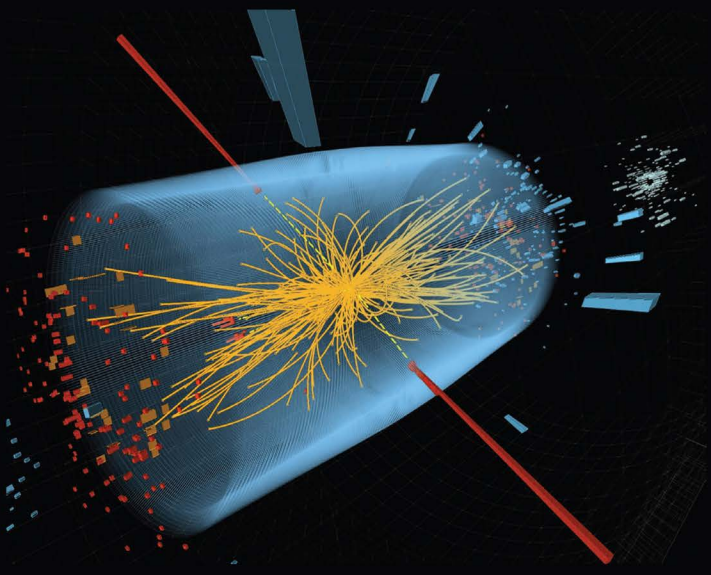
FOURTH EDITION

GAUGE THEORIES IN PARTICLE PHYSICS

A PRACTICAL INTRODUCTION

VOLUMES 1&2



Ian J.R. Aitchison • Anthony J.G. Hey

CRC Press
Taylor & Francis Group

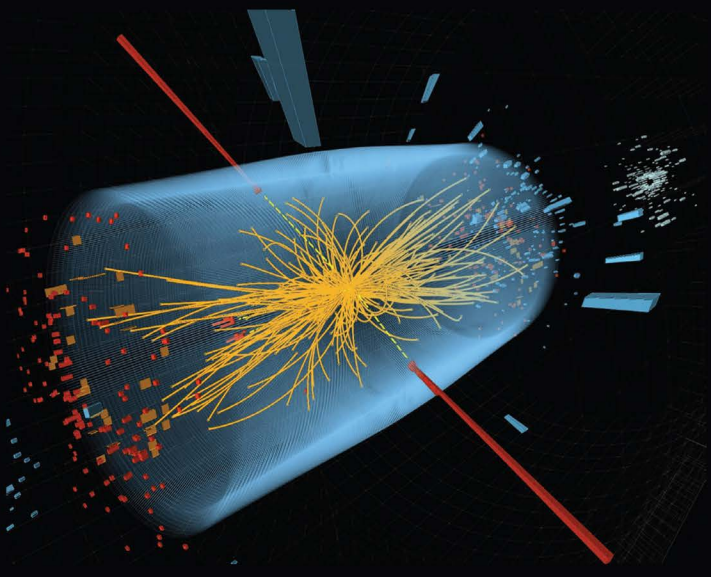
FOURTH EDITION

GAUGE THEORIES IN PARTICLE PHYSICS

A PRACTICAL INTRODUCTION

VOLUME 1

From Relativistic Quantum Mechanics to QED



Ian J.R. Aitchison • Anthony J.G. Hey

CRC Press
Taylor & Francis Group

FOURTH EDITION

GAUGE THEORIES IN PARTICLE PHYSICS

A PRACTICAL INTRODUCTION

VOLUME 1

From Relativistic Quantum Mechanics to QED


Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

FOURTH EDITION

GAUGE THEORIES IN PARTICLE PHYSICS

A PRACTICAL INTRODUCTION

VOLUME 1

From Relativistic Quantum Mechanics to QED

Ian J.R. Aitchison • Anthony J.G. Hey

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business

CRC Press
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

Printed in the United States of America on acid-free paper
Version Date: 2012912

International Standard Book Number: 978-1-4665-1299-3 (Hardback)

Library of Congress Cataloging-in-Publication Data

Aitchison, Ian Johnston Rhind, 1936-
Gauge theories in particle physics : a practical introduction / Ian J R Aitchison, Anthony J.G. Hey. -- 4th ed.
v. cm.
Includes bibliographical references and index.
Contents: v. 1. From relativistic quantum mechanics to QED -- v. 2. Non-Abelian gauge theories : QCD and the electroweak theory.
ISBN 978-1-4665-1299-3 (v. 1 : hardback) -- ISBN 978-1-4665-1307-5 (v. 2 : hardback)
1. Gauge fields (Physics) 2. Particles (Nuclear physics) 3. Weak interactions (Nuclear physics) 4. Quantum electrodynamics. 5. Feynman diagrams. I. Hey, Anthony J. G. II. Title.

QC793.3.F5A34 2012
539.7'21--dc23 2012031181

Visit the Taylor & Francis Web site at
http://www.taylorandfrancis.com

and the CRC Press Web site at
http://www.crcpress.com

To Jessie
and to
Jean, Katherine and Elizabeth


Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

Contents

Preface to the Fourth Edition

In the Preface to the first edition of this book, published thirty years ago, we wrote that our aim was to help the reader to acquire a 'reasonable understanding of gauge theories that are being tested by contemporary experiments in high-energy physics'; and we stressed that our approach was intended to be both practical and accessible.

We have pursued the same aim and approach in later editions. Shortly after the appearance of the first edition, a series of major discoveries at the CERN $\bar{\text{p}}$p collider confirmed the existence of the W and Z bosons, with properties predicted by the Glashow-Salam-Weinberg electroweak gauge theory; and also provided further support for quantum chromodynamics, or QCD. Our second edition followed in 1989, expanded so as to include discussion, on the experimental side, of the new results; and, on the theoretical side, a fuller treatment of QCD, and an elementary introduction to quantum field theory, with limited applications. Subsequently, experiments at LEP and other laboratories were precise enough to test the Standard Model beyond the first order in perturbation theory ('tree level'), being sensitive to higher order effects ('loops'). In response, we decided it was appropriate to include the basics of 'one-loop physics'. Together with the existing material on relativistic quantum mechanics, and QED, this comprised volume 1 (2003) of our two-volume third edition. In a natural division, the non-Abelian gauge theories of the Standard Model, QCD and the electroweak theory, formed the core of volume 2 (2004). The progress of research on QCD, both theoretical and experimental, required new chapters on lattice quantum field theory, and on the renormalization group. The discussion of the central topic of spontaneous symmetry breaking was extended, in particular so as to include chiral symmetry breaking.

This new fourth edition retains the two-volume format, which has been generally well received, with broadly the same allocation of content as in the third edition. The principal new additions are, once again, dictated by substantial new experimental results – namely, in the areas of CP violation and neutrino oscillations, where great progress was made in the first decade of this century. Volume 2 now includes a new chapter devoted to CP violation and oscillations in mesonic and neutrino systems. Partly by way of preparation for this, volume 1 also contains a new chapter, on Lorentz transformations and discrete symmetries. We give a simple do-it-yourself treatment of Lorentz transformations of Dirac spinors, which the reader can connect to the group theory approach in appendix M of volume 2; the transformation properties of

bilinear covariants are easily managed. We also introduce Majorana fermions at an early stage. This material is suitable for first courses on relativistic quantum mechanics, and perhaps should have been included in earlier editions (we thank a referee for urging its inclusion now).

To make room for the new chapter in volume 1, the two introductory chapters of the third edition have been condensed into a single one, in the knowledge that excellent introductions to the basic facts of particle physics are available elsewhere. Otherwise, apart from correcting the known minor errors and misprints, the only other changes in volume 1 are some minor improvements in presentation, and appropriate updates on experimental numbers. Volume 2 contains significantly more in the way of updates and additions, as will be detailed in the Preface to that volume. But we have continued to omit discussion of speculations going beyond the Standard Model; after all, the crucial symmetry-breaking (Higgs) sector has only now become experimentally accessible.

Acknowledgements

Many people helped us with each of the previous editions, and their input remains an important part of this one. Colleagues at Oxford and Southampton, and elsewhere, read much – or in some cases all – of our drafts; these include especially Jack Paton and the late Gary McEwen and Euan Squires. The coverage of the discoveries at the CERN $\bar{\text{p}}$p collider in the 1980s was based on superb material generously made available to us by Luigi DiLella. Much of our presentation of quantum field theory was developed in our lectures at various Summer Schools, and we thank Roger Cashmore, John Dainton, David Saxon and John March-Russell for these opportunities. Paolo Strolin and Peter Williams each provided full lists of misprints, and valuable suggestions for improvements, for volume 1 of the third edition. IJRA has enjoyed a lively correspondence with John Colarusso, originally about volume 1, but ranging far beyond it; John also spotted a number of typos. A special debt is owed by IJRA to the late George Emmons, who contributed so much to the production of the second and third editions, but who died before plans began for the fourth; he is greatly missed.

For this new edition, we are grateful to Frank Close for helpful comments on chapter 4 of volume 1. Others who assisted with volume 2 are acknowledged in the Preface to that volume.

Ian J R Aitchison
Anthony J G Hey
September 2012

Part I

Introductory Survey, Electromagnetism as a Gauge Theory, and Relativistic Quantum Mechanics


Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

1

The Particles and Forces of the Standard Model

1.1 Introduction: the Standard Model

The traditional goal of particle physics has been to identify what appear to be structureless units of matter and to understand the nature of the forces acting between them; all other entities are then to be successively constructed as composites of these elementary building blocks. The enterprise has a two-fold aspect: matter on the one hand, forces on the other. The expectation is that the smallest units of matter should interact in the simplest way; or that there is a deep connection between the basic units of matter and the basic forces. The joint matter/force nature of the enquiry is perfectly illustrated by Thomson's discovery of the electron and Maxwell's theory of the electromagnetic field, which together mark the birth of modern particle physics. The electron was recognized both as the 'particle of electricity' – or as we might now say, as an elementary source of the electromagnetic field, with its motion constituting an electromagnetic current – and also as an important constituent of matter. In retrospect, the story of particle physics over the subsequent one hundred years or so has consisted in the discovery and study of two new (non-electromagnetic) forces – the *weak* and the *strong* forces – and in the search for 'electron-figures' to serve both as constituents of the new layers of matter which were uncovered (first nuclei, and then hadrons) and also as sources of the new force fields. In the last quarter of the twentieth century, this effort culminated in decisive progress: the identification of a collection of matter units which are indeed analogous to the electron; and the highly convincing experimental verification of theories of the associated strong and weak force fields, which incorporate and generalize in a beautiful way the original electron/electromagnetic field relationship. These theories are collectively called 'the Standard Model' (or SM for short), to which this book is intended as an elementary introduction.

In brief, the picture is as follows. The matter units are fermions, with spin-$\frac{1}{2}$ (in units of $\hbar$). They are of two types, *leptons* and *quarks*. Both are structureless at the smallest distances currently probed by the highest-energy accelerators. The leptons are generalizations of the electron, the term denoting particles which, if charged, interact both electromagnetically and weakly; and

if neutral, only weakly. By contrast, the quarks – which are the constituents of hadrons, and thence of nuclei – interact via all three interactions, strong, electromagnetic and weak. The weak and electromagnetic interactions of both quarks and leptons are described in a (partially) unified way by the electroweak theory of Glashow, Salam and Weinberg (GSW), which is a generalization of quantum electrodynamics or QED; the strong interactions of quarks are described by quantum chromodynamics or QCD, which is also analogous to QED. The similarity with QED lies in the fact that all three interactions are types of *gauge theories*, though realized in different ways. In the first volume of this book, we will get as far as QED; QCD and the electroweak theory are treated in volume 2.

The reader will have noticed that the most venerable force of all – gravity – is absent from our story. In practical terms this is quite reasonable, since its effect is very many orders of magnitude smaller than even the weak force, at least until the interparticle separation reaches distances far smaller than those we shall be discussing. Conceptually also, gravity still seems to be somewhat distinct from the other forces which, as we have already indicated, are encouragingly similar. There are no particular fermionic sources carrying 'gravity charges': it seems that *all* matter gravitates. This of course was a motivation for Einstein's geometrical approach to gravity. Despite the lingering promise of *string theory* (Green *et al.* 1987, Polchinski 1998, Zwiebach 2004), it is fair to say that the vision of the unification of all the forces, which possessed Einstein, is still some way from realization. Gravitational interactions are not part of the SM.

This book is not intended as a completely self-contained textbook on particle physics, which would survey the broad range of observed phenomena and outline the main steps by which the picture described here has come to be accepted. For this we must refer the reader to other sources (e.g. Perkins 2000, Bettini 2008). We proceed with a brief review of the matter (fermionic) content of the SM.

1.2 The fermions of the Standard Model

1.2.1 Leptons

Forty years after Thomson's discovery of the electron, the first member of another *generation* of leptons (as it turned out) – the muon – was found independently by Street and Stevenson (1937), and by Anderson and Neddermeyer (1937). Following the convention for the electron, the μ^- is the particle and the μ^+ the antiparticle. At first, the muon was identified with the particle postulated by Yukawa only two years earlier (1935) as the field quantum of the 'strong nuclear force field', the exchange of which between two nucleons

would account for their interaction (see section 1.3.2). In particular, its mass (105.7 MeV) was nicely within the range predicted by Yukawa. However, experiments by Conversi *et al.* (1947) established that the muon could not be Yukawa's quantum since it did not interact strongly; it was therefore a lepton. The μ^- seems to behave in exactly the same way as the electron, interacting only electromagnetically and weakly, with interaction strengths identical to those of an electron.

In 1975 Perl *et al.* (1975) discovered yet another 'replicant' electron, the τ^- with a mass of 1.78 GeV. Once again, the weak and electromagnetic interactions of the τ^- (τ^+) are identical to those of the e^- (e^+).

At this stage one might well wonder whether we are faced with a 'lepton spectroscopy', of which the e^-, μ^- and τ^- are but the first three states. Yet this seems not to be the correct interpretation. First, no other such states have (so far) been seen. Second, all these leptons have the same spin ($\frac{1}{2}$), which is certainly quite unlike any conventional excitation spectrum. And third, no γ-transitions are observed to occur between the states, though this would normally be expected. For example, the branching fraction for the process

$$\mu^- \to e^- + \gamma \qquad (\textit{not} \text{ observed}) \tag{1.1}$$

is currently quoted as less than 1.2×10^{-11} at the 90% confidence level (Nakamura *et al.* 2010). Similarly there are (much less stringent) limits on $\tau^- \to \mu^- + \gamma$ and $\tau^- \to e^- + \gamma$.

If the e^- and μ^- states in (1.1) were, in fact, the ground and first excited states of some composite system, the decay process (1.1) would be expected to occur as an electromagnetic transition, with a relatively high probability because of the large energy release. Yet the experimental upper limit on the rate is very tiny. In the absence of any mechanism to explain this, one *systematizes* the situation, empirically, by postulating the existence of a selection rule forbidding the decay (1.1). In taking this step, it is important to realize that 'absolute forbidden-ness' can never be established experimentally: all that can be done is to place a (very small) upper limit on the branching fraction to the 'forbidden' channel, as here. The possibility will always remain open that future, more sensitive, experiments will reveal that some processes, assumed to be forbidden, are in fact simply extremely rare.

Of course, such a proposed selection rule would have no physical content if it only applied to the one process (1.1); but it turns out to be generally true, applying not only to the electromagnetic interaction of the charged leptons, but to their weak interactions also. The upshot is that we can consistently account for observations (and non-observations) involving e's, μ's and τ's by assigning to each a new additive quantum number (called 'lepton flavour') which is assumed to be conserved. Thus we have electron flavour L_e such that $L_e(e^-) = 1$ and $L_e(e^+) = -1$; muon flavour L_μ such that $L_\mu(\mu^-) = 1$ and $L_\mu(\mu^+) = -1$; and tau flavour L_τ such that $L_\tau(\tau^-) = 1$ and $L_\tau(\tau^+) = -1$. Each is postulated to be conserved in all leptonic processes. So (1.1) is then

forbidden, the left-hand side having $L_e = 0$ and $L_\mu = 1$, while the right-hand side has $L_e = 1$ and $L_\mu = 0$.

The electromagnetic interactions of the mu and the tau leptons are the same as for the electron. In weak interactions, each charged lepton (e, μ, τ) is accompanied by its 'own' neutral partner, a neutrino. The one emitted with the e^- in β-decay was originally introduced by Pauli in 1930, as a 'desperate remedy' to save the conservation laws of four-momentum and angular momentum. In the Standard Model, the three neutrinos are assigned lepton flavour quantum numbers in such a way as to conserve each lepton flavour separately. Thus we assign $L_e = -1, L_\mu = 0, L_\tau = 0$ to the neutrino emitted in neutron β-decay

$$n \to p + e^- + \bar{\nu}_e, \tag{1.2}$$

since $L_e = 0$ in the initial state and $L_e(e^-) = +1$; so the neutrino in (1.2) is an antineutrino 'of electron type' (or 'of electron flavour'). The physical reality of the antineutrinos emitted in nuclear β-decay was established by Reines and collaborators in 1956 (Cowan *et al.* 1956), by observing that the antineutrinos from a nuclear reactor produced positrons via the inverse β-process

$$\bar{\nu}_e + p \to n + e^+. \tag{1.3}$$

The neutrino partnering the μ^- appears in the decay of the π^-:

$$\pi^- \to \mu^- + \bar{\nu}_\mu \tag{1.4}$$

where the $\bar{\nu}_\mu$ is an antineutrino of muon type ($L_\mu(\bar{\nu}_\mu) = -1$, $L_e(\bar{\nu}_\mu) = 0 = L_\tau(\bar{\nu}_\mu)$). How do we know that $\bar{\nu}_\mu$ and $\bar{\nu}_e$ are not the same? An important experiment by Danby *et al.* (1962) provided evidence that they are not. They found that the neutrinos accompanying muons from π-decay always produced muons on interacting with matter, never electrons. Thus, for example, the lepton flavour conserving reaction

$$\bar{\nu}_\mu + p \to \mu^+ + n \tag{1.5}$$

was observed, but the lepton flavour violating reaction

$$\bar{\nu}_\mu + p \to e^+ + n \qquad (\textit{not} \text{ observed}) \tag{1.6}$$

was not. As with (1.1), 'non-observation' of course means, in practice, an upper limit on the cross section. Both types of neutrino occur in the β-decay of the muon itself:

$$\mu^- \to \nu_\mu + e^- + \bar{\nu}_e, \tag{1.7}$$

in which $L_\mu = 1$ is initially carried by the μ^- and finally by the ν_μ, and the L_e's of the e^- and $\bar{\nu}_e$ cancel each other out.

In the same way, the ν_τ is associated with the τ^-, and we have arrived at *three generations* of charged and neutral *lepton doublets*:

$$(\nu_e, e^-) \qquad (\nu_\mu, \mu^-) \qquad \text{and} \qquad (\nu_\tau, \tau^-) \tag{1.8}$$

together with their antiparticles.

TABLE 1.1
Properties of SM leptons.

Generation	Particle	Mass (MeV)	Q/e	L_{e}	L_μ	L_τ
1	ν_{e}	$< 2 \times 10^{-6}$	0	1	0	0
	e^-	0.511	- 1	1	0	0
2	ν_μ	< 0.19	0	0	1	0
	μ^-	105.658	- 1	0	1	0
3	ν_τ	< 18.2	0	0	0	1
	τ^-	1777	- 1	0	0	1

We should at this point note that another type of weak interaction is known, in which – for example – the $\bar{\nu}_\mu$ in (1.5) scatters elastically from the proton, instead of changing into a μ^+:

$$\bar{\nu}_\mu + \mathrm{p} \rightarrow \bar{\nu}_\mu + \mathrm{p}. \tag{1.9}$$

This is an example of what is called a 'neutral current' process, (1.5) being a 'charged current' one. In terms of the Yukawa-like exchange mechanism for particle interactions, to be described in the next section, (1.5) proceeds via the exchange of charged quanta ($\mathrm{W}^\pm$), while in (1.9) a neutral quantum (Z^0) is exchanged.

As well as their flavour, one other property of neutrinos is of great interest, namely their mass. As originally postulated by Pauli, the neutrino emitted in β-decay had to have very small mass, because the maximum energy carried off by the e^- in (1.2) was closely equal to the difference in rest energies of the neutron and proton. It was subsequently widely assumed (perhaps largely for simplicity) that all neutrinos were strictly massless, and it is fair to say that the original Standard Model made this assumption. Yet there is, in fact, no convincing reason for this (as there is for the masslessness of the photon – see chapter 6), and there is now clear evidence that neutrinos do indeed have very small, but non-zero, masses. It turns out that the question of neutrino masslessness is directly connected to another one: whether neutrino flavour is, in fact, conserved. If neutrinos are massless, as in the original Standard Model, neutrinos of different flavour cannot 'mix', in the sense of quantum-mechanical states; but mixing can occur if neutrinos have mass. The phenomenon of neutrino flavour mixing (or 'neutrino oscillations') is now well established, and is a subject of intense research. In this book we shall simply regard non-zero neutrino masses as part of the (updated) Standard Model.

The SM leptons are listed in table 1.1, along with some relevant properties. Note that the limits on the neutrino masses, which are taken from Nakamura

et al. 2010, do not include the results obtained from analyses of neutrino oscillations. These oscillations, to which we shall return in chapter 21 in volume 2, are sensitive to the differences of squared masses of the neutrinos, not to the absolute scale of mass.

We now turn to the other fermions in the SM.

1.2.2 Quarks

Quarks are the constituents of hadrons, in which they are bound by the strong QCD forces. Hadrons with spins $\frac{1}{2}, \frac{3}{2}, \frac{5}{2}, \ldots$ (i.e. fermions) are baryons, those with spins 0, 1, 2, ... (i.e. bosons) are mesons. Examples of baryons are nucleons (the neutron n and the proton p), and hyperons such as Λ^0 and the Σ and Ξ states. Evidence for the composite nature of hadrons accumulated during the 1960s and 1970s. Elastic scattering of electrons from protons by Hofstadter and co-workers (Hofstadter 1963) showed that the proton was not pointlike, but had an approximately exponential distribution of charge with a root mean square radius of about 0.8 fm. Much careful experimentation in the field of baryon and meson spectroscopy revealed sequences of excited states, strongly reminiscent of those well-known in atomic and nuclear physics.

The conclusion would now seem irresistible that such spectra should be interpreted as the energy levels of systems of bound constituents. A specific proposal along these lines was made in 1964 by Gell-Mann (1964) and Zweig (1964). Though based on somewhat different (and much more fragmentary) evidence, their suggestion has turned out to be essentially correct. They proposed that baryons contain three spin-$\frac{1}{2}$ constituents called quarks (by Gell-Mann), while mesons are quark-antiquark systems. One immediate consequence is that quarks have fractional electromagnetic charge. For example, the proton has two quarks of charge $+\frac{2}{3}$, called 'up' (u) quarks, and one quark of charge $-\frac{1}{3}$, the 'down' (d) quark. The neutron has the combination ddu, while the π^+ has one u and one anti-d ($\bar{\mathrm{d}}$) and so on.

Quite simple quantum-mechanical bound state *quark models*, based on these ideas, were remarkably successful in accounting for the observed hadronic spectra. Nevertheless, many physicists, in the 1960s and early 1970s, continued to regard quarks more as useful devices for systematizing a mass of complicated data than as genuine items of physical reality. One reason for this scepticism must now be confronted, for it constitutes a major new twist in the story of the structure of matter.

Gell-Mann ended his 1964 paper with the remark: 'A search for stable quarks of charge $-\frac{1}{3}$ or $+\frac{2}{3}$ and/or stable di-quarks of charge $-\frac{2}{3}$ or $+\frac{1}{3}$ or $+\frac{4}{3}$ at the highest energy accelerators would help to reassure us of the nonexistence of real quarks'. Indeed, with one possible exception (La Rue *et al.* 1977, 1981), this 'reassurance' has been handsomely provided! *Unlike* the constituents of atoms and nuclei, quarks have *not* been observed as stable isolated particles. When hadrons of the highest energies currently available are smashed into each other, what is observed downstream is only lots more

hadrons, not fractionally charged quarks. The explanation for this novel behaviour of quarks is now believed to lie in the nature of the interquark force (QCD). We shall briefly discuss this force in section 1.3.6, and treat it in detail in volume 2. The consensus at present is that QCD does imply the '*confinement*' of quarks – that is, they do not exist as isolated single particles[1], only as groups confined to hadronic volumes.

When Gell-Mann and Zweig made their proposal, three types of quark were enough to account for the observed hadrons: in addition to the u and d quarks, the 'strange' quark s was needed to describe the known strange particles such as the hyperon Λ^0 (uds), and the strange mesons like $K^0(d\bar{s})$. In 1964, Bjorken and Glashow (1964) discussed the possible existence of a fourth quark on the basis of quark–lepton symmetry, but a strong theoretical argument for the existence of the c ('charm') quark, within the framework of gauge theories of electroweak interactions, was given by Glashow, Iliopoulos and Maiani (1970), as we shall discuss in volume 2. They estimated that the c quark mass should lie in the range 3–4 GeV. Subsequently, Gaillard and Lee (1974) performed a full (one-loop) calculation in the then newly-developed renormalizable electroweak theory, and predicted $m_c \approx 1.5$ GeV. The prediction was spectacularly confirmed in November of the same year with the discovery (Aubert *et al.* 1974, Augustin *et al.* 1974) of the J/ψ system, which was soon identified as a $c\bar{c}$ composite (and dubbed 'charmonium'), with a mass in the vicinity of 3 GeV. Subsequently, mesons such as $D^0(c\bar{u})$ and $D^+(c\bar{d})$ carrying the c quark were identified (Goldhaber *et al.* 1976, Peruzzi *et al.* 1976), consolidating this identification.

The *second generation* of quarks was completed in 1974, with the two quark doublets (u, d) and (c, s) in parallel with the lepton doublets (ν_e, e^-) and (ν_μ, μ^-). But even before the discovery of the c quark, the possibility that a completely new third-generation quark doublet might exist was raised in a remarkable paper by Kobayashi and Maskawa (1973). Their analysis focused on the problem of incorporating the known violation of **CP** symmetry (the product[2] of particle-antiparticle conjugation **C** and parity **P**) into the quark sector of the renormalizable electroweak theory. **CP**-violation in the decays of neutral K-mesons had been discovered by Christenson *et al.* (1964), and Kobayashi and Maskawa pointed out that it was very difficult to construct a plausible model of **CP**-violation in weak transitions of quarks with only two generations. They suggested, however, that **CP**-violation could be naturally accommodated by extending the theory to three generations of quarks. Their description of **CP**-violation thus entailed the very bold prediction of two entirely new and undiscovered quarks, the (t, b) doublet, where t ('top') has charge $\frac{2}{3}$ and b ('bottom') has charge $-\frac{1}{3}$.

In 1975, with the discovery of the τ^- mentioned earlier, there was already evidence for a third generation of leptons. The discovery of the b quark

[1] With the (fleeting) exception of the t quark, as we shall see in a moment.
[2] We shall discuss these symmetries in chapter 4.

in 1977 resulted from the observation of massive mesonic states generally known as Υ ('upsilon') (Herb *et al.* 1977, Innes *et al.* 1977), which were identified as $b\bar{b}$ composites. Subsequently, b-carrying mesons were found. Finally, firm evidence for the expected t quark was obtained by the CDF and D0 collaborations at Fermilab in 1995 (Abe *et al.* 1995, Abachi *et al.* 1995); see Bettini 2008, section 4.10, for details about the discovery of the top quark. The full complement of *three generations* of *quark doublets* is then

$$(\mathrm{u}, \mathrm{d}) \qquad (\mathrm{c}, \mathrm{s}) \qquad \text{and} \qquad (\mathrm{t}, \mathrm{b}) \tag{1.10}$$

together with their antiparticles, in parallel with the three generations of lepton doublets (1.8).

One particular feature of the t quark requires comment. Its mass is so large that, although it decays weakly, the energy release is so great that its lifetime is some two orders of magnitude shorter than typical strong interaction timescales; this means that it decays before any t-carrying hadrons can be formed. So when a t quark is produced (in a p-$\bar{\mathrm{p}}$ collision, for example), it decays as a free (unbound) particle. Its mass can be determined from a kinematic anaysis of the decay products.

We must now discuss the quantum numbers carried by quarks. First of all, each quark listed in (1.10) comes in three varieties, distinguished by a quantum number called 'colour'. It is precisely this quantum number that underlies the dynamics of QCD (see section 1.3.6). Colour, in fact, is a kind of generalized charge, for the strong QCD interactions. We shall denote the three colours of a quark by 'red', 'blue', and 'green'. Thus we have the triplet $(\mathrm{u_r}\ ,\ \mathrm{u_b}\ ,\ \mathrm{u_g})$, and similarly for all the other quarks.

Secondly, quarks carry flavour quantum numbers, like the leptons. In the quark case, they are as follows. The two quarks which are familiar in ordinary matter, 'u' and 'd', are an isospin doublet (see chapter 12 in volume 2) with $\mathrm{T}_3 = +1/2$ for 'u' and $\mathrm{T}_3 = -1/2$ for 'd'. The flavour of 's' is strangeness, with the value $S = -1$. The flavour of 'c' is charm, with value $C = +1$, that of 'b' has value $\tilde{B} = -1$ (we use $\tilde{B}$ to distinguish it from baryon number B), and the flavour of 't' is $T = +1$. The convention is that the sign of the flavour number is the same as that of the charge.

The strong and electromagnetic interactions of quarks are independent of quark flavour, and depend only on the electromagnetic charge and the strong charge, respectively. This means, in particular, that flavour cannot change in a strong interaction among hadrons – that is, flavour is conserved in such interactions. For example, from a zero strangeness initial state, the strong interaction can only produce pairs of strange particles, with cancelling strangeness. This is the phenomenon of 'associated production', known since the early days of strange particle physics in the 1950s. Similar rules hold for the other flavours: for example, the t quark, once produced, cannot decay to a lighter quark via a strong interaction, since this would violate T-conservation.

TABLE 1.2
Properties of SM quarks.

Generation	Particle	Mass	Q/e	S	C	$\tilde{B}$	T
1	$u_r\ u_b\ u_g$	1.7 to 3.1 MeV	2/3	0	0	0	0
	$d_r\ d_b\ d_g$	4.1 to 5.7 MeV	- 1/3	0	0	0	0
2	$c_r\ c_b\ c_g$	1.15 to 1.35 GeV	2/3	0	1	0	0
	$s_r\ s_b\ s_g$	80 to 130 MeV	- 1/3	- 1	0	0	0
3	$t_r\ t_b\ t_g$	172 to 174 GeV	2/3	0	0	0	1
	$b_r\ b_b\ b_g$	4 to 5 GeV	- 1/3	0	0	- 1	0

In weak interactions, by contrast, quark flavour is generally not conserved. For example, in the semi-leptonic decay

$$\Lambda^0(\mathrm{uds}) \to \mathrm{p(uud)} + \mathrm{e}^- + \bar{\nu}_\mathrm{e}, \tag{1.11}$$

an s quark changes into a u quark. The rather complicated flavour structure of weak interactions, which remains an active field of study, will be reviewed when we come to the GSW theory in volume 2. However, one very important, though technical, point must be made about the weak interactions of quarks and leptons. It is natural to wonder whether a new generation of quarks might appear, *unaccompanied* by the corresponding leptons – or vice versa. Within the framework of the Standard Model interactions, the answer is no. It turns out that subtle quantum field theory effects called 'anomalies', to be discussed in chapter 18 of volume 2, would spoil the renormalizability of the weak interactions (see section 1.4.1), unless there are equal numbers of quark and lepton generations.

We end this section with some comments about the quark masses; the values listed in Table 1.2 are based on those given in Nakamura *et al.* (2010). As we have already noted, the t quark is the only one whose mass can be directly measured. All the others are (it would appear) permanently confined inside hadrons. It is therefore not immediately obvious how to define – and measure – their masses. In a more familiar bound state problem, such as a nucleus, the masses of the constituents are those we measure when they are free of the nuclear binding forces – i.e. when they are far apart. For the QCD force, the situation is very different. There it turns out that the force is very weak at *short* distances, a property called *asymptotic freedom* – see section 1.3.6; this important property will be treated in section 15.3 of volume 2. We may think of the force as very roughly analogous to that of a spring joining two constituents. To separate them, energy must be supplied to the system. So

when the constituents are no longer close, the energy of the system is greater than the sum of the short distance (free) quark masses. In potential models (see section 1.3.6), the effect is least pronounced for the 'heavy' quarks (m_q greater than about 1 GeV). For example, the ground state of the $\Upsilon(b\bar{b})$ lies at about 9.46 GeV, which is close to the average value of $2m_b$ as given in Table 1.2. For $\psi(c\bar{c})$ the ground state is at about 3 GeV, somewhat greater than $2m_c$. For the three lightest quarks, and especially for the u and d quarks, the position is quite different: for example, the proton (uud) with a mass of 938 MeV is far more massive than $2m_u + m_d$. Here the 'spring' is responsible for about 300 MeV per quark.

While this picture is qualitatively useful, it is clearly model dependent, as would be even a more sophisticated quark model. To do the job properly, we have to go to the actual QCD Lagrangian, and use it to calculate the hadron masses with the Lagrangian masses as input. This can be done through a lattice simulation of the field theory, as will be described in chapter 16 of volume 2. Independently, another handle on the Lagrangian masses is provided by the fact that the QCD Lagrangian has an extra symmetry ('chiral symmetry') which is exact when the quark masses are zero. This is, in fact, an excellent approximation for the u and d quarks, and a fair one for the s quark. The symmetry is, however, dynamically ('spontaneously') broken by QCD, in such a way as to generate (in the case $m_u = m_d = 0$) the nucleon mass entirely dynamically, along with a massless pion. The small Lagrangian masses can then be treated perturbatively in a procedure called 'chiral perturbation theory'. These essential features of QCD will be treated in chapter 18 of volume 2. For the moment, we accept the values in Table 1.2; Nakamura *et al.* (2010) contains a review of quark masses.

1.3 Particle interactions in the Standard Model

1.3.1 Classical and quantum fields

In the world of the classical physicist, matter and force were clearly separated. The nature of matter was intuitive, based on everyday macroscopic experience; force, however, was more problematical. Contact forces between bodies were easy to understand, but forces which seemed capable of acting at a distance caused difficulties.

> That gravity should be innate, inherent and essential to matter, so that one body can act upon another at a distance, through a vacuum, without the mediation of anything else, by and through which action and force may be conveyed from one to the other, is to me so great an absurdity, that I believe no man who has in philosophical matters

> a competent faculty of thinking can ever fall into it. (Letter from Newton to Bentley)

Newton could find no satisfactory mechanism or physical model, for the transmission of the gravitational force between two distant bodies; but his dynamical equations provided a powerful predictive framework, given the (unexplained) gravitational force law; and this eventually satisfied most people.

The 19th century saw the precise formulation of the more intricate force laws of electromagnetism. Here too the distaste for action-at-a-distance theories led to numerous mechanical or fluid mechanical models of the way electromagnetic forces (and light) are transmitted. Maxwell made brilliant use of such models as he struggled to give physical and mathematical substance to Faraday's empirical ideas about lines of force. Maxwell's equations were indeed widely regarded as describing the mechanical motion of the ether – an amazing medium, composed of vortices, gear wheels, idler wheels and so on. But in his 1864 paper, the third and final one of the series on lines of force and the electromagnetic field, Maxwell himself appeared ready to throw away the mechanical scaffolding and let the finished structure of the *field equations* stand on its own. Later these field equations were derived from a Lagrangian (see chapter 7), and many physicists came to agree with Poincaré that this 'generalized mechanics' was more satisfactory than a multitude of different ether models; after all, the same mathematical equations can describe, when suitably interpreted, systems of masses, springs and dampers, or of inductors, capacitors and resistors. With this step, the concepts of mechanics were enlarged to include a new fundamental entity, the *electromagnetic field*.

The action-at-a-distance dilemma was solved, since the electromagnetic field permeates all of space surrounding charged or magnetic bodies, responds locally to them, and itself acts on other distant bodies, propagating the action to them at the speed of light: for Maxwell's theory, besides unifying electricity and magnetism, also predicted the existence of electromagnetic waves which should travel with the speed of light, as was confirmed by Hertz in 1888. Indeed, light *was* a form of electromagnetic wave.

Maxwell published his equations for the dynamics of the electromagnetic field (Maxwell 1864) some forty years before Einstein's 1905 paper introducing special relativity. But Maxwell's equations are fully consistent with relativity as they stand (see chapter 2), and thus constitute the first relativistic (classical) field theory. The Maxwell Lagrangian lives on, as part of QED.

It seems almost to be implied by the local field concept, and the desire to avoid action at a distance, that the fundamental carriers of electricity should themselves be point-like, so that the field does not, for example, have to interact with different parts of an electron simultaneously. Thus the point-like nature of elementary matter units seems intuitively to be tied to the local nature of the force field via which they interact.

Very soon after the successes of classical field physics, however, another world began to make its appearance – the quantum one. First the photoelectric effect and then – much later – the Compton effect showed unmistakeably

that electromagnetic *waves* somehow also had a *particle*-like aspect, the photon. At about the same time, the intuitive understanding of the nature of matter began to fail as well: supposedly *particle*-like things, like electrons, displayed *wave*-like properties (interference and diffraction). Thus the conceptual distinction between matter and forces, or between particle and field, was no longer so clear. On the one hand, electromagnetic forces, treated in terms of fields, now had a particle aspect; and on the other hand, particles now had a wave-like or field aspect. 'Electrons', writes Feynman (1965a) at the beginning of volume 3 of his *Lectures on Physics*, 'behave just like light'.

How can we build a theory of electrons and photons which does justice to all the 'point-like', 'local', 'wave/particle' ideas just discussed? Consider the apparently quite simple process of spontaneous decay of an excited atomic state in which a photon is emitted:

$$A^* \to A + \gamma. \tag{1.12}$$

Ordinary non-relativistic quantum mechanics cannot provide a first-principles account of this process, because the degrees of freedom it normally discusses are those of the '*matter*' units alone – that is, in this example, the electronic degrees of freedom. However, it is clear that something has changed radically in the *field* degrees of freedom. On the left-hand side, the matter is in an excited state and the electromagnetic field is somehow not manifest; on the right, the matter has made a transition to a lower-energy state and the energy difference has gone into creating a quantum of electromagnetic radiation. What is needed here is a quantum theory of the electromagnetic field – a *quantum field theory*.

Quantum field theory – or qft for short – is the fundamental formal and conceptual framework of the Standard Model. An important purpose of this book is to make this core twentieth century formalism more generally accessible. In chapter 5 we give a step-by-step introduction to qft. We shall see that a free classical field – which has infinitely many degrees of freedom – can be thought of as mathematically analogous to a vibrating solid (which has merely a very large number). The way this works mathematically is that the Fourier components of the field act like independent harmonic oscillators, just like the vibrational 'normal modes' of the solid. When quantum mechanics is applied to this system, the energy eigenstates of each oscillator are quantized in the familiar way, as $(n_r + 1/2)\hbar\omega_r$ for each oscillator of frequency ω_r: we say that such states contain 'n_r quanta of frequency ω_r'. The state of the entire field is characterized by how many quanta of each frequency are present. These 'excitation quanta' are the particle aspect of the field. In the ground state there are no excitations present – no field quanta – and so that is the *vacuum* state of the field.

In the case of the electromagnetic field, these quanta are of course photons (for the solid, they are phonons). In the process (1.12) the electromagnetic field was originally in its ground (no photon) state, and was raised finally to an excited state by the transfer of energy from the electronic degrees of freedom.

The final excited field state is defined by the presence of one quantum (photon) of the appropriate energy.

We obviously cannot stop here ('Electrons behave just like light'). All the particles of the SM must be described as excitation quanta of the corresponding quantum fields. But of course Feynman was somewhat overstating the case. The quanta of the electromagnetic field are *bosons*, and there is no limit on the number of them that can occupy a single quantum state. By contrast, the quanta of the electron field, for example, must be *fermions*, obeying the exclusion principle. In chapter 7 we shall see what modifications to the quantization procedure this requires. We must also introduce interactions between the excitation quanta, or equivalently between the quantum fields. This we do in chapter 6 for bosonic fields, and in chapter 7 for the Dirac and Maxwell fields thereby arriving at QED, our first quantum gauge field theory of the SM.

One reason the Lagrangian formulation of classical field (or particle) physics is so powerful is that *symmetries* can be efficiently incorporated, and their connection with *conservation laws* easily exhibited. The same is even more true in qft. For example, only in qft can the symmetry corresponding to electric charge conservation be simply understood. Indeed, all the quantum gauge field theories of the SM are deeply related to symmetries, as will become clear in the subsequent development.

In some cases, however, the symmetry – though manifest in the Lagrangian – is not visible in the usual empirical ways (conservation laws, particle multiplets, and so on). Instead, it is 'spontaneously (or dynamically) broken'. This phenomenon plays a crucial role in both QCD and the GSW theory. An aid to understanding it physically is provided by the analogy between the vacuum state of an interacting qft and the ground state of an interacting quantum many-body system – an insight due to Nambu (1960). We give an extended discussion of spontaneously broken symmetry in Part VII of volume 2. We shall see how the neutral bosonic (Bogoliubov) superfluid, and the charged fermionic (BCS) superconductor, offer instructive working models of dynamical symmetry breaking, relevant to chiral symmetry breaking in QCD, and to the generation of gauge boson masses in the GSW theory.

The road ahead is a long one, and we begin our journey at a more descriptive and pictorial level, making essential use of Yukawa's remarkable insight into the quantum nature of force. In due course, in chapter 6, we shall begin to see how qft supplies the precise mathematical formulae associated with such pictures.

1.3.2 The Yukawa theory of force as virtual quantum exchange

Yukawa's revolutionary paper (Yukawa 1935) proposed a theory of the strong interaction between a proton and a neutron, and also considered its possible extension to neutron β-decay. He built his theory by analogy with electromag-

netism, postulating a new field of force with an associated new field quantum, analogous to the photon. In doing so, he showed with particular clarity how, in quantum field theory, *particles interact by exchanging virtual quanta*, which *mediate* the force.

Before proceeding, we should emphasize that we are not presenting Yukawa's ideas as a viable candidate theory of strong and weak interactions. Crucially, Yukawa assumed that the nucleons and his quantum (later identified with the pion) were point-like, but in fact both nucleons and pions are quark composites with spatial extension. The true 'strong' interaction relates to the quarks, as we shall see in section 1.3.6. There are also other details of his theory which were (we now know) mistaken, as we shall discuss. Yet his approach was profound, and – as happens often in physics – even though the initial application was ultimately superseded, the ideas have broad and lasting validity.

Yukawa began by considering what kind of static potential might describe the n–p interaction. It was known that this interaction decreased rapidly for interparticle separation $r \geq 2$ fm. Hence, the potential could not be of coulombic type $\propto 1/r$. Instead, Yukawa postulated an n–p potential energy of the form

$$U(\boldsymbol{r}) = \frac{-g_{\mathrm{N}}^2}{4\pi} \frac{\mathrm{e}^{-\mathrm{r/a}}}{r} \tag{1.13}$$

where 'g_{N}' is a constant analogous to the electric charge e, $r = |\boldsymbol{r}|$ and 'a' is a range parameter (~ 2 fm). This static potential satisfies the equation

$$\left(\boldsymbol{\nabla}^2 - \frac{1}{a^2}\right) U(\boldsymbol{r}) = g_{\mathrm{N}}^2 \delta(\boldsymbol{r}) \tag{1.14}$$

(see appendix G) showing that it may be interpreted as the mutual potential energy of one point-like test nucleon of 'strong charge' g_{N} due to the presence of another point-like nucleon of equal charge g_{N} at the origin, a distance r away. Equation (1.14) should be thought of as a finite range analogue of Poisson's equation in electrostatics (equation (G.3))

$$\boldsymbol{\nabla}^2 V(r) = -\rho(r)/\epsilon_0, \tag{1.15}$$

the delta function in (1.14) (see appendix E) expressing the fact that the 'strong charge density' acting as the source of the field is all concentrated into a single point, at the origin.

Yukawa now sought to generalize (1.14) to the non-static case, so as to obtain a field equation for $U(\boldsymbol{r}, t)$. For $r \neq 0$, he proposed the free-space equation (we shall keep factors of c and $\hbar$ explicit for the moment)

$$\left(\boldsymbol{\nabla}^2 - \frac{\partial^2}{c^2 \partial t^2} - \frac{1}{a^2}\right) U(\boldsymbol{r}, t) = 0 \tag{1.16}$$

which is certainly relativistically invariant (see appendix D). Thus far, U is still a classical field. Now Yukawa took the decisive step of treating U quantum

mechanically, by looking for a (de Broglie-type) *propagating wave solution* of (1.16), namely

$$U \propto \exp(\mathrm{i}\boldsymbol{p}\cdot\boldsymbol{r}/\hbar - \mathrm{i}Et/\hbar). \tag{1.17}$$

Inserting (1.17) into (1.16) one finds

$$\frac{E^2}{c^2\hbar^2} = \frac{\boldsymbol{p}^2}{\hbar^2} + \frac{1}{a^2} \tag{1.18}$$

or, taking the positive square root,

$$E = \left[c^2\boldsymbol{p}^2 + \frac{c^2\hbar^2}{a^2}\right]^{1/2}.$$

Comparing this with the standard E–$\boldsymbol{p}$ relation for a massive particle in special relativity (appendix D), the fundamental conclusion is reached that the quantum of the finite-range force field U has a mass m_U given by

$$m_\mathrm{U}^2c^4 = \frac{c^2\hbar^2}{a^2} \qquad \text{or} \qquad m_\mathrm{U} = \frac{\hbar}{ac}. \tag{1.19}$$

This means that the *range parameter* in (1.13) is related to the *mass of the quantum* m_U by

$$a = \frac{\hbar}{m_\mathrm{U}c}. \tag{1.20}$$

Inserting $a \approx 2$ fm gives $m_\mathrm{U} \approx 100$ MeV, Yukawa's famous prediction for the mass of the nuclear force quantum.

Next, Yukawa envisaged that the U-quantum would be emitted in the transition $\mathrm{n} \to \mathrm{p}$, via a process analogous to (1.12):

$$\mathrm{n} \to \mathrm{p} + \mathrm{U}^- \tag{1.21}$$

where charge conservation determines the U^- charge. Yet there is an obvious difference between (1.21) and (1.12): (1.21) violates energy conservation since $m_\mathrm{n} < m_\mathrm{p} + m_\mathrm{U}$ if $m_\mathrm{U} \approx 100$ MeV, so it cannot occur as a real emission process. However, Yukawa noted that if (1.21) were combined with the inverse process

$$\mathrm{p} + \mathrm{U}^- \to \mathrm{n} \tag{1.22}$$

then an n–p interaction could take place by the mechanism shown in figure 1.1(a); namely, by the emission and subsequent absorption – that is, by the *exchange* – of a U^- quantum. He also included the corresponding U^+ exchange, where U^+ is the antiparticle of the U^-, as shown in figure 1.1(b).

An energy-violating transition such as (1.21) is known as a 'virtual' transition in quantum mechanics. Such transitions are routinely present in quantum-mechanical time-dependent perturbation theory and can be understood in terms of an 'energy–time uncertainty relation'

$$\Delta E \Delta t \geq \hbar/2. \tag{1.23}$$

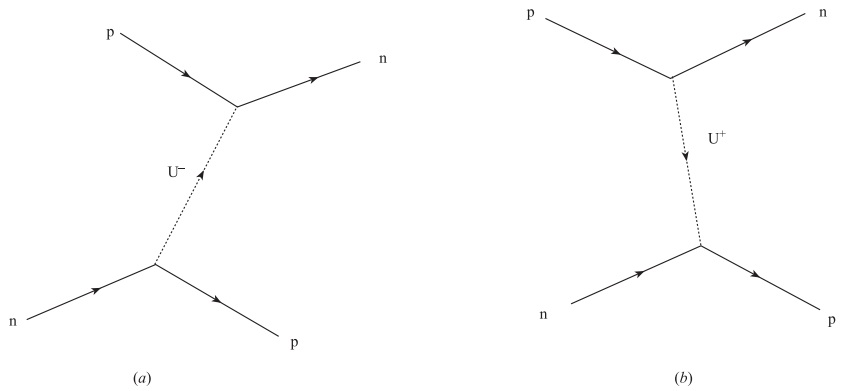


FIGURE 1.1
Yukawa's single-U exchange mechanism for the n–p interaction. (a) U^- exchange. (b) U^+ exchange.

The relation (1.23) may be interpreted as follows (we abridge the careful discussion in section 44 of Landau and Lifshitz (1977)). Imagine an 'energy-measuring device' set up to measure the energy of a quantum system. To do this, the device must interact with the quantum system for a certain length of time Δt. If the energy of a sequence of identically prepared quantum systems is measured, only in the limit $\Delta t \to \infty$ will the same energy be obtained each time. For finite Δt, the measured energies will necessarily fluctuate by an amount ΔE as given by (1.23); in particular, the shorter the time over which the energy measurement takes place, the larger the fluctuations in the measured energy.

Wick (1938) applied (1.23) to Yukawa's theory, and thereby shed new light on the relation (1.20). Suppose a device is set up capable of checking to see whether energy is, in fact, conserved while the $U^{\pm}$ crosses over in figure 1.1. The crossing time t must be at least r/c, where r is the distance apart of the nucleons. However, the device must be capable of operating on a time scale smaller than t (otherwise it will not be in a position to detect the $U^{\pm}$), but it need not be very much less than this. Thus the energy uncertainty in the reading by the device will be[3]

$$\Delta E \sim \frac{\hbar c}{r}. \tag{1.24}$$

As r decreases, the uncertainty ΔE in the measured energy increases. If we

[3]In this kind of argument, the '$\sim$' sign should be understood as meaning that numerical factors of order 1 (such as 2 or π) are not important. The coincidence between (1.25) and (1.20) should not be taken too literally. Nevertheless, the physics of (1.25) is qualitatively correct.

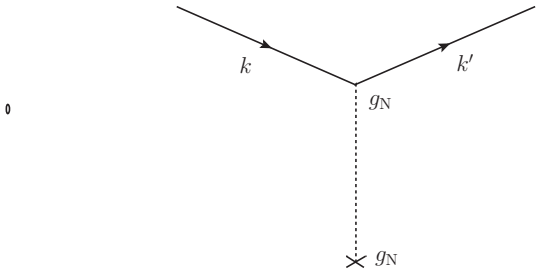


FIGURE 1.2
Scattering by a static point-like U-source.

require $\Delta E = m_U c^2$, then

$$r \sim \frac{\hbar}{m_U c} \tag{1.25}$$

just as in (1.20). The 'r' in (1.25) is the extent of the separation allowed between the n and the p, such that – in the time available – the $U^{\pm}$ can 'borrow' the necessary energy to come into existence and cross from one to the other. In this sense, r is the effective range of the associated force, as in (1.20).

Despite the similarity to virtual intermediate states in ordinary quantum mechanics, the Yukawa–Wick process is nevertheless truly revolutionary because it postulated an energy fluctuation ΔE great enough to create an as yet unseen new particle, a new state of matter.

We proceed to explore further aspects of Yukawa's force mechanism. The reader should note that throughout the remainder of this book we shall generally (unless otherwise stated) use units such that $\hbar = c = 1$: see Appendix B.

1.3.3 The one-quantum exchange amplitude

Consider a particle, carrying 'strong charge' g_N, being scattered by an infinitely massive (static) point-like U-source also of 'charge' g_N as pictured in figure 1.2. From the previous section, we know that the potential energy in the Schrödinger equation for the scattered particle is precisely the $U(\boldsymbol{r})$ from (1.13). Treating this to its lowest order in $U(\boldsymbol{r})$ ('Born Approximation' – see appendix H), the scattering amplitude is proportional to the Fourier transform of $U(\boldsymbol{r})$:

$$f(\boldsymbol{q}) = \int \mathrm{e}^{\mathrm{i}\boldsymbol{q}\cdot\boldsymbol{r}} U(\boldsymbol{r})\, \mathrm{d}^3\boldsymbol{r} \tag{1.26}$$

where $\boldsymbol{q}$ is the momentum (or wavevector, since $\hbar = 1$) transfer $\boldsymbol{q} = \boldsymbol{k} - \boldsymbol{k}'$. The transform is evaluated in appendix G equation (G.24), or in problem 1.1,

with the result

$$f(\boldsymbol{q}) = -\frac{g_{\mathrm{N}}^2}{\boldsymbol{q}^2 + m_{\mathrm{U}}^2}. \tag{1.27}$$

This implies that the amplitude (in this static case) for the one-U exchange amplitude is proportional to $-1/(\boldsymbol{q}^2 + m_{\mathrm{U}}^2)$, where $\boldsymbol{q}$ is the momentum carried by the U-quantum.

In this scattering by an infinitely massive source of potential, the energy of the scattered particle cannot change. In a real scattering process such as that in figure 1.1, both energy and momentum can be transferred by the U-quantum – that is, $\boldsymbol{q}$ is replaced by the four-momentum $q = (q_0, \boldsymbol{q})$, where $q_0 = k_0 - k_0'$. Then, as indicated in appendix G, the factor $-1/(\boldsymbol{q}^2 + m_U^2)$ is replaced by $1/(q^2 - m_{\mathrm{U}}^2)$ and the amplitude for figure 1.1 is, in this model,

$$\frac{g_{\mathrm{N}}^2}{q^2 - m_{\mathrm{U}}^2}. \tag{1.28}$$

It will be the main burden of chapters 5 and 6 to demonstrate just how this formula is arrived at, using the formalism of quantum field theory. In particular, we shall see in detail how the *propagator* $(q^2 - m_{\mathrm{U}}^2)^{-1}$ arises. For the present, we can already note (from appendix G) that such propagators are, in fact, momentum–space Green functions.

In chapter 6 we shall also discuss other aspects of the physical meaning of the propagator, and we shall see how diagrams which we have begun to draw in a merely descriptive way become true 'Feynman diagrams', each diagram representing by a precise mathematical correspondence a specific expression for a quantum amplitude, as calculated in perturbation theory. The expansion parameter of this perturbation theory is the dimensionless number $g_{\mathrm{N}}^2/4\pi$ appearing in the potential $U(\boldsymbol{r})$ (cf (1.13)). In terms of Feynman diagrams, we shall learn in chapter 6 that one power of g_{N} is to be associated with each 'vertex' at which a U-quantum is emitted or absorbed. Thus successive terms in the perturbation expansion correspond to exchanges of more and more quanta. Quantities such as g_{N} are called 'coupling strengths', or 'coupling constants'.

It is not too early to emphasize one very important point to the reader: true Feynman diagrams are *representations of momentum–space amplitudes*. They are not representations of space–time processes: *all* space–time points are integrated over in arriving at the formula represented by a Feynman diagram. In particular, the two 'intuitive' diagrams of figure 1.1, which carry an implied 'time-ordering' (with time increasing to the right), are *both* included in a single Feynman diagram with propagator (1.28), as we shall see in detail (for an analogous case) in section 7.1.

We now indicate how these general ideas of Yukawa apply to the actual interactions of quarks and leptons.

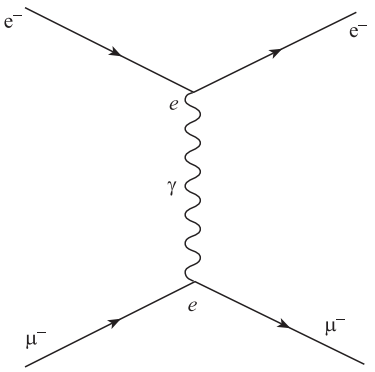


FIGURE 1.3
One photon exchange mechanism between charged leptons.

1.3.4 Electromagnetic interactions

From the foregoing viewpoint, electromagnetic interactions are essentially a special case of Yukawa's picture, in which g_{N}^2 is replaced by the appropriate electromagnetic charges, and $m_{\mathrm{U}} \to m_\gamma = 0$ so that $a \to \infty$ and the potential (1.13) returns to the Coulomb one, $-e^2/4\pi r$. A typical one-photon exchange scattering process is shown in figure 1.3, for which the generic amplitude (1.28) becomes

$$e^2/q^2. \tag{1.29}$$

Note that we have drawn the photon line 'vertically', consistent with the fact that both time-orderings of the type shown in figure 1.1 are included in (1.29). In the case of electromagnetic interactions, the coupling strength is e and the expansion parameter of perturbation theory is $e^2/4\pi \equiv \alpha \sim 1/137$ (see appendix C).

We can immediately use (1.29) to understand the famous $\sim \sin^{-4}\theta/2$ angular variation of Rutherford scattering. Treating the target muon as infinitely heavy (so as to simplify the kinematics), the electron scatters elastically so that $q_0 = 0$ and $q^2 = -(\boldsymbol{k} - \boldsymbol{k}')^2$ where $\boldsymbol{k}$ and $\boldsymbol{k}'$ are the incident and final electron momenta. So $q^2 = -2\boldsymbol{k}^2(1 - \cos\theta) = -4\boldsymbol{k}^2 \sin^2\theta/2$ where we have used the elastic scattering condition $\boldsymbol{k}^2 = \boldsymbol{k}'^2$. Inserting this into (1.29) and remembering that the cross section is proportional to the square of the amplitude (appendix H) we obtain the distribution $\sin^{-4}\theta/2$. Thus, such a distribution is a clear signature that *the scattering is proceeding via the exchange of a massless quantum.*

Unfortunately, the detailed implementation of these ideas to the electromagnetic interactions of quarks and leptons is complicated, because the electromagnetic potentials are the components of a 4-vector (see chapter 2), rather than a scalar as in (1.29), and the quarks and leptons all have spin-$\frac{1}{2}$, necessi-

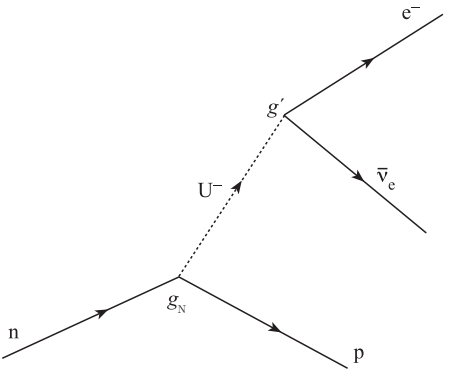


FIGURE 1.4
Yukawa's U-exchange mechanism for neutron β-decay.

tating the use of the Dirac equation (chapter 3). Nevertheless, (1.29) remains the essential 'core' of electromagnetic amplitudes.

As far as the electromagnetic field is concerned, its 4-vector nature is actually a fundamental feature, having to do with a *symmetry* called *gauge invariance*, or (better) *local phase invariance*. As we shall see in chapters 2 and 7, the form of the electromagnetic interaction is very strongly constrained by this symmetry. In fact, turning the argument around, one can (almost) understand the *necessity* of electromagnetic interactions as being due to the requirement of gauge invariance. Most significantly, we shall see in section 7.3.1 how the *masslessness of the photon* is also related to gauge invariance.

In chapter 8 a number of elementary electromagnetic processes will be fully analysed, and in chapter 11 we shall discuss higher-order corrections in QED.

1.3.5 Weak interactions

In a bold extension of his 'strong force' idea, Yukawa extended his theory to describe neutron β-decay as well, via the hypothesized process shown in figure 1.4 (here and in figure 1.5 we revert to the more intuitive 'time-ordered' picture – the reader may supply the diagrams corresponding to the other time-ordering). As indicated on the diagram, Yukawa assigned the strong charge $g_{\rm N}$ at the n–p end, and a different 'weak' charge g' at the lepton end. Thus the same quantum mediated both strong and weak transitions, and he had an embryonic 'unified theory' of strong and weak processes! If we take $\rm U^-$ to be the π^-, Yukawa's mechanism predicts the existence of the weak decay $\pi^- \to e^- + \bar{\nu}_e$.

This decay does indeed occur, though at a much smaller rate than the main mode which is $\pi^- \to \mu^- + \bar{\nu}_\mu$. But – apart from the now familiar problem with the compositeness of the nucleons and pions – this kind of unification is not

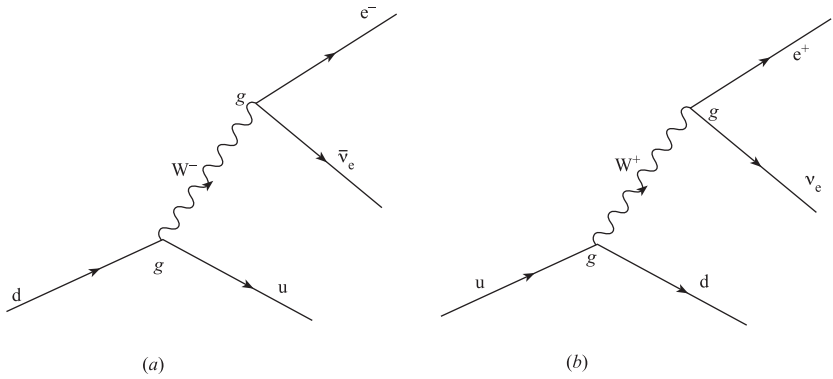


FIGURE 1.5
(a) β-decay and (b) e^+ emission at the quark level, mediated by $W^{\pm}$.

chosen by Nature. Not unreasonably in 1935, Yukawa was assuming that the range $\sim m_{\mathrm{U}}^{-1}$ of the strong force in n–p scattering (figure 1.1) was the same as that of the weak force in neutron β-decay (figure 1.4); after all, the latter (and more especially positron emission) was viewed as a nuclear process. But this is now known not to be the case: in fact, the range of the weak force is much smaller than nuclear dimensions – or, equivalently (see (1.19)), the masses of the mediating quanta are much greater than that of the pion.

β-decay is now understood as occurring at the quark level via the W^--exchange process shown in figure 1.5(a). Similarly, positron emission proceeds via figure 1.5(b). Other 'charged current' processes all involve $W^{\pm}$-exchange, generalized appropriately to include flavour mixing effects (see volume 2). 'Neutral current' processes involve exchange of the Z^0-quantum; an example is given in figure 1.6. The quanta $W^{\pm}, Z^0$ therefore mediate these weak interactions as does the photon for the electromagnetic one. Like the photon, the W and Z fields are the quanta of 4-vector fields[4]and have spin 1, but unlike the photon, the masses of the W and Z are far from zero – in fact $M_{\mathrm{W}} \approx 80$ GeV and $M_{\mathrm{Z}} \approx 91$ GeV. So the range of the force is $\sim M_{\mathrm{W}}^{-1} \sim 2.5 \times 10^{-18}$ m, much less than typical nuclear dimensions ($\sim$ few $\times 10^{-15}$ m). This, indeed, is one way of understanding why the weak interactions appear to be so weak: this range is so tiny that only a small part of the hadronic volume is affected.

Thus Nature has not chosen to unify the strong and weak forces via a common mediating quantum. Instead, it has turned out that the weak and strong forces (see section 1.3.6) are both *gauge theories*, generalizations of electromagnetism, as will be discussed in volume 2. This raises the possibility that it may be possible to 'unify' all three forces.

[4]This is dictated by the phenomenology of weak interactions – see chapter 20 in volume 2.

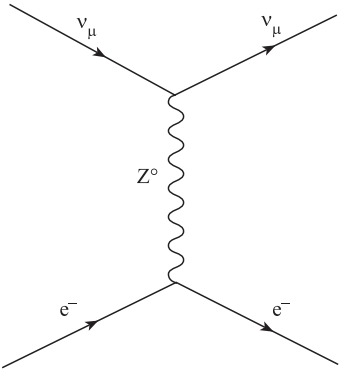


FIGURE 1.6
Z^0-exchange process.

Some initial idea of how this works in the 'electroweak' case may be gained by considering the amplitude for figure 1.5(a) in the low $-q^2$ limit. In a simplified version analogous to (1.29) which ignores the spin of the W and of the leptons, this amplitude is

$$g^2/(q^2 - M_{\mathrm{W}}^2) \tag{1.30}$$

where g is a 'weak charge' associated with W-emission and absorption. In actual β-decay, the square of the 4-momentum transfer q^2 is tiny compared to M_{W}^2, so that (1.30) becomes independent of q^2 and takes the constant value $-g^2/M_{\mathrm{W}}^2$. This corresponds, in configuration space, to a point-like interaction (the Fourier transform of a delta function is a constant). Just such a point-like interaction, shown in figure 1.7, had been postulated by Fermi (1934a, b) in the first theory of β-decay: it is a 'four-fermion' interaction with strength G_{F}. The value of G_{F} can be determined from measured β-decay rates. The dimensions of G_{F} turn out to be energy $\times$ volume, so that $G_{\mathrm{F}}/(\hbar c)^3$ has dimension (energy^{-2}). In our units $\hbar = c = 1$, the numerical value of G_{F} is

$$G_{\mathrm{F}} \sim (300\ \mathrm{GeV})^{-2}. \tag{1.31}$$

If we identify this constant with g^2/M_{W}^2 we obtain

$$g^2 \sim M_{\mathrm{W}}^2/(300\ \mathrm{GeV})^2 \sim 0.064 \tag{1.32}$$

a value quite similar to that of the electromagnetic charge e^2 as determined from $e^2 = 4\pi\alpha \sim 0.09$. Though this is qualitatively correct, we shall see in volume 2 that the actual relation, in the electroweak theory, between the weak and electromagnetic coupling strengths is somewhat more complicated than the simple equality '$g = e$'. (Note that a corresponding connection with Fermi's theory was also made by Yukawa!)

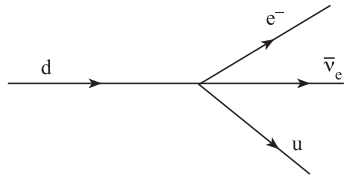


FIGURE 1.7
Point-like four-fermion interaction.

We can now understand the 'weakness' of the weak interactions from another viewpoint. For $q^2 \ll M_{\mathrm{W}}^2$, the ratio of the electromagnetic amplitude (1.29) to the weak amplitude (1.30) is of order q^2/M_{W}^2, given that $e \sim g$. Thus despite having an intrinsic strength similar to that of electromagnetism, weak interactions will appear very weak at low energies such that $q^2 \ll M_{\mathrm{W}}^2$. At energies approaching M_{W}, however, weak interactions will grow in importance relative to electromagnetic ones and, when $q^2 \gg M_{\mathrm{W}}^2$, weak and electromagnetic interactions will contribute roughly equally.

'Similar' coupling strengths are still not 'unified', however. True unification only occurs after a more subtle effect has been included, which goes beyond the one-quantum exchange mechanism. This is the *variation* or 'running' of the coupling strengths as a function of energy (or distance), caused by higher-order processes in perturbation theory. This will be discussed more fully in chapter 11 for QED, and in volume 2 for the other gauge couplings. It turns out that the possibility of unification depends crucially on an important difference between the weak interaction quanta $\mathrm{W}^{\pm}$ (to take the present example) and the photons of QED, which has not been apparent in the simple β-decay processes considered so far. The W's are themselves 'weakly charged', acting as both carriers and sources of the weak force field, and they therefore interact directly amongst themselves even in the absence of other matter. By contrast, photons are electromagnetically neutral and have no direct self-interactions. In theories where the gauge quanta self-interact, the coupling strength decreases as the energy increases, while for QED it increases. It is this differing 'evolution' that tends to bring the strengths together, ultimately.

Even granted similar coupling strengths and the fact that both are 4-vector fields, the idea of any electroweak unification appears to founder immediately on the markedly *different* ranges of the two forces or, equivalently, of the masses of the mediating quanta ($m_\gamma = 0$, $M_{\mathrm{W}} \sim 80$ GeV!). This difficulty becomes even more pointed when we recall that, as previously mentioned, the masslessness of the photon is related to gauge invariance in electrodynamics: how then can there be any similar kind of gauge symmetry for weak interactions, given the distinctly non-zero masses of the mediating quanta? Nevertheless, in one of the great triumphs of 20th century theoretical physics, it *is* possible to see the two theories as essentially similar gauge theories, the

gauge symmetry being 'spontaneously broken' in the case of weak interactions. This is a central feature of the GSW electroweak theory. An indication of how gauge quanta might acquire mass will be given in section 11.4 but a fuller explanation, with application to the electroweak theory, is reserved for volume 2. We will have a few more words to say about it in section 1.4.1.

1.3.6 Strong interactions

We turn to the contemporary version of Yukawa's theory of strong interactions, now viewed as occurring between quarks rather than nucleons. Evidence that the strong interquark force is in some way similar to QED comes from nucleon-nucleon (or nucleon-antinucleon) collisions. Regarding the nucleons as composites of point-like quarks, we would expect to see prominent events at large scattering angles corresponding to 'hard' q–q collisions (recall Rutherford's discovery of the nucleus). Now the result of such a hard collision would normally be to scatter the quarks to wide angles, 'breaking up' the nucleons in the process. However, quarks (except for the t quark) are *not* observed as free particles. Instead, what appears to happen is that, as the two quarks separate from each other, their mutual potential energy increases – so much so that, at a certain stage in the evolution of the scattering process, the energy stored in the potential converts into a new $\mathrm{q\bar{q}}$ pair. This process continues, with in general many pairs being produced as the original and subsequent pairs pull apart. By a mechanism which is still not quantitatively understood in detail, the produced quarks and antiquarks (and the original quarks in the nucleons) bind themselves into hadrons within an interaction volume of order 1 fm^3, so that no free quarks are finally observed, consistent with 'confinement'. Very strikingly, these hadrons emerge in quite well-collimated 'jets', suggesting rather vividly their ancestry in the original separating qq pair. Suppose, then, that we plot the angular distribution of such '*two jet events*': it should tell us about the dynamics of the original interaction at the quark level.

Figure 1.8 shows such an angular distribution from proton–antiproton scattering, so that the fundamental interaction in this case is the elastic scattering process $\mathrm{\bar{q}q} \to \mathrm{\bar{q}q}$. Here θ is the scattering angle in the $\mathrm{\bar{q}q}$ centre of mass system (CMS). Amazingly, the θ-distribution follows almost exactly the 'Rutherford' form $\sin^{-4}\theta/2$.

We saw how, in the Coulomb case, this distribution could be understood as arising from the propagator factor $1/q^2$, which itself comes from the $1/r$ potential associated with the massless quantum involved, namely the photon. In the present case, the same $1/q^2$ factor is responsible: here, in the $\mathrm{\bar{q}q}$ centre of mass system, $\boldsymbol{k}$ and $-\boldsymbol{k}$ are the momenta of the initial $\mathrm{\bar{q}}$ and q, while $\boldsymbol{k}'$ and $-\boldsymbol{k}'$ are the corresponding final momenta. Once again, for elastic scattering there is no energy transfer, and $q^2 = -\boldsymbol{q}^2 = -(\boldsymbol{k} - \boldsymbol{k}')^2 = -4\boldsymbol{k}^2 \sin^2\theta/2$ as before, leading to the $\sin^{-4}\theta/2$ form on squaring $1/q^2$. Once again, such a

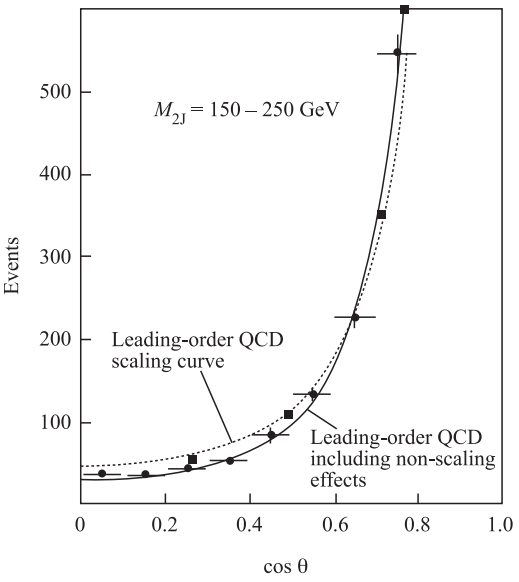


FIGURE 1.8
Angular distribution of two-jet events in $p\bar{p}$ collisions (Arnison *et al.* 1985) as a function of $\cos\theta$, where θ is the CMS scattering angle. The broken curve is the prediction of QCD, obtained in the lowest order of perturbation theory (one-gluon exchange); it is virtually indistinguishable from the Rutherford (one-photon exchange) shape $\sin^{-4}\theta/2$. The full curve includes higher order QCD corrections.

distribution is a clear signal that a massless quantum is being exchanged – in this case, the *gluon*.

It might then seem to follow that, as in the case of QED, the QCD interaction has infinite range. But this cannot be right: the strong forces do not extend beyond the size of a typical hadron, which is roughly 1 fm. Indeed, the QCD force is mediated by the massless spin-1 gluon, and QCD is also a gauge theory; but the form of the QCD interaction, though somewhat analogous to QED, is more complicated, and the long range behaviour of the force is very different.

As we have seen, each quark comes in three colours, and the QCD force is sensitive to this colour label: the gluons effectively 'carry colour' back and forth between the quarks, as shown in the one-gluon exchange process of figure 1.9. Because the gluons carry colour, they can interact with themselves, like the W's and Z's of the GSW theory. As in that case, these gluonic self-interactions cause the QCD interaction strength to decrease at short distances (or high energies), ultimately tending to zero, the property known as

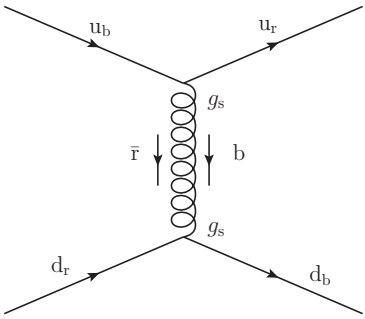


FIGURE 1.9
Strong scattering via gluon exchange. At the top vertex, the 'flow' of colour is b (quark) → r (quark) + r̄b (gluon); at the lower vertex the flow is r̄b (gluon) + r (quark) → b (quark).

asymptotic freedom. So in 'hard' collisions occurring at short inter-particle distances, the one-gluon exchange mechanism gives a good first approximation to the data. But the force grows much stronger as the quarks separate from each other, and perturbation theory is no longer a reliable guide. In fact, it seems that a new, non-perturbative, effect occurs – namely *confinement*. Once again, a gauge theory, with formal similarity to QED, has very different physical consequences.

A phenomenological qq (or $q\bar{q}$) potential which is often used in quark models has the form

$$V = -\frac{a}{r} + br \tag{1.33}$$

where the first term, which dominates at small r, arises from a single-gluon exchange so that $a \sim g_s^2$, where the strong (QCD) charge is g_s. The second term models confinement at larger values of r. Such a potential provides quite a good understanding of the gross structure of the $c\bar{c}$ and $b\bar{b}$ systems (see problem 1.5). A typical value for b is 0.85 GeV fm^{-1} (which corresponds to a constant force of about 14 tonnes!). Thus at $r \sim 2$ fm, there is enough energy stored to produce a pair of the lighter quarks. This 'linear' part of the potential cannot be obtained by considering the exchange of one, or even a finite number of, gluons: in other words, not within an approach based on perturbation theory.

It is interesting to note that the linear part of the potential may be regarded as the solution of the *one-dimensional* form of $\boldsymbol{\nabla}^2 V = 0$, namely $\mathrm{d}^2V/\mathrm{d}r^2 = 0$; this is in contrast to the Coulombic $1/r$ part, which is a solution (except at $r = 0$) to the full three-dimensional Laplace equation. This suggests that the colour field lines connecting two colour charges spread out into all of space when the charges are close to each other, but are somehow 'squeezed' into an elongated one-dimensional 'string' as the distance between

the charges becomes greater than about 1 fm. In the second volume, we shall see that numerical simulations of QCD, in which the space–time continuum is represented as a discrete lattice of points, indicate that such a linear potential does arise when QCD is treated non-perturbatively. It remains a challenge for theory to demonstrate that confinement follows from QCD.

It is believed that gluons too are confined by QCD, so that – like quarks – they are not seen as isolated free particles. But they too 'hadronize' after being produced in a primitive short-distance collision process, as happens in the case of q's and $\bar{\mathrm{q}}$'s. Such 'gluon jets' provide indirect evidence for the existence and properties of gluons, as we shall see in volume 2.

This is an appropriate moment at which to emphasize what appears to be a crucial distinction between the three 'charges' (electromagnetic, weak and strong) on the one hand, and the various flavour quantum numbers on the other. The former have a dynamical significance, whereas the latter do not. In the case of electric charge, for example, this means simply that a particle carrying this property responds in a definite way to the presence of an electromagnetic field and itself creates such a field. No such force fields are known for any of the flavour numbers, which are (at present) purely empirical classification devices, without dynamical significance.

1.3.7 The gauge bosons of the Standard Model

We can now gather together the mediators of the SM forces. They are all *gauge bosons*, meaning that they are the quanta of various 4-vector gauge fields. For example, the photon is the quantum of the electromagnetic (Maxwell) 4-vector potential $A^\mu(x)$ (see chapter 2 and section 6.3.1), which is the simplest gauge field. The gluon is the quantum of the QCD potential $A^\mu_a(x)$, where the colour index a runs from 1 to 8. The reason there are 8 of them may be guessed from figure 1.9: each gluon can be thought of as carrying one colour-anticolour combination, such as $\bar{\mathrm{r}}\mathrm{b}$, $\bar{\mathrm{b}}\mathrm{g}$, and so on; the symmetric combination $\bar{\mathrm{r}}\mathrm{r} + \bar{\mathrm{b}}\mathrm{b} + \bar{\mathrm{g}}\mathrm{g}$ is totally colourless and is discarded (see section 12.2 in volume 2). In the GSW electroweak theory, there are four gauge fields, $W^\mu_i(x)$ where i runs from 1 to 3, and $B^\mu(x)$ which is analogous to $A^\mu(x)$. One linear combination of $W^\mu_3(x)$ and $B^\mu(x)$ is associated with the photon field $A^\mu(x)$; the orthogonal combination is associated with the $Z^\mu(x)$ field whose quantum is the Z^0. The charged carriers $\mathrm{W}^\pm$ are associated with the $W^\mu_1(x)$ and $W^\mu_2(x)$ components of the $W^\mu_i(x)$ field.

We shall assume that the mass of the photon and of the gluon is exactly zero. This can never be established experimentally, of course: the current experimental limit on the photon mass is that it is less than 1×10^{-18} (Nakamura *et al.* 2010). All gauge fields have spin 1 (in units of $\hbar$). Ordinarily, a spin-1 particle would be expected to have three polarization states, according to quantum mechanics. However it is a general result that in the massless case the quanta have only two polarization states, both transverse to the direction of motion; the longitudinally polarized state is absent (this property,

TABLE 1.3
Properties of SM gauge bosons.

Particle	Polarization states	Mass	Width/Lifetime
γ (photon)	2	0 (theoretical)	stable
g (gluon)	2	0 (theoretical)	stable
$W^{\pm}$	3	80.399 ± 0.023 GeV	$\Gamma_W = 2.085 \pm 0.042$ GeV
Z^0	3	91.187 ± 0.0021 GeV	$\Gamma_Z = 2.4952 \pm 0.0023$ GeV

familiar for the corresponding classical fields which are purely transverse, will be discussed in section 7.3.1). By contrast, all three polarization states are present for the massive gauge bosons.

The photon and the gluon are stable particles. The $W^{\pm}$ and Z^0 particles decay with total widths of the order of 2 GeV (lifetimes $\sim 0.3 \times 10^{-24}$ s). Although this is significantly shorter than typical strong interaction decay lifetimes, these are of course weak decays, the rate being enhanced by the large energy release.

Table 1.3 lists the properties of the SM gauge bosons; the masses and widths are taken from Nakamura *et al.* (2010).

1.4 Renormalization and the Higgs sector of the Standard Model

1.4.1 Renormalization

So far we have been discussing processes in which only one particle is exchanged. These will generally be the terms of lowest order in a perturbative expansion in powers of the coupling strength. But we must clearly go beyond lowest order, and include the effects of multi-particle exchanges. We shall explain how to do this in chapter 10, for a simple scalar field theory. Such multi-particle exchange amplitudes are given by integrals over the momenta of the exchanged particles, constrained only by four-momentum conservation (no integral arises in the case of the exchange of a single particle, because its four-momentum is fixed in terms of the momenta of the scattering particles, as in section 1.2.3). It turns out that the integrals nearly always *diverge* as the momenta of the exchanged particles tend to infinity. Nevertheless, as we shall explain in chapter 10, this theory can be reformulated, by a process called

renormalization, in such a way that all multi-particle (higher-order) processes become finite and calculable – a quite remarkable fact, and one that is of course an absolutely crucial requirement in the case of the Standard Model interactions, where the relevant data are precise enough to test the accuracy of the theory well beyond lowest order, particularly in the case of QED (see chapter 11). The price to be paid for this taming of the divergences is just that the basic parameters of the theory, such as masses and coupling constants, have to be treated as parameters to be determined by comparison to the data, and cannot themselves be calculated.

But some theories cannot be reformulated in this way – they are *non-renormalizable*. A simple test for whether a theory is renormalizable or not will be discussed in section 11.8: if the coupling constant has dimensions of a mass to an inverse power, the theory is non-renormalizable. An example of such a theory is the original four-Fermi theory of weak interactions, where the coupling constant $G_{\rm F}$ has the dimensions of an inverse square mass (or energy) as we saw in (1.31). We will look at this theory again in section 11.8, but the essential point for our purpose now is that the dimensionful coupling constant introduces an *energy scale* into the problem, namely ${G_{\rm F}}^{-1/2} \sim 300$ GeV. It seems reasonable to infer that a more relevant measure of the interaction strength will be given by the dimensionless number $EG_{\rm F}^{1/2}$, where E is a characteristic physical energy scale of any weak process under consideration – for example, the energy in the centre of momentum frame in a two-particle scattering process, at least at energies much greater than the particle masses. Then, for energies very much less than ${G_{\rm F}}^{-1/2}$ the effective strength will be very weak, and the lowest order term in perturbation theory will work fine; this is how the Fermi theory was used, for many years. But as the energy increases, what happens is that more and more parameters have to be taken from experiment, in order to control the divergences; as the energy approaches $G_{\rm F}^{-1/2}$, the theory becomes totally non-predictive and breaks down. Thus renormalizability is regarded as highly desirable in a theory.

One might hope to come up with a renormalizable theory of weak interactions by replacing the four-fermion interaction by a Yukawa-like mechanism, with exchange of a quantum of mass M and dimensionless coupling y, say. Then just as in (1.32) we would identify $G_{\rm F} \sim y^2/M^2$ at low energies. However, as we have seen, phenomenology implies that the massive exchanged quantum must have spin 1. Unfortunately, this type of straightforward massive spin-1 theory is not renormalizable either, as we shall discuss in chapter 22 (in volume 2). The trouble can be traced directly to the existence of the *longitudinal* polarization state which, as noted previously, is present for a massive spin-1 particle. If the exchanged spin-1 quantum were massless, as in QED, it would lack that third polarization state, and the theory would be renormalizable. But weak interaction facts dictate both non-zero mass and spin-1.

In the case of QED, there is a symmetry principle behind both the zero mass of the photon and the absence of the longitudinal polarization state:

this symmetry is *gauge invariance* as we shall explain in section 7.3.1. It turns out that this symmetry is vital in rendering QED renormalizable. It is natural then to ask whether in the case of QED, a situation ever arises where the photon acquires mass, while retaining fully gauge-invariant interactions – and hence renormalizability (we would hope). If so, we would then have an analogue of what is needed for a renormalizable theory of weak interactions. The answer is that this can indeed happen, but it requires some *extra dynamics* to do it. Nature has actually provided us with a working model of what we want, in the phenomenon of superconductivity. There, the Meissner effect can be interpreted as implying that the photons propagating in a thin surface layer of the material have non-zero mass (see section 19.2). The dynamics behind this is subtle, and required many years of theoretical efforts before it was finally understood by Bardeen, Cooper and Schrieffer (1957). In simple terms, the mechanism is a two-step process. First, lattice interactions cause electrons to bind into pairs; then these pairs undergo Bose-Einstein condensation. This 'condensate' is the BCS superconducting ground state. The essential point is that although the electromagnetic interactions are fully gauge invariant, the ground state is not. When a symmetry is broken by the ground state, it is said to be 'spontaneously' broken. We shall provide an introduction to the BCS ground state in chapter 17 of volume 2.

The BCS theory is an example of spontaneous symmetry breaking occurring dynamically (through the particular lattice interactions). Many of the physically important phenomena can, however, be very satisfactorily described in terms of an *effective theory*, which treats only the electrodynamics of the condensate. Such a description was proposed by Ginzburg and Landau (1950), well before the BCS paper, in fact.

How can this be applied in particle physics? Recall the idea, mentioned in section 1.3.1, that the analogue of the many-body ground state is the qft vacuum (Nambu 1961). In the Standard Model, the weak interactions are indeed described by a gauge-invariant theory, and the *assumption* is made that the vacuum breaks the gauge symmetry. The simplest way this idea can be implemented is along the lines of the Ginzburg-Landau theory, as suggested by Weinberg (1967) and by Salam (1968), and their proposal is embodied in the Glashow-Salam-Weinberg electroweak theory, which is part of the SM. It requires the introduction of four new spin-0 fields, which are called Higgs fields (Higgs 1964, Englert and Brout 1964, Guralnik *et al.* 1964), and which we may think of as playing the role of the BCS condensate (but not for electromagnetism, of course). The combined theory of quarks, leptons, electroweak gauge fields, and Higgs fields is gauge invariant, but one of the Higgs fields is supposed to have a non-zero average value in the physical vacuum, which breaks the gauge symmetry. The other three Higgs fields effectively become the longitudinal parts of the massive spin-1 $W^{\pm}$ and Z^0 fields, while the quantized excitations of the fourth Higgs field away from its vacuum value appear physically as neutral spin-0 particles, called Higgs bosons (Higgs 1964).

Apart from giving mass to the $W^{\pm}$ and Z^0, the Higgs fields have more work to do. The electroweak gauge symmetry is exact only if all the fermion masses are zero; this is because it is a *chiral* symmetry (similar to, but not the same as, the chiral symmetry of QCD mentioned in section 1.2.2). Once again, this chiral gauge symmetry is essential to the renormalizability of the theory: if the fermion masses are incorporated in the usual way as parameters in the Lagrangian, the latter is no longer gauge invariant and the theory is non-renormalizable. In the SM, this problem is solved by having no fermion masses in the Lagrangian, and by postulating gauge-invariant Yukawa interactions between the fermions and the Higgs fields, which are arranged in such a way that, when the Higgs field gets a vacuum expectation value, the interaction terms yield just the fermion masses. So again, the symmetry breaking is economically blamed on the same property of the vacuum. When the Higgs field oscillates away from its vacuum value, the result will be residual interactions between the fermions and the Higgs boson, which will have the defining characteristic that each fermion will interact with the Higgs boson with a strength proportional to its (i.e. the fermion's) mass. This is clearly a testable prediction, once the Higgs boson is found.

We have emphasized the role that the Higgs fields play in the renormalizability of the GSW theory. The all-important proof of that renormalizability was given by 't Hooft (1971b), and he also proved the renormalizability of QCD (1971a); see also 't Hooft and Veltman (1972).

The SM Higgs sector is the simplest one that will do the job; more complicated versions are possible. Perhaps the Higgs field is a composite formed in some new heavy fermion-antifermion dynamics, reminiscent of BCS pairing. In any case, the SM Higgs sector is there to be tested experimentally. In the following section we shall discuss briefly what is presently known about the SM Higgs boson, postponing a fuller discussion until we present the GSW theory in chapter 22 in volume 2.

Before ending this section we must note that modern renormalization theory is concerned with more than perturbative calculability. The *renormalization group* and related ideas provide powerful tools for 'improving' perturbation theory, by systematically resumming terms which (in the particle physics case) dominate at short distances. Prominent among the results of this analysis (see chapters 15 and 16) are the concepts of energy-dependent ('running') masses and coupling strengths, and the calculation of QCD corrections to parton-model predictions.

1.4.2 The Higgs boson of the Standard Model

According to the SM, just one neutral spin-0 Higgs boson is expected; its mass m_{H} is not predicted by the theory. The experimental discovery of the SM Higgs boson has been a major goal of several generations of accelerators: the LEP e^+e^- collider at Cern, the Tevatron $p\bar{p}$ collider at Fermilab, and most recently the LHC pp collider at Cern. Experimentally, bounds on the

Higgs mass can be obtained directly, through searching for its production and subsequent decay; non-observation will lead to a lower bound for m_{H}. There are also indirect constraints, coming from fits to precision measurements of electroweak observables. The latter are sensitive to higher order corrections which involve the Higgs boson as a virtual particle; these depend logarithmically on the unknown parameter m_{H} and give upper bounds on m_{H}, assuming, of course, that the SM is correct.

A lower bound

$$m_{\mathrm{H}} > 114.4 \ \ \mathrm{GeV} \quad (95\% \ \ \mathrm{C.L.}) \tag{1.34}$$

was set at LEP (LEP 2003) by combining data on direct searches. Combining this with a global fit to precision electroweak data, an upper bound

$$m_{\mathrm{H}} < 186 \ \ \mathrm{GeV} \quad (95\% \ \ \mathrm{C.L.}) \tag{1.35}$$

was obtained (Nakamura *et al.* 2010).

By early 2012, the combined results of the CDF and D0 experiments at the Tevatron, and the ATLAS and CMS experiments at the LHC, excluded an m_{H} value in the interval (approximately) 130 GeV to 600 GeV, at 95 % C.L. Finally, in July 2012 the ATLAS (Aad *et al.* 2012) and CMS (Chatrchyan *et al.* 2012) collaborations announced the discovery, with a significance of 5σ, of a neutral boson with a mass in the range 125–126 GeV, its production and decay rates being broadly compatible with the predictions for the SM Higgs boson. The existence of the measured decay to two photons implies that the particle is a boson with spin different from 1 (Landau 1948, Yang 1950), but spin-0 has not yet been confirmed. Nevertheless, it is probable that this is the (or perhaps a) Higgs boson. Its long-anticipated discovery opens a new era in particle physics: the experimental exploration of the symmetry-breaking sector of the SM.

1.5 Summary

The Standard Model provides a relatively simple picture of quarks and leptons and their non-gravitational interactions. The quark colour triplets are the basic source particles of the gluon fields in QCD, and they bind together to make hadrons. The weak interactions involve quark and lepton doublets – for instance the quark doublet (u, d) and the lepton doublet (ν_e, e^-) of the first generation. These are sources for the $\mathrm{W}^{\pm}$ and Z^0 fields. Charged fermions (quarks and leptons) are sources for the photon field. All the mediating force quanta have spin-1. The weak and strong force fields are generalizations of electromagnetism; all three are examples of gauge theories, but realized in subtly different ways.

In the following chapters our aim will be to lead the reader through the mathematical formalism involved in giving precise quantitative form to what we have so far described only qualitatively and to provide physical interpretation where appropriate. In the remainder of part I of the present volume, we first show how Schrödinger's quantum mechanics and Maxwell's electromagnetic theory may be combined as a gauge theory – in fact the simplest example of such a theory. We then introduce relativistic quantum mechanics for spin-0 and spin-$\frac{1}{2}$ particles, and include electromagnetism via the gauge principle. Lorentz transformations and discrete symmetries are also covered. In part II, we develop the formalism of quantum field theory, beginning with scalar fields and moving on to QED; this is then applied to many simple ('tree level') QED processes in part III. In the final part IV, we present an introduction to renormalization at the one-loop level, including renormalization of QED. The more complicated gauge theories of QCD and the electroweak theory are reserved for volume 2.

Problems

1.1 Evaluate the integral in (1.26) directly. [*Hint*: Use spherical polar coordinates with the polar axis along the direction of $\boldsymbol{q}$, so that $\mathrm{d}^3\boldsymbol{r} = r^2\mathrm{d}r\ \sin\theta\ \mathrm{d}\theta\ \mathrm{d}\phi$, and $\exp(\mathrm{i}\boldsymbol{q}\cdot\boldsymbol{r}) = \exp(\mathrm{i}|\boldsymbol{q}|r\cos\theta)$. Make the change of variable $x = \cos\theta$, and do the ϕ integral (trivial) and the x integral. Finally do the r integral.]

1.2 Using the concept of strangeness conservation in strong interactions, explain why the threshold energy (for π^- incident on stationary protons) for

$$\pi^- + \mathrm{p} \to \mathrm{K}^0 + \text{anything}$$

is less than for

$$\pi^- + \mathrm{p} \to \bar{\mathrm{K}}^0 + \text{anything}$$

assuming both processes proceed through the strong interaction.

1.3 Note: the invariant square p^2 of a 4-momentum $p = (E, \boldsymbol{p})$ is defined as $p^2 = E^2 - \boldsymbol{p}^2$. We remind the reader that $\hbar = c = 1$ (see Appendix B).

(a) An electron of 4-momentum k scatters from a stationary proton of mass M via a one-photon exchange process, producing a final hadronic state of 4-momentum p', the final electron 4-momentum being k'. Show that

$$p'^2 = q^2 + 2M(E - E') + M^2$$

where $q^2 = (k - k')^2$, and E, E' are the initial and final electron energies in this frame (i.e. the one in which the target proton is

at rest). Show that if the electrons are highly relativistic then $q^2 = -4EE' \sin^2 \theta/2$, where θ is the scattering angle in this frame. Deduce that for elastic scattering E' and θ are related by

$$E' = E \Big/ \left(1 + \frac{2E}{M} \sin^2 \theta/2\right).$$

(b) Electrons of energy 4.879 GeV scatter elastically from protons, with $\theta = 10°$. What is the observed value of E'?

(c) In the scattering of these electrons, at 10°, it is found that there is a peak of events at $E' = 4.2$ GeV; what is the invariant mass of the produced hadronic state (in MeV)?

(d) Calculate the value of E' at which the 'quasi-elastic peak' will be observed, when electrons of energy 400 MeV scatter at an angle $\theta = 45°$ from a He nucleus, assuming that the struck nucleon is at rest inside the nucleus. Estimate the broadening of this final peak caused by the fact that the struck nucleon has, in fact, a momentum distribution by virtue of being localized within the nuclear size.

1.4

(a) In a simple non-relativistic model of a hydrogen-like atom, the energy levels are given by

$$E_n = \frac{-\alpha^2 Z^2 \mu}{2n^2}$$

where Z is the nuclear charge and μ is the reduced mass of the electron and nucleus. Calculate the splitting in eV between the $n = 1$ and $n = 2$ states in positronium, which is an e^+e^- bound state, assuming this model holds.

(b) In this model, the e^+e^- potential is the simple Coulomb one

$$-\frac{e^2}{4\pi\epsilon_0 r} = -\frac{\alpha}{r}.$$

Suppose that the potential between a heavy quark Q and an antiquark $\bar{Q}$ was

$$-\frac{\alpha_s}{r}$$

where α_s is a 'strong fine structure constant'. Calculate values of α_s (different in (i) and (ii)) corresponding to the information (the quark masses are phenomenological 'quark model' masses)

(i) the splitting between the $n = 2$ and $n = 1$ states in charmonium ($c\bar{c}$) is 588 MeV, and $m_c = 1870$ MeV;

(ii) the splitting between the $n = 2$ and $n = 1$ states in the upsilon series ($b\bar{b}$) is 563 MeV, and $m_b = 5280$ MeV.

(c) In positronium, the $n = 1\,^3S_1$ and $n = 1\,^1S_0$ states are split by the hyperfine interaction, which has the form $\frac{7}{48}\alpha^4 m_e \boldsymbol{\sigma}_1 \cdot \boldsymbol{\sigma}_2$ where m_e is the electron mass and $\boldsymbol{\sigma}_1, \boldsymbol{\sigma}_2$ are the spin matrices for the e^- and e^+ respectively. Calculate the expectation value of $\boldsymbol{\sigma}_1 \cdot \boldsymbol{\sigma}_2$ in the $^3S_1$ and $^1S_0$ states, and hence evaluate the splitting between these levels (calculated in lowest order perturbation theory) in eV. [*Hint*: the total spin $\boldsymbol{S}$ is given by $\boldsymbol{S} = \frac{1}{2}(\boldsymbol{\sigma}_1 + \boldsymbol{\sigma}_2)$. So $\boldsymbol{S}^2 = \frac{1}{4}(\boldsymbol{\sigma}_1^2 + \boldsymbol{\sigma}_2^2 + 2\boldsymbol{\sigma}_1 \cdot \boldsymbol{\sigma}_2)$. Hence the eigenvalues of $\boldsymbol{\sigma}_1 \cdot \boldsymbol{\sigma}_2$ are directly related to those of $\boldsymbol{S}^2$.]

(d) Suppose an analogous 'strong' hyperfine interaction existed in the $c\bar{c}$ system, and was responsible for the splitting between the $n = 1\,^3S_1$ and $n = 1\,^1S_0$ states, which is 116 MeV experimentally (i.e. replace α by α_s and m_e by $m_c = 1870$ MeV). Calculate the corresponding value of α_s.

1.5 The potential between a heavy quark Q and an antiquark $\bar{Q}$ is found empirically to be well represented by

$$V(r) = -\frac{\alpha_s}{r} + br$$

where $\alpha_s \approx 0.5$ and $b \approx 0.18$ GeV^2. Indicate the origin of the first term in $V(r)$, and the significance of the second.

An estimate of the ground-state energy of the bound $Q\bar{Q}$ system may be made as follows. For a given r, the total energy is

$$E(r) = 2m - \frac{\alpha_s}{r} + br + \frac{p^2}{m}$$

where m is the mass of the Q (or $\bar{Q}$) and p is its momentum (assumed non-relativistic). Explain why p may be roughly approximated by $1/r$, and sketch the resulting $E(r)$ as a function of r. Hence show that, in this approximation, the radius of the ground state, r_0, is given by the solution of

$$\frac{2}{mr_0^3} = \frac{\alpha_s}{r_0^2} + b.$$

Taking $m = 1.5$ GeV as appropriate to the $c\bar{c}$ system, verify that for this system

$$(1/r_0) \approx 0.67 \text{ GeV}$$

and calculate the energy of the $c\bar{c}$ ground state in GeV, according to this model.

An excited $c\bar{c}$ state at 3.686 GeV has a total width of 278 keV, and one at 3.77 GeV has a total width of 24 MeV. Comment on the values of these widths.

1.6 The Hamiltonian for a two-state system using the normalized base states $|1\rangle, |2\rangle$ has the form

$$\begin{pmatrix} \langle 1|H|1\rangle & \langle 1|H|2\rangle \\ \langle 2|H|1\rangle & \langle 2|H|2\rangle \end{pmatrix} = \begin{pmatrix} -a\cos 2\theta & a\sin 2\theta \\ a\sin 2\theta & a\cos 2\theta \end{pmatrix}$$

where a is real and positive. Find the energy eigenvalues E_+ and E_-, and express the corresponding normalized eigenstates