $|P| < R_P$ is proportional to R_P^3. We therefore end up with the strong suppression factor of Eq. (45), in the amplitude for a momentum non-conserving process at low energy.

In conclusion the method of dimensional analysis does not really succeed in deriving energy conservation as a long distance symmetry. We have only shown that the energy non-conserving amplitude is suppressed, provided we assume that the 4-momenta for both incoming and outgoing particles are small. However, if exact energy conservation is assumed, 3-momentum conservation results as a low energy symmetry, without explicitly assuming that the final state 4-momenta are small. Thus the method of dimensional analysis is successful in deriving space translational invariance as a low energy symmetry.

6.2.4. Lorentz invariance from string theory

Lorentz invariance can be derived from string theory, in the sense that it is a consequence of the other postulated symmetries of the underlying two dimensional conformal field theory.[8] In particular, the following four assumptions are made for the bosonic string:

(i) The 2 dimensional fields describing the string consist of a symmetric second order metric tensor $g_{\alpha\beta}(\sigma)$ and a set of real scalar fields $X^\mu(\sigma)$, which are interpreted as position coordinates in the exterior space-time.

(ii) Reparameterisation invariance under transformations of the two coordinates $\sigma = (\sigma^0, \sigma^1)$; this is the usual general relativity diffeomorphism symmetry in two dimensions.

(iii) Translational invariance in the exterior space-time under the following transformations:

$$X^\mu(\sigma) \longrightarrow X^\mu(\sigma) + a^\mu \ , \tag{46}$$

$$g_{\alpha\beta}(\sigma) \longrightarrow g_{\alpha\beta}(\sigma) \ . \tag{47}$$

(iv) Weyl invariance under the local rescaling of lengths:

$$g_{\alpha\beta}(\sigma) \longrightarrow f(\sigma) g_{\alpha\beta}(\sigma) \,, \tag{48}$$

$$X^{\mu}(\sigma) \longrightarrow X^{\mu}(\sigma) \,. \tag{49}$$

Under the reparameterisation corresponding to a scaling of the two σ co-ordinates

$$\sigma^{\alpha} \longrightarrow \lambda \sigma^{\alpha} \tag{50}$$

we have the transformation properties

$$d^2\sigma \longrightarrow \lambda^2 d^2\sigma \tag{51}$$

and

$$\frac{\partial}{\partial \sigma^{\alpha}} \longrightarrow \lambda^{-1} \frac{\partial}{\partial \sigma^{\alpha}} \,. \tag{52}$$

So the string action can only remain invariant under the scaling transformation Eq. (50), while satisfying the combined requirements of reparameterisation and Weyl invariance, if the Lagrangian density contains precisely two derivatives.

The above argument, for the presence of just two derivatives, tacitly assumes the absence of derivatives in the denominator of any term in the Lagrangian density. The justification for disallowing such denominators, consisting of a non-trivial homogeneous polynomial in derivatives, is as follows. A positive definite term in the Hamiltonian density having derivatives in the denominator, such as

$$(X')^2 = \frac{\partial X^{\mu}}{\partial \sigma^1} \frac{\partial X^{\nu}}{\partial \sigma^1} \hat{\eta}_{\mu\nu} \tag{53}$$

with $\hat{\eta}_{\mu\nu}$ some constant real symmetric matrix, gives an infinite contribution to the energy for $X' = 0$. The presence of such a term makes it energetically unfavourable to have zero derivatives, $X' = 0$, in the ground state. So the scalar fields $X^{\mu}(\sigma)$ take on σ-dependent values $X^{\mu}_{\rm vac}(\sigma)$ in the vacuum. The Hamiltonian density can then be re-expressed in terms of new shifted scalar fields.

$$\tilde{X}^{\mu}(\sigma) = X^{\mu}(\sigma) - X^{\mu}_{\rm vac}(\sigma) \tag{54}$$

which describe the deviation from the ground state. This new form of the Hamiltonian density no longer contains a denominator which is a homogeneous polynomial in derivatives. So either scale invariance is spontaneously broken violating assumption (ii), or any denominator just becomes equal to a constant.

It follows that the most general form of action, allowed under the above assumptions, is

$$\begin{aligned} I(g,\, X) = -\frac{1}{2\pi} \int d^2\sigma \Bigg[\sqrt{|\det g|} \Big(& g^{\alpha\beta}(\sigma) \frac{\partial X^{\mu}}{\partial \sigma^{\alpha}} \frac{\partial X^{\nu}}{\partial \sigma^{\beta}} \eta_{\mu\nu} \\ & + kR + g^{\alpha\beta}(\sigma) D_{\alpha} D_{\beta} X^{\mu} \xi_{\mu} \Big) \\ & + \varepsilon^{\alpha\beta} \frac{\partial X^{\mu}}{\partial \sigma^{\alpha}} \frac{\partial X^{\nu}}{\partial \sigma^{\beta}} \chi_{\mu\nu} \Bigg] \,. \end{aligned} \tag{55}$$

Here R is the two-dimensional curvature scalar, D_α is the two-dimensional gravitational covariant derivative and $\varepsilon^{\alpha\beta}$ is the antisymmetric symbol. The real number sets $\eta_{\mu\nu}, k, \xi_\mu$ and $\chi_{\mu\nu}$ are constant parameters in the two-dimensional field theory. For a Lorentz invariant action, $\eta_{\mu\nu}$ would be the metric tensor and the parameters ξ_μ and $\chi_{\mu\nu}$ would be zero.

We now argue that the non-Lorentz invariant terms in the action are of no consequence. In fact all the terms in Eq. (55), apart from the first, are total divergences and can be transformed into boundary terms by partial integration. The last term gives rise to the boundary term

$$-\frac{1}{2\pi}\int d\sigma t^\beta \frac{\partial X^\mu}{\partial \sigma^\beta}\chi_{\mu\nu}X^\nu(\sigma) \tag{56}$$

which can be interpreted as the action for a string having electric charges at its ends and moving in a constant background electromagnetic field. The boundary term arising from the third term in the action leads to inconsistent equations of motion, and this term is therefore forbidden. The boundary term arising from the second term, which is just the two dimensional Einstein action, is a topological invariant proportional to the Euler characteristic χ of the string world sheet. It contributes a factor to the functional integral for a string amplitude equal to a constant raised to the χ-th power and, therefore, it effectively renormalises the string interaction coupling constant.

We have now essentially reduced the action to the single term

$$I(g,\ X) = -\frac{1}{2\pi}\int d^2\sigma\sqrt{|\det g|}g^{\alpha\beta}(\sigma)\frac{\partial X^\mu}{\partial \sigma^\alpha}\frac{\partial X^\nu}{\partial \sigma^\beta}\eta_{\mu\nu} \tag{57}$$

where $\eta_{\mu\nu}$ is a constant metric tensor in the exterior space-time. By a linear change of basis for the scalar fields $X^\mu(\sigma)$, the metric tensor $\eta_{\mu\nu}$ can be brought into a standard diagonal form

$$\eta_{\mu\nu} = \begin{bmatrix} +1 & & & & & & \\ & \ddots & & & & & \\ & & +1 & & & & \\ & & & 0 & & & \\ & & & & \ddots & & \\ & & & & & 0 & \\ & & & & & & -1 & \\ & & & & & & & \ddots & \\ & & & & & & & & -1 \end{bmatrix} \tag{58}$$

with just $+1$'s, zeros and -1's along the diagonal. The zeros along the diagonal are easily eliminated, by discarding those linear combinations of $X^\mu(\sigma)$ fields that do not occur in the action. The last step is to argue that, in order to avoid negative norm modes on the string, there can be at most one $+1$ on the diagonal.

So, using two-dimensional reparameterisation and Weyl invariance and exterior space-time translational invariance, Weinberg[8] derives Lorentz invariance for the string. There remains just the single ambiguity in the metric, which must be either Euclidean or Minkowskian with one time co-ordinate and several space co-ordinates.

6.3. Local Gauge Invariance

We have already discussed, in Chapter 4.6.2, the origin of gauge symmetry in string theory. In this section we consider two other methods of deriving local gauge invariance.

6.3.1. Formal appearance of gauge symmetry

The gauge symmetry of Yang-Mills theory and of electrodynamics is very reminiscent of diffeomorphism symmetry in general relativity. It should therefore not be too surprising to learn that gauge invariance can be introduced as a formal symmetry, purely by definition. Quantum fluctuations may then reveal it to be a real, physically significant, gauge symmetry, for some finite region in the space of coupling constants [Paper 26]. We have already discussed examples of this phenomenon, in Chapter 4.2 and 4.3, for the G/H valued non-linear sigma model [Papers 17–19]. For the type of gauge symmetry derivation to be discussed in the present subsection, one starts with a fundamental lattice action,[9] which contains the same degrees of freedom as a gauge theory but is not gauge invariant. There can, for instance, be a latticised gauge boson mass term in the action.

As a simple example, we consider a Euclidean action for U(1) lattice electrodynamics of the form

$$S = \beta \sum_{\square} \mathrm{Re} U_{\square} + \alpha \sum_{-} \mathrm{Re} U(-) \; . \tag{59}$$

The action S is defined on a regular hypercubic Euclidean space-time lattice, with lattice spacing a. The fundamental variables are the link variables $U(-)$ which are defined on each link $-$ of the lattice and take complex values of unit norm

$$U(-) \in U(1) = \{ z | z \in \mathbb{C}, \; |z| = 1 \} \; . \tag{60}$$

The plaquette or flux variable $U_{\square}$ is not an independent variable, but is given by the product

$$U_{\square} = U(\square) U(\square) U(\square) U(\square) \tag{61}$$

of four link variables associated with the plaquette $\square$, made up of four neighbouring links. The summations $\sum_{\square}$ and $\sum_{-}$ respectively, run over all plaquettes and all links on the lattice.

An exactly invariant pure U(1) gauge theory corresponds to retaining just the first term in Eq. (59), where β is twice the inverse gauge coupling constant squared. In fact this first term is invariant under the lattice gauge transformation

$$U(x \longrightarrow x + a\delta_{\mu}) \longrightarrow \Lambda(x) U(x \longrightarrow x + a\delta_{\mu}) \Lambda^{-1}(x + a\delta_{\mu}) \; . \tag{62}$$

Here μ refers to the direction of the link $\overrightarrow{x\ x+a\delta_\mu}$ connecting the sites with coordinates x^ρ and $x^\rho + a\delta^\rho_\mu$. The other term in the action S corresponds to a photon mass term, which is not gauge invariant. The remarkable result of [Paper 26] is that, provided β is sufficiently large and α is sufficiently small but still non-zero, quantum fluctuations can generate an effectively exact gauge symmetry in this model. This gauge invariance is obtained via an inverse Higgs mechanism [Paper 19]: the massive photon is converted, by quantum effects, into a massless photon plus a massive scalar particle.

Gauge invariance is initially introduced into the model as a purely formal symmetry, obtained by expressing the action in terms of a new superfluously large set of field variables $(U_h(\bullet\longrightarrow), H(\bullet))$. These new "human" fields, $U_h(\bullet\longrightarrow)$ and $H(\bullet)$, are defined on the set of links and sites respectively and take values in the group U(1):

$$U_h(\bullet\longrightarrow) \in \mathrm{U}(1), \quad H(\bullet) \in \mathrm{U}(1)\ . \tag{63}$$

They are related to the original fundamental, or "God's", variables $U(\bullet\longrightarrow)$ by the definition

$$U(\overrightarrow{x\ x+a\delta_\mu}) = U_h^H(\overrightarrow{x\ x+a\delta_\mu}) \tag{64}$$

$$= H^{-1}(x)U_h(\overrightarrow{x\ x+a\delta_\mu})H(x+a\delta_\mu)\ . \tag{65}$$

The link variables $U(\bullet\longrightarrow)$ and $U_h(\bullet\longrightarrow)$ are strictly speaking functions of oriented links satisfying

$$U_h(\overleftarrow{x\ x+a\delta_\mu}) = U_h^{-1}(\overrightarrow{x\ x+a\delta_\mu}) \tag{66}$$

$$= U_h^*(\overrightarrow{x\ x+a\delta_\mu})\ . \tag{67}$$

The new site variable $H(\bullet)$ may be regarded as a Higgs field.

Clearly we have introduced more "human" field variables $(U_h(\bullet\longrightarrow), H(\bullet))$ than there were "God's" variables $U(\bullet\longrightarrow)$. There is thus the possibility of transforming the "human" variables around without changing $U(\bullet\longrightarrow)$. In fact the field variables $(U_h(\bullet\longrightarrow), H(\bullet))$ can be transformed under the artificial gauge transformation $\Lambda(x) \in \mathrm{U}(1)$:

$$U_h(\overrightarrow{x\ y}) \longrightarrow U_h^\Lambda(\overrightarrow{x\ y})$$

$$= \Lambda^{-1}(x)U_h(\overrightarrow{x\ y})\Lambda(y)\ , \tag{68}$$

$$H(x) \longrightarrow \Lambda^{-1}(x)H(x) \tag{69}$$

$$y = x + a\delta_\mu\ , \tag{70}$$

without changing the original field

$$U(\overrightarrow{x\ y}) = H^{-1}(x)U_h(\overrightarrow{x\ y})H(y) \tag{71}$$

$$\longrightarrow (\Lambda^{-1}(x)H(x))^{-1}\Lambda^{-1}(x)U_h(\overrightarrow{x\ y})\Lambda(y)\Lambda^{-1}(y)H(y) \tag{72}$$

$$= U(\overrightarrow{x\ y})\ . \tag{73}$$

It follows that any action $S[U(\bullet\longrightarrow\bullet)]$, which is a functional of the fundamental field $U(\bullet\longrightarrow\bullet)$ only, is automatically invariant under the formal gauge transformation of Eqs. (68)–(70). In particular the action of Eq. (59), for U(1) lattice electrodynamics, manifests this formal gauge invariance when expressed in terms of the new variables:

$$S[U_h, H] = S[U = U_h^H] \tag{74}$$

$$= \beta \sum_{\square} \mathrm{Re} U_{h\square} + \alpha \sum_{-} \mathrm{Re}[H^{-1}(x) U_h(\overset{\bullet}{x}\longrightarrow\overset{\bullet}{x} + a\delta_\mu) H(x + a\delta_\mu)] \; . \tag{75}$$

Here we have used the fact that

$$U_{h\square} = U_h(\square)U_h(\square)U_h(\square)U_h(\square) \tag{76}$$

$$= U_{\square} \tag{77}$$

which is invariant under the U(1) gauge transformation of Eq. (62).

The surprising result is that the apparently purely formal gauge symmetry, Eqs. (68)–(69), is physical enough to give rise to a phase having a massless photon.[10,11] Monte Carlo studies of the lattice gauge-Higgs action, Eq. (75), have mapped out the phase boundaries. Here we shall be content to demonstrate the existence of the Coulomb phase, and its massless photon, in a mean field approximation.[11] In mean field theory,[9] link and site variables are assumed to have expectation values, which are then determined by self-consistency conditions. These conditions are obtained by allowing a single link, or site, variable to fluctuate quantum mechanically, while replacing all fields on neighbouring links and sites by their average values. Since only fluctuations of one link or site are treated at a time, it is possible to obtain non-zero average values and to break local gauge symmetry spontaneously, without the necessity of working in a fixed gauge.[12]

We use the translational and rotational invariance of the vacuum to make the gauge-Higgs mean field ansatz:

$$\langle U_h((\overset{\bullet}{x}\longrightarrow\overset{\bullet}{x} + a\delta_\mu)\rangle = V_{Uh} \tag{78}$$

and

$$\langle H(x)\rangle = V_H \tag{79}$$

where V_{Uh} and V_H take on the same real value for each link and site respectively. The real expectation value V_{Uh} is independent of the orientation of the link, as follows from Eq. (67). The real expectation value of $H^{-1}(x)$ is V_H, as follows from the equality of $H^{-1}(x)$ and $H^*(x)$. It also follows that V_{Uh} and V_H should both have a norm less than or equal to unity.

We now introduce an effective single-link action by replacing all the variables, except for just the single-link variable $U_h(\overset{\bullet}{x}\overset{\bullet}{\;\;x} + a\delta_\mu)$ itself, by their average values in the action of Eq. (75). In d spacetime dimensions, each link has $2(d-1)$ neighbouring plaquettes and the effective single-link action

$$S_{\rm eff}[U_h(\overset{\bullet}{x}\overset{\bullet}{\;\;x} + a\delta_\mu)] = [2(d-1)\beta V_{Uh}^3 + \alpha V_H^2]{\rm Re}U_h(\overset{\bullet}{x}\overset{\bullet}{\;\;x} + a\delta_\mu) \tag{80}$$

is easily obtained. Similarly each site has $2d$ neighbouring links and we obtain the effective single-site action

$$S_{\rm eff}[H(x)] = \alpha d V_{Uh} V_H {\rm Re}\ H(x)\ . \tag{81}$$

Using the effective single-link action in the defining functional integral expression for V_{Uh}, we obtain the self-consistency condition for the mean value of the link variable

$$V_{U_h} = \frac{\int DU_h(x + a\delta_\mu)\exp\{S_{\rm eff}[U_h(\overset{\bullet}{x}\overset{\bullet}{\;\;x} + a\delta_\mu)]\}U_h(\overset{\bullet}{x}\overset{\bullet}{\;\;x} + a\delta_\mu)}{\int DU_h(\overset{\bullet}{x}\overset{\bullet}{\;\;x} + a\delta_\mu)\exp\{S_{\rm eff}[U_h(\overset{\bullet}{x}\overset{\bullet}{\;\;x} + a\delta_\mu)]\}}\ . \tag{82}$$

We now make the substitutions

$$U_h(\overset{\bullet}{x}\overset{\bullet}{\;\;x} + a\delta_\mu) = e^{i\theta} \tag{83}$$

for the link variable and

$$DU_h(\overset{\bullet}{x}\overset{\bullet}{\;\;x} + a\delta_\mu) = \frac{d\theta}{2\pi} \tag{84}$$

for the Haar measure in the U(1) group integration.[9] The self-consistency condition then takes the form

$$V_{Uh} = \frac{I_1[2(d-1)\beta V_{Uh}^3 + \alpha V_H^2]}{I_0[2(d-1)\beta V_{Uh}^3 + \alpha V_H^2]} \tag{85}$$

where $I_n(X)$ denotes the modified Bessel function

$$I_n(X) = \int_0^{2\pi} \frac{d\theta}{2\pi} e^{X\cos\theta}\cos^n\theta \tag{86}$$

and d is the dimension of spacetime ($d = 4$). Analogously the self-consistency condition for the mean value of the Higgs field becomes

$$V_H = \frac{I_1(\alpha d V_{Uh} V_H)}{I_0(\alpha d V_{Uh} V_H)}\ . \tag{87}$$

We must now look for self-consistent solutions of Eqs. (85) and (87) in the plane of the two coupling constants (α, β). Since $I_1(0) = 0$ and $I_0(0) = 1$, it immediately follows that the trivial solution

$$(V_{Uh}, V_H) = (0, 0) \tag{88}$$

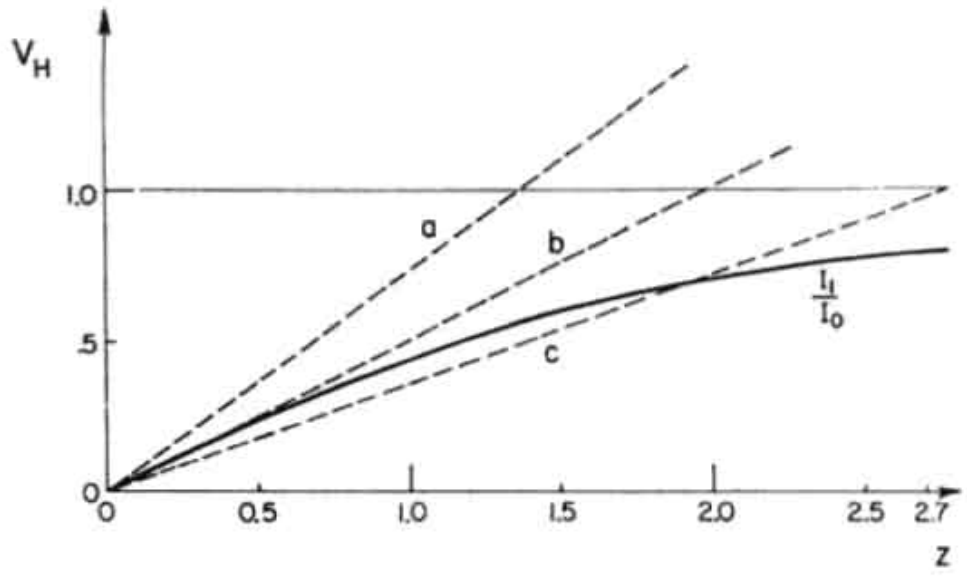


Fig. 6.5. Graphical solution of Eqs. (89) and (90) for (a) $\alpha d\ V_{Uh} < 1$, (b) $\alpha d\ V_{Uh} = 1$ and (c) $\alpha d\ V_{Uh} > 1$.

satisfies the equations for any values of α and β. In some regions of the (α, β) plane, another non-trivial solution can be obtained graphically. This is illustrated in Fig. 6.5 for the second self-consistency condition, Eq. (87), written in the form of two simultaneous equations

$$V_H = \frac{I_1(z)}{I_0(z)} , \tag{89}$$

and

$$V_H = \frac{z}{\alpha d V_{Uh}} . \tag{90}$$

The function $I_1(z)/I_0(z) \simeq z$ for small z and its slope then decreases monotonically to zero as $z \longrightarrow \infty$, where the function approaches its asymptotic value of unity. Clearly there is a solution of the simultaneous equations satisfying

$$V_H \neq 0 \quad \text{iff} \quad |\alpha d V_{Uh}| > 1 . \tag{91}$$

In particular, since $d = 4$ and $|V_{Uh}| \leq 1$, only the solution $V_H = 0$ exists when $\alpha < \frac{1}{4}$.

Similarly, when $V_H = 0$ and β is small enough, the first self-consistency condition, Eq. (85), only has the solution $V_{Uh} = 0$. Thus, for $\alpha < \frac{1}{4}$ and small β, there is only the trivial self-consistent solution of Eqs. (85) and (87) with $V_H = 0$ and $V_{Uh} = 0$. For $V_{Uh} \neq 0$, the right-hand side of Eq. (85) tends to unity as $\beta \longrightarrow \infty$. It follows that, for sufficiently large β, there is a solution with V_{Uh} a little below unity, $V_{Uh} < 1$. This non-trivial, $V_{Uh} \neq 0$, solution of Eqs. (85) and (87) will have $V_H = 0$ for $\alpha < \frac{1}{4}$ and $V_H \neq 0$ for sufficiently large α.

For each combination (α, β) of the action parameters, there is just one physical solution of the self-consistency conditions. This physical solution is the one having

the lowest free energy, i.e. having the largest value of $\log Z$ where Z is the partition function

$$Z = \int DU_h DH \exp[S(U_h, H)] \,. \tag{92}$$

The value of $\log Z$, corresponding to a certain self-consistent solution (V_{Uh}, V_H), is obtained by estimating the "entropy corrections" to the action $S(V_{Uh}, V_H)$. These "entropy-corrections" correspond to the logarithm of the $DU_h DH$-volume, which contributes significantly to Z in the mean field approximation. It turns out that the "entropy corrections" essentially depend logarithmically on α and β. For large α or β values, the action $S(V_{Uh}, V_H)$ for a non-trivial self-consistent solution will dominate the entropy contribution to $\log Z$, since $S(V_{Uh}, V_H)$ depends linearly on α or β. The non-trivial solution will therefore be realised physically for large α or β, rather than the trivial solution with zero action $S(0,0)$. In particular, for large β the physical mean field solution has $V_{Uh} \neq 0$, with $V_H = 0$ for $\alpha < \frac{1}{4}$ and $V_H \neq 0$ for large α. For small α and β there is only the trivial mean field solution, $V_{Uh} = 0$ and $V_H = 0$.

The complete mean field phase diagram for the gauge-Higgs action can be determined numerically.[11] It turns out that there are three phases in mean field approximation:

1. *Strong coupling or confined phase*, which includes the region of small α and β values and corresponds to the trivial mean field solution with $V_H = V_{Uh} = 0$.
2. *Higgs phase*, which includes the region of large α and β values and corresponds to a solution with $V_H \neq 0$ and $V_{Uh} \neq 0$.
3. *Coulomb phase*, which includes the region of large β values with $\alpha < \frac{1}{4}$ and corresponds to a solution with $V_H = 0$ but $V_{Uh} \neq 0$.

These phases are illustrated in the phase diagram of Fig. 6.6. In reality the first two phases are analytically connected;[10] the boundary separating the confinement and Higgs phases, shown as a dashed line in Fig. 6.6, terminates in the interior of the diagram and does not extend down to $\beta = 0$ in Monte Carlo lattice calculations.[11] However our main point is the existence of the Coulomb phase: a finite region in the (α, β) plane with $V_H = 0$ and $V_{Uh} \neq 0$.

We now want to argue that, in the Coulomb phase, there exist long range correlations corresponding to a massless photon, due to the formal gauge invariance. The appearance of a massless gauge particle in a theory is determined by the gauge properties of the vacuum. For the discussion of these properties, it is necessary to fix the gauge[12] and we choose the Lorentz gauge ($\partial_\mu A^\mu = 0$). In this gauge, we consider the behaviour of the vacuum expectation values V_{Uh} and V_H under (a) global gauge transformations with a constant gauge function α

$$\Lambda_G(x) = e^{i\alpha} \tag{93}$$

and (b) local gauge transformations with a linear gauge function $\alpha^\mu x_\mu$

$$\Lambda_L(x) = e^{i\alpha^\mu x_\mu} \,. \tag{94}$$

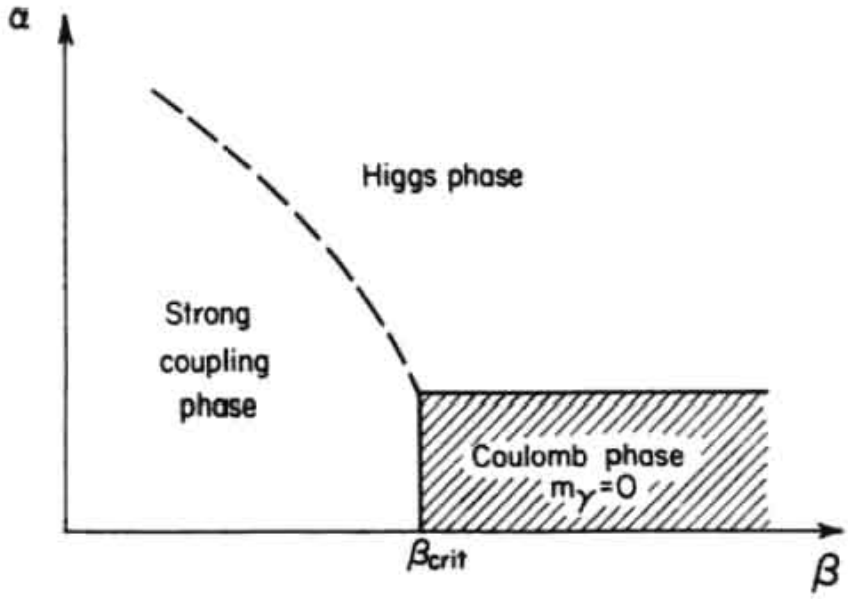


Fig. 6.6. Schematic phase diagram for the U(1) gauge-Higgs model.

It follows from Eqs. (68)–(70) that under the global gauge transformation

$$\Lambda = \Lambda_G : \quad V_{Uh} \longrightarrow V_{Uh}, \quad V_H \longrightarrow e^{-i\alpha} V_H \tag{95}$$

and under the gauge transformation with a linear gauge function

$$\Lambda = \Lambda_L : V_{Uh} \longrightarrow e^{i\alpha^\mu a_\mu} V_{Uh}, \quad V_H \longrightarrow e^{-i\alpha^\mu x_\mu} V_H \ . \tag{96}$$

Here a_μ is the appropriate lattice link vector.

The gauge symmetry properties of the vacuum, under Λ_G and Λ_L, are readily deduced from Eqs. (95)–(96) for the phases found in mean field approximation, and are presented in Table 6.2.

Table 6.2. Gauge symmetry properties of vacuum phases in the U(1) gauge-Higgs model.

Gauge Transformation / phase	Constant Gauge Function $\Lambda_G(x) = e^{i\alpha}$	Linear Gauge Function $\Lambda_L(x) = e^{i\alpha^\mu x_\mu}$
Confined	Invariant Vacuum $V_H = 0$	Invariant Vacuum $V_{Uh} = V_H = 0$
Higgs	Non-invariant Vacuum $V_H \neq 0$	Non-invariant Vacuum $V_{Uh} \neq 0$, $V_H \neq 0$
Coulomb	Invariant Vacuum $V_H = 0$	Non-invariant Vacuum $V_{Uh} \neq 0$

Massless gauge particles (photons) can be considered as Nambu-Goldstone bosons, accompanying the spontaneous breakdown of the symmetry under gauge transformations with a linear gauge function.[13] However the proof of the Nambu-Goldstone theorem makes essential use of translation invariance. It requires the spontaneously broken symmetry generator to effectively commute with the momentum operator P_μ: the commutator must annihilate the vacuum. The commutator in question is

$$[P_\mu, Q_\nu] = -ig_{\mu\nu}Q \tag{97}$$

where Q_ν is the generator of gauge transformations with a linear gauge function $\Lambda_L(x)$ and Q is the generator of global gauge transformations Λ_G. The generators Q and Q_μ can be defined in terms of an expansion of the generator $Q[\lambda]$ of a general infinitesimal gauge transformation; $Q[\lambda]$ is a functional of the corresponding infinitesimal gauge function $\lambda(x)$. Expanding the arbitrary infinitesimal gauge function $\lambda(x)$ in a Taylor series

$$\lambda(x) = \alpha + \alpha^\mu x_\mu + \alpha^{\mu\nu} x_\mu x_\nu + \dots \tag{98}$$

we obtain the required expansion of $Q[\lambda]$:

$$Q[\lambda] = \alpha Q + \alpha^\mu Q_\mu + \alpha^{\mu\nu} Q_{\mu\nu} + \dots \ . \tag{99}$$

The existence of a massless photon as a Nambu-Goldstone boson thus requires the generator Q to annihilate the vacuum. The following two conditions must therefore be satisfied by a phase containing a massless photon:

(i) The gauge symmetry for linear gauge functions must be spontaneously broken, i.e. the vacuum should not be invariant under gauge transformations with a linear gauge function $\Lambda_L(x)$.

(ii) The global part of the gauge symmetry must not be spontaneously broken, i.e. the vacuum should be invariant under gauge transformations with a constant gauge function $\Lambda_G(x)$.

It is immediately seen from Table 6.2 that the Coulomb phase, with $V_{Uh} \neq 0$ and $V_H = 0$, satisfies both conditions and therefore must contain a massless photon.

We have thus seen that the dynamics of the gauge non-invariant action, Eq. (59), for U(1) lattice electrodynamics can generate an exact massless gauge symmetric theory, without fine-tuning of parameters. We will now briefly discuss the generalisation of this result to a non-Abelian lattice gauge theory, with explicit gauge symmetry breaking terms in the action. A very weakly coupled non-Abelian gauge theory is approximately equivalent to a field theory with several Abelian types of gauge particle. So, in this approximation, we should expect a phase of the theory to exist possessing an exact gauge symmetry. However we can hardly expect the existence of a Coulomb phase with physical massless gauge particles; a non-Abelian theory should exhibit confinement,[9] unless the number of matter fields is sufficiently

large to violate asymptotic freedom. In such a confining phase there will, ignoring matter fields, be no massless states but only glueball states, i.e. gauge group singlet bound states of the gauge field. The glueball mass scale arises by dimensional transmutation,[9] in an analogous way to the strong interaction mass scale Λ_{QCD} in the standard model. This mass scale is exponentially suppressed as a function of the inverse gauge coupling constant squared, β, defined at some fundamental lattice scale.

A non-Abelian generalisation of the gauge symmetry violating lattice action of Eq. (59) is easily constructed, by requiring the link variable $U(-)$ to take values in the non-Abelian group. As for the U(1) case, a new set of "human" variables can be introduced, including a Higgs-like field $H(\bullet)$ which takes values in the group manifold. Block-spinning[14] leads to a more coarse-grained field theory, where $H(\bullet)$ can presumably be approximated by a field taking values in a small faithful representation of the group, such as the fundamental representation. The action expressed in terms of the "human" gauge-Higgs variables automatically possesses a formal gauge invariance.

In the non-Abelian theory there is no region, in (α, β) parameter space, corresponding to a Coulomb phase. The phase diagram for a lattice non-Abelian gauge-Higgs model is expected to have the form shown in Fig. 6.7, which has been verified by Monte Carlo studies for SU(2) and SU(3). At large values of β, there is a phase boundary between the small α confinement phase and the large α spontaneously broken Higgs phase. The phase boundary terminates in the interior of the diagram: the confinement and Higgs phases are analytically connected[10] and really form just one big phase without any massless particles. However as β becomes larger and larger, for small enough α, we enter what would be the Coulomb phase in the Abelian case. In this region, shaded in Fig. 6.7, the glueball mass tends towards zero in an exponential way:

$$m_g \approx \frac{1}{a} \exp(-c\beta) \tag{100}$$

as β tends to infinity, where a is the lattice spacing and c is a positive constant of order unity. Long distance physics in the lattice theory is then well approximated by a conventional gauge symmetric continuum Yang-Mills theory. In this sense the appearance of gauge symmetry "out of nothing" also works for a non-Abelian group.

The assumption of a regular lattice is not essential for the success of the inverse Higgs mechanism. We may equally well consider a random discretised model, with a basic set of fixed sites distributed randomly in four-dimensional space-time. Indeed the mechanism is expected to work for a very general type of field theory. For instance it should work in a field theory glass[15] where the parameters, such as coupling constants, and even the type and number of degrees of freedom vary from place to place in space-time in a quenched random way: they are chosen once and for all and not varied in the functional integral implementing quantum mechanics. It is because of these frozen-in parameters, being analogous to the atomic binding

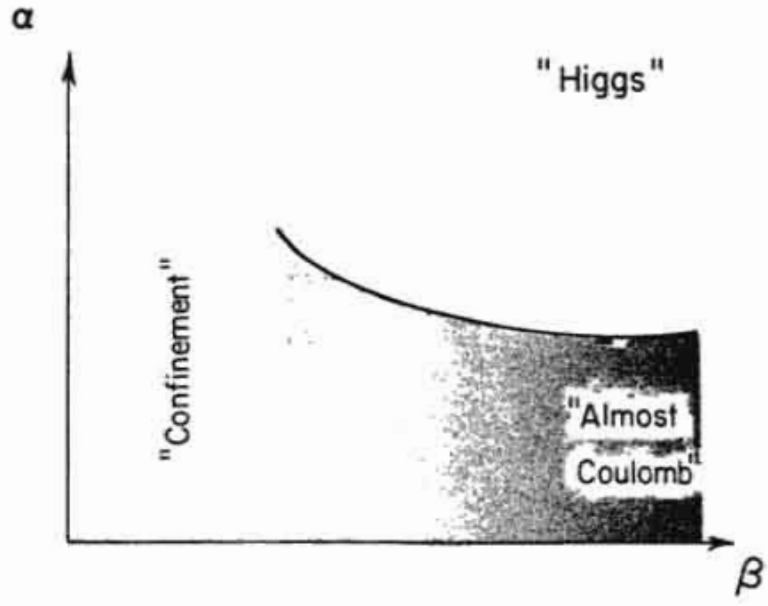


Fig. 6.7. Schematic phase diagram for the non-Abelian gauge-Higgs model.

structure in a real glass, that we use the term "field theory glass". The action for a field theory glass is not even strictly local at the fundamental lattice scale, which we take to be the Planck length. The field theory glass idea is used in the formulation of the random dynamics project discussed in Chapter 7. Here we use it to demonstrate the very general nature of the inverse Higgs mechanism and the spontaneous appearance of gauge symmetry.

The construction of a field theory glass model is most easily visualised in terms of an imagined Monte Carlo computer simulation. The computer program first selects a randomly distributed set of points, or sites, $\{i\}$ in four-dimensional space-time. Next it follows an algorithm for constructing, by means of random numbers, a manifold M_i at each of the points i on the random lattice. The manifold M_i, somewhat unusually, changes randomly from point to point. The fundamental dynamical variable is a generalised quantum field $\phi(i)$ which maps each point i onto the corresponding manifold M_i. The parameters of a very general semilocal action S are in turn chosen as random numbers. Thereby one finally has a field theory glass model, with random degrees of freedom $\phi(i)$ and random action, whose long wavelength properties can be studied by Monte Carlo methods. Our claim is that the resulting physical degrees of freedom, with the most long range correlations, are similar to the gauge degrees of freedom obtained in the translational invariant lattice models discussed above. This means that it should be possible to introduce some new "human" variables, analogous to U_h and H of Eqs. (64) and (65), and hence an exact formal gauge invariance. Then some of this formal gauge symmetry is realised physically, in the sense that it gives rise to massless gauge particles or to low energy confinement. We outline below how such an exact gauge symmetry may be generated in a very general type of field theory glass.

We take the field theory glass semilocal action to be a sum over contributions S_r of quenched random form, from very many overlapping small regions r of space-

time, of the order of the Planck size:

$$S[\phi] = \sum_r S_r(\phi(i)), \quad i \in r \,. \tag{101}$$

This action is of course not Poincaré invariant. We assume here that Poincaré symmetry arises as a consequence of one of the mechanisms discussed in Sect. 6.2, and concentrate on the origin of gauge symmetry in the field theory glass. For this purpose, we imagine that a small demon goes through the random theory, small region after small region, looking for approximate local symmetries of the action. These accidental local symmetries correspond to transformations of just the degrees of freedom in some limited part of space-time, in general not coinciding with a region r of Eq. (101), which leave all contributions to the action approximately invariant. Such a small part or "ball" of space-time, which in general is not spherical and is of the order of the Planck size, will be called a "gauge ball". The demon selects a site s as the centre of the gauge ball and identifies the approximate local gauge symmetry group $G(s)$ within the gauge ball $B(s)$. He passes through all of space-time and, in general, finds a large density of strongly overlapping gauge balls; at least let us imagine so.

The approximate local symmetry groups $G(s)$, within each gauge ball $B(s)$, are now converted into formally exact gauge symmetries by our procedure of introducing a superfluously large set of "human" variables (ϕ_h, H). The generalised Higgs field H is defined on the centres s of the gauge balls and $H(s)$ belongs to the local symmetry group $G(s)$. The field ϕ_h is a new human variable, with the same structure as the fundamental field ϕ, and is defined by the equation

$$\phi(i) = \phi_h^H(i) \,. \tag{102}$$

Here ϕ^Ω denotes the result of transforming ϕ with an element Ω from the direct product of all the local symmetry groups $G(s)$. At a particular site i, of course, the field $\phi_h(i)$ can only transform non-trivially under those groups $G(s)$ for which the gauge ball $B(s)$ contains i.

The original action $S[\phi]$ can be written in terms of the new variables

$$S[\phi_h, H] = S[\phi = \phi_h^H] \,. \tag{103}$$

This new action $S[\phi_h, H]$ is automatically invariant under the formal gauge transformation

$$\phi_h \longrightarrow \phi_h^\Omega \tag{104}$$

$$H \longrightarrow \Omega^{-1} H \tag{105}$$

since the fundamental field ϕ is unaltered by this artificial transformation

$$\phi = \phi_h^H \longrightarrow (\phi_h^\Omega)^{\Omega^{-1}H} = \phi_h^H = \phi \,. \tag{106}$$

The field theory glass, expressed in terms of (ϕ_h, H), thus manifests a formal gauge symmetry of a chaotic nature, in that the symmetry group changes randomly from place to place.

In general, for any group K, we expect to find gauge balls distributed throughout space-time, within which K is an approximate gauge symmetry of the fundamental action $S[\phi]$. We are, of course, interested in the situation where the associated Higgs field H, in the new action $S[\phi_h, H]$, has no long range correlations but strong local fluctuations; so that its "average value" is zero in the vacuum. (In order to define such an "average value", the group K has to be embedded in its convex closure.) Thus we restrict our attention to gauge balls within which the symmetry breaking terms of the fundamental action are rather small, corresponding to the small α situation in our earlier lattice gauge-Higgs examples. The symmetry breaking terms in the fundamental notation are transformed into correlation terms, between the Higgs field at different points, in the human variable notation. For sufficiently small symmetry breaking, it follows from general decay correlation theorems[16] that the correlation function between the Higgs field at two points, $H(x)$ and $H(y)$, does indeed fall off exponentially with the separation. So we now assume that the vacuum expectation value $\langle H \rangle = 0$, as is needed for the invariance of the vacuum under global transformations.

We now wish to introduce a continuum gauge field $A^a_\mu(x)$ for the group K, with generators $\lambda^a/2$, describing small deviations from the globally gauge invariant vacuum state. Therefore we need a rule to implement the modification of the vacuum state described by a smoothly varying continuum Yang-Mills field $A^a_\mu(x)$. So let us consider the human field variable $\phi_h(i)$ at an arbitrary site i in the vacuum state. We now locate all the gauge balls which contain i and belong to the restricted set, with sufficiently small K-symmetry violating terms in the fundamental action to satisfy the conditions of the decay correlation theorems. We may consider the continuum field $A^a_\mu(x)$ to be essentially constant over all the gauge balls containing the site i, having coordinates x^μ_i. The modification of the field $\phi_h(i)$, relative to its vacuum value, is then constructed from the vacuum value, by applying the linear gauge transformation which would set up a constant continuum field with the value $A^a_\mu(x_i)$. In other words, we apply a series of gauge transformations

$$\Lambda_s(x_i) = \exp\left[iA^a_\mu(x_i)\frac{\lambda^a}{2}(x^\mu_s - x^\mu_i)\right] \tag{107}$$

to $\phi_h(i)$, one after the other; one for each gauge ball containing x_i. The coordinates x^μ_s are those of the gauge ball centre s. In this way we modify the field ϕ_h at each site and set up a new configuration corresponding to the continuum gauge field $A^a_\mu(x)$.

The gauge transformation $\Lambda_s(x_i)$ depends on the choice of gauge ball centre x^μ_s. However, this dependence is simply equivalent to a gauge transformation corresponding to the gauge ball at s *independent* of x_i, i.e. it is the same for every field variable $\phi_h(i)$. Here again we make use of the fact that the continuum Yang-Mills

field $A_\mu^a(x)$ is smoothly varying. So, in the absence of any physical dependence upon the choice of x_s, the modification corresponding to the continuum gauge field $A_\mu^a(x)$, as given by Eq. (107), is well defined up to a microscopic gauge transformation. The latter has only formal significance. In any event the centre x_s should of course be close to the sites involved in the gauge transformation $\Lambda_s(x_i)$ and is therefore rather well defined. It can happen that two or more isomorphic groups K are found as subgroups in $G(s)$. In order to avoid ambiguity in this case, it is necessary to apply the same gauge transformation $\Lambda_s(x_i)$ for them all; thus we only consider a continuum gauge field for the diagonal subgroup. There is a similar ambiguity in setting up a continuum gauge field for any group with an outer automorphism;[17] complex conjugation is an outer automorphism for SU(N) groups with $N > 2$. Our construction of a continuum gauge field fails for a group with an outer automorphism; it can however be rescued if the field theory glass representations of the group are not symmetric under the automorphism. We note that, although the standard model group $S(\mathrm{U}(2)\times\mathrm{U}(3))$ is symmetric under complex conjugation, the chiral quark and lepton representations are not.

By the above construction, for each smoothly varying continuum field configuration $A_\mu^a(x)$, we assign a new configuration of the field theory glass. The action can be evaluated for each new configuration and it must be invariant under the gauge group K. This follows from the natural correspondence between a gauge transformation on the continuum field $A_\mu^a(x)$ and a gauge transformation on the field theory glass variables $\phi_h(i)$. The effective action for the continuum field $A_\mu^a(x)$ is therefore invariant under K gauge group transformations. In this way we have apparently set up a Yang-Mills field theory for an arbitrary group K without outer automorphisms. With outer automorphisms, our "demon" would have to make arbitrary choices all over space-time, or else one would have to rely on matter fields to specify which way to implement the group.

In order to avoid spontaneous symmetry breaking of the gauge group K by the Higgs field H, it was necessary to make a careful selection of gauge balls; so there may be only rather few gauge balls selected. In this case the majority of sites i do not lie inside any gauge ball, let alone inside several. It follows that most of the field variables $\phi_h(i)$ transform trivially under gauge transformations and are unchanged by setting up a continuum gauge field $A_\mu^a(x)$. The action is then very slowly varying as a function of $A_\mu^a(x)$; this means that the vacuum is in a strong coupling phase, corresponding to the small β situation in our earlier lattice gauge-Higgs examples. The gauge field $A_\mu^a(x)$ is thus confined at the Planck scale and the gauge symmetry is not realised physically in the long wavelength properties of the theory.

Clearly several conditions must be satisfied before a formal gauge symmetry group K of a field theory glass manifests itself physically, as a conventional gauge theory at large distances. In particular the fundamental action $S[\phi]$ must manifest the symmetry to a good approximation over a large fraction, locally, of the field theory glass, and the coefficients of the non-trivial symmetry conserving terms should be large. The chances of a field theory glass satisfying these requirements,

by accident, are better the smaller the number of degrees of freedom involved. Thus invariance groups K with a few generators and small representations are favoured to show up with massless gauge bosons or low energy confinement. The representations of the standard model gauge group $S(\mathrm{U}(2)\times\mathrm{U}(3))$ are indeed small. For the purposes of this discussion, we may consider the $W^{\pm}$ and Z^0 particles to be massless gauge bosons, since their masses are negligible at the Planck scale. We postpone, to Chapter 7, further discussion of the conditions which a gauge group should satisfy, in order to survive from the fundamental scale down to the low energy scale.

We conclude that low energy gauge symmetry can arise spontaneously in a very general class of theories, although we do not have a complete proof of the effect. In principle we should make real Monte Carlo computer calculations for a field theory glass, in order to verify our speculations and to find out which gauge groups survive at long wavelengths.

It has been emphasized [Paper 27] that the gauge symmetries of a theory in the infrared must also be symmetries in the ultraviolet, in order to avoid non-renormalisable ultraviolet divergences. If a gauge symmetry is broken at the fundamental cut-off scale, the divergences arising from loop diagrams with internal vector bosons will not cancel. For instance, the phenomenological successes of the standard model would be destroyed, if the gauge invariance of the model were broken at the fundamental lattice scale. At first sight, this infrared-ultraviolet connection is a disaster for the inverse Higgs mechanism as the origin of low energy gauge symmetries. However the physical gauge symmetry, arising from the inverse Higgs mechanism, is a low energy manifestation of a formal gauge invariance, which is exact in both the infrared and the ultraviolet; the formal gauge invariance is purely a consequence of the definition of the human variables (ϕ_h, H). Thus the formal appearance of gauge symmetry is not in conflict with Veltman's result.

The infrared-ultraviolet connection does, however, pose a threat to a renormalisation group derivation of gauge invariance at an infrared fixed point. The renormalisation group approach to the origin of gauge symmetry is the subject of the next subsection.

6.3.2. Gauge symmetry from the renormalisation group

The perturbative renormalisation group method has already been used to derive Lorentz invariance as a low energy symmetry in Sect. 6.2.2. Here we consider its application to gauge invariance [Paper 28] and also to combined gauge and Lorentz invariance [Paper 29].

It is clear, on dimensional grounds, that the vanishing of the photon mass term cannot be derived perturbatively in the infrared limit. The coefficient, $\frac{1}{2}\mu^2$ say, of any photon mass term $\frac{1}{2}\mu^2 A_\mu^2$ in the Lagrangian density has the dimension of $(\mathrm{mass})^2$, and actually gets more and more important in the infrared limit. It is necessary to appeal to a non-perturbative renormalisation group calculation, in order to derive the zero photon mass. Indeed the U(1) lattice electrodynamics Monte Carlo calculation,[11] mentioned in the previous subsection, can be interpreted,

in terms of blockspinning,[14] as successively producing variables at longer and longer distance scales. In this subsection we shall, therefore, set the photon mass equal to zero and study other gauge symmetry violating terms perturbatively. Actually we shall set all mass terms to zero, in order to investigate the far infrared

Abelian gauge symmetry is found to be an infrared attractive fixed point for quantum electrodynamics (QED) with a spinor field (the electron say); while for QED with a scalar field it is not [Paper 28]. In order to avoid the difficulty with renormalisability [Paper 27], the gauge invariant theory is approached from a class of theories quantised in a Hilbert space with an indefinite metric. This use of an indefinite metric Hilbert space is analogous to that in the Gupta-Bleuler formulation[18] of gauge invariant QED. However it is introduced at the cost of losing the physical interpretation of the theories away from the gauge symmetric limit, since the negative norm squared states do not then decouple. An alternative way of avoiding the problem with renormalisability, without introducing negative norm squared states, is to consider a class of theories which violate Lorentz invariance as well as gauge invariance [Paper 29]. It can then be arranged that $A_0 = 0$, so that the theories are formulated in a Hilbert space with a positive definite metric.

The Lagrangian density for gauge symmetry violating, but Poincaré, parity and charge conjugation symmetry conserving, spinor QED is

$$\begin{aligned}\mathcal{L} =&\overline{\psi}(i\partial - m)\psi - \frac{1}{4}(\partial_\mu A_\nu - \partial_\nu A_\mu)^2 + \frac{1}{2}\mu^2(A_\mu A^\mu) \\ &+ \xi(\partial_\mu A^\mu)^2 - e\overline{\psi}\gamma_\mu\psi A^\mu - \frac{1}{4}g(A_\mu A^\mu)^2 \ . \end{aligned} \tag{108}$$

As mentioned above, we set $m = \mu = 0$ and consider the behaviour of the dimensionless coupling parameters e, g and ξ in the infrared. Gauge invariant QED corresponds to the limit $g = \xi = 0$, and so we work to lowest order in the gauge symmetry breaking parameters g and ξ. The one-loop renormalisation group β functions then come entirely from the wave function renormalisation of the electromagnetic potential A_μ and are given by

$$\pi^2\beta_{e^2} = \frac{1}{6}e^4 \ , \tag{109}$$

$$\pi^2\beta_g = \frac{1}{3}e^2 g \ , \tag{110}$$

$$\pi^2\beta_\xi = \frac{1}{6}e^4\xi \ , \tag{111}$$

These expressions imply that the running coupling parameters g and ξ get smaller and smaller towards the infrared; gauge invariant spinor QED is an infrared attractive fixed point. The dimensionless coupling parameters approach their gauge symmetric values logarithmically; the approach is too slow to realistically claim that it may account for the observed high accuracy of gauge symmetry in quantum electrodynamics. This problem with the accuracy of the derived low energy symmetry, in the renormalisation group method, is discussed in Sect. 6.2.2.

We note here that in [Paper 28] the term $\xi(\partial_\mu A^\mu)^2$ is ignored, on the grounds of it just fixing the gauge in the symmetric limit, and a term of order g^2 is included in β_g. The term $\xi(\partial_\mu A^\mu)^2$ is considered in [Paper 29], using a Lorentz non-covariant formalism. It is pointed out that ξ and the coefficient ζ of a Lorentz symmetry breaking term go to zero in the infrared; gauge symmetry and Lorentz symmetry are simultaneously recovered at the infrared attractive fixed point.

Scalar field QED however does not have a gauge symmetric infrared fixed point. In fact it is shown in [Paper 28] that the running coefficient α of the seagull term, $\alpha A_\mu A^\mu \phi^+ \phi^-$, does not go towards the gauge symmetric value, e^2, in the infrared limit. Rather the parameter $\delta_1 = \alpha/e^2 - 1$, measuring the deviation from this gauge symmetric value, has the one-loop β function

$$\beta_{\delta_1} = -e^2 \left(3\delta_1 + \frac{9}{2}\delta_2 \right) \tag{112}$$

in the extreme infrared limit. Here δ_2 is the measure of another gauge symmetry violating term which does go to zero in the infrared, as does the coefficient λ of the term $-\frac{1}{4}\lambda(\phi^+\phi^-)^2$ in the Lagrangian density. It follows, from the sign of the β function in Eq. (112), that δ_1 does not go to zero and that the gauge symmetric theory is infrared unstable.

It is well known that non-Abelian gauge theories are asymptotically free,[18] unless a sufficiently large number of matter fields are included. Asymptotic freedom means that the wave function renormalisation group effect of the precursor gauge field, in a non-Abelian theory with gauge symmetry breaking terms, has the opposite sign to that in an Abelian theory. The vector boson wave function renormalisation is the dominant effect, responsible for achieving gauge symmetry in the infrared for spinor QED. So it should not be surprising that gauge symmetry is an infrared repulsor for all non-Abelian groups.

A general non-Abelian theory, with both gauge and Lorentz symmetry violating terms, is considered in [Paper 29]. The Lorentz and gauge invariant fixed point is explicitly shown to be infrared repulsive, in the absence of matter fields. If sufficiently many fermion species are introduced that the β function for the precursor Yang-Mills gauge coupling constant changes sign, then both Lorentz and gauge symmetry are obtained in the low energy limit.

By approaching the infrared fixed point via a class of theories which violate Lorentz invariance, the threat posed to renormalisability by the infrared-ultraviolet connection is avoided. However asymptotic freedom is a property of QCD in the standard model (with up to 16 quark flavours). So we must conclude that the renormalisation group approach fails to explain the origin of Yang-Mills gauge symmetry in the standard model.

(*Note added in proof*: We should also mention an alternative attempt[21] to derive gauge symmetry based on the idea that some scalar fields — the so-called "spoiler fields" — would adjust themselves so as to make the effective gauge symmetry breaking terms vanish.)

6.4. Supersymmetry

The renormalisation group method has also been used to study the stability of supersymmetry as a low energy symmetry. The infrared stability of $N = 1$ global supersymmetry was first investigated by Curtright and Gandour,[19] who found that in general a supersymmetric infrared attractive fixed point does not exist.

A simple example, for which global $N = 1$ supersymmetry is an infrared attractor, is presented in [Paper 28]. The model involves a Majorana spinor $\psi(x)$, a scalar $A(x)$ and a pseudoscalar $B(x)$. The class of theories considered is described by the Lagrangian density

$$\mathcal{L} = -\frac{1}{2}(\partial_\mu A)^2 - \frac{1}{2}(\partial_\mu B)^2 - \frac{1}{2}i\overline{\psi}\partial\!\!\!/\psi - if\overline{\psi}\psi A$$
$$- if\overline{\psi}\gamma_5\psi B - \frac{1}{2}\lambda(A^2 - B^2)^2 - gA^2B^2 \ . \tag{113}$$

This, of course, is not the most general form possible for the Lagrangian density; there could have been different coefficients multiplying the scalar $\overline{\psi}\psi A$ and pseudoscalar $\overline{\psi}\gamma_5\psi B$ Yukawa terms, and also different coefficients on the A^4 and B^4 terms. As in our previous applications of the perturbative renormalisation group, we have ignored mass terms. The supersymmetric limit is then obtained for $\lambda = g = f^2$.

It is convenient to use the parameters

$$\delta_1 = \frac{\lambda - g}{f^2} \tag{114}$$

and

$$\delta_2 = \left(\frac{\lambda}{f^2} - 1\right) + \frac{1}{5}\left(\frac{g}{f^2} - 1\right) \tag{115}$$

to measure the deviations from supersymmetry. The corresponding β functions behave like:

$$\beta_{\delta_1} \sim \frac{3}{2}\delta_1 \tag{116}$$

and

$$\beta_{\delta_2} \sim \frac{9}{2}\delta_2 \tag{117}$$

close to the supersymmetric situation. The signs of the β functions imply that the supersymmetric fixed point is an infrared attractor. Thus we have here an example of deriving $N = 1$ global supersymmetry as a low energy approximation.

The renormalisation group method cannot be applied to locally supersymmetric theories, due to the requirement of renormalisability. The largest number N of spinorial supersymmetry generators, for which a globally supersymmetric theory can be constructed is $N = 4$. It is shown in [Paper 29] how an $N = 4$ global supersymmetry may appear as an infrared attractor, once $N = 1$ supersymmetry is imposed. A generic $N = 1$ supersymmetric Yang-Mills theory, with gauge group

G, is investigated, which has the following field content: a gauge superfield V and three chiral superfields $\Phi^i, i = 1, 2, 3$, each being in the adjoint representation of G.

If the gauge group G is a so-called safe group, for which

$$\mathrm{Tr}(\Phi^i\{\Phi^j, \Phi^k\}) = 0 \tag{118}$$

is an identity, then $N = 4$ supersymmetric Yang-Mills theory is an infrared attractor. If, however, G is one of the groups $\mathrm{SU}(n), n \geq 3$, or E_6, which are not safe, then in general $N = 4$ supersymmetry is not infrared stable. Infrared stability can be obtained by imposing an SO(3) symmetry, under which the chiral superfields rotate into each other. This SO(3) symmetry serves to prevent the presence of a $\mathrm{Tr}(\Phi^i\{\Phi^j, \Phi^k\})$ term in the Lagrangian density. However it seems unreasonable to introduce an SO(3) symmetry solely for this purpose, without any physical justification.

It is perhaps satisfactory that $N = 4$ supersymmetry is infrared unstable, since its presence as a low energy symmetry would forbid the existence of the chiral quark and lepton representations of the standard model.[20] Indeed there is no direct phenomenological evidence for the existence of even $N = 1$ supersymmetry, although it might be technically useful in understanding why the mass scale of the standard model is so much less than the Planck mass.[20] The example derivation of $N = 1$ supersymmetry given above is really only a very partial derivation, due to the special form of Lagrangian density assumed. It is tempting to interpret this failure of the renormalisation group method, to really derive any supersymmetry in the infrared limit, as evidence against the existence of supersymmetry as a low energy symmetry. However this would be dangerous, since the renormalisation group method also failed to derive non-Abelian gauge symmetry as a low energy symmetry.

References

1. C.W. Misner, K.S. Thorne and J.A. Wheeler, *Gravitation*, (W.H. Freeman, 1973).
2. J.L. Anderson, *Principles of Relativity Physics*, (Academic Press, 1967); H.C. Ohanian, *Gravitation and Spacetime*, (W.W. Norton, 1976).
3. M. Lehto, H.B. Nielsen and M. Ninomiya, *Nucl. Phys.* **B272**, 213 and 228 (1986).
4. A.B. Borisov and V.I. Ogievetskii, *Teor. Mat. Fiz.* **21**, 329 (1974).
5. M. Lehto, H.B. Nielsen and M. Ninomiya, *Phys. Lett.* **219B**, 87 (1989).
6. J. Ellis, M.K. Gaillard, D.V. Nanopolous and S. Rudaz, *Nucl. Phys.* **B176**, 61 (1980).
7. C. Kittel, *Introduction to Solid State Physics*, (Wiley, 1986).
8. S. Weinberg, *Proc. of the XXIII Int. Conf. on High Energy Physics*, Berkeley, 1986, (World Scientific, 1987), p. 271.
9. M. Creutz, *Quarks, Gluons and Lattices*, (Cambridge University Press, 1983).
10. E. Fradkin and S.H. Shenker, *Phys. Rev.* **D19**, 3682 (1979).
11. J. Ranft, J. Kripfganz and G. Ranft, *Phys. Rev.* **D28**, 360 (1983).
12. S. Elitzur, *Phys. Rev.* **D12**, 3978 (1975).
13. R. Ferrari and L.E. Picasso, *Nucl. Phys.* **B31**, 316 (1971); E.A. Ivanov and V.I. Ogievetskii, *JETP Lett.* **23**, 606 (1976).
14. K.G. Wilson and J. Kogut, *Phys. Reports* **12**, 75 (1974).
15. H.B. Nielsen and N. Brene, *Workshop on Skyrmions and Anomalies*, (World Scientific, 1987), p. 493.
16. L. Gross, *Commun. Math. Phys.* **68**, 9 (1979); M. Lehto, H.B. Nielsen and N. Ninomiya, *Commun. Math. Phys.* **93**, 483 (1984).

17. R. Gilmore, *Lie Groups, Lie Algebras and some of their Applications*, (Wiley, 1974).
18. G. Itzykson and J.B. Zuber, *Quantum Field Theory*, (McGraw-Hill, 1980).
19. T. Curtright and G. Ghandour, *Ann. Phys.* **106**, 209 (1977); **112**, 237 (1978).
20. E. Witten, *Nucl. Phys.* **B188**, 513 (1981).
21. K. Cahill and P. Denes, *Nuovo Cim. Lett.* **33**, 184 (1982).

Chapter VII

CONCLUSION

In this book we have collected together several examples of cases in which symmetries of the laws of nature can be derived; the symmetries are deduced under some assumptions not involving, too openly at least, the assumption of the symmetry in question. We begin this chapter with a status report on how many of the known symmetries can be derived. This provides a motivation for the random dynamics project reviewed in the second section. In the final section we attempt to classify the mechanisms responsible for producing symmetries.

7.1. Conclusion on the Origin of Symmetries

We review here our progress in understanding the origin of symmetries in (i) macroscopic physics and (ii) the standard $S(U(2)\times U(3))$ model of microscopic physics.

The well-known conservation laws of classical dynamics are associated with the symmetries of Euclidean geometry and time translational invariance. These 'geometrical' symmetries are rather fundamental and any derivation must be somewhat speculative, appearing near the top of the quantum staircase. Rotational symmetry and space-time translational symmetry are just part of the diffeomorphism symmetry of general relativity, which suggests that they have their origin in gravitation theory. The theory of general relativity was of course constructed to have these symmetries, as part of an overall Poincaré symmetry, built into it. So it is necessary to go beyond general relativity and consider a pregeometric theory, in order to obtain a proper derivation of physical Poincaré invariance. It is then possible to introduce a purely formal diffeomorphism symmetry; the crucial point is to show that the corresponding formal Poincaré symmetry is not spontaneously broken. Quantum fluctuations have been proposed, in Chapter 6.2.1, as a mechanism for stabilising the Poincaré symmetry.

A more convincingly derived macroscopic symmetry is that of scale symmetry, since we know it is not present at the atomic scale. Scaling symmetry arises when a very large number of molecules is present.

The parity symmetry of the macroscopic elasticity properties of, say, sugar is another good example of a derived symmetry. In fact, for this system, parity sym-

metry appears and disappears twice as a function of scale. At very small sub-nuclear distances, where weak interactions are relevant, there is no parity symmetry; then at nuclear and atomic scales, where only strong and electromagnetic interactions are relevant, parity symmetry is valid. In the case of sugar parity symmetry is spontaneously broken at the molecular level, as a consequence of biological effects; only species producing one handedness of sugar have survived. Finally, at macroscopic scales, parity symmetry reappears as a consequence of the presence of a very large number of sugar molecules.

Parity symmetry is now understood as arising naturally from the low energy limit, or long distance behaviour, of the standard model. The same is true of time reversal invariance; however the initial conditions in the world violate time reversal symmetry so much that the second law of thermodynamics is valid.

The macroscopic effects of electrodynamics manifest gauge symmetry and electric charge conservation. Any derivation of gauge symmetry has its origin beyond the standard model and must be considered speculative; suggested derivations are discussed in Chapter 6.3.

At the macroscopic level, we have considered symmetries under the following seven types of transformation: 1) Rotation, 2) Space Translation, 3) Time Translation, 4) Parity, 5) Time Reversal, 6) Scale and 7) Gauge. Although we have presented derivations for all seven symmetries, four of the derivations are rather speculative while three of the symmetries (parity, time reversal and scale symmetry) have rather well-established origins. In these statistics we only include symmetries valid for the whole field of macroscopic physics and not those specific to particular dynamical situations, such as the O(4) symmetry of planetary motion in a gravitational $1/r$ potential.

At the standard model level, we can also ask how many symmetries are derived safely, i.e. from the standard model itself, and how many are derived more speculatively, i.e. higher up on the quantum staircase. We consider eight symmetries at the standard model level: 1) Space-Time Translational Invariance, 2) Lorentz Invariance, 3) Gauge Invariance, 4) Charge Conjugation Invariance, 5) Parity Invariance, 6) Time Reversal Invariance, 7) Flavour Conservation and 8) Chiral Symmetry. In this list, baryon and lepton number conservation is subsumed under 7) and Gell-Mann SU(3) flavour symmetry is subsumed under 8). As summarised in Fig. 3.4, only the first three of the listed symmetries require a speculative explanation outside the model itself. The remaining five symmetries are essentially derived within the standard model, although with the help of some mild auxiliary assumptions (see Fig. 3.4); it is necessary to assume the existence of small quark masses and the absence of the QCD topological θ-term. It is reasonable to suppose that these auxiliary assumptions will be understood in terms of physics beyond the standard model. We can therefore claim that five out of eight symmetries have been derived. It may even be possible to explain the other symmetries, Poincaré invariance and gauge invariance, near the top of the quantum staircase.

So what do we now learn from the above statistics? There appear to be two main conclusions:

I. An experimentally well-established symmetry is not necessarily fundamental.

II. Approximate symmetries appear naturally as derived symmetries.

From conclusion I, we are warned of the danger of imposing symmetries on a fundamental theory. The development of weak interaction theory was delayed by the general assumption, prior to 1956, that parity invariance was a fundamental symmetry of nature. Another example is provided by Heisenberg's unified field theory of elementary particles[1]; this theory assumed isospin symmetry (and also scale symmetry) to be fundamental, but spontaneously broken by an asymmetric vacuum. Isospin symmetry is now known not to be fundamental, see Chapter 3.2, but to be just a consequence of SU(3) colour symmetry and of the existence of two light quarks. Similarly the five symmetries, listed above and derived from the standard model, should no longer be considered fundamental.

Conclusion II follows from the observation that many of the derived symmetries considered in this book only appear as approximate symmetries. For instance, of the symmetries derived from the standard model, just CPT invariance, baryon number conservation and the conservation of the lepton numbers for each generation are exact (apart from the effects of the weak anomaly, which are negligible at normal temperatures[2]). C, P and T invariance, chiral symmetry (including Gell-Mann SU(3) symmetry) and flavour conservation are all approximate symmetries.

The examples taken from macroscopic physics and atomic physics are also only approximate symmetries. The scale symmetry of macroscopic physics is valid to the extent that the number of molecules is very large; the decoupling of spin in atomic physics is only true ignoring relativistic effects; the $O(4)$ symmetry of the hydrogen atom, or of the sun-planet system, is only valid in the presence of an exact non-relativistic $1/r$ potential. The SU(n) symmetry for an n-dimensional harmonic oscillator is exact but does not correspond to a real physical system.

We have also considered, in Chapter IV, the derivation of some symmetries which have been suggested as relevant somewhere beyond the standard model. These symmetries turn out to be exact within the models considered, but many of them must be spontaneously broken if they are to be relevant for nature. For the speculative derivations, in Chapter VI, of Poincaré invariance and gauge invariance, we have considered both exact and approximate derivations (the formal approach and the renormalisation group method).

So we see, from the above examples, that weakly broken symmetries have a natural understanding when symmetries are derived. As long as the symmetries were supposed to be a fundamental property of nature and treated as input to a theory, it seemed rather mysterious why, say, parity or strangeness conservation was violated by just the weak interactions. Now that parity and strangeness conservation are derived as consequences of the theory of strong and electromagnetic interactions, there is no *a priori* reason why they should be conserved by the weak interactions.

There are essentially two views on the role of symmetries in physics. As summarized by Weinberg[3]: it is a question of whether or not you think it is the job of physics to explain symmetries or to explain their absence. Weinberg's point of view is that symmetries have to be explained and if there is no good explanation for them you should not believe them. The alternative view, expressed by Pati[4] at the same meeting, is that you should look for the maximum possible symmetry and then, whatever you do not see in nature, explain how it got broken. In this latter approach, symmetries are provided with a kind of "holiness". An example is the

aesthetic desire to restore parity invariance as a fundamental symmetry at short distances, despite its violation by the weak interaction, in the left-right symmetric $SU(2)_L \times SU(2)_R \times SU(4)_c$ model.[5]

We consider that the above statistics on symmetry derivations show that symmetries are really not so holy. The origin of so many symmetries is now understood that they should not be considered as fundamental or mysterious. So we support the point of view, expressed by Weinberg, that symmetries ought to be explained and not just imposed as *a priori* fundamental principles. We therefore conclude that there is no good philosophical basis for introducing left-right symmetric models purely on the grounds of imposing parity invariance as a fundamental symmetry; there may of course be other motivations for considering such models.

Indeed, encouraged by the success in understanding symmetries, we have gone further than Weinberg and suggested that all the symmetries of the laws of nature may be derivable; in Chapter VI we questioned the fundamental nature of translational invariance, Lorentz invariance and gauge invariance. These three symmetries are conventionally assumed to be fundamental, reflecting the geometrical structure of space-time and the dynamics needed for a consistent relativistic quantum field theory.

We have understood five out of eight of the symmetries of microscopic physics within the standard model, which in all likelihood is the low energy limit of some more fundamental theory. There is therefore a hope that the three remaining symmetries may be explained by new physics on the way up the quantum staircase, from the electroweak $W^{\pm}, Z^0$ scale of 10^{-18} m to the presumably fundamental Planck scale of 10^{-35} m. The seventeen orders of magnitude change in scale is similar to that in passing from macroscopic physics, at the 1 m scale, to the electroweak scale. We see, from Fig. 1.1, that the latter change in scale corresponds to around six steps, or the development of six fields of physics, on the quantum staircase. There is thus plenty of room, up to the Planck mass or even beyond, for the development of new physics and the derivation of the fundamental symmetries of Poincaré invariance and gauge invariance, even if the speculations of Chapter VI are not taken seriously.

This scenario, where all the symmetries of the laws of nature can be derived, leads us naturally to a discussion of random dynamics.

7.2. Random Dynamics

As discussed in the previous section, the progress made in deriving so many of the observed symmetries of nature is very suggestive of the idea that, as we learn more about physics beyond the standard model, we shall learn the origin of all the known symmetries. It is then even natural to go one step further and speculate that all the regularities, i.e. all the laws of physics known today, should similarly have an explanation. This is the random dynamics hypothesis: all phenomenologically found regularities have some explanation, i.e. there is some mechanism responsible for making the known laws of nature valid in at least some domain. Typically this domain is the regime of low energy physics defined as, say, the experimentally accessible energies below 1 TeV.

The existence of a mechanism for deriving some symmetry or other regularity, such as locality or the principle of superposition, implies that almost any theory

proposed at random will possess the regularity in, say, the low energy domain. If indeed there is such a mechanism for each of the known regularities, one could take almost any model and all these regularities would appear in one or more limits, such as at low energy. However it would then hardly matter what the fundamental 'theory' is; it would anyway give rise to the observed regularities. This would mean that almost any theory could explain 'everything' known and thus be good enough as a theory of everything (TOE). There would then be no reason to accept any special model or TOE as the truth; it would be better to imagine that the most fundamental physics is chaotic — a random model — or not to assume anything about it. This is really the content of the random dynamics hypothesis: the fundamental physics, or TOE, does not matter, since almost all models at the fundamental level will have sufficient structure that they agree with the phenomenologically observed regularities.

We may take the above random dynamics hypothesis as a replacement for a fundamental TOE. In practice all one would get out of a fundamental theory would be the low energy regularities; the rest of a TOE cannot be realistically tested anyway. If it is indeed true that we do not need any special fundamental model, in order to obtain the low energy regularities, the existence of such a TOE might just be a superfluous hypothesis. We could just as well choose a random theory from a very large class of dynamical models. This is really a rather simple idea and random dynamics provides a realistic alternative to superstring theory or any other TOE.

If, for any reason, the fundamental dynamical model was very complicated, random dynamics would seem to be the only practical approach. We would have to rely on the hypothesis that the observed physical laws are stable under changes of the fundamental model. This basic philosophy[6] was proposed[7] several years ago [Paper 30]. A similar philosophy can be found in some other approaches to the laws of nature.[8] In particular, recent developments in quantum gravity lead to a remarkably similar picture; this baby universe theory is discussed in the next subsection.

7.2.1. Baby universe theory suggesting random dynamics

The idea that physics is chaotic at the fundamental level is supported by recent progress in quantum gravity: baby universe theory.[9] When quantising Einstein's theory of gravity, using the Feynman path integral prescription,[10] one functionally integrates over all possible four-dimensional Riemann spaces. It is then natural, at least in the Euclidean formulation where a topology change is not accompanied by a singularity in the metric, to include Riemann spaces of all possible topologies; quantum mechanically even the topology of space-time is expected to fluctuate. In this way one can argue for the existence of a 'spacetime foam', which has a complicated topological structure at the Planck scale $G_N^{1/2} \sim 10^{-33}$ cm.

The topological structures relevant to baby universe theory are wormholes, which provide a microscopic connection between two different, otherwise weakly curved, space-time manifolds; alternatively a wormhole may connect two well separated regions of the same manifold. Such a baby universe wormhole may be visualised, by imagining that two small 4-volumes of space-time are removed from the weakly

curved Riemann manifold(s) and a tube of the topological shape $I \times S_3$ attached in their place. Here I is an interval of the real axis and S_3 is the sphere in a four-dimensional space (but S_3 could be replaced by another compact 3-dimensional closed space). It is the S_3-shaped 3-space which is known as the baby universe and it is assumed to have a size of order the Planck length. The wormhole corresponds to the exchange of a baby universe from one 'footprint' to the other on the weakly curved space-time manifold(s).

So we are led to consider contributions to the Euclidean path integral from configurations consisting of many weakly curved space-time manifolds, interconnected by a large number of wormholes, as in Fig. 7.1; one of these manifolds represents the space-time development of our own universe. This means that 'our universe' is being bombarded by baby universes from 'outside space-time'. Baby universes, just like our own universe, have no position; position only makes sense inside our universe or inside a baby universe. In order to describe, quantum mechanically, the effects of these baby universe degrees of freedom, it is convenient to introduce operators, ${a_i}^+$ and a_i, for creating and annihilating whole baby universes. Here the label i denotes the interior state of a baby universe, specified by its size, geometry, contents etc. These 'third quantised' degrees of freedom, a_i^+ and a_i, are not functions of the usual quantum fields $\phi(x)$ defined at space-time points x: they are degrees of freedom from 'outside space-time'. We shall see, in Sec. 7.2.2, that the more speculative ideas in random dynamics also suggest the existence of degrees of freedom outside of space-time.

There is an interaction between the usual quantum field theory degrees of freedom and the third quantised degrees of freedom, where a baby universe unites with our own universe or with another weakly curved universe. Since the third quantised degrees of freedom have no location in space-time, it is not surprising that the effect of this interaction turns out to be the same over all of space-time. In fact it turns out to renormalise the parameters, i.e. the coupling constants λ_i, of the quantum field theory in our universe.

The general form of a 'fundamental' field theoretical Lagrangian density for our universe is

$$\mathcal{L}_0(x) = \sum_i \lambda_i \mathcal{L}_i(\phi(x), g(x)) \ . \tag{1}$$

The summation is over all possible local functions $\mathcal{L}_i$, constructed from the metric $g(x)$, the fields $\phi(x)$ and their derivatives, which are consistent with the gauge symmetries of the theory. The label i runs over infinitely many terms, since we include non-renormalisable interactions which are irrelevant to low energy physics; of course many of the coupling constants λ_i could vanish. This fundamental Lagrangian density $\mathcal{L}_0$ is modified, by the inclusion of the effects of baby universe emission and absorption, into the effective Lagrangian density

$$\mathcal{L}_{\text{eff}}(x) = \mathcal{L}_0(x) + \sum_i (a_i^+ + a_{i^*}) \mathcal{L}_i(\phi(x), g(x)) \tag{2}$$

for distances large compared with the wormhole scale. We may here think of the operator ${a_i}^+ + a_{i^*}$, where i^* denotes the CPT conjugate of a baby universe of type i, as just some dynamical variable representing a third quantised degree of freedom.

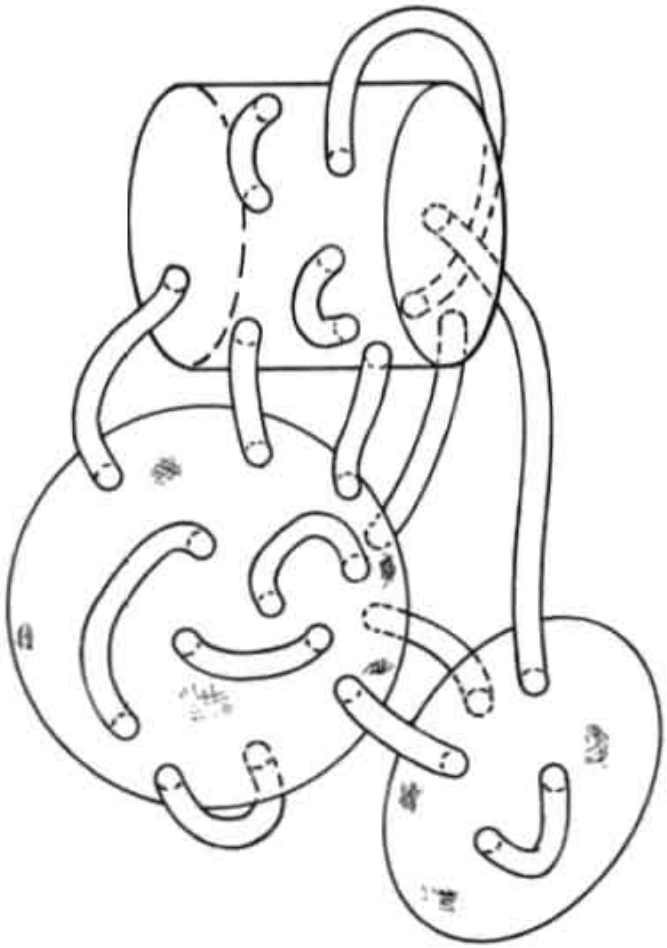


Fig. 7.1. A typical configuration of wormhole connections between weakly curved space-time manifolds.

The dynamical variable $a_i^+ + a_i$ is independent of space and time; once its physical value α_i is measured, it will have the measured value over all space forever. Thus the effect of baby universes is just to renormalise all the fundamental coupling constants λ_i, so that the observed coupling constants become $\lambda_i + \alpha_i$.

The blatant nonlocal behaviour of wormholes is therefore effectively absorbed into shifts of the fundamental coupling constants; this is because the Einstein field equations guarantee that a baby universe cannot carry away any 4-momentum from a weakly curved universe. For example, the Newtonian gravitational field emanating from a region keeps track of the amount of energy within it; so if some energy suddenly disappeared down a wormhole, the corresponding Newtonian field would disappear violating the Einstein equations. Since a baby universe does not take up any 4-momentum, it must enter our universe with the same amplitude at any space-time point, as a consequence of Heisenberg's uncertainty relation. So a baby universe couples with the same amplitude all over space-time, and this is simulated by a change in the coupling constants in the effective Lagrangian density $\mathcal{L}_{\text{eff}}$.

The fact that the wormhole corrections α_i to the coupling constants λ_i appear as quantum mechanical dynamical variables means that, in general, their measured values are unpredictable. Before a measurement of the coupling constants, the initial state of the universe is not expected to be an eigenstate of the operators

$a_i^+ + a_{i^*}$. For instance the Hartle-Hawking initial state condition[10] corresponds to assuming that there are no baby universes in the initial state; this means that the third quantised degrees of freedom are at first in harmonic oscillator ground states $|0\rangle$, satisfying $a_i|0\rangle = 0$. The effective coupling constants $\lambda_i + \alpha_i$ become *a priori* random variables.

This predicted randomness of the effective coupling constants $\lambda_i + \alpha_i$ has several interesting consequences:

1. Random dynamics may be considered to be a consequence of baby universe theory. A theory having random couplings is a good step in the direction of being a random theory. According to this attitude, random dynamics would be an effective theory applicable below the wormhole energy scale. It would not be the fundamental underlying theory; the gauge group and matter degrees of freedom may not be random.
2. Purely global conservation laws are violated by baby universes carrying a non-zero global charge. Then *a priori* vanishing constants, $\lambda_j = 0$, which measure the strength of the gobal symmetry violating interactions, are renormalised by the impact of such charged baby universes; the effective symmetry violating coupling constants $\lambda_j + \alpha_j = \alpha_j$ are non-vanishing in general. However gauged charges remain conserved even after wormhole renormalisation; a gauged charge cannot escape into a baby universe without upsetting the Coulomb field left behind. Thus a baby universe is forbidden to carry a gauged-charge just as it is forbidden to carry 4-momentum. It follows that, according to quantum gravity, we do not expect any of the observed global conservation laws to be fundamental. A global conservation law must have some explanation at an energy lower than the wormhole scale, as the baryon and lepton number conservation laws, discussd in Chapter 3.3, are understood in the standard model.
3. Discrete symmetries are similarly violated by the effects of baby universes. As an example we may consider parity in a left-right symmetric model, which *a priori* has no parity violation:
   Let us imagine we have one Higgs field for the right-handed gauge field and another Higgs field for the left-handed gauge field, each having the same Higgs potential. This means that one superposition $\phi_s(x)$ of Higgs fields is a scalar and the orthogonal superposition $\phi_p(x)$ is a pseudoscalar. The branching off of a baby universe, containing one scalar and one pseudoscalar Higgs (or anti-Higgs) particle, induces a term of the form $\alpha_{ps}\phi_p(x)\phi_s(x)$ in the effective Lagrangian density $\mathcal{L}_{\text{eff}}$. This effective interaction violates left-right symmetry; the 'tachyonic' mass degeneracy of the Higgs particles is broken and parity is not conserved. So baby universes will, in general, lead to violation of all discrete symmetries. Any observed discrete symmetry must be understood as the result of some mechanism, operating below the wormhole energy scale, which forbids the presence of symmetry violating couplings. An example is the explanation of charge conjugation and other discrete symmetries in the strong and electromagnetic interactions provided by the standard model, as discussed in Chapter 3.2.
4. Any theory of everything, TOE, disappears down the wormhole, in the sense

that it is impossible to test its coupling constant predictions due to their random renormalisation, $\lambda_i \to \lambda_i + \alpha_i$. It may still be possible to partially test, say, some superstring theory via its gauge group and matter field gauge group representations. However the inability to predict the constants of nature is a severe blow to any theory claiming to be a fundamental TOE.

5. Random dynamics may be considered to be the fundamental theory beyond the wormhole energy scale. According to point 4 the theory beyond the wormhole scale becomes rather insignificant. Here we go further and suggest that, for physics observed at experimental scales, it does not matter at all what the fundamental theory may be; we might as well imagine it to be a random model. According to this attitude, random dynamics would be more fundamental than baby universe theory. This should be contrasted to the attitude of point 1, where random dynamics is just an effective model approximating the effects of quantum gravity. The derivations of gauge symmetry and the field theory glass scenario, discussed in Chapters 6.3.1 and 7.2.3, are only relevant in this second attitude, where random dynamics is operative beyond the wormhole scale. Of course diffeomorphism symmetry and quantum gravity must then be hoped to arise automatically from a random pregeometric model, as speculated in Chapter 6.2.1.

The renormalisation and consequent intrinsic uncertainty of all physical parameters seems to be the most reliable result derived from baby universe theory. A more controversial result is the claim[11,12] that the probability distributions for the values of some, and possibly all, of the physical parameters $\lambda_i + \alpha_i$ are very sharply peaked and thereby effectively fixed. In particular the cosmological constant is predicted to vanish and potential solutions are suggested for other fine tuning problems: the value of the topological θ parameter in QCD[13] and the smallness of the weak interaction mass scale, and Higgs mass, compared to the Planck mass.[12] If successful in predicting the correct value of the θ parameter, baby universe theory would explain the origin of CP symmetry in strong interactions without an axion.

The probability distributions for the values of the physical parameters are calculated using the euclidean path integral approach to quantum gravity. The summation of wormhole effects introduces a sharply peaked extra factor into the probability distribution for the parameters α_i, in addition to the factor arising from the ansatz for the third quantized wave function of the universe. This extra factor in the measure for the integration over the third quantised variables $\alpha = \{\alpha_i\}$ dominates the partition function

$$Z(\alpha) \approx \exp\left[\exp\left(\int d^4x\sqrt{g}\tilde{\mathcal{L}}(g,\alpha)\right)\right] \tag{3}$$

$$= \exp\left[\exp\left(\frac{3}{8G_N{}^2(\alpha)\Lambda(\alpha)}\right)\right] . \tag{4}$$

Here $\tilde{\mathcal{L}}(g,\alpha)$ is the effective Lagrangian density as seen at very large, i.e. astronomical, distances where the geometry of space-time is measured. $\tilde{\mathcal{L}}(g,\alpha)$ is supposed to include quantum corrections from all the different interactions to all orders in a loop expansion. Expanding $\tilde{\mathcal{L}}(g,\alpha)$ in powers of the Riemann curvature, the two

leading terms are

$$\tilde{\mathcal{L}}(g,\alpha) = \Lambda(\alpha) - \frac{1}{16\pi G_N(\alpha)} R \tag{5}$$

where $\Lambda(\alpha)$ is the cosmological constant and $G_N(\alpha)$ is Newton's gravitational constant. (We note that the usual cosmological constant of general relativity is defined to be $\lambda = 8\pi G_N \Lambda$.) The integration in Eq. (3) is over the space-time of a large smooth universe, with the topology of a 4-sphere, connected to our own universe by wormholes.

The partition function $Z(\alpha)$ clearly blows up when $G_N^2(\alpha)\Lambda(\alpha) = 0$. Thus values of the α_i for which the long-distance or 'dressed' cosmological constant $\Lambda(\alpha)$ vanishes are overwhelmingly favoured. This is how baby universe theory solves the cosmological constant problem: why is space-time so flat that the cosmological constant satisfies the astronomical bound $G_N^2\Lambda < 10^{-120}$? The wormhole solution is to provide a distribution of 'dynamical' coupling constants $\lambda_i + \alpha_i$, which is strongly peaked for $G_N^2(\alpha)\Lambda(\alpha)$ very small and positive; so the bound $G_N^2(\alpha)\Lambda(\alpha) < 10^{-120}$ is exceedingly likely to be satisfied.

In addition to a strongly peaked distribution for $G_N^2\Lambda$, the rapid variation of $Z(\alpha)$ should in principle determine all the other physical coupling constants. This 'big fix' of the physical parameters would effectively remove their randomness; the version of random dynamics considered under point 1 above would disappear down the wormhole. However the attitude to random dynamics taken in point 5 would still survive; random dynamics at the fundamental level could lead to the appearance of an effective diffeomorphism gauge group and hence of quantum gravity and baby universe theory. Random dynamics might also predict the gauge groups and their matter field representations.

The above 'big fix' of the cosmological constant and the other physical coupling constants has been subjected to a number of serious criticisms. A careful re-analysis[14] of the calculation of $Z(\alpha)$ reveals a missing minus sign, so that the previously dominant factor in the partition function, Eq. (4), now becomes

$$\exp\left[-\exp\left(\frac{3}{8G_N^2(\alpha)\Lambda(\alpha)}\right)\right] \tag{6}$$

spoiling the argument for $G_N^2\Lambda = 0$. It is important for the derivation of this result that the inner exponent

$$\int_{S_4} d^4x\sqrt{g}\tilde{\mathcal{L}}(g,\alpha) \approx \frac{3}{8G_N^2(\alpha)\Lambda(\alpha)} \tag{7}$$

should be real. This exponent is a large number and just a small phase, from say a tiny CP violating θ-term, might easily compensate the minus sign in Eq. (6). In fact such terms would occur in complex conjugate pairs in the path integral, corresponding to the two orientations of the 4-sphere,[11] and would introduce a real prefactor $f(\alpha)$ multiplying the inner exponential in Eq. (6). If $f(\alpha)$ could be negative, the values of α for which $f(\alpha)$ is negative would be favoured and the argument for $G_N^2\Lambda = 0$ reinstated. However there is a general argument,[14] based on the hermiticity of the Hamiltonian, that matter fields can only give a positive

prefactor $f(\alpha)$. The gravitational field has to evade this general argument, in order to justify the introduction of the minus sign itself in Eq. (6), and can only do so due to the instability of the Euclidean action.

Another criticism is the claimed presence of wormholes of arbitrarily large scale,[15] which would imply strong nonlocal interactions over arbitrarily large distances. There are arguments that this catastrophe could be averted.[15]

We must conclude that the status of the predictions of the values of physical parameters from baby universe theory is very uncertain at present. Indeed, apart from the value of the cosmological constant, the predicted values are not quite correct. The effective topological $\bar{\theta}$-parameter (see Eq. (92) of Chapter 3) is predicted[13] to take the CP-conserving value $\bar{\theta} = \pi$, up to weak interaction corrections, rather than the phenomenologically preferred value $\bar{\theta} = 0$. Again although scalars by themselves may become massless according to baby universe theory,[12] the origin of the small electroweak scale relative to the wormhole scale is not yet understood. Nonetheless baby universe theory may well carry a good deal of truth, since it should be a consequence of almost any theory of quantum gravity. Here we mainly consider it, see point 5 above, as a motivation for random dynamics as the fundamental underlying theory.

7.2.2. The first steps in random dynamics

The starting point of random dynamics should be the consideration of as large a class of conceivable 'theories' or 'models' as possible. A probability measure should then be chosen on the set of models. The random dynamics hypothesis [Paper 30] is that a model chosen at random from this set will almost always contain the empirically known laws of nature; i.e. all the presently known laws of physics follow from almost any model, which is complicated enough, provided we go to some limit, generally the low energy limit. In order to verify this hypothesis, one should in principle progress by developing a series of physical models, successively approximating one another, as one goes down the quantum staircase of Fig. 1.1. One such intermediate model is the so-called quantum field theory glass introduced in Chapter 6.3.1. In this subsection we discuss how the quantum field theory glass might be derived from random dynamics, while we consider the consequences of the quantum field theory glass in the next subsection. The content of this subsection is therefore the part of random dynamics furthest away from experiment and, consequently, the least studied and most speculative part of the subject. So it contains difficult material and may safely be omitted by the reader.

At first, it may seem rather hopeless to try to derive low energy physics, or even the quantum field theory glass, out of a totally random mathematical model. However we believe some concepts are so general that we can hope to find them relevant in almost any sufficiently rich and complicated 'model'. For example it should be possible to introduce a crude concept of distance, or of some sort of topology, on almost any mathematical structure S which contains a huge number of 'elements' having relations between themselves: the longer the chains of intermediate elements needed to establish a relation between two 'elements', the further apart the two 'elements' are defined to be. For such a mathematical structure, rich in the sense of having a huge number of elements and also being rather repetitive, it must be

possible to define the concept of small modifications $a, b \ldots$ of the structure S. This should be a modification which only changes the number of repeated substructures by a relatively small amount. Assuming that the random model obeys some axioms, which are then of course also random, the system of small modifications will approximately make up some Abelian group; indeed we expect that the most important small modifications could be naturally organised as a vector space V.

Using the approximate topology suggested above, we may consider the restriction of an allowed small modification to some neighbourhood with respect to the topology. We want to interpret such small modifications as wave functions of the Wheeler-DeWitt type,[16] which describe the situation for all times, acting as follows:

$$\psi : \; S_T \rightarrow \mathbb{C}^N \text{ or } R^N \; . \tag{8}$$

Here N is an integer which possibly depends on where we are in the topological space S_T of the structure S. The components of ψ measure the various amounts of small modifications of different type in the 'region' in question. The axioms will impose linear relations between the simultaneously allowed small modifications in neighbouring regions of the structure.

We define a formal restriction $b|C$ of an allowed and meaningful modification b to a neighbourhood $C \subseteq S$ as follows: $b|C$ is that small modification which equals b for the part of the structure inside C and equals no modification outside C. This formal restriction of a small modification will of course introduce problems, e.g. disagreement with the axioms on the boundary of C. Nonetheless these formal restrictions $b|C$ may be used to generate an extended vector space $\tilde{V}$. This extended vector space $\tilde{V}$ essentially has the character of a space of functions ψ on the topological space S_T. The truly allowed small modifications now appear as a subspace V allowed by the 'axioms'; the subspace can be characterized by a set of constraint equations

$$\phi_i \psi = 0 \tag{9}$$

where the ϕ_i are linear functionals.

We required above that the modifications be small enough to guarantee linearity and the existence of a vector space $\tilde{V}$. There may therefore be a problem in restricting such a modification to a very small region C: if C is itself supposed to be a relatively simple substructure the effect of any modification on it is likely to be quite severe. A remedy for this problem is to choose regions, such as C, that are sufficiently large as to contain a fair amount of repetitive structure. The region C could still be extremely small compared to, say, all of S and be approximately a point in S_T. The physically relevant states ψ should be superpositions over so many local contributions, like $b|C$, that the need to modify the same region C twice practically never arises.

We have now argued that one can construct, out of almost any random mathematical structure S, a vector space or Abelian group of 'wave functions' ψ. These wave functions ψ are identified with small modifications and constraints ensuring compatibility with the axioms not only locally but also globally. The wave functions are defined on a topological space S_T constructed from the structure S. The constraints are local, since the topology is defined by means of the relations between the elements.

Although we have not yet introduced space-time, we may obtain a very general form of quantum mechanics from the mathematical structure provided we can interpret some of the constraint equations, $\phi_j \psi = 0$, as a sort of Wheeler-deWitt equation. The Wheeler-deWitt equation of quantum gravity is not an equation of time evolution in the usual sense; in fact time does not appear at all and it takes the form of a Hamiltonian constraint equation.[16]

$$\mathcal{H}\psi = 0 \,. \tag{10}$$

Time must be interpreted as a description of the correlations between variables in the Wheeler-deWitt wave function. In this interpretation[17] one postulates the existence of a weakly interacting, more or less classical, variable which plays the role of a physical clock. One then essentially takes the clock Hamiltonian density $\mathcal{H}_{\text{clock}}$ to be conjugate to the variable h describing the position of the 'hand' of the clock:

$$\mathcal{H}_{\text{clock}} = p_h = \frac{1}{i}\frac{\partial}{\partial h} \,. \tag{11}$$

The Hamiltonian constraint operator is decomposed into the form

$$\mathcal{H} = \mathcal{H}_{\text{clock}} + \mathcal{H}_{\text{rest}} \tag{12}$$

and the Wheeler-deWitt equation becomes

$$\mathcal{H}\psi = \left(\frac{1}{i}\frac{\partial}{\partial h} + \mathcal{H}_{\text{rest}}\right)\psi = 0 \,. \tag{13}$$

The constraint Eq. (13) can now be interpreted as the Schrödinger equation for the rest of the variables. In the same way we want to interpret some of the constraint equations of our system, $\phi_j \psi = 0$, to be a generalised Schrödinger equation.

The next step in such a random mathematical model is to introduce a space-time concept. The identification of physical objects in the general mathematical structure S is of course essential, if there is to be a physical interpretation of the structure. These interpretive assumptions are to a high degree arbitrary, particularly in the first steps of random dynamics. We have already assumed that the vector space of small modifications of the mathematical structure is to be identified with the Hilbert space of a generalised quantum mechanics. As a further tentative interpretation, let us now assume that the topological space S_T is the configuration space for a somewhat extended quantum field theory describing nature. In other words, we imagine that each point, i.e. each element, in S_T corresponds to a field configuration.

We would now like to consider the quantum field as a function mapping all of space-time into a target space. In this way we treat space and time on an equal footing and the equation of motion for the field will take the form of a constraint.

The interpretation of the configuration space as a space of functions requires that S_T is a very large, or infinite, dimensional space. The field values over all space-time are then coordinates for the topological space S_T, which must therefore essentially be a differentiable manifold. So we assume that it is possible to construct a tangent vector space to S_T. For this purpose, it is necessary to have an additive

composition law on the tiny neighbourhood of elements in S_T near to any element A of the mathematical structure; this neighbourhood is then approximately the tangent plane. The composition law $AB + AC = AD$ is naturally defined by the requirement that the same small operation is needed to go from A to B as from C to D. When the points A, B, C and D are close to each other, the operation needed to go from one of these points to another has so little effect that the order of the operations should not matter. This means that the composition law is commutative.

We now need to introduce a topological structure on the basis vectors or 'directions' in the tangent space, so that we can identify the coordinates in S_T with points or small regions in a pregeometrical space P. We can think of the set of all the various 'directions' or 'vectors', in the tangent space of the configuration space S_T, as another example of a complicated mathematical structure. So we argue, as we did for S_T, that an approximate topology can be introduced, by saying that two 'elements' or 'directions' in the tangent space are close together when there is a relatively simple operation connecting one to the other. Once we have introduced such a topology, the 'directions' in the tangent space can be grouped into neighbourhoods or small overlapping subspaces. These selected subspaces are then to be associated with small regions in a pregeometrical space-time P. If the above construction is possible, we have associated the field values, i.e. the coordinates of S_T, with small regions of space-time, ensuring a local field theory in the long wavelength limit.

The next step is to assume that it is possible to identify some configuration as the vacuum state and find an approximate symmetry, under some group of gauge transformations, in the configuration space close to the vacuum. It is then hoped that the mechanism described in Chapter 6.3.1 can be used to promote this approximate symmetry group to an exact formal symmetry which, via quantum fluctuations, is realised physically as a local gauge symmetry. Among these gauge symmetries there may be some which have the structure of a diffeomorphism group, transforming some of the points of the pregeometric space P into each other. In addition we might hope to find some Yang-Mills type gauge symmetry as a subgroup at each point in P.

We can only give a rather weak argument for the presence of the diffeomorphism group. It is based on the remark, in Chapter 6.3.1 and the following subsection, that the field theory glass transformations, which survive as low energy symmetries, should be rather uniquely implemented with as few outer automorphisms as possible. We suppose the Yang-Mills gauge transformation group appears in the pregeometrical space as a cross product of extremely many isomorphic factors, one for each point in P. The presence of the diffeomorphism group is then favoured, in the sense that it converts many of the permutations of these factors from outer to inner automorphisms.

If the diffeomorphism part of the gauge group transforms the pregeometric points in the usual way, we can identify these points with the points of space-time observed phenomenologically. The orbits of these points under the diffeomorphism group would then be identified as space-time manifolds.

It is not necessary that all the points in the pregeometric space-time are associated with an orbit for the diffeomorphism part of the group of gauge transforma-

tions; it could happen that some of the points are invariant under diffeomorphisms. Such trivially transforming points would have their associated degrees of freedom outside of space-time, in the same sense as the third quantised degrees of freedom in baby universe theory. We therefore expect these diffeomorphism invariant degrees of freedom to act in the same way on all space-time points and, for practical purposes, just to influence the values of the parameters in the effective laws of nature. So this feature of exterior degrees of freedom is expected in random dynamics even if, for some reason, genuine wormholes should not form. Of course we expect that random dynamics should lead to quantum gravity and hence to the existence of baby universes. As discussed in Chapter 6.2.1, general relativity results from the spontaneous breakdown of the above diffeomorphism gauge symmetry to an exact Poincaré gauge symmetry.

We have given above some preliminary ideas on how physically relevant conclusions might be deduced from a random complicated mathematical model. There are clearly many gaps still to be filled in the argument. Our main motivation, in presenting such an incomplete discussion, was to convince the reader that it is not a forlorn hope to derive physical laws from a general mathematical model. Indeed if something like the above construction is made to work on a general mathematical structure, we would be close to deriving the field theory glass model of Chapter 6.3.1, with some gauge symmetries and some extra degrees of freedom outside of space-time. However we also introduced a diffeomorphism gauge symmetry and thus, in the sense of Chapter 6.2.1, gravity and translational invariance. Of course the field theory glass is not translational invariant and we therefore seem to have derived too much symmetry at this stage. But this is presumably harmless since, in the end, the field theory glass is supposed to be smoothed out, by quantum fluctuations, into a continuum model with reparameterisation symmetry. We had to appeal to similar quantum fluctuations to obtain diffeomorphism gauge symmetry, and hence translational invariance, from the pregeometrical space-time P. So we actually expect to obtain the continuum version of the field theory glass and the diffeomorphism gauge symmetry from quantum fluctuations, on the same steps of the quantum staircase.

The above discussion of the origin of space-time did not address the question of its dimensionality and how the diffeomorphism group of $3+1$ dimensions is selected in a random theory. We speculate below that fermionic degrees of freedom might play an important role in singling out $3+1$ dimensions.

So far our discussion has been very abstract and we have only outlined how, in a very general model, we might identify some fundamental physical notions: a) linearity in quantum mechanics, b) time, c) space-time and d) locality. We would now like to present some slightly more concrete derivations of some of these notions. These derivations will be based on different assumptions, which deviate in detail from the above exceedingly speculative discussion. This is consistent with the spirit of random dynamics, according to which the final physical results should be insensitive to the input assumptions.

A. Quantum Mechanics

First we shall consider a derivation of quantum mechanics or rather the linearity of the Schrödinger equation. A linearity rule is generically explained by assuming

analyticity and using a Taylor expansion; it is then necessary to argue that the constant term in the expansion is zero or should be subtracted away, and that all the terms beyond the linear term are negligibly small. It was precisely the smallness of the small modifications, which was crucial for the speculative identification of quantum mechanics in the random general mathematical structure discussed above.

As the starting point for our derivation of quantum mechanics, we assume the existence of a time variable t and that the evolution of the world is described by a general autonomous differential equation

$$\frac{dX(t)}{dt} = F(X(t)) \, . \tag{14}$$

Here $X(t)$ has a huge number of components describing the state of the physical world. Time-translational invariance has been assumed so that the velocity field $F(X(t))$ only depends on $X(t)$. In the spirit of random dynamics, $F(X(t))$ can now be taken to be a randomly chosen analytic field, in a computer simulation which looks for the typical behaviour of the solution to Eq. (14). With a measure chosen so that a random number generator creates fields $F(X(t))$ with many zeros, the typical behaviour for the time development of $X(t)$ turns out[18] to be the approach to a fixed point X_0:

$$X(t) \to X_0 \text{ as } t \to \infty \, . \tag{15}$$

It follows that, after a long time, $X(t)$ is very close to the fixed point X_0 and the Taylor expansion

$$\begin{aligned} \frac{dX}{dt} &= F(X(t)) \\ &= F(X_0) + (X(t) - X_0)\frac{\partial F}{\partial X}\Big|_{X_0} + \dots \end{aligned} \tag{16}$$

is a good approximation, retaining just the first two terms. At the fixed point X_0

$$F(X_0) = 0 \tag{17}$$

and we can rewrite Eq. (16) in the form

$$i\frac{d\psi(t)}{dt} = H\psi(t) \tag{18}$$

when we define

$$\psi(t) = X(t) - X_0 \tag{19}$$

and

$$H = i\frac{\partial F}{\partial X}\Big|_{X_0} \, . \tag{20}$$

This notation suggests that we should interpret $\psi(t)$ as the wave function for the world and H as the Hamiltonian operator, so that Eq. (18) can be identified as the Schrödinger equation.

However there is no reason why the matrix H should be hermitean and its eigenvalues λ will in general be complex. The stability of the fixed point X_0 requires that

$$\mathrm{Im}\,\lambda \leq 0\ . \tag{21}$$

The time development of the wave function can be expanded in terms of the eigenvectors ψ_λ of H:

$$\psi(t) = \sum_\lambda e^{-i\lambda t}\psi_\lambda\ . \tag{22}$$

As time passes, all the components in the expansion will die out, except for those with eigenvalues having a numerically very small imaginary part $\mathrm{Im}\ \lambda \approx 0$. It follows that only the almost real eigenvalues are relevant after a long time and then the Hamiltonian, restricted to this subspace of wave functions, may appear to be hermitean to a very good approximation. However, in addition to the reality of the eigenvalues, hermiticity requires the eigenvectors to form a complete orthonormal basis. It may therefore be necessary to adjust the definition of the Hilbert space inner product, so as to make the Hamiltonian H more closely hermitean with respect to it.

The age of the universe is indeed very large in Planck units. Thus we can claim to have derived some of the general principles, i.e. linearity and hermiticity, of quantum mechanics from random dynamics; although hermiticity is not fully derived. It is actually a prediction of random dynamics that the hermiticity of the Hamiltonian is not exact, and thereby 'the probability of living at different moments of time' is not exactly constant. Experimental tests looking for a possible small antihermitean part of the Hamiltonian,[6,19] and also for small deviations from linearity,[20] are therefore important.

B. 3 + 1 Dimensions of Space-Time

We have already suggested that space-time could result from the appearance of a diffeomorphism symmetry in a rather general pregeometric space. This mechanism, however, would not *a priori* explain why the dimensionality of space-time should be $3 + 1$. In a slightly different scheme, there is a random dynamics argument for space-time having precisely $3 + 1$ dimensions[21] [Paper 30]. This argument presumes the existence of very general fermion quantum fields in a random non-Lorentz invariant, but space-time translational invariant, model. This model would therefore not be viable if indeed Lorentz invariance and translational invariance appeared together, as special diffeomorphism symmetries in a pregeometric model.

Assuming space-time translational invariance, but not rotational or Lorentz invariance, the general free fermion field equation in d-dimensions takes the form

$$D(p)\psi(p) = 0 \tag{23}$$

in momentum representation. The assumption analogous to that of hermiticity for the Hamiltonian, in the case of a single particle, is that, say, the energy p_0 is required to be real if the other components of the d-momentum vector p_μ are real. This suggests that $D(p)$ should be a hermitean matrix, when the components of p_μ are taken to be real. In the spirit of random dynamics, we assume $D(p)$ to be a

random or generic matrix function of p_μ. In other words we look for properties of $D(p)$ that would be true for whole regions in the space of matrix functions satisfying $D(p) = D(p)^+$ for real p_μ.

The condition

$$\det D(p) = 0 \tag{24}$$

must be satisfied in order to have a solution of the free fermion field equation, Eq. (23). This condition is generically satisfied on submanifolds in momentum space of codimension 1. In d-dimensional space a submanifold of dimension $d - c$ is said to have a codimension c. We shall be interested in degenerate solutions of Eq. (23). Generically one finds q solutions, with the same value of p_μ, on submanifolds of codimension q^2. This can be seen by taking a basis for $\psi(p)$-space, in which the q degenerate solutions are taken as basis vectors. In order to stay on the q-degeneracy submanifold, the $q \times q$ submatrix of $D(p)$ corresponding to the q solutions must be zero. Thus q^2 real conditions must be satisfied, since $D(p)$ is assumed to be hermitean.

For a low-energy physicist, the only accessible fermion states are those near the Fermi surface, which separates empty and filled levels. We now want to consider the possible relationships between the Fermi surface and the q degeneracy surfaces. We will first show that the Fermi surface cannot consist solely of points in momentum space on degeneracy surfaces with $q \geq 3$. Around a point in p_μ-space with $q \geq 3$ degenerate solutions of Eq. (23), there will generically be solution branches with lower q-values. In the neighbourhood of a point in p_μ-space with q solutions, the relevant part of $D(p)$ is a $q \times q$ hermitean matrix. The manifold in this neighbourhood with $\hat{q}$ solutions corresponds to such matrices having rank $q - \hat{q}$. If $q - \hat{q} = 1$, the $\hat{q}$-solution manifold is cut into two disconnected pieces when the q-solution submanifold is removed. This result follows because the single non-zero eigenvalue, of the rank 1 $q \times q$ matrix, cannot change sign without passing through zero. However if $(q - \hat{q}) > 1$, there are at least two non-zero eigenvalues and they can easily change sign, without passing through a p_μ-point having two of them zero simultaneously. So, for example, removing the 3-solution points splits the 2-solution manifold into pieces, but the 1-solution manifold remains connected. We therefore conclude that the Fermi surface, which must separate empty and filled states, cannot consist of $q \geq 3$ points alone.

The Fermi surface might, however, consist solely of points in momentum space with $q = 2$ solutions. Indeed this seems to be the situation in nature when we consider the fermions of the standard model to be massless Weyl particles, as is suggested by the fact that the left-handed and right-handed fermions have different gauge quantum numbers. In nature it appears that the Dirac sea of negative energy states is just filled and the Fermi surface for a Weyl particle lies at the isolated zero energy-momentum point $p_\mu = 0$. For all Weyl states with $p_\mu \neq 0$, the helicity constraint that the particle should be, say, left-handed allows only one polarisation state, satisfying the condition

$$p_0 = -\underline{\sigma} \cdot \underline{p} \,. \tag{25}$$

However for the zero energy-momentum state, with $p_0 = \underline{p} = 0$, the helicity constraint, Eq. (25), does not restrict the spin $\underline{\sigma}$ and two states of polarisation are allowed. The fact that the Fermi energies of quarks and leptons now essentially

coincide with the double degeneracy point, $p_\mu = 0$, is in large part due to the Hubble expansion of the universe attenuating the fermion number densities. The actual values of the baryon and lepton numbers of the universe are of course not yet understood.

Let us now assume that, even in the hypothetical d-dimensional non-Lorentz invariant theory, the Hubble expansion drives the Fermi surface towards the submanifold of d-momentum space having $q = 2$ solutions of Eq. (23). A physicist performing experiments at energies low compared to the Planck mass (here meaning the fundamental mass), would only have access to states very close to the Fermi surface. So the observed energy-momentum dispersion relation should be obtained from a Taylor expansion of $D(p)$ around values of p_μ on the Fermi surface, which we have assumed are points with $q = 2$ solutions. Near a $q = 2$ point we may effectively take $D(p)$ to be a 2×2 matrix, since the solutions in this neighbourhood are dominantly composed from the two degenerate states. Denoting the $q = 2$ point about which we expand by $p_{(0)\mu}$, the expansion of the effective 2×2 matrix takes the form

$$D(p) = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix} + (p_\mu - p_{(0)\mu}) \cdot \frac{\partial D(p)}{\partial p_\mu}\Big|_{p_\mu = p_{(0)\mu}} + \dots \ . \tag{26}$$

The requirement that $D(p)$ be hermitean for real values of p_μ implies that $\frac{\partial D}{\partial p_\mu}$ in Eq. (26) is hermitean. Since there are only four independent 2×2 hermitean Pauli matrices $\sigma^0 = I, \sigma^1 = \sigma_x$, $\sigma^2 = \sigma_y$ and $\sigma^3 = \sigma_z$, we may write

$$\frac{\partial D(p)}{\partial p_\mu}\Big|_{p_\mu = p_{(0)\mu}} = V_\alpha^\mu \sigma^\alpha \ . \tag{27}$$

Thus the equation of motion for the free fermion field, Eq. (23), in the neighbourhood of the $q = 2$ point becomes

$$(p_\mu - p_{(0)\mu}) V_\alpha^\mu \sigma^\alpha \psi(p) = 0 \ . \tag{28}$$

We now renormalise the d-momentum and define the practical momentum to be

$$\tilde{p}_\mu = p_\mu - p_{(0)\mu} \ . \tag{29}$$

The equation of motion then becomes the Weyl equation

$$\tilde{p}_\mu V_\alpha^\mu \sigma^\alpha \psi(\tilde{p}) = 0 \ . \tag{30}$$

The vierbein V_α^μ can easily be transformed away, by defining an affine basis in which the momentum components are $\tilde{p}'_\alpha = V_\alpha^\mu \tilde{p}_\mu$.

The signature of the metric associated with the Weyl equation, Eq. (30), may be deduced from the dispersion relation condition

$$\begin{aligned}
0 &= \det D(p) && (31)\\
&= \det(\tilde{p}_\mu V_\alpha^\mu \sigma^\alpha) && (32)\\
&= \det \begin{pmatrix} \tilde{p}_\mu (V_0{}^\mu + V_3{}^\mu) & \tilde{p}_\mu (V_1{}^\mu - i V_2{}^\mu) \\ \tilde{p}_\mu (V_1{}^\mu + i V_2{}^\mu) & \tilde{p}_\mu (V_0{}^\mu - V_3{}^\mu) \end{pmatrix} && (33)\\
&= (\tilde{p}_\mu V_0{}^\mu)^2 - (\tilde{p}_\mu V_1{}^\mu)^2 - (\tilde{p}_\mu V_2{}^\mu)^2 - (\tilde{p}_\mu V_3{}^\mu)^2 && (34)\\
&= \tilde{p}_\mu \tilde{p}_\nu V_\alpha{}^\mu V_\beta{}^\nu \eta^{\alpha\beta} \ . && (35)
\end{aligned}$$

Here

$$\eta^{\alpha\beta} = \mathrm{diag}(1, -1, -1, -1, 0, 0, \ldots 0) \tag{36}$$

is a metric with the usual $(1, -1, -1, -1)$ signature, except for some degeneracy.

We conclude that the rank of the effective metric

$$g^{\mu\nu} = V_\alpha{}^\mu V_\beta{}^\nu \eta^{\alpha\beta} \tag{37}$$

for the fermion is only four, even when the dimension d of the energy-momentum space is higher than four! The fermion "moves" with zero velocity components in the directions for which the metric vanishes. Any material made from such fermions cannot move in those directions. So the directions in which the metric vanishes are effectively unobservable for purely fermionic matter. In this way the appearance of $3+1$-dimensional space-time is favoured.

The boson fields might still, *a priori* move in all d-dimensions. Similarly different types of fermion field might move in different sets of $3+1$ dimensions. However there is a mechanism for selecting just one $3+1$-dimensional space-time as relevant at low energy, if the fermion fields couple to a gauge field. The fermion fields then contribute to the renormalisation of the non-Lorentz invariant coupling parameters $\eta^{\mu\nu\rho\sigma}$ in the d-dimensional non-covariant Yang-Mills action, analogous to Eq. (20) of Chapter 6.2.2. The contribution of each Weyl field to the renormalisation group beta function $\beta_\eta^{\mu\nu\rho\sigma}$ is non-vanishing in the appropriate $3+1$ dimensions, and the corresponding components of $\eta^{\mu\nu\rho\sigma}$ in these $3+1$ dimensions grow relatively stronger towards the infrared. Provided the $3+1$-dimensional contributions from the various Weyl fields are not too unparallel in the d-dimensional space-time, they should add up to give a major term in the renormalisation group equation along some 'average' $3+1$-dimensional space and minor terms in the other dimensions. The components of $\eta^{\mu\nu\rho\sigma}$ in this average $3+1$-dimensional space will then grow large in the infrared, compared to the components in the other $d-4$ space-time directions.

The coupling parameters $\eta^{\mu\nu\rho\sigma}$ are essentially inversely proportional to the square of the effective gauge coupling constant. Hence the above renormalisation group behaviour implies that the effective Yang-Mills coupling strength will become greater in the extra $d-4$ dimensions than in the average $3+1$-dimensional space. This favours the formation of a so-called layer phase[22] with confinement, at relatively high energy, in all directions except in those of the average $3+1$-dimensional space. In such a situation, there would be four-dimensional layers in the higher d-dimensional space-time and particles with non-trivial gauge quantum numbers would be confined to their layer. Thus the $d-4$ excess dimensions would be unobservable for gauge-charged matter.

As discussed in Chapter 6.2.2, the renormalisation group scale dependence will also make the effective $3+1$-dimensional metrics $g^{\mu\nu}$ for the different Weyl fields approach each other, and a Lorentz covariant theory, towards the infrared.

Under the influence of the low energy Higgs vacuum for the electroweak interactions, neutrinos and Z^0 particles might be able to escape from a $3+1$-dimensional space-time layer out into the d-dimensional world. However this would be avoided if the diffeomorphism gauge invariant gravity theory is also in a layer phase; energy and momentum are then confined to the $3+1$-dimensional layers and particles would not be observed disappearing from our $3+1$-dimensional space-time.

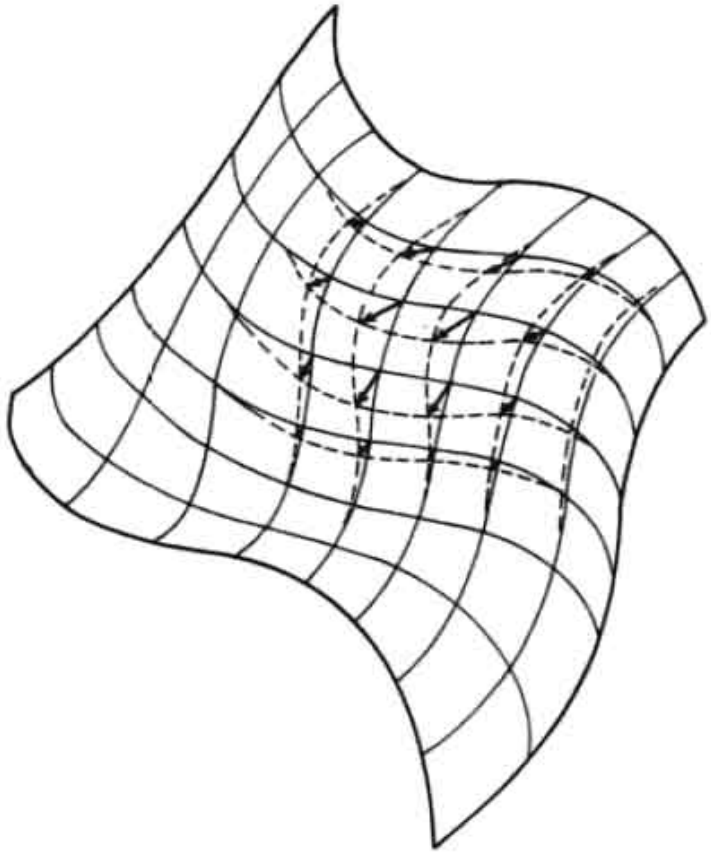


Fig. 7.2. A diffeomorphism which only transforms points in a small neighbourhood.

We conclude that it is possible to obtain $3+1$ effective space-time dimensions in a natural way from random dynamics.

C. Locality

As a third example we consider how the principle of locality might be derived in random dynamics. For this purpose we assume that general relativity in $3+1$-dimensional space-time has already been derived, together with its characteristic diffeomorphism symmetry and a space-time metric $g_{\mu\nu}(x)$.

The simplest type of non-local, but still reparameterisation invariant, action S_{nl} take the form of a sum of products of ordinary allowed local action expressions:

$$S_{nl} = \sum_{n=1}^{\infty} \int \sqrt{g(y_1)} d^4 y_1 \mathcal{L}_n^1(\phi(y_1), \partial_\mu \phi(y_1)) \cdot \int \sqrt{g(y_2)} d^4 y_2 \mathcal{L}_n^2(\phi(y_2), \partial_\mu \phi(y_2))$$
$$\ldots \int \sqrt{g(y_m)} d^4 y_m \mathcal{L}_n^m(\phi(y_m), \partial_\mu \phi(y_m)) \,. \tag{38}$$

Here $g = |\det g_{\mu\nu}|$ and $\mathcal{L}_n^i(\phi(y_i), \partial_\mu \phi(y_i))$, for $i = 1, \ldots, m$, are ordinary reparameterisation invariant Lagrangian density expressions, dependent on a set of fields $\phi(y_i)$ and their derivatives. Terms formed from an integrand depending on two points x and y in a combined way are not diffeomorphism invariant, because it is possible to make a reparameterisation which translates one of the points but not the other. A diffeomorphism that only makes a change close to one point is illustrated in Fig. 7.2.

In order to obtain a diffeomorphism invariant action, involving two or more points in a combined way, it is necessary to consider constructions involving the metric field $g_{\mu\nu}(x)$ along curves connecting the two points. An example of such a non-local term is

$$S'_{nl} = \int \sqrt{g(x)}d^4x\sqrt{g(y)}d^4y\mathcal{L}_1(\phi(x),\partial_\mu\phi(x))\mathcal{L}_2(\phi(y),\partial_\mu\phi(y)) \cdot \left[\int DX(\tau)\exp\left(-b\int_{\tau_i}^{\tau_f} d\tau\sqrt{g_{\mu\nu}(X(\tau))\frac{dX^\mu}{d\tau}\frac{dX^\nu}{d\tau}}\right)\right] . \tag{39}$$

This integral over pairs of space-time points x and y, has an integrand which itself contains a functional integral over the set of all curves $X(\tau)$ connecting the points $x = X(\tau_i)$ and $y = X(\tau_f)$. The functional integral measure is denoted by $DX(\tau)$ and b is a constant. The expressions $\mathcal{L}_1$ and $\mathcal{L}_2$ are ordinary reparameterisation invariant Lagrangian density functions of the fields at x and y.

Let us now consider the physical consequences of the actions S_{nl} and S'_{nl} respectively.

A non-local action of the form S_{nl} does not give any truly non-local physical effects. All the separate integrals

$$I_j^n = \int \sqrt{g(y_j)}d^4y_j\mathcal{L}_n^j(\phi(y_j),\partial_\mu\phi(y_j)) \tag{40}$$

in Eq. (38) must, when evaluated, have some numerical values; they are constant in time and over space. The equations of motion derived from S_{nl} can thus be exactly simulated by a local action

$$S = \int \sqrt{g(x)}d^4x\mathcal{L}(\phi(x),\partial_\mu\phi(x)) . \tag{41}$$

Here the effective Lagrangian density is given by the double sum

$$\mathcal{L}(\phi(x),\partial_\mu\phi(x)) = \sum_{n=1}^{\infty}\sum_{i=1}^{m}\mathcal{L}_n^i(\phi(x),\partial_\mu\phi(x))C_n^i \tag{42}$$

and the constants C_n^i are given in terms of the integrals I_j^n of Eq. (40) by

$$C_n^i = I_1^n I_2^n \ldots I_{i-1}^n I_{i+1}^n \ldots I_m^n . \tag{43}$$

The action S_{nl} can therefore be replaced by S and only gives rise to local effects.

Non-local actions of the form S'_{nl} involve the path lengths of space-time curves between the interacting points x and y. Integrating out the functional integral in euclidean space, one obtains the approximate distance behaviour $\exp(-b_{\text{ren}} \cdot d_{xy})$ for the action term S'_{nl}. Here b_{ren} is the 'renormalised' value of the constant b in Eq. (39) and d_{xy} is the space-time separation of the points x and y. On dimensional grounds the constant b_{ren} is expected to be of order the inverse Planck length (again meaning the fundamental length) or zero. If b_{ren} vanishes the action S'_{nl} degenerates

to the product form S_{nl} of Eq. (38), which we have just seen only gives rise to local effects. If b_{ren} is of order the inverse Planck length, the non-local effects fall off exponentially for distances greater than the Planck length. So we would effectively obtain locality at experimentally accessible scales.

7.2.3. Field theory glass and gauge glass

We have indicated, in the previous subsection, how the quantum field theory glass model of Chapter 6.3.1, with its rather chaotic set of degrees of freedom, might arise from a random complicated mathematical structure. In particular we assume here that the rudimentary physical concepts of a $3+1$ dimensional space-time, locality and linearity in quantum mechanics have been derived at some fundamental scale l_{P}, which we usually take to be the Planck length. More specifically we assume:

1. The existence of an underlying $3+1$ dimensional space-time.
2. Translational and Lorentz invariance are broken at the fundamental scale l_{P}, but are present statistically over larger distances.
3. The action and degrees of freedom, in each small region of space-time, are chosen in a quenched random way; so the model has an amorphous structure, like a glass.
4. The action is semi-local: there is no direct interaction between degrees of freedom associated with points separated by distances greater than a few times l_{P}.
5. The action is taken to be that of a discretized, lattice-like, quantum field theory, rather than a genuine continuum theory.
6. The resulting quantum field theory glass is formulated as a Feynman path integral, in which the quenched random parameters are kept constant.

Quantum fluctuations are expected to smooth out the field theory glass model into a continuum theory, as part of the presumed mechanism for the appearance of diffeomorphism symmetry and quantum gravity at distances somewhat larger than l_{P}. The resulting quantum gravity theory would resemble a "molten" field theory glass, containing small field theory glass pieces locally: it would be analogous to a fluid containing small pieces of approximate crystals but, on the large scale, having statistical mechanical fluctuations everywhere. This continuum theory has not been developed, but it is expected that the more intuitive discrete field theory glass model will give similar results in the long wavelength limit.

The construction of a field theory glass model was described in Chapter 6.3.1. Its degrees of freedom are incorporated in a generalised quantum field $\phi(i)$, defined on a set of quenched random space-time points $\{i\}$, where it takes values in quenched random manifolds M_i. Quenched random parameters are chosen randomly but are then not varied in the Feynman path integral. The field theory glass action

$$S[\phi] = \sum_{r} S_r(\phi(i)) \ , \quad i \in r \tag{44}$$

is semilocal, being a sum over contributions $S_r(\phi(i))$ from very many small overlapping regions r in space-time.

We argued in Chapter 6.3.1 that the inverse Higgs mechanism, promoting approximate gauge symmetries to exact physical gauge symmetries, should work even

for a field theory glass. With the action and field value manifold varying from point to point, it is likely that some degrees of freedom will, by accident, approximate one or another discretised gauge theory in various regions. We therefore introduced a formal gauge symmetry

$$\phi_h \to \phi_h^{\Omega} \tag{45}$$

$$H \to \Omega^{-1} H \tag{46}$$

by defining a new so-called "human" field variable ϕ_h, and the gauge group valued "Higgs field" H, related to the original fundamental field by the equation

$$\phi(i) = \phi_h^H(i) \ . \tag{47}$$

We thereby formally constructed a field theory glass endowed with gauge symmetries of a random nature, i.e. for which the gauge symmetry group varies randomly from region to region. This is what we call a gauge glass.[23]

We also described how continuum gauge field degrees of freedom $A_\mu^a(x)$ might be implemented on the gauge glass, by a suitable modification of the formal gauge glass degrees of freedom $\phi_h(i)$. *A priori* we might attempt to implement a Yang-Mills theory for an arbitrary gauge group K; we would probably find gauge glass degrees of freedom corresponding to it, here and there, throughout space-time. However it may only be possible to implement the gauge field $A_\mu^a(x)$, for the group K, very sparsely in space-time. The effective continuum gauge field Lagrangian density then acquires a term of the form

$$\mathcal{L}_{\text{eff}} = -\frac{1}{4g^2} F_{\mu\nu}{}^a(x) F^{\mu\nu a}(x) \tag{48}$$

with a very small coefficient, $-1/4g^2$, because $A_\mu^a(x)$ represents relatively few gauge glass degrees of freedom. This means the gauge coupling constant, g^2, is strong; so the gauge degrees of freedom are confined close to the Planck scale and become irrelevant to experimentally accessible low energy physics. In order for a gauge group K to survive at low energies, it is necessary for the group to be richly represented, as a good approximate symmetry of the fundamental field theory glass action $S[\phi]$, throughout space-time. This is more likely to happen, by chance, when the number of degrees of freedom involved is small, i.e. for gauge groups of small dimension and for matter fields belonging to small representations. There are also technical reasons, which we discuss below, why many gauge symmetry groups would spontaneously break down near the Planck scale.

There is a particular problem in implementing a continuum Yang-Mills field $A_\mu^a(x)$ on a gauge glass, when the gauge group K has outer automorphisms.[24] This technical difficulty seems to be of some importance, in random dynamics, for understanding the origin of the standard model group and of the number of quark and lepton generations. The group of outer automorphisms of a group K is defined[24] as the factor group of all automorphisms f, calculated modulo the inner automorphisms h which are just the similarity transformations:

$$\text{Out}(K) = \frac{\{f : K \to K | f \text{ automorphism}\}}{\{h : K \to K | \exists b \forall g \in K [h(g) = bgb^{-1}]\}} \ . \tag{49}$$

If the gauge group K has a nontrivial outer automorphism, there is a problem in deciding, locally, which degrees of freedom of the gauge glass to identify with which degrees of freedom of the continuum field $A^a_\mu(x)$. For example, suppose a given shift $\Delta\phi_h(i)$ of the human gauge glass field, from the vacuum configuration, is made to correspond to a continuum gauge field configuration $A^a_\mu(x)$. Then we could equally well make the correspondence in which the shift $\Delta\phi_h(i)$ is associated with the continuum field $A^{fa}_\mu(x)$, obtained by applying the automorphism f to $A^a_\mu(x)$. So, in every small space-time region, there is an ambiguity in deciding how to represent the continuum gauge field.

In the case of an inner automorphism,

$$f(g) = bgb^{-1},\ b \in K \text{ and } g \in K\ , \tag{50}$$

the transformed field $A^{fa}_\mu(x)$ is related to $A^a_\mu(x)$ by a gauge transformation, with a constant gauge function:

$$\frac{\lambda^a}{2}A^{fa}_\mu(x) = U(x)\frac{\lambda^a}{2}A^a_\mu(x)U^{-1}(x) - i\partial_\mu U(x)U^{-1}(x) \tag{51}$$

where $U(x) = b$ and $\frac{\lambda^a}{2}$ are group generators. So we recognise the ambiguity in identifying the continuum field, due to inner automorphisms, as the usual gauge ambiguity in defining the gauge field. It does not give rise to any ambiguity in the physics, but just adds a background field.

For a nontrivial outer automorphism however, the identification ambiguity is more than just a gauge ambiguity. How can this ambiguity be resolved? The requirements of continuity and of gauge invariance for the effective action of the continuum gauge field $A^a_\mu(x)$ will, in general , enforce a definite extension of the identification convention from one neighbourhood to the next. The semilocal contribution $S_r[\phi_h, H]$ to the action, from a small region r overlapping both neighbourhoods, will not be invariant under a convention shift in only one of the neighbourhoods. In this sense a relative convention is fixed. So it is possible to extend the identification convention along some closed curved Γ in space-time. But now we have a consistency condition: the convention on returning to the starting point must be the same as the starting convention. In a gauge glass model it seems inevitable that, among the many closed curves, some of them will lead to an inconsistency between the initial convention and the final convention. The gauge group on return is said to get 'confused' with its own automorphic image.[23]

The fact that inconsistencies are likely to arise, in the identification convention for the gauge glass degrees of freedom, may be seen by considering a change in the gauge glass structure, in some small region r along a closed curve Γ. A change in the quenched random parameters, occurring in the semilocal contribution $S_r[\phi_h, H]$ to the action, may simulate the effect of making a relative convention shift between two successive neighbourhoods in r along Γ. The interaction between the human fields $\phi_h(i)$ and $\phi_h(j)$, in two successive neighbourhoods, and the Higgs field $H(s)$ is then effectively replaced by a similar interaction between the fields $f(\phi_h(i)), \phi_h(j)$ and $H(s)$, where $f(\phi_h(i))$ is the appropriate automorphic image of $\phi_h(i)$. This change in the relative convention, between two successive neighbourhoods along Γ, requires

a corresponding change in convention along the rest of Γ. So a switch in convention is needed on return to the starting point of Γ. Since the quenched parameters are randomly chosen in all of the small regions along a closed curve Γ, the probability for any specific convention being needed, on return to the starting point, should be the same for all identification conventions. There are many possible closed curves Γ; so it seems overwhelmingly likely that some of them will lead to an inconsistency in the choice of convention.

The inconsistency can be gauge transformed away in the case of an inner automorphism, but not in the case of an outer automorphism. An inconsistency corresponding to the inner automorphism $f(g) = bgb^{-1}$ is transformed away, by a gauge transformation with a gauge function varying from I to b along the closed loop Γ. This transformation is achieved at the cost of setting up a magnetic flux through the loop: the "inner" inconsistencies are transformed into background gauge fields. The lowest energy state will, of course, tend to adjust itself so as to cancel any background fields if it can.

It appears that a continuum Yang-Mills theory, for a gauge group with outer automorphisms, is unlikely to be implemented consistently on a gauge glass. There is one way of avoiding this conclusion if fermion degrees of freedom, or other matter degrees of freedom, are present throughout the gauge glass: the matter degrees of freedom can be used to specify a consistent identification convention. For example the cross product group SU(2)×SU(2) has an outer automorphism, corresponding to a permutation of the two isomorphic SU(2) groups. It is possible to characterise one SU(2) factor, in an SU(2)×SU(2) gauge group, by having a particle in, say, its doublet representation, but no particle in the doublet representation of the other SU(2) factor. Confusion of the two isomorphic SU(2) groups is thereby avoided.

Random dynamics therefore predicts that the continuum gauge group must not have outer automorphisms that can be extended to true discrete symmetries. For instance some left-right symmetric models[4,5] have a discrete symmetry of the gauge group and matter field representations, corresponding to a parity symmetry. These left-right symmetric models do not arise as effective theories out of random dynamics.

Some simple groups have outer automorphisms, corresponding to the permutation symmetries of the Dynkin diagram for the associated Lie algebra.[24] The SU(N) group, for $N > 2$, has an outer automorphism corresponding to complex conjugation. Similarly the standard model group S(U(2)×U(3)), discussed in Chapter III, has complex conjugation as an outer automorphism: complex conjugation of the elements of the 5 × 5 matrix **U** defined in Eq. (1) of Chapter 3.1, corresponds to charge conjugation. Charge conjugation symmetry is broken in the standard model, since only left-handed fermions and right-handed anti-fermions couple to the $W^{\pm}$ intermediate vector boson. If the fermion degrees of freedom that break charge conjugation symmetry are present in all the considered small regions of the gauge glass, they may be used to specify a convention for the identification of the continuum gauge field $A^a_\mu(x)$ with the gauge glass degrees of freedom.

Non-semi-simple groups have a U(1) factor in their Lie algebra and at least one outer automorphism: a complex conjugation, which changes the sign on the U(1) subalgebra. The standard model group thus has the smallest number of

outer automorphisms possible for a non-semi-simple group. It is also distinguished, among the non-semi-simple groups up to dimension 18 with only complex conjugation as an outer automorphism, by having the fewest so-called generalised outer automorphisms.[25] Generalised outer automorphisms of a non-semi-simple group are isomorphisms between two of its factor groups, obtained by dividing out invariant subgroups with a small number of elements. In order to find a generalised outer automorphism for the standard model group $S(U(2)\times U(3))$, one of the invariant subgroups divided out has to have at least five elements; whereas for the other abovementioned non-semi-simple groups of dimension less than 19, the invariant subgroups can have just two or three elements. The generalised outer automorphism of the standard model group, obtained by dividing out the invariant subgroup with five elements, corresponds to scaling the U(1) hypercharge up by a factor five combined with complex conjugation.

Generalised outer automorphisms act as ordinary outer automorphisms of the Lie algebra and, thus, of the continuum gauge fields $A^a_\mu(x)$, which we seek to implement on the gauge glass. There is again a danger of inconsistencies arising in the identification convention for the gauge glass degrees of freedom, when the gauge group K has generalised outer automorphisms. It seems that inconsistencies, arising from such an outer automorphism of the Lie algebra, are more likely the closer this automorphism is to being a symmetry of the group structure and of the matter fields. Inconsistencies may be avoided, if the group structure and the matter fields can be used to keep order in the identification convention, connecting the gauge glass variables to the continuum fields $A^a_\mu(x)$.

An inconsistency in the identification convention for the gauge glass degrees of freedom, corresponding to a generalised outer automorphism, involves the identification of a factor group of the gauge group K with the continuum gauge field. The identification of the factor group[24] with the continuum field means that the set of group elements, belonging to a given coset of the factor group, must correspond to the same continuum field $A^a_\mu(x)$. This requires that these group elements should all correspond to approximately the same value of the gauge glass action. In other words, there must be an approximate symmetry of the action under the permutation of the group elements inside a coset. The larger the number of elements inside each coset, the less likely it is for the gauge glass to manifest such a symmetry by accident. Thus the threat of inconsistencies, due to a generalised outer automorphism, is expected to become less significant, as the number of elements in the associated invariant subgroup, divided out of the gauge group K, increases. As remarked above, this number of elements has to be at least five in the case of the standard model group.

Identification ambiguities, due to generalised outer automorphisms, may also be avoided in the presence of gauge glass fermion degrees of freedom. If the fermion spectrum present in the gauge glass is not invariant under the generalised outer automorphism, it may be used to specify an identification convention. In general the fermion spectrum would have to consist of infinitely many particles, in order to be invariant under the scaling of a U(1) "hypercharge" subalgebra by a factor S. If there exists a particle with hypercharge Y_0, invariance of the spectrum under scaling requires the existence of particles with hypercharge $Y = Y_0 S^n$, where n runs

through all integer values both positive and negative; thus the number of particles is required to be infinite, with accumulation points at hypercharge values of zero and infinity, unless $S = -1$. However one can say there is an approximate symmetry under scaling by a factor S, if there exist some fermions that transform into other existing fermions under such a scaling; i.e. if there exist some pairs of fermions with hypercharges differing by a factor S,but otherwise having identical gauge quantum numbers.

The fermion spectrum of the standard model is almost completely devoid of such an approximate symmetry under scaling of the weak hypercharge. In most cases, Weyl fermions with a given combination of non-Abelian gauge quantum numbers belong to exactly the same weak hypercharge representation. The fermion representations are repeated, in each of the three quark-lepton generations, without varying the weak hypercharge values. There is only one combination of non-Abelian quantum numbers, for which more than one weak hypercharge value is found phenomenologically: the left-handed antiquarks, see Table 3.1, have $\frac{1}{2}y = -\frac{2}{3}$ and $\frac{1}{2}y = \frac{1}{3}$ respectively. So there is a very approximate symmetry under scaling weak hypercharge by a factor $S = -2$ in the standard model. Apart from this exception, which has to be present in order to avoid gauge anomalies, nature appears to avoid even approximate symmetries under weak hypercharge scaling. Again the standard model seems well equipped to avoid inconsistencies due to generalised outer automorphisms.

The confusion mechanism, causing the collapse of a gauge group K with outer automorphisms near the Planck scale, can be further clarified, by considering the simulation of a gauge glass using a generalisation of ordinary lattice gauge theory.[23] This approach also allows us to identify other collapse mechanisms,[26] which may prevent a gauge group K surviving from the Planck scale down to the relatively low energy scale of present-day experimental particle physics. The modifications of ordinary lattice gauge theory are made in a quenched random way, i.e. they are chosen randomly but then kept fixed in the Feynman path integral or its Monte Carlo simulation. We list the set of modifications below:

i) The regular hypercubic lattice is replaced by an amorphous glass-like structure, by letting the lattice sites take random positions separated by distances of order the Planck length l_P.

ii) The gauge group is allowed to vary from site to site.

iii) The link variables $U(\bullet\!\!\longrightarrow\!\!\bullet)$ for the gauge group K are coupled together to form plaquette variables $U_\Box$, defined in Eq. (61) of Chapter 6.3, and plaquette action $S_\Box$ contributions. The plaquette variables and action contributions are invariant under the "confused gauge transformation"

$$U(\overset{\bullet}{x}\!\!\longrightarrow\!\!\overset{\bullet}{y}) \to f_{x,y}(\Lambda(x))^{-1} U(\overset{\bullet}{x}\!\!\longrightarrow\!\!\overset{\bullet}{y}) f_{y,x}(\Lambda(y)) \tag{52}$$

rather than simply under the usual gauge transformation

$$U(\overset{\bullet}{x}\!\!\longrightarrow\!\!\overset{\bullet}{y}) \to \Lambda(x)^{-1} U(\overset{\bullet}{x}\!\!\longrightarrow\!\!\overset{\bullet}{y}) \Lambda(y) \ . \tag{53}$$

Here $f_{x,y}$ are quenched randomly chosen outer automorphisms of the gauge group K.

iv) The plaquette action $S_\square$ is taken to be a quenched random function of the plaquette variable $U_\square$. Gauge invariance requires $S_\square$ to be a function defined over the space of conjugacy classes[27] for the group K, i.e. $S_\square$ takes the same value for two group elements $U_\square$ related by an inner automorphism. We therefore expand $S_\square$ on characters for the various representation r of the group K

$$S_\square = \sum_r \beta_r \chi_r(U_\square) \tag{54}$$

where the coefficients β_r are quenched random variables satisfying

$$\beta_r = \beta_{\bar{r}}^* \tag{55}$$

and $\bar{r}$ denotes the complex conjugate representation to r. The character for representation r is

$$\chi_r(U_\square) = \mathrm{Tr}\{\rho_r(U_\square)\} \tag{56}$$

where $\rho_r(U_\square)$ is the representation matrix for $U_\square$ in the representation r. The coefficients β_r are chosen, independently for each plaquette and for each pair of conjugate representations, from a gaussian distribution. The distribution is chosen to have zero average and a width which decreases with the dimension of the representation, in order to obtain a convergent and reasonably smooth plaquette action $S_\square$.

v) The link variables $U(\bullet\longrightarrow\bullet)$ are allowed to take values on a group space and the group structure varies randomly from link to link, as in Fig. 7.3. The plaquette variables $U_\square$ are then constructed from common factor groups for the groups of the surrounding links. Outer automorphism ambiguities then naturally arise in identifying factor groups of different link groups that just happen to be isomorphic. It is therefore essential to consider the confused gauge transformations of modification iii).

By introducing the above modifications of ordinary lattice gauge theory, we simulate the action for a gauge glass. We now consider the consequences of these various modifications and how their glassy nature may cause a breakdown, at the Planck scale, of many of the gauge groups that are present *a priori* in this gauge glass.

Spontaneous symmetry breakdown occurs when the vacuum state is not invariant under the global gauge symmetry. This is expected by analogy with the Higgs mechanism, in which gauge bosons acquire a nonzero mass when the vacuum state is not globally gauge invariant. The gauge glass plaquette action of Eq. (54) does not contain an explicit Higgs field, but it is still possible to investigate the vacuum state of the theory, in the classical approximation, by considering the energetically preferred value of each plaquette variable $U_\square$ separately. If the vacuum value of a link variable $U(\bullet\longrightarrow\bullet)$, or a plaquette variable $U_\square$, does not commute with all the elements of the gauge group K, the vacuum will not be invariant under global gauge transformations. Hence we require all the link variables, and consequently all the plaquette variables, in the vacuum state to belong to the centre of the gauge group

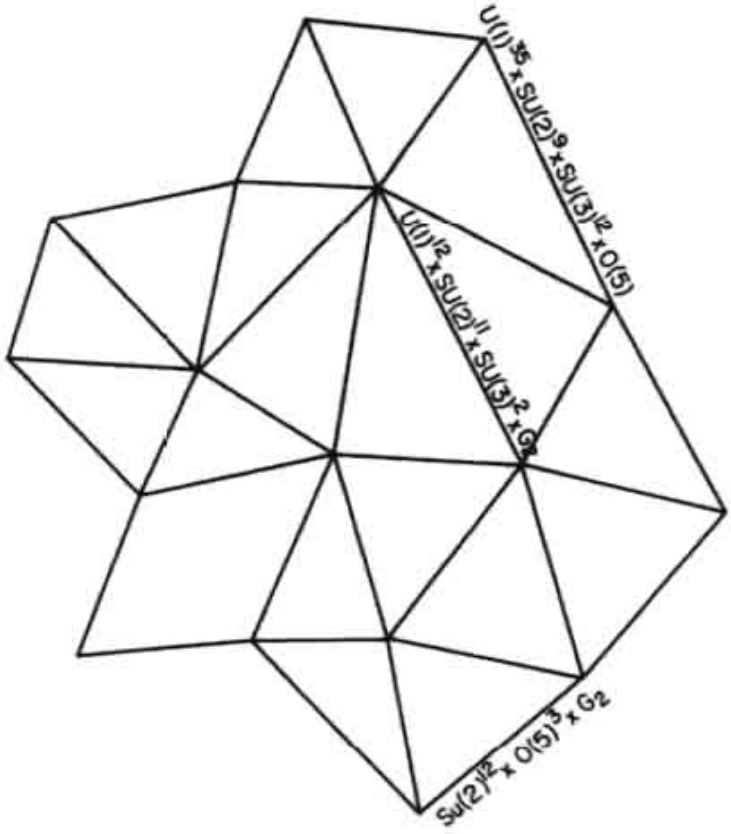


Fig. 7.3. A lattice gauge model with an irregular lattice and random groups on the various links.

K; otherwise the gauge symmetry will spontaneously collapse. We review below the conditions a gauge group K should satisfy in order to resist such a collapse:[23,26]

I. *Absence of outer automorphisms.* The gauge glass action, for a gauge group K having outer automorphisms, is invariant under the confused transformation of Eq. (52). We assume that the vacuum values of the corresponding confused link variables $U(\overset{\bullet}{x}\!\!-\!\!\overset{\bullet}{y})$ take values in the centre of the group. These vacuum values will not be invariant under global gauge transformations, $\Lambda(x) = \Lambda$, unless they obey

$$f_{x,y}(\Lambda) = \Lambda \tag{57}$$

for all the outer automorphisms $f_{x,y}$ randomly chosen throughout the lattice. It follows that the vacuum state is only symmetric under the subgroup, which is left invariant by all the outer automorphisms of K. So the gauge group K spontaneously breaks down to a subgroup without outer automorphisms (or at least with very few).

The gauge particles corresponding to directions in the Lie algebra of K, which are not left invariant by the outer automorphisms, acquire masses of the order of the Planck mass.[23] Inner automorphisms can be transformed into background gauge fields. These background fields may give masses to some of the gauge particles via a Higgs-like mechanism, which is essentially equivalent to the mechanisms to be discussed below associated with frustration of the gauge glass and with other vacuum fields in the gauge glass.

The confusion mechanism discussed here is, of course, essentially the problem of identification inconsistencies for gauge glass degrees of freedom in another guise.

Consequently it is possible to circumvent the confusion mechanism and allow a gauge group with a few outer automorphisms to survive, by including matter degrees of freedom.

II. *A non-trivial centre.* If the vacuum state is to be invariant under global gauge transformations of a group K with a trivial centre, all the plaquette variables $U_{\square}$ must take on the value of the unit element in K. However the coefficients β_r in the gauge glass plaquette action of Eq. (54) may have either sign. This means that the unit element is as likely to correspond to an energy maximum as to an energy minimum. Thus minimising the plaquette energy will, in about one out of two cases, bring the plaquette variable away from the unit element and hence out of the centre, for a group with a trivial centre. This, in turn, means that the vacuum is not invariant under the global part of the gauge group K, which is thereby broken. So, in a gauge glass, groups with a trivial centre are highly susceptible to collapse and only groups with a non-trivial centre survive.

III. *A connected centre.* A similar argument to the foregoing, applied to a set of neighbouring plaquettes instead of to a single plaquette, suggests that gauge groups with a topologically disconnected centre tend to collapse. The plaquettes, which together constitute the surface of a generalised three-cube of the amorphous lattice, must satisfy the Bianchi identity. This identity arises from the fact that each link is contained in two plaquettes of the cell and the plaquette variables $U_{\square}$ are therefore not independent. If the $U_{\square}$ lie in the centre of the group K, i.e. the Abelian subgroup which commutes with all elements of the group, the product of the plaquette variables forming the surface of a cell must equal the unit element:

$$\prod_{\square \in \boxtimes} U_{\square} = I \, . \tag{58}$$

The Bianchi identity, Eq. (58), is not automatically satisfied when each individual plaquette variable is chosen to minimise its energy. The energetically preferred individual plaquette variables $U_{\square \mathrm{pref}}$ must belong to the centre to prevent collapse of the gauge group K but, due to the random form of the plaquette action, we expect the product

$$F = \prod_{\square \in \boxtimes} U_{\square \,\mathrm{pref}} \tag{59}$$

to be a random centre element. We define F to be the frustration of the cell $\boxtimes$. If $F \neq I$, the true vaccum values $U_{\square \,\mathrm{vac}}$ of the plaquette variables cannot equal $U_{\square \,\mathrm{pref}}$: there must be some compromise between the various plaquettes forming the cell.

It is possible a compromise can be achieved by small adjustments of the plaquette variables, without leaving the centre of the group, if the centre is continuous. However if the centre of the group is not also topologically connected, the frustration F for some cells will belong to a different connectedness component to that of the unit element. It is then probably necessary to adjust all the plaquette variables forming such a cell, leading to deviations $U_{\square \,\mathrm{pref}} U_{\square \,\mathrm{vac}}{}^{-1}$ between $U_{\square \,\mathrm{pref}}$ and $U_{\square \,\mathrm{vac}}$. These deviations correspond to "magnetic fluxes" leaving the frustrated cell, which

add up to a path in group space leading from I to F. Such a path must leave the centre if I and F belong to different connectedness components. It follows that some of the deviations $U_{\Box\,\mathrm{pref}}U_{\Box\mathrm{vac}}{}^{-1}$ are non-central. Since the $U_{\Box\,\mathrm{pref}}$ are assumed to belong to the centre, this means that some of the vacuum plaquette variables $U_{\Box\mathrm{vac}}$ are non-central and the gauge group K collapses. In order to avoid such a gauge symmetry breakdown, it is necessary that the gauge group K has a connected centre. However it is not sufficient.[28]

IV. *Conjugacy class space with only central singularities.* The plaquette action $S_\Box$ is defined over conjugacy class space, taking the same value for two group elements related by a similarity transformation. The minimum of the plaquette energy therefore tends to occur at the extremities, or singularities, of conjugacy class space. The vacuum plaquette variables are thus likely to take up these singular values, which must then belong to the centre of the group in order to avoid the spontaneous breakdown of the gauge symmetry.

The singularities are typically corners or edges of conjugacy class space. For example the conjugacy class spaces for the rank one groups SU(2) and SO(3) are line segments and their singularities are the end points. The conjugacy classes for SU(2) and SO(3) are characterized by the angle of rotation about some direction in the three dimensions. The end points correspond to angles of 0 and 2π for SU(2) and to angles of 0 and π for SO(3). For SU(2) both end points belong to the centre, corresponding to the group elements $I$and -I; while for SO(3) only the end point corresponding to the unit element I is a centre element. The non-central end point singularity of SO(3) conjugacy class space represents the class of rotations by an angle π. If a plaquette variable in the vacuum takes on a value in this singular π-class, the SO(3) gauge group spontaneously breaks down. SU(2) is favoured to survive as a gauge group, since it has only central corners.

Another type of singularity is a so-called "top hat" point, which lies in the interior of conjugacy class space. For example the group of rotations in four dimensions

$$\mathrm{SO}(4) = \mathrm{SU}(2)\times\mathrm{SU}(2)/\{(I,I),(-I,-I)\} \tag{60}$$

has a conjugacy class space with the geometry of a paper hat. The class space for the covering group SU(2)×SU(2) in the square formed by the cross product of two line segments, representing the conjugacy class spaces of the two SU(2) groups. The conjugacy class space of SO(4) corresponds to the paper hat, formed by folding the square along a diagonal and then gluing the two halves of the diagonal together: the midpoint of the diagonal corresponds to the top of the hat. This singular point at the top of the hat represents the class of 4×4 orthogonal matrices with two eigenvalues equal to -1 and two eigenvalues equal to $+1$. It is non-central and, if a plaquette takes on its value in the vacuum, the SO(4) gauge group collapses.

The above four conditions favour the standard model group

$$\mathrm{SMG} = \mathrm{S}(\mathrm{U}(2)\times\mathrm{U}(3)) \tag{61}$$

as being particularly well suited to avoid spontaneous collapse in the gauge glass and to survive as a low energy symmetry group. The SMG is the one group with the Lie algebra U(1)×SU(2)×SU(3) which has a non-trivial connected centre:

$$S(U(2)\times U(3)) = U(1)\times SU(2)\times SU(3)/D_3 \tag{62}$$

where the discrete group divided out

$$D_3 = \{h^n | n \in Z_6\} \tag{63}$$

is generated from the central element

$$h = \{\exp(i2\pi/6), -I_2, \exp(i2\pi/3)I_3\} \tag{64}$$

of the group U(1)×SU(2)×SU(3). Here Z_N denotes the additive group of integers modulo N and I_N denotes the unit element of SU(N). There are no non-central corners or top hat singularities in the conjugacy class space of the SMG. Also it has only one outer automorphism, i.e. complex conjugation, under which the chiral quark-lepton representations are anyway not symmetric.

Let us now consider more systematically the type of group favoured by the four requirements I-IV. In general, the necessity for a continuous centre requires that the gauge group have an Abelian invariant subgroup or U(1) factor. For the SU(N) groups the extremities of conjugacy class space all correspond to central elements $\exp(i2\pi m/N), m \in Z_N$. This is not true for other simple groups, which are therefore expected to spontaneously break down in the gauge glass. So we are led to consider groups with a Lie algebra consisting of a direct product of various SU(N) factors and an abelian U(1) factor. Multiple occurrences of the same subalgebra do not survive, due to the confusion mechanism of condition I.

The centre of the covering group for a Lie algebra, consisting of a direct product of SU(N) factors and a U(1) factor, is not connected. In order to satisfy the connected centre condition III, it is necessary to divide out some discrete central subgroup, which has a non-empty intersection with all the connectedness components of the centre of the covering group. Some elements in all connectedness components of the centre of the covering group are thereby identified with the unit element. If the N values for the various SU(N) factors are all mutually prime, it is possible to generate the required discrete central subgroup from a single centre element, as the discrete group D_3 is constructed, in Eq. (63), for the SMG from the centre element of Eq. (64). However if the various N values have common divisors, such a discrete subgroup cannot be generated from a single centre element. An extra generator, with zero component along the U(1) group, is then needed. This extra generator is identified with the unit element, when the discrete subgroup is divided out of the covering group. In this case, the resulting factor group must have a non-central singularity in conjugacy class space, analogous to the top hat singularity for the SO(4) group of Eq. (60). Condition IV is therefore violated if any of the N values have common divisors.

Since gauge groups with a small number of dimensions are most likely to arise by chance in a gauge glass, we expect the preferred N values to be as small as possible, as well as mutually prime. We therefore conclude that the gauge glass model, and hence random dynamics, favours a gauge group of the form:

$$G^{(P)} = \mathrm{U}(1)\times\mathrm{SU}(2)\times\mathrm{SU}(3)\times\mathrm{SU}(5)\times\ldots\times\mathrm{SU}(P)/D_P \tag{65}$$

where the SU(M) groups are taken for M running through all prime numbers greater than or equal to 2 and less than or equal to P. The discrete group divided out

$$D_P = \{h_P^n | n \in Z_{N_P}\} \tag{66}$$

is generated from the centre element

$$\begin{aligned} h_P = &\{\exp(i2\pi/N_P), -I_2, \exp(i2\pi/3)I_3\ , \\ &\exp(i2\pi m_5/5)I_5, \ldots, \exp(i2\pi m_P/P)I_P\}\ . \end{aligned} \tag{67}$$

Here

$$N_P = 2\times 3\times 5\times\ldots\times P \tag{68}$$

is the product of all prime numbers less than or equal to P and the parameters m_n for $n > 3$, are integers not divisible by n. For $P > 3$, there are several different versions of $G^{(P)}$ corresponding to the different values of the integers $m_5, m_7, \ldots, m_P$.

The first three members of the series of $G^{(P)}$ groups are

$$G^{(1)} = \mathrm{U}(1) \tag{69}$$

$$G^{(2)} = \mathrm{U}(2) \tag{70}$$

$$G^{(3)} = \mathrm{SMG} = \mathrm{S}(\mathrm{U}(2)\times\mathrm{U}(3))\ . \tag{71}$$

The groups with $P > 3$ contain SU(N) subalgebras with numerically larger renormalisation group beta functions, in the usual notation,[27] than those of the SMG; therefore they might be expected to confine at a relatively high energy scale. We make a more careful estimate of this confinement energy scale in the next subsection.

Chiral fermion fields are required to avoid spontaneous collapse of the gauge group $G^{(P)}$, by confusion under charge conjugation, in the gauge glass. However, in general, we expect the matter fields to belong to as small representations as possible, if the gauge symmetry is to appear accidentally in a field theory glass. States belonging to matter field representations of $G^{(P)}$ must obey $h_P\psi = \psi$, leading to the quantisation rule for weak hypercharge:

$$\begin{aligned} &\frac{y}{2} + \frac{1}{2}\text{“duality”} + \frac{1}{3}\text{“triality”} \\ &+ \frac{m_5}{5}\text{“quintality”} + \ldots, + \frac{m_P}{P}\text{“}P\text{-ality”} = 0(\mathrm{mod}\ 1)\ . \end{aligned} \tag{72}$$

Here the "N-ality" is the rank index[27] of the representation ψ under the SU(N) subgroup: it counts, modulo N, the number of $\underline{\mathrm{N}}$ representations used to build up the representation ψ.

In the case of the SMG, Eq. (72) expresses the usual electric charge quantisation rule, Eq. (1) of Chapter III, of the standard model. As we have already seen in Chapter 3.2, the experimentally observed pattern of SMG matter field representations is the smallest one, consistent with the requirements of cancellation of gauge and mixed anomalies and fermion mass protection. This minimal fermion spectrum is just repeated a number of times, corresponding to the number of quark-lepton generations. We consider the number of generations, predicted by fitting the fine structure constants of the standard model using random dynamics, in the following subsection.

7.2.4. Numerical predictions from random dynamics

In addition to the many features of random dynamics which favour the survival of the SMG as a low energy symmetry group, one of the most suggestive results is the prediction[29] of the number of quark and lepton generations: $N_{\mathrm{gen}} = 3$.

As discussed in the previous subsection practically any gauge group may be found locally, at the fundamental scale l_{P}, in a field theory glass. In general, however, we only expect gauge groups of relatively small dimension to survive at larger distance scales. A randomly selected compact connected group K, having a dimension of a given order of magnitude, will typically have a simply connected covering group $\tilde{K}$ equal to the direct product of a number of repeated factors of U(1), SU(2), SU(3), SO(5) etc:

$$\begin{aligned}\tilde{K} =& \mathrm{U}(1) \times \mathrm{U}(1) \times \ldots \times \mathrm{U}(1) \times \mathrm{SU}(2) \times \ldots \times \mathrm{SU}(2) \\ & \times \mathrm{SU}(3) \times \ldots \times \mathrm{SU}(3) \times \mathrm{SO}(5) \times \ldots .\end{aligned} \tag{73}$$

The randomly selected gauge group K will in general be obtained by dividing out some discrete subgroup D of the centre from $\tilde{K}$:

$$K = \tilde{K}/D\,. \tag{74}$$

On the one hand gauge groups K which have a charge conjugation like automorphism, but no chiral fermions to protect them, will collapse by confusion with their own charge conjugate. On the other hand it is most likely that the surviving gauge groups have as few matter fields as possible, in small chiral representations. The chiral fermions must, however, satisfy the conditions for the cancellation of gauge and mixed anomalies. In order to obtain a minimal set of chiral fermions, one is therefore led to consider fermions belonging to non-trivial representations of several of the subgroup factors in $\tilde{K}$:
1) The same chiral fermions can thereby protect several of the factors against charge conjugation confusion, and 2) the division of $\tilde{K}$ by the discrete subgroup D may

require the smallest allowed representations of K to couple to more than one gauge subgroup factor.

As an approximation, motivated by the random dynamics expectation of small chiral fermion representations, we shall ignore the possible existence of fermions involving higher dimensional gauge groups than U(1), SU(2) and SU(3). Even the smallest representations for, say, the SU(P) groups, with P prime and greater than 3, have $P = 5, 7, \ldots$ particles and quickly become rather large compared to the irreducible representations of the standard model. Furthermore if the representations are non-trivial under both SU(P) and SU(2) or SU(3), even more particles are required, i.e. $2P$ or $3P$. As discussed in Chapter 3.2, the anomaly cancellation conditions for chiral fermions can be fulfilled in a very particle economic way, using just the U(1), SU(2) and SU(3) groups of the standard model: in fact one quark-lepton generation constitutes the smallest anomaly free and mass protected representation of the standard model group. This principle of small chiral representations then favours the situation in which any higher dimensional SU(P) groups have no fermion representations at all. We will estimate later the confinement energy scale for such higher dimensional SU(P) gauge groups, should they exist.

Consequently we may imagine reducing the gauge group K to a direct product of several factors of U(1), SU(2) and SU(3), each combined to form the standard model group

$$\mathrm{SMG} = \mathrm{S}(\mathrm{U}(2) \times \mathrm{U}(3)) . \tag{75}$$

Each SMG factor is provided with just one generation of chiral fermions; the minimum number required to avoid collapse by confusion. Any left over uncombined factors of U(1), SU(2) or SU(3) are expected to collapse by one of the mechanisms discussed in the previous subsection. We are thus led to consider the direct product of N_{gen} copies of the standard model group:

$$K_{N_{\mathrm{gen}}} = (\mathrm{SMG})_1 \times (\mathrm{SMG})_2 \times \ldots \times (\mathrm{SMG})_{N_{\mathrm{gen}}} . \tag{76}$$

This direct product group $K_{N_{\mathrm{gen}}}$ has outer automorphisms, corresponding to the permutations of the isomorphic SMG factors. It therefore, in turn, collapses by confusion to the diagonal subgroup:

$$\mathrm{SMG}_{\mathrm{diag}} = \{(h, h, \ldots, h) | h \in \mathrm{SMG}\} \tag{77}$$

which has no outer automorphisms unprotected by chiral fermions. It is the 'massless' gauge bosons associated with this diagonal subgroup, which are to be identified with the phenomenologically relevant gauge particles. The remaining gauge bosons all acquire a mass of the order of the confusion breakdown energy scale.

The breakdown of the gauge group $K_{N_{\mathrm{gen}}}$ to $\mathrm{SMG}_{\mathrm{diag}}$ is expected to occur just below the Planck scale, where the three gauge couplings $(g_i^2)_{\mathrm{diag}}$ of the diagonal subgroup are related to the gauge couplings $(g_i^2)_1, (g_i^2)_2, \ldots, (g_i^2)_{N_{\mathrm{gen}}}$ of the parent group factors $(\mathrm{SMG})_1, (\mathrm{SMG})_2, \ldots, (\mathrm{SMG})_{N_{\mathrm{gen}}}$:

$$\frac{1}{(g_i^2)_{\mathrm{diag}}} = \frac{1}{(g_i^2)_1} + \frac{1}{(g_i^2)_2} + \ldots + \frac{1}{(g_i^2)_{N_{\mathrm{gen}}}} . \tag{78}$$

This result is easily obtained by using the fact that for an oscillation in the U(1), SU(2) or SU(3) component of the diagonal subgroup field $A^a_{\mu_{\rm diag}}(x)$, all the parent fields $A^a_{\mu 1}(x), A^a_{\mu 2}(x), \ldots, A^a_{\mu_{N_{\rm gen}}}(x)$ oscillate in the same way:

$$A^a_{\mu 1}(x) = A^a_{\mu 2}(x) = \ldots = A^a_{\mu_{N_{\rm gen}}}(x) = A^a_{\mu_{\rm diag}}(x)\,. \tag{79}$$

The Lagrangian density for such a diagonal subgroup oscillation becomes

$$\begin{aligned}\mathcal{L}_{\rm diag}(x) = &-\frac{1}{4(g^2)_1}\left(F^a_{\mu\nu 1}(x)\right)^2 - \frac{1}{4(g^2)_2}\left(F^a_{\mu\nu 2}(x)\right)^2 \\ &- \ldots - \frac{1}{4(g^2)_{N_{\rm gen}}}\left(F^a_{\mu\nu_{N_{\rm gen}}}(x)\right)^2\end{aligned} \tag{80}$$

which, using Eq. (79), may be rewritten:

$$\mathcal{L}_{\rm diag}(x) = -\frac{1}{4}\left(\frac{1}{(g^2)_1} + \frac{1}{(g^2)_2} + \ldots + \frac{1}{(g^2)_{N_{\rm gen}}}\right)\left(F^a_{\mu\nu_{\rm diag}}(x)\right)^2. \tag{81}$$

This is of the usual form

$$\mathcal{L}_{\rm diag}(x) = -\frac{1}{4(g^2)_{\rm diag}}\left(F^a_{\mu\nu_{\rm diag}}(x)\right)^2 \tag{82}$$

provided Eq. (78) is satisfied for the U(1), SU(2) and SU(3) gauge coupling constants.

It follows from Eq. (78) that the gauge coupling constants $(g_i^2)_{\rm diag}$ for the diagonal subgroup are smaller than the gauge coupling constants $(g_i^2)_1, (g_i^2)_2, \ldots, (g_i^2)_{N_{\rm gen}}$ for the parent group. In particular, if the latter are all taken to be equal we would have

$$(g_i)_{\rm diag} = \frac{(g_i)_1}{\sqrt{N_{\rm gen}}}\,. \tag{83}$$

There is thus a relationship between the standard model gauge coupling constants, identified with $(g_i)_{\rm diag}$, and the gauge coupling constants of the parent group $K_{N_{\rm gen}}$ at the confusion breakdown energy scale μ_c. The confusion scale μ_c is presumably of the same order of magnitude but less than the fundamental cut-off, which we associate with the Planck mass $\mu_{\rm P}$. The requirement that the parent gauge group $K_{N_{\rm gen}}$ should not be in a confining phase, in the small energy gap between μ_c and $\mu_{\rm P}$, is a crucial ingredient of the random dynamics prediction relating $N_{\rm gen}$ to the phenomenologically measured gauge coupling constants. Usually confinement is defined by the presence, at sufficiently large distances, of a confining force which prevents the separation of the confined particles. In the following we shall use the concept of confinement at a certain scale of energy or of length. If the running gauge coupling constant, at a certain energy scale μ, is strong enough that a confining force is already active at distances of order μ^{-1}, we shall say that there is confinement at the scale μ. In fact we are really not so much concerned with

confining particles, as with the correlation length being less than or equal to μ^{-1}. So we shall use the word 'confinement' to describe such a strong coupling phase, even when we ignore matter fields.

If the parent direct product gauge group $K_{N_{gen}}$ is in a confining phase, there are no correlations over distances as large as the confusion length scale μ_c^{-1}. It is then hard to imagine that any confusion breakdown of $K_{N_{gen}}$ can make such a field theory function as a Yang-Mills theory, which could be identified with the standard model at large distances. We therefore expect that the parent gauge group coupling constants $(g_i)_1, (g_i)_2, \cdots, (g_i)_{N_{gen}}$ must have upper limits $(g_i)_{\rm crit}$:

$$(g_i)_1 \leq (g_i)_{\rm crit}, \ \ (g_i)_2 \leq (g_i)_{\rm crit}, \ \ \cdots \ , \ \ (g_i)_{N_{gen}} \leq (g_i)_{\rm crit} \, . \tag{84}$$

The critical values $(g_i)_{\rm crit}$ are determined from the requirement that the direct product gauge group $K_{N_{gen}}$ must not confine at energies above the confusion scale μ_c, which is close to the Planck mass: $\mu_c \sim 10^{19}$ GeV. Otherwise all the gauge bosons would be replaced by glueballs with masses of order 10^{19} GeV and not even the diagonal subgroup could survive as a low energy gauge group. These critical couplings are calculated in the mean field approximation (MFA) of lattice gauge theory, where they correspond to a phase transition between confined and Coulomb-like phases. The MFA seems an appropriate way to characterise the physical behaviour of the fields *at the lattice scale*. It is, namely, a very local approximation that does not take into account correlations over distances much larger than the lattice spacing; in this sense the MFA totally ignores long wavelength effects. We thus obtain the upper bounds

$$(g_i(\mu_c))_{\rm diag} \leq \frac{(g_i)_{\rm crit}}{\sqrt{N_{\rm gen}}} \tag{85}$$

for the standard model gauge coupling constants at the confusion scale μ_c.

There are even reasons to believe that the inequality, Eq. (85), should be saturated. The field theory glass random action, Eq. (44), should be imagined as chosen from a probability distribution, leading dominantly to action contributions $S_r(\phi(i))$ which are roughly constant, for continuity reasons, over small regions in the space of $\phi(i)$ configurations. The definition of the probability measure for the functions $S_r(\phi(i))$ thus requires a metric to be introduced on the $\phi(i)$ configuration space, so that the typical chosen function is approximately constant over a distance of order unity in this metric. In random dynamics the chances of a gauge symmetry appearing accidentally are the better, the smaller are the orbits, as measured in this metric, of any starting configuraton $\phi(i)$ under the approximate gauge transformations. Whenever an approximate Yang-Mills theory is present in a field theory glass, it should be possible, at least crudely, to identify some co-ordinates of the field theory glass configuration space with the link variables of a lattice gauge theory. The approximate gauge symmetry is most likely to occur by chance, when the variation of a link variable over the gauge group volume corresponds to a variation over a relatively small volume in configuration space. In this case one also expects that the

variation of a plaquette variable, constructed from the link variables, over its range will correspond to a relatively small variation in configuration space. This, in turn, means a slow variation of the randomly selected action contributions $S_r(\phi(i))$ as a function of the plaquette variables. Since the effective gauge field action is proportional to $\frac{1}{g^2}$, we conclude that gauge symmetries are most likely to appear by chance for strong gauge couplings g^2. As discussed above, such a gauge symmetry cannot survive at long wavelengths, if the coupling is strong enough to cause confinement in the small energy gap between the fundamental scale μ_P and the confusion scale μ_c. So the most likely coupling resulting from the field theory glass, which is still able to contribute to low energy physics, must be the largest possible consistent with the existence of a Coulomb-like phase in the energy gap. This means saturation of our inequality, Eq. (85).

We have argued above that small orbits of the gauge group are favoured in the field theory glass configuration space, by giving a better chance for the accidental appearance of the gauge symmetry. This, in turn, suggests that the gauge group itself should be small in some physical sense, in order to favour such accidental symmetry. Our conclusion that strong gauge group couplings are most likely to appear in a field theory glass actually corresponds to the gauge group being small, in terms of a group metric normalised by the inverse squares of the gauge coupling constants. The gauge coupling constants then act as units of distance in the gauge group manifold; there are different such units along the different invariant subgroups. Normalising in this way, the Yang-Mills Lagrangian density $\mathcal{L}_{\mathrm{YM}}(x)$ is naturally expressed in terms of a metric tensor g_{ab} defined by

$$\mathcal{L}_{\mathrm{YM}}(x) = -\frac{1}{4} g_{ab} F^a_{\mu\nu}(x) F^{\mu\nu b}(x) . \tag{86}$$

Here then

$$g_{ab} = \frac{1}{g_i^2} \delta_{ab} \tag{87}$$

for a and b denoting basis vectors in the ith invariant subalgebra.

Assuming that the inequality of Eq. (85) is saturated, it is possible to express the standard model gauge coupling constants $(g_i(\mu_c))_{\mathrm{diag}}$, for $\mu_c \sim 10^{19}$ GeV, in terms of the MFA critical values[30]:

$$(g_1^2)_{\mathrm{crit}} = (g^2_{\mathrm{U}(1)})_{\mathrm{crit}} = 1.98 , \tag{88}$$

$$(g_2^2)_{\mathrm{crit}} = (g^2_{\mathrm{U}(2)})_{\mathrm{crit}} = 1.14 , \tag{89}$$

$$(g_3^2)_{\mathrm{crit}} = (g^2_{\mathrm{U}(3)})_{\mathrm{crit}} = 0.82 . \tag{90}$$

In first approximation we here ignore the difference between U(N) and SU(N) critical couplings. The predicted values of $g_i(\mu_c)_{\mathrm{diag}}$ may then be extrapolated, by means of the renormalisation group equations [Paper 16], from μ_c down to the experimental scale $\mu \sim 100$ GeV. Including just the standard model fields for N_{gen} generations in the renormalisation group extrapolation, rather good agreement is

obtained with the experimentally measured gauge coupling constants for $N_{\rm gen} \approx 3$ to 4.

The concept of an MFA critical coupling $(g_1)_{\rm crit}$ for the U(1) group is ambiguous by a factor of 6, depending on the choice of the weak hypercharge quantum $\frac{1}{2}y_{\rm quantum}$. The standard model particles which are neutral under the non-Abelian groups SU(2) and SU(3), such as the right-handed electron, have even integer weak hypercharge: $\frac{1}{2}y_{\rm quantum} = 1$, corresponding to the critical value $(g_1)_{\rm crit}$ given in Eq. (88). However generally the standard model particles have weak hypercharge values, see Table 3.1, which are multiples of one-third: $\frac{1}{2}y_{\rm quantum} = \frac{1}{6}$, corresponding to a six times larger value for $(g_1)_{\rm crit}$. It is not *a priori* obvious whether to use $\frac{1}{2}y_{\rm quantum} = 1$ or $\frac{1}{6}$ as the U(1) quantum. A speculative random dynamics analysis of the critical couplings for SMG $= S(\mathrm{U}(2) \times \mathrm{U}(3))$ lattice gauge theory resolves the ambiguity in favour of $\frac{1}{2}y_{\rm quantum} = 1$ and, hence, the smaller value for $(g_1)_{\rm crit}$ given in Eq. (88).

In this way fits are obtained of $N_{\rm gen}$ to the experimental values of the U(1), SU(2) and SU(3) gauge coupling constants:

$$N_{\rm gen} = 3.4 \quad \text{from} \quad \mathrm{U}(1) \tag{91}$$

$$N_{\rm gen} = 3.5 \quad \text{from} \quad \mathrm{SU}(2) \tag{92}$$

$$N_{\rm gen} = 3.1 \quad \text{from} \quad \mathrm{SU}(3). \tag{93}$$

An earlier version[29] of this model similarly predicted $N_{\rm gen} \approx 3$ quite accurately, prior to the LEP measurement of the number of light neutrino species.[31]

If the random dynamics argument for the saturation of the inequality Eq. (85) is not accepted, our predictions for $N_{\rm gen}$ become upper limits:

$$N_{\rm gen} \leq 3.5 \quad \text{from SU(2)} \tag{94}$$

$$N_{\rm gen} \leq 3.1 \quad \text{from SU(3)}. \tag{95}$$

Almost any further particles, which couple to the standard model SU(2) or SU(3) gauge fields, populating the 'desert' between $\mu \approx 100$ GeV and $\mu_c \approx 10^{19}$ GeV would effectively violate these inequalities, due to their modification of the renormalisation group beta functions. In particular supersymmetric particles much below the 10^{19} GeV scale are not allowed by random dynamics. Also technicolour models are ruled out, unless there are very few techni-particles. Random dynamics essentially predicts the existence of the 'desert'. The crucial ingredient in these predictions is the assumed embedding of the standard model gauge group, as the diagonal subgroup, in a direct product of $N_{\rm gen}$ copies of the SMG = S(U(2)×U(3)) group at the confusion scale $\mu_c \sim 10^{19}$ GeV.

We will now briefly consider the possible physical existence of the higher dimensional SU(5), SU(7),...,SU(P) gauge groups, which appear as factors in the random dynamics favoured series of groups $G^{(P)}$ of Eq. (65). Let us first estimate at what energy scales they might confine in the 'desert'. For this purpose, we assume that

the above confusion scheme generalises to a direct product gauge group of N_{gen} factors of $G^{(P)}$, which breaks down to its diagonal subgroup at the confinement scale $\mu_c \approx 10^{19}$ GeV. It turns out that the MFA critical coupling constants $(g_N)_{\text{crit}}$, for SU(N) gauge groups, are given approximately by

$$N(g_N^2)_{\text{crit}} \approx 2.5\,. \tag{96}$$

Assuming critical couplings for the parent direct product factor groups at the confusion scale leads, by saturation of Eq. (85), to the result

$$(Ng_N^2(\mu_c)) = \frac{N(g_N^2)_{\text{crit}}}{N_{\text{gen}}} \approx \frac{2.5}{N_{\text{gen}}}\,. \tag{97}$$

The renormalisation group beta function for the coupling parameter $(Ng_N^2)^{-1}$ is N-independent:

$$\frac{d}{d\ln\mu}\left(\frac{1}{Ng_N^2(\mu)}\right) = \beta_{1/Ng_N^2} = \frac{11}{24\pi^2} \tag{98}$$

ignoring the contribution of any matter fields. Thus, in the absence of matter fields, the SU(N) gauge couplings would all diverge at approximately the same confinement energy scale μ_{conf}:

$$\mu_{\text{conf}} = \mu_c \exp\left(-\frac{24\pi^2}{11}\cdot\frac{N_{\text{gen}}}{2.5}\right) \tag{99}$$

$$= 6\times 10^{-12}\mu_c \tag{100}$$

$$\simeq 6\times 10^{7}\ \text{GeV}\,. \tag{101}$$

The inclusion of fermion matter fields coupled to the SU(N) gauge fields would lower the confining energy scale, as happens in QCD. The contribution of, say, an $\underline{N}$ representation of fermion fields to the beta function becomes smaller, relative to the gauge field contributions, as N increases; so the higher dimensional SU(N) groups should confine at higher energy scales, as intuitively expected.

According to the quantisation rule Eq. (72), the matter fields for the SU(P) groups, with P prime, should at least carry some weak hypercharge. We must therefore consider whether such matter fields, populating the 'desert', are consistent with the above random dynamics results for N_{gen}. In order to construct an anomaly free and mass protected fermion representation, using $\underline{P}$ and $\underline{P}^*$ irreducible representations, one is led to include at least one representation with $\left|\frac{Y}{2}\right| > 1$. The contribution of such a P-dimensional representation with $\left|\frac{Y}{2}\right| > 1$ to the beta function, for the U(1) gauge coupling constant, is at least equivalent to that of $\frac{3P}{10}$ ordinary quark-lepton generations. It happens that approximately half of the N_{gen}-dependence, of the above random dynamics predicted standard model gauge couplings, is due to the renormalisation group extrapolation; the other half is due to the number of SMG factors in Eq. (76) being N_{gen}. Thus $\frac{3P}{10}$ extra generations contributing to the beta function would lower our prediction Eq. (91), for the number of ordinary

quark-lepton generations, by roughly $\frac{3P}{20} \approx \frac{P}{7}$. For, say, SU(5) this would lower the U(1) prediction from $N_{\rm gen} = 3.4$ to $N_{\rm gen} \approx 2.7$, which is just consistent with three quark-lepton generations.

However, if the confusion mechanism is to work properly, the same matter field representations should be associated with each of the $N_{\rm gen}$ factors of SU(P) in the parent direct product factor group. Then, for three generations, the correction to the fitted value, Eq. (91), for $N_{\rm gen}$ would be increased by a factor of three to $\frac{3P}{7}$: so the prediction, Eq. (91), would be lowered to $N_{\rm gen} \approx 1.3$ or less, in strong disagreement with three quark-lepton generations. It would be possible to avoid this disagreement, if the weak hypercharge quantum could be taken to be $\frac{1}{2}y_{\rm quantum} = \frac{1}{6}$ instead of the preferred value $\frac{1}{2}y_{\rm quantum} = 1$. Weyl fermions coupling as both a P-dimensional representation of SU(P), and as a doublet of SU(2) or a triplet of SU(3), would lower the predictions of Eqs. (92) and (93) for $N_{\rm gen}$ by approximately $\frac{P}{8}$. This could be tolerated for one representation, but not for the $N_{\rm gen} = 3$ copies required to make confusion work.

Thus, if matter fields are assigned to the higher dimensional SU(P) groups, random dynamics does not really fit with three quark-lepton generations; at least, if $\frac{1}{2}y_{\rm quantum} = 1$ is used. We conclude that there are no such matter fields populating the 'desert'. The higher dimensional SU(P) gauge fields — should they exist — would then be hard to detect: SU(P) glueballs might only function as dark matter in the Universe.

Another numerical result from random dynamics, but with only order of magnitude accuracy, is a computer simulation of the quark-lepton mass spectrum and weak mixing angles, using approximately conserved random 'horizontal' U(1) chiral charges [Paper 11]. Good qualitative agreement is obtained with the experimental masses and mixing angles, but with evidence for some unifying interaction beyond the standard model:

a) The similarity between the mass spectra for the three charged fermion families, i.e. the $Q = \frac{2}{3}$ quarks, $Q = -\frac{1}{3}$ quarks and $Q = -1$ leptons, indicates a need for correlations between right-handed quarks and between quarks and leptons. These correlations are most easily introduced by postulating the existence of right-handed currents and leptoquark currents. Without such correlations between the horizontal quantum numbers on right-handed quarks, our computer simulation predicts too large fluctuations in the logarithm of a fermion mass inside a given generation. In other words, the masses of the electron, up quark and down quark are much more similar than expected from just the standard model, when they are treated as the lightest members of three independent fermion families. This is also true of the second (c, s, μ) generation. The predicted spread in the logarithm of fermion mass is smallest for the heaviest or third generation: phenomenologically the third generation has the largest spread and, for a top quark mass $m_t \approx 100 - 200$ GeV, is in agreement with the model prediction without using any right-handed or leptoquark symmetry. This would suggest that right-handed currents and leptoquark currents are needed for just the two

lighter generations.

b) The smallness of the off-diagonal elements of the quark weak coupling matrix[32] also suggests the existence of correlations between the right-handed quark components. The large weak interaction transition matrix elements between quarks in the same generation, $u \to d, c \to s$ and $t \to b$, and small inter-generation transition matrix elements would only occur accidentally, if not enforced by a correlation between the random horizontal quantum numbers inside a generation.

There are two free parameters in the simulation, which can be fitted to give the overall mass scale, set by the W-mass, and the strength of the horizontal symmetry breaking interaction respectively. The results of the computer simulation are not very sensitive to the number of types of horizontal chiral charge introduced, nor to the possible existence of further generations.

After elimination of the two fitted parameters, our main prediction for the fermion masses [Paper 11] becomes:

$$\frac{\ln(m_3/m_2)}{\ln(m_2/m_1)} \approx 0.6 \pm 0.3 \qquad (102)$$

where m_i denotes the ith generation mass within a family of quarks or leptons of a given electric charge. It agrees rather well with experiment, but is of course not a very strong prediction. If right-handed currents are introduced, the quark mixing matrix elements are naturally reproduced, within the large statistical fluctuations of our model,[32] but with a somewhat low average value for the Cabibbo angle which mixes d and s quarks. The phase angle, in the standard particle data group parameterisation[33] of the weak mixing matrix of [Paper 14], is of order unity in our model. The correct order of magnitude is therefore obtained for the CP violating parameter ϵ in kaon decay.

Another general conclusion from our statistical analysis is that grand unified models, such as SU(5) or SO(10), which put some left-handed fermions in the same representation as their antiparticles, tend to give one or more cases of families with a pair of almost mass degenerate particles.[34] The lack, phenomenologically, of such inter-generation approximate mass degeneracies disfavours grand unified models like SU(5) or SO(10).

Grand unified models based on simple gauge groups like SU(5) or SO(10) are, in any case, not expected to occur in random dynamics, according to the general field theory glass scenario discussed in the previous subsection. However gauge fields coupling to the right-handed components of quarks and leptons, e.g. an $SU(2)_R$ gauge group, are also *not* expected to survive the confusion breakdown mechanism of random dynamics. Thus there is an apparent conflict between random dynamics and the suggested existence of right-handed currents from our mass matrix simulation.

It is assumed in our mass matrix simulation model that the chiral horizontal symmetry group commutes with the standard model gauge group. However, random

dynamics suggests the existence of an intermediate gauge group

$$K_3 = (\mathrm{SMG})_1 \times (\mathrm{SMG})_2 \times (\mathrm{SMG})_3 \tag{103}$$

at the confusion scale; with one direct product factor $(\mathrm{SMG})_i$ for each quark-lepton generation. With such a direct product unified gauge group, it is possible that the separate gauge quantum numbers of $(\mathrm{SMG})_1$, $(\mathrm{SMG})_2$ and $(\mathrm{SMG})_3$ could play the role of horizontal quantum numbers and be responsible for the different mass scales of the three generations.

Let us now summarize the results obtained from random dynamics:

1) The standard model gauge group SMG = S(U(2)× U(3)) is strongly favoured.
2) The number of generations, $N_{\mathrm{gen}} = 3$, is successfully related to the phenomenological values of the gauge coupling constants for the three invariant subgroups of SMG. This result essentially requires a 'desert' between presently available energies and the Planck scale.
3) The successful but rather crude prediction, Eq. (102), for the ratio of logarithms of quark or lepton mass ratios. Also the weak mixing matrix is statistically predicted, in qualitative agreement with all phenomenology: but the suggested existence of an $\mathrm{SU}(2)_R$ gauge group would violate the confusion mechanism of random dynamics.

It is of course rather dangerous, in general, to conclude backwards from a successful prediction that the underlying theory is correct. This is especially true of random dynamics predictions, because they should follow from almost all models: so if they turn out to agree with experiment, they may not be suggestive of any particular theory. On the other hand if a random dynamics prediction fails, one should really learn something.

It would indeed be most unreliable to conclude anything about the truth of random dynamics from the successful result 3) above. The computer simulation of quark and lepton masses uses a type of mass matrix, which could crudely arise in many models that would seem realistic from more conventional points of view.

The results 1) and 2) seem more promising evidence in favour of random dynamics, but they are far from conclusive. At first sight it may seem that one uses many unconventional assumptions built into the field theory glass, such as discretized space-time and many gauge groups of the same type at the fundamental level. However, at the end, it is the confusion breakdown mechanism which is the important ingredient in deriving both 1) and 2). So the main supported element of random dynamics, if anything at all, may be 'confusion' and the lack of outer automorphisms in the gauge group of nature. As previously emphasized, the fact that so many of the symmetries found in the laws of nature can be derived is a strong argument in favour of the random dynamics model.

A priori the main evidence against random dynamics is provided by the well-known fine-tuning problems: 1) the hierarchy problem of why the $W^{\pm}$ and Z^0 masses are so small compared to the Planck mass, 2) the vanishing cosmological constant problem and 3) the problem of CP symmetry in strong interactions.

Baby universe theory may resolve some of these problems, but this subject is still controversial. Fine-tuning problems are of course especially worrying for random dynamics, which is based on the principle of no fine-tuning taken to its extreme.

7.3. Classification of Symmetry Derivations

How can it happen that symmetries appear in physical models without having been put in? We attempt to answer this question here, by classifying the symmetry derivations collected in this book.

We first consider classifying our examples of symmetry appearance, according to whether or not the symmetry is really present in the language of the original model, in which it is to be derived. For example the quark and lepton flavour — strangeness, charm, muon number etc. — conservation laws are true symmetries which are really present in the strong and electromagnetic interactions [Paper 8]. We note here that the quark flavour conservation laws appear as symmetries of the standard model in the low energy limit, where the weak interactions can be neglected; nonetheless the symmetry operations really leave the model invariant in this limit. However, in the case of the macroscopic scaling symmetry discussed in Chapter II, the scaling transformations cannot really be performed, in an unambiguous way, on the molecules of the underlying model of molecular physics. Similarly the formal derivation of gauge symmetry [Paper 26] is an example in which the symmetry is *not* present in the original model.

For convenience we shall refer to the original model, in which the symmetry, is to be derived, as the "fundamental" model or theory. We shall then divide the symmetry derivations into two main groups:

I. Symmetry really present in the fundamental theory.

II. Symmetry not really present in the fundamental theory.

It must be admitted that in some of our examples it is a slight exaggeration to talk about a symmetry 'derivation', when there is already a larger symmetry manifestly built into the fundamental model considered. A symmetry derived in this way will be called a 'transformed symmetry'. For example, if Poincaré invariance is considered to arise from general relativity as a special case of diffeomorphism symmetry, see Chapter 6.2.1, it is clearly a rather mildly transformed symmetry. A more genuine transformed symmetry is the case of the Kaluza-Klein model [Paper 20], in which diffeomorphism symmetry turns into gauge symmetry. In cases of transformed symmetry, the symmetry is really there in the fundamental model and belongs to group I; we subclassify transformed symmetries as category I.0.

For some symmetries of group I, it could simply be that there are so many assumptions, principles or restrictions, built into the fundamental model under consideration, that it is not possible for the structure in different parts of the model to behave in a different manner. When the restrictions imposed on a model are so strong as to imply some symmetry, we say there is a 'symmetry by restriction' and subclassify it as category I.1. This would be analogous to what happens in, say, group theory: Many groups have automorphisms, i.e. symmetry maps compatible

with the group composition law. The axioms of group theory are so 'restrictive' that, very often, they lead to different parts of the group becoming so similar that it is possible to find automorphisms, which map different such structurally identical pieces into each other.

For example, in QCD there are several quark flavours in colour triplet representations, whose interactions are restricted by the following conditions: (a) Poincaré invariance, (b) colour gauge invariance, (c) renormalisability, (d) cancellation of gauge anomalies and (e) smallness of representations. These restrictions are so strong that, when the quark masses can be neglected, a symmetry results under unitary transformations independently on the left-handed and on the right-handed components of the quark fields: there is a $U(N_F)_L \times U(N_F)_R$ chiral symmetry, but with an axial U(1) anomaly [Paper 10].

In addition to having restrictions, it might be that the vacuum, or some other important structure, adjusts itself in such a way that some symmetry results. We then say that we have a 'symmetry by adjustment', which we subclassify as category I.2. An example of this phenomenon is the Peccei-Quinn mechanism [Paper 9], leading to CP invariance for the strong interactions. In this case it is really two scalar fields with non-zero vacuum expectation values, which adjust themselves to minimise the vacuum energy density and whose relative phase then cancels the effects of the SU(3) topological θ term.

It is only acceptable to consider symmetries that are not really present in the fundamental theory as examples of symmetry derivation, if they exist as symmetries for a new language describing the physics in terms of degrees of freedom relevant to the limit of conditions under consideration. Such a new language is often introduced when stepping down from one level to the next on the quantum staircase (Fig. 1.1). A symmetry may then happen to be present in the language appropriate to the latter but not in that of the former level. In the case of macroscopic scaling symmetry, for example, we pass from the language of molecular dynamics to the language of macroscopic physics. For group II symmetry derivations, we therefore have a relation R corresponding to translation between the two languages. This relation R between the 'fundamental language' of the original 'model' and the 'effective language' of the new field of physics will, in general, not be one to one.

If many statements in the effective language correspond to the same statement in the fundamental language, it may be possible to identify a symmetry of the former simply by transforming the effective language, without changing anything in the fundamental language. We shall call an example of this type a 'formal symmetry derivation' and subclassify it as category II.1. Many of the gauge symmetry derivations in this book are of this category.

On the other hand, if many statements in the fundamental language correspond to a single statement in the effective language, it means that much of the content in the fundamental theory is ignored in the effective language. Such a procedure of ignoring structure opens up the possibility of finding more symmetries; these 'symmetries by ignoring structure' are subclassified as category II.2. Scaling symmetry

in macroscopic physics is of this type, since one has to ignore a large amount of detailed information, about the positions etc. of the molecules, in order that such a scaling symmetry may be implemented. A huge number of molecular configurations are described, approximately of course, by a given density and velocity field for the fluid in the effective macroscopic language.

The above suggested classification scheme is not entirely free of ambiguity: it turns out to be a matter of degree whether or not a symmetry can be considered as really present in the fundamental model. It namely depends on how much structure is considered to be part of the fundamental model and is required to be respected by the symmetry. A class of examples of such doubtful classifications is what we could call the 'hidden symmetries'. Taken literally, hidden symmetries should mean those that are really present in the fundamental model but are not immediately recognised. So we should expect them to be classified under category I.1: symmetry by restriction. (We here use the term 'hidden symmetries' in a narrower sense than in Chapter 4.2 and 4.3 where it was taken to include local symmetries, such as SU(8) symmetry in $N = 8$ supergravity; local gauge symmetries are more properly classified above as formal symmetries.)

A typical example of a hidden symmetry is the O(4) symmetry of the hydrogen atom or of a planet-sun system [Paper 4]. For an exact non-relativistic Coulomb force system, there is in one sense an exact O(4) symmetry: the Hamiltonian is invariant under the four dimensional rotation group O(4). However it is really only a true symmetry if certain structure is abstracted from the system. There is, for instance, no underlying isometry of the geometrical space of positions of the electron, or the planet, corresponding to the O(4) transformations. An O(4) transformation acts on the full position and momentum phase space, in general mixing up position with momentum, and is not compatible with the spatial metric, i.e., it maps a pair of points into a pair with a different spatial separation. This means that the position space metric is part of the structure that must be abstracted from the fundamental model, if the O(4) transformations are to be considered a symmetry group of the model. If this structure — the metric — is taken to be an integral part of the fundamental model, we must reclassify the hidden O(4) symmetry derivation as II.2: symmetry by ignoring structure. The O(4) transformations only constitute a symmetry of the effective model obtained by ignoring some structure, including the position space and its metric.

The SO(32) and $E_8 \times E_8$ symmetries of the heterotic string [Paper 22] are also hidden symmetries, which can quite analogously be classified as category I.1 or category II.2. In this case it is necessary to ignore the structural difference between a soliton and a phonon mode of excitation of the string. Similarly the hidden non-compact $E_{7(+7)}$ symmetry of the $N = 8$ supergravity model [Paper 18] may be classified under I.1 or II.2, depending on the point of view.

So we shall classify the symmetry derivations, in Tables 7.1–7.5, under the following five headings:

I.0. Transformed symmetries

I.1. Symmetries by restriction
I.2. Symmetries by adjustment
II.1. Formal symmetry derivations
II.2. Symmetries by ignoring structure

The columns of the tables will be used to indicate (1) the role played by any limit in the symmetry derivation, and (2) the suggesting principle motivating the idea that the symmetry might exist.

We consider first the possibility that a) no limit is needed in order to derive the symmetry. For instance the hidden symmetries discussed above, such as the O(4) symmetry of the hydrogen atom, the SU(N) symmetry of the N-dimensional harmonic oscillator etc., are valid exactly inside the idealised models considered: the hydrogen atom is taken to be purely non-relativistic, etc. It is not necessary to restrict the region of study to, say, low energy inside the models considered and the hidden symmetries are placed in column a): no limit taken.

The other possibility b) is, of course, that a limit is significant either by bα) simply causing the validity of the symmetry in that limit, or bβ) suggesting a new point of view or language appropriate to the limit. The two cases bα) and bβ) apply to group I and group II symmetry derivations respectively, i.e. according to whether or not the symmetry is truly present in the fundamental model. Typical symmetry derivations of class bα) are those for which the symmetry appears at low energy, due to the suppression of symmetry breaking terms towards the infrared by the renormalisation group equations. The best working renormalisation group symmetry derivation is that of Lorentz invariance in the infrared, for non-covariant gauge theories with non-covariant chiral fermions [Papers 24, 25]. Macroscopic scaling symmetry, as discussed above, is an example in which the significance of the limit is rather to influence the choice of language and, thereby, to make the symmetry manifest itself, as far as the relevant degrees of freedom are concerned.

When we here talk about 'limits', we mainly have in mind a corner in the space of possible physical conditions; typically where the particles have low enough energy and momentum so that, for example, they can be produced under available experimental conditions. However, in a few of the symmetry derivations presented in this book, it is crucial that some parameters of the fundamental model tend to a special value, say zero. Such a limit cannot be reached simply by restricting the experimental conditions; one must rather hope that nature has chosen, for some as yet to be understood reason, to select parameter values close to the limit. An example of a parameter, relevant to understanding the origin of symmetries, is the ratio of a quark mass m_q to the QCD mass scale parameter Λ_{QCD}. It is, for instance, the smallness of this ratio for some quark flavours that is responsible for the chiral symmetry of the strong interactions [Paper 10].

We therefore distinguish between 'experimental' limits and 'parameter' limits in the tables, by introducing separate columns labelled bαe) and bαp), for symmetries which are truly present in some working limit of the physical conditions, e.g. the low energy limit, and some parameter limit respectively. We have found no examples

of a parameter limit in the group II symmetry derivations and we, therefore, have a single column labelled bβ) in Tables 7.4 and 7.5.

The first column of Tables 7.1–7.5 is used to give a suggesting principle, where appropriate, from which the idea of the symmetry group might be said to originate. For instance, many of the fundamental models considered contain the geometry of Minkowski space-time, or of just three-dimensional space, as an ingredient in their structure. True symmetries of such a model must transform, although perhaps trivially, this geometrical structure in an automorphic way. This leads to the idea that symmetries of the geometry are especially promising candidate symmetries for the full model. Derived symmetries which can be interpreted in this way will be called geometrically suggested symmetries. We have found several symmetry proposal mechanisms useful in classifying symmetry derivations and we list them below:

Geometry
Fermion number
Analytic continuation
Repetition or permutation
Stability subgroup
Approximate symmetry

The first four of the above symmetry suggesting principles are used to subclassify the restriction symmetries of Table 7.2. Some of these symmetry derivation examples were difficult for us to subclassify and, for such 'left-over' cases, we left the first column blank. The principle of 'Analytic continuation' is, in practice, applied to the Lorentz group, in order to derive the CPT theorem and to motivate the related idea of particle-antiparticle or charge conjugation symmetry. The other two restriction symmetry subclassifications — 'Repetition or permutation' and 'Fermion number' — represent rather general ideas. In the case of 'Repetition or permutation' inspired symmetries, some objects in the model — e.g. particles — are restricted so strongly that they have identical behaviour in almost all respects; thus they become related to each other by the permutation symmetry operation. In quantum mechanics, even superpositions of such identically behaving objects are considered; so the permutation symmetry is extended to a symmetry under the transformations of a unitary group. Similarly the idea of various types of 'Fermion number' suggests the existence of generalised charge conservation laws and symmetry under the associated phase transformations.

The subclassification of the formal symmetry derivations in Table 7.4 required us to introduce the last two symmetry proposal mechanisms listed above. Several of the formal symmetry examples are suggested by considering a field taking values in a homogeneous space. This homogeneous space is identifiable with the space of cosets G/H, where G is the group of transformations of this space and H is a subgroup in G: the subgroup H is the 'Stability subgroup' for a point in the homogeneous space. The stability subgroup H becomes the gauge symmetry group in the examples considered, which we therefore identify as 'Stability subgroup' suggested. In the

formal gauge symmetry derivations of [Paper 26] and of the field theory glass, an approximate gauge symmetry of the fundamental model is shown to manifest itself as an exact symmetry physically, by the inverse Higgs mechanism. The gauge symmetry group derived is of course suggested by the 'Approximate symmetry' of the fundamental model.

The tranformed symmetries of Table 7.1 are already naturally suggested by the original symmetry group of the fundamental model considered, which simply gets transformed into the derived symmetry. So we have crossed out the first column of Table 7.1 as redundant. However it must be admitted that sometimes, as in the case of general relativity, the symmetry group of the fundamental model is so huge that the choice of subgroup to be transformed requires further motivation. The Poincaré subgroup is suggested as the physically relevant symmetry, since it is the only surviving unbroken symmetry in the usual flat space-time vacuum. Our so-called derivation of Poincaré invariance from general relativity could therefore be said to be 'suggested by the flat universe'.

We now present Tables 7.1–7.5 which constitute a classification of all of the symmetry derivations discussed in this book, according to the above principles. The number in front of each entry refers to Table 7.6, which is a complete list of the symmetry derivations given in the order in which they occur in our commentary. Table 7.6 also contains cross-references to the classifications given in Tables 7.1–7.5 and indicates when a symmetry derivation does not work or only works rather inaccurately.

Table 7.1. Classification of symmetry derivations: Transformed symmetry derivations.

I.0 Transformed symmetries			
Suggesting principle	a) No limit taken	bαe) Symmetry truly present in working limit	bαp) Symmetry truly present in parameter limit
	25 Kaluza-Klein gauge symmetry	32 Translational invariance of glass	
	27 Gauge symmetry from strings		
	28 Diffeomorphism symmetry from strings		
	30 Poincaré invariance from general relativity		

Table 7.2. Classification of symmetry derivations: Symmetries by restriction.

I.1	Symmetries by restriction		
Suggesting principle	a) No limit taken	b∘c) Symmetry truly present in working limit	b∘p) Symmetry truly present in parameter limit
Geometry	19 Conformal invariance from scaling	2 Parity of macroscopic elasticity	
	33 Lorentz invariance from strings	5 Spin decoupling in atomic physics	
		18 Bjorken scaling	
		31 Lorentz invariance from renormalisation group	
Fermion number	8 Quark flavours in QCD and QED	8 Quark flavours in Standard Model	
	13 Baryon number B (excluding weak anomaly)	13 Baryon number B (including weak anomaly)	
	14 Lepton flavours L_e, L_μ, L_τ (excluding weak anomaly)	14 Lepton flavours L_e, L_μ, L_τ (including weak anomaly)	
	15 Lepton flavour differences $L_e - L_\mu, L_e - L_\tau$		
	16 B-L from Standard Model		
	20 B-L from SU(5) GUT		
Analytic continuation	9 Charge conjugation C from QCD and QED	9 Charge conjugation C from Standard Model	
	29 CPT		
Repetition permutation	26 SO(32) and $E_8 \times E_8$ from the heterotic string		12 Chiral symmetry $SU(N_F)_L \times SU(N_F)_R$
			17 Flavour conservation for neutral currents (GIM)
	3 O(4)-symmetry of hydrogen atom	35 Gauge symmetry from renormalisation group	
	4 SU(n)-symmetry of n-dimensional harmonic oscillator		
	24 $E_{7(+7)}$-symmetry from $N=8$ supergravity		

Table 7.3. Classification of symmetry derivations: Symmetries by adjustment.

I.2	Symmetries by adjustment		
Suggesting principle	a) No limit taken	bαϵ) Symmetry truly present in working limit	bαβ) Symmetry truly present in parameter limit
Analytic continuation and geometry of time and space		10 & 11 *CP, T, P* from PQ-extended Standard Model	
	37 *CP, T, P* from baby universe extended QCD	37 *CP, T, P* from baby universe extended Standard Model	

Table 7.4. Classification of symmetry derivations: Formal symmetries.

II.1	Formal symmetry derivation	
Suggesting principle	a) No limit taken	bβ) Limit suggests language giving symmetry
Stability subgroup		21 U(1)-gauge symmetry in CP^{n-1}-model
		22 *H*-gauge symmetry in *G/H* sigma model
		23 SU(8)-gauge symmetry in $N = 8$ supergravity
Approximate symmetry		34 Formal appearance of gauge symmetry

Table 7.5. Classification of symmetry derivations: Ignoring structure.

II.2	Symmetries by ignoring structure	
Suggesting principle	a) No limit taken	bβ) Limit suggests language giving symmetry
Geometry		1 Macroscopic scaling
Repetition and permutation	26 SO(32) and $E_8 \times E_8$ from the heterotic string	
	3 O(4)-symmetry of hydrogen atom	
	4 SU(n)-symmetry of n-dimensional harmonic oscillator	
	24 $E_{7(+7)}$-symmetry from $N=8$ supergravity	

Table 7.6. List of symmetry derivations.

Chapter II. Symmetries from non-relativistic physics		
1.	Macroscopic scaling from molecular physics	II.2.bβ geometry
2	Parity invariance of the elastic properties of sugar	I.1.bαc geometry
3.	O(4)-symmtry of planetary motion and the hydrogen atom [Paper 4]	I.1.a & II.2.a
4.	SU(n)-symmetry for the n-dimensional harmonic oscillator [Paper 5]	I.1.a & II.2.a
5.	Separate conservation of spin and orbital angular momentum in non-relativistic atomic physics (Spin decoupling)	I.1.bαc geometry
6.	Wigner SU(4)-symmetry of nuclear physics [Paper 6]	rather inaccurate
7.	SU(6)-symmetry of strong interactions [Paper 7]	rather inaccurate
Chapter III. Symmetries from the Standard Model		
3.2. Strong and electromagnetic interactions		
8.	Flavour symmetry for quarks and leptons from QCD and QED [Paper 8]	I.1.a fermion number
9.	Charge conjugation C from QCD and QED from Standard Model [Paper 2]	I.1.a I.1.bαc analytic continuation
10.	Parity P from Peccei-Quinn extended Standard Model [Papers 8 & 9]	I.2.bαc geometry
11.	Time reversal T and CP from Peccei-Quinn extended Standard model [Paper 9]	I.2.bαc analytic continuation and geometry
12.	Chiral symmetry $SU(N_F)_L \times SU(N_F)_R$ [Paper 10]	I.1.bαp Repetition and permutation

Table 7.6. Cont'd.

	3.3. Full Standard Model	
13.	Baryon number conservation B [Paper 12]	I.1.boc (incl. anomaly) I.1.a (excl. anomaly) fermion number
14.	Lepton flavour conservation L_e, L_μ, L_τ	I.1.boc (incl. anomaly) I.1.a (excl. anomaly) fermion number
15.	Lepton flavour differences L_e-L_μ, L_e-L_τ (even including weak anomaly)	I.1.a fermion number
16.	Baryon number minus lepton number B-L (even including weak anomaly)	I.1.a fermion number
17.	Flavour conservation for neutral currents (GIM-mechanism) [Paper 13]	I.a.boβ repetition
18.	Bjorken scaling	I.1.boc geometry
19.	Conformal invariance from scaling	I.1.a geometry
	Chapter IV. Beyond the Standard Model	
	4.1. Grand unification	
20.	Baryon number minus lepton number B-L from SU(5)-GUT	I.1.a fermion number
	4.2. Hidden local symmetry and dynamical gauge bosons in non-linear sigma models	
21.	U(1)-gauge symmetry in CP^{n-1} sigma model [Paper 17]	II.1.bβ stability subgroup
22.	H-gauge symmetry from general G/H-sigma model	II.1.bβ stability subgroup
	4.3. Hidden symmetries in $N = 8$ supergravity	
23.	SU(8)-gauge symmetry from $N = 8$ supergravity [Paper 18]	II.1.bβ unlikely to work stability subgroup
24.	$E_{7(+7)}$ global symmetry from $N = 8$ supergravity [Paper 18]	I.1.a or I.2.a
	4.4. Kaluza-Klein theory	
25.	Gauge symmetry in Kaluza-Klein model [Paper 20]	I.0.a
	4.6. Strings	
26.	SO(32) and $E_8 \times E_8$ symmetry from heterotic string [Paper 22]	I.1.a or II.2.a repetition
27.	Gauge symmetry from strings	I.0.a
28.	Diffeomorphism symmetry of external general relativity from strings	I.0.a

Table 7.6. Cont'd.

Chapter V. The CPT-theorem.		
29.	*CPT*-symmetry in quantum field theory [Paper 23]	I.1.a analytic continuation
Chapter VI. The fundamental symmetries		
6.2. Poincaré invariance		
30.	Poincaré invariance from diffeomorphism symmetry of general relativity	I.0.a
31	Lorentz invarince from renormalisation group [Papers 24, 25, 29]	I.1.b∘c geometry
32.	Translational invariance in the long wavelength limit in a glass-like vacuum	I.0.b∘c
33.	Lorentz invariance from string theory [Weinberg]	I.1.a geometry
6.3. Local gauge invariance		
34.	Formal appearance of gauge symmetry [Paper 26]	II.1.b$\not\beta$ approximate symmetry
35.	Gauge symmetry from renormalisation group [Papers 28, 29, 27]	I.1.b∘c Does not work for asymptotically free theory
6.4. Supersymmetry		
36.	Supersymmetry from renormalisation group [Paper 29]	In general it does not work
Chapter VII. 7.2. Random Dynamics		
37.	*CP*, *T*, *P* for strong interactions from baby inverse extended QCD; from baby universe extended Standard Model	I.2.a I.2.b∘c(p) analytic continuation and geometry

Some of the placements of the symmetry derivations in the tables deserve further explanation:

Table 7.1: Transformed symmetries

As mentioned above, some very special diffeomorphism symmetries of general relativity are transformed into Poincaré invariance. Similarly special diffeomorphism symmetries, associated with i) the n extra compact space dimensions in Kaluza-Klein theory and ii) the two-dimensional space-time of the string world sheet, are transformed into gauge symmetries. The translational symmetry of a glassy theory at large distances is inherited from the large scale translational invariance present statistically in the assumed glass-like structure of the vacuum. This derivation of translational invariance comes close to transforming a symmetry into itself and it is a moot point whether it should really be counted as a symmetry derivation.

Table 7.2: Symmetries by restriction

All the 'experimental' limits of column bαe) correspond to the low energy or long wavelength limit, except for the case of Bjorken scaling in QCD. Bjorken scaling symmetry appears in the opposite limit, when the energy is large compared to the mass scale parameter Λ_{QCD} and the relevant quark masses.

The quark flavour symmetries and also charge conjugation symmetry have been classified under the 'no limit' column a), as well as under column bαe). These double placements are made, since the symmetries may be derived either from the standard model, or from a model composed from just QCD and QED. If the symmetries are considered to be derived from the standard model, it is necessary to take a low energy limit in order to avoid weak interaction effects. However, if the "fundamental" model is taken to consist of QCD and QED no such limit is needed.

Baryon number and lepton flavour conservation laws are also presented twice in the table, according to whether or not the effects of the weak interaction anomaly are taken into account. The baryon and lepton number conservation laws are violated by sphaleron-like transitions[35] between the topologically distinct vacua of the standard electroweak theory. In order to eliminate these baryon and lepton number violating processes, it is necessary to take the limit of low temperature[2] relative to the electroweak phase transition temperature, $T \ll 300$ GeV; it may also be necessary to take the limit of low energy[36] relative to the electroweak SU(2) sphaleron energy, $E \ll 10$ TeV. However, the differences between lepton flavour numbers, $L_e - L_\mu$ etc., and between baryon number and lepton number, $B - L$, are anomaly free and exactly conserved in the standard model; no limit is required.

We felt a little doubtful about how to classify the symmetries resulting, in the infrared limit, from the use of the renormalisation group equations. It is not sufficient just to make use of the general restrictions on a renormalisable theory, in order to identify these symmetries. It is also necessary to calculate the β-functions for the various coupling constants, in order to see how the latter develop towards the infrared. There seems to be no *a priori* principle, as to when such a calculation will result in the running coupling constants tending to values which respect some symmetry. In fact, it seems that most of the discussed cases of renormalisation group symmetry derivation only work under somewhat special conditions. The derivation of Lorentz invariance is the most robust example. It is apparently rather accidental whether or not a symmetry is derivable from the renormalisation group.

In order to shed more light on these renormalisation group derived symmetries, the following brief discussion of the physical mechanism underlying the renormalisation group may be helpful. The renormalisation group effects are quantum effects due to loop diagrams. This means that they reflect the effects of particles being virtually split into the 'bare' particles of the fundamental fields. More and more of these splittings are taken into account, as we move towards the infrared limit. This means that, in the far infrared, the observed particles are composed of very many 'bare' constituents, as illustrated in Fig. 7.4.

As an example of the effects of these many constituents, we may consider the

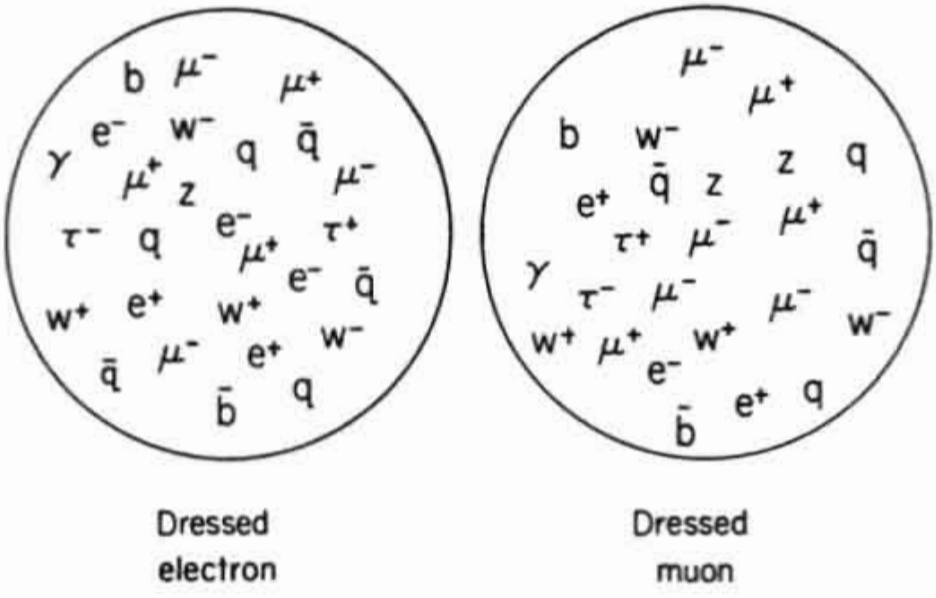


Fig. 7.4. Illustration of the composite structure of the dressed electron and the dressed muon; this structure can be revealed by high energy probes.

derivation of Lorentz invariance in non-covariant electrodynamics [Paper 24]. The different metric tensors for the photon, and for the left-handed and right-handed components of the electron, must be averaged over all their respective constituent particles. Consequently, in the infrared limit, all the particles have the same — namely the average metric tensor.

In a similar way we can intuitively understand why any gauge symmetry violating terms, introduced into the pure Yang-Mills particle sector of a non-asymptotically free theory, tend to zero in the infrared limit [Paper 29]. The presence of several fermion multiplets in a non-asymptotically free theory implies that the Yang Mills particles become more and more composed out of fermions in the infrared. The Yang-Mills self-energy contributions, in the infrared, then come dominantly from loop diagrams consisting of two fermion lines. Due to the global non-Abelian symmetry of the fermion kinetic energy and interaction terms in the Lagrangian density, these fermionic contributions to the renormalised Yang-Mills kinetic energy term are automatically locally gauge invariant. So, as these contributions become more and more dominant in the infrared limit, the self-interactions of the Yang-Mills particles tend to their gauge symmetric form.

Conformal invariance has been listed as geometrically inspired. This may be considered a slight exaggeration, since the conformal transformations do not form an isometry group of Minkowski space-time, but rather rescale the metric by a space-time dependent conformal factor $\Omega(x)$:

$$g_{\mu\nu} \rightarrow \Omega^2(x) g_{\mu\nu} \,. \tag{104}$$

However conformal transformations do act on space-time and preserve the important lightcone structure, mapping timelike vectors into timelike vectors, null vectors into null vectors and spacelike vectors into spacelike vectors. So we feel justified in listing geometry as suggestive of conformal symmetry.

There are just two entries in column bαp) corresponding to symmetry derivations relying on a parameter limit: (i) chiral symmetry, and (ii) flavour conservation for neutral currents (the GIM mechanism). Both examples rely on the relevant quark masses being small, compared to (i) Λ_{QCD} and (ii) the W and Z gauge boson masses respectively.

Table 7.3. Symmetries by adjustment

There are just the discrete symmetries P, CP and T in this class; CP and T are of course essentially equivalent according to the safely derived CPT theorem. They are entered three times in the table: In the Peccei-Quinn two Higgs doublet extension of the standard model, the vacuum adjusts itself to cancel out the effects of the QCD topological θ term; the low energy limit must be taken to avoid P and CP violation effects from the weak interactions. In baby universe theory the effective coupling constants become dynamical variables and, more controversially, may naturally adjust themselves to have very sharply peaked probability distributions; in the case of the topological θ parameter of QCD, it is predicted to peak at a P and CP symmetric value. P and CP symmetries are then derived, if QCD is taken to be the fundamental model, without the need for any limit. However if baby universe theory is applied to the full standard model, it is again necessary to take the low energy limit in order to avoid P and CP violating weak interaction effects.

Table 7.4. Formal Symmetry Derivation

This table comprises the various derivations of local gauge symmetry for field theory models which, in their original formulation, *a priori* have no such symmetry. In all the cases listed, the local gauge symmetry is introduced by a rather trivial extension of the notation used: the symmetry is initially purely one of formalism. It is of course possible to introduce practically any symmetry by a careful choice of formalism; however usually such a symmetry would be of no physical significance. In order for a formally introduced symmetry to be considered a true symmetry derivation, there must exist some reason why precisely the formalism chosen is relevant in some physical conditions. There has to be a physical limit under which the new language or formalism is enforced on us, before the formal symmetry can be accepted as being a physical symmetry; we have therefore crossed out column a), which is reserved for symmetry derivations where no limit is involved. The local gauge symmetric formalism is needed in the low energy limit of models where the inverse Higgs mechanism operates.

The formal gauge symmetry under consideration would be broken by the non-zero vacuum expectation value of a scalar field, analogous to that of the Higgs model, were it not for quantum fluctuations. Radiative corrections can renormalise the parameters; in particular the would-be negative mass squared Higgs scalar field may be renormalised to have zero vacuum expectation value and turn out to describe an ordinary positive mass squared spin zero particle. The inverse Higgs mechanism then operates and the quantum fluctuations are so large as to wash out all physical

gauge symmetry breaking effects. The massive vector bosons, which lose their mass by this mechanism, may either already be present as elementary particles or arise as composite particles in the fundamental model. In the first case vector bosons are present as elementary particles and the fundamental model possesses a manifest approximate gauge symmetry; in the second case they are absent but the model contains, for example, fundamental scalar fields.

The successful operation of the inverse Higgs mechanism and the consequent existence of an unbroken gauge symmetry is a matter of dynamics. The parameters of the fundamental model must lie in the appropriate range. In the CP^{n-1} model the coupling constant f has to be greater than some critical value[37] f_c; in $d = 2$ space-time dimensions $f_c = 0$. For the general G/H sigma model, the inverse Higgs mechanism is not expected to work if G is a non-compact group, as is the case for $E_{7(+7)}$ in the $N = 8$ supergravity model. In general, for the compact group models considered, there is a whole phase where zero mass gauge particles exist: it is not necessary to take particular limiting values of the coupling constants, but just values lying within some extended region of parameter space. We therefore do not classify the gauge symmetry as arising in a parameter limit, but rather as needed to describe the degrees of freedom surviving in the low energy limit.

At low energy, the massless gauge particles can be produced but the massive scalar particles of the would-be Higgs field H cannot. The formalism should reflect this physical property and the H field should be removed from the effective field theory in the low energy limit. This requirement that the would-be Higgs field H should disappear from the formalism in the low energy limit is not satisfied in the fundamental models considered: the gauge is fixed in the original formalism by requiring the phase of the H field to vanish, $\arg H = 0$, which clearly makes no sense once the H fields is left out. In this way the low energy limit drives us to use a new formalism, which corresponds to choosing another gauge and is thus strongly suggestive of a gauge invariant formalism.

Table 7.5. Symmetries by ignoring structure

Macroscopic scaling symmetry arises in the limit of a macroscopically large number of molecules, when the details of how the individual molecules distribute themselves locally is not important. A description in terms of macroscopic hydrodynamic variables is then more appropriate than in terms of the more fundamental molecular variables. In this new effective language of hydrodynamics, the detailed molecular configurations are ignored and the scaling can be an exact symmetry.

The other entries in the table are what we call 'hidden symmetries', which have also been included in Table 7.2. The classification chosen depends on how much structure is included in the fundamental model under consideration. For instance, in the heterotic string model, the objects that are transformed around by the $E_8 \times E_8$ or SO(32) symmetry operations are not all of the same kind. We may consider the transformations acting on the massless excitations of the string. These quasiparticles are all connected with the 16 extra dimensions in the model and are

of two types: 1) phonons and 2) solitons. The quasiparticles form a basis for the adjoint reresentation of the symmetry group, associated with the root diagram used in the Cartan classification of the Lie algebra. Those basis vectors corresponding to zero weight are physically the phonons, while those corresponding to the non-zero weights are the solitons. The twisting of the string in the 16-dimensional torus space, whenever such a soliton is present, is given by the weight of the corresponding basis vector. The $E_8 \times E_8$ transformations do not constitute an exact symmetry, if the structural difference between a phonon and a soliton is considered to be an integral part of the model; the symmetry must then be classified as derived by ignoring structure.

The many entries in the above discussed tables emphasize again that the origin of a large number of the symmetries of the laws of nature can, to a greater or lesser extent, be explained. In a way, this main conclusion of our book is in conflict with Einstein's well-known quotation: "Subtle is the Lord, but malicious He is not". By this statement, Einstein meant that although nature might have a sophisticated structure, she would not mislead us. However we would claim that some famous physicists have been misled about the fundamental physical laws by the experimentally observed symmetries. For example, Heisenberg enshrined isospin symmetry as one of the fundamental principles underlying his unified field theory of elementary particles.[1] As we have already emphasized, isospin symmetry is now understood as an 'accidental' consequence of QCD and the smallness of the up and down quark masses. The moral of our book is that symmetries are not so holy.

References

1. W. Heisenberg, *Introduction to the Unified Field Theory of Elementary Particles* (Wiley, 1966).
2. V. Kuzmin, V. Rubakov and M. Shaposhnikov, *Phys. Lett.* **155B**, 36 (1985).
3. S. Weinberg, in *Second Workshop on Grand Unification*, (Birkhäuser, 1981), p. 297.
4. J. C. Pati, in *Second Workshop on Grand Unification*, (Birkhäuser, 1981), p. 195.
5. J. C. Pati and A. Salam, *Phys. Rev.* **D10**, 275 (1974); R. N. Mohapatra and J. C. Pati, *Phys. Rev.* **D11**, 566, 2558 (1975).
6. H. B. Nielsen, in *Gauge Theories of the Eighties*, eds. R. Raitio and J. Lindfors, (Springer-Verlag, 1983), p. 288; H. B. Nielsen, D. L. Bennet and N. Brene, in *Recent Developments in Quantum Field Theory*, eds. J. Ambjørn, B. Durhuus and J. L. Petersen, (Elsevier Science Publishers, 1985), p. 263.
7. D. Weingarten, private communication.
8. J. A. Wheeler, "Law without Law", in *Quantum Theory and Measurement*, eds. J. A. Wheeler and W. H. Zurek, (Princeton University Press, 1983), p. 182; J. A. Wheeler, "Beyond the black hole", in *Some Strangeness in the Proportion*, ed. H. Woolf, (Addison-Wesley, 1980), p. 341; R. P. Feynman, *The Character of Physical Law* (MIT Press); G. F. Chew, in *Properties of the Fundamental Interactions*, ed. E. Zichichi, (Editrice Compositori, 1973), p. 3; C. H. Woo, "Mission Impossible? A Look at Past Setbacks in the Search for Elementary Matter and for Universal Symmetries", Zentrum für Interdisziplinäre Forschung, University of Bielefeld Preprint.
9. S. W. Hawking, *Phys. Rev.* **D37**, 904 (1988); S. W. Hawking and R. Laflamme, *Phys. Lett.* **209B** 39 (1988); S. Coleman, *Nucl. Phys.* **B307**, 867 (1988); S. B. Giddings and A. Strominger, *Nucl. Phys.* **B307**, 854 (1988).
10. J. B. Hartle and S. W. Hawking, *Phys. Rev.* **D28**, 2960 (1983).

11. S. Coleman, *Nucl. Phys.* **B310**, 643 (1988); J. Preskill, *Nucl. Phys.* **B323**, 141 (1989).
12. B. Grinstein and M. B. Wise, *Phys. Lett.* **212B**, 407 (1988); B. Grinstein and C. T. Hill, *Phys. Lett.* **220B**, 520 (1989).
13. H. B. Nielsen and M. Ninomiya, *Phys. Rev. Lett.* **62**, 429 (1989); J. Preskill, S. P. Trivedi and M. B. Wise, *Phys. Lett.* **223B**, 26 (1989).
14. J. Polchinski, *Phys. Lett.* **219B**, 251 (1989).
15. W. Fischler and L. Susskind, *Phys. Lett.* **217B**, 48 (1989); S. Coleman and K. Lee, *Phys. Lett.* **221B**, 242 (1989).
16. J. A. Wheeler, in *Battelle Rencontres: 1967 Lectures in Mathematics and Physics*, eds. C. DeWitt and J. A. Wheeler, (Benjamin, 1968), p. 242; B. S. DeWitt, *Phys. Rev.* **160**, 1113 (1967).
17. T. Banks, *Nucl. Phys.* **B249**, 332 (1985).
18. S. Chadha, C. Litwin and H. B. Nielsen, unpublished.
19. H. B. Nielsen, *Acta Physica Polonica* **B20**, 427 (1989).
20. S. Weinberg, *Phys. Rev. Lett.* **62**, 485 (1989); J. J. Bollinger *et al.*, *Phys. Rev. Lett.* **63**, 1031 (1989).
21. S. Chadha and H. B. Nielsen, "Naturalness of Weyl Equation and 3 + 1 Dimensionality", in A Report on the Research Activities at the Niels Bohr Institute and Nordita, 1984, p. 117.
22. Y. K. Fu and H. B. Nielsen, *Nucl. Phys.* **B236**, 167 (1984).
23. H. B. Nielsen and N. Brene, in *Proc. of the XVIII International Symposuim*, (Academie der Wissenschaften der DDR, Berlin-Zeuthen, 1985).
24. R. Gilmore, *Lie Groups, Lie Algebras and Some of their Applications*, (Wiley, 1974).
25. H. B. Nielsen and N. Brene, *Phys. Lett.* **223B**, 399 (1989).
26. H. B. Nielsen and N. Brene, *Nucl. Phys.* **B224**, 396 (1983).
27. L. O' Raifeartaigh, *Group Structure of Gauge Theories* (Cambridge University Press).
28. K. Hansen, Niels Bohr Institute Preprint NBI-HE-88-60 (1988).
29. D. L. Bennett, H. B. Nielsen and I. Picek, *Phys. Lett.* **208B**, 275 (1988).
30. J. M. Drouffe and J. B. Zuber, *Phys. Rep.* **102**, 96 (1983).
31. J. Dydak, in *Proc. of the 25th Int. Conf. on High Energy Physics*, p. 3.
32. A. Conkie, C. D. Froggatt and H. B. Nielsen, *Phys. Lett.* **161B**, 347 (1985).
33. Particle Data Group, *Phys. Lett.* **239B** (1990).
34. C. D. Froggatt and H. B. Nielsen, *Phys. Lett.* **106B**, 487 (1981).
35. F. R. Klinkhamer and N. S. Manton, *Phys. Rev.* **D30**, 2212 (1984).
36. V. A. Rubakov, in *Proc. of the 25th Int. Conf. on High Energy Physics*, p. 309; L. B. Okun, *ibid.*, p. 319.
37. I. Ya. Aref'eva and S. I. Azakov, *Nucl. Phys.* **B162**, 298 (1980); T. Kugo, H. Terao and S. Uehara, *Prog. Theor. Phys. Suppl.* **85**, 122 (1985).

REPRINTED PAPERS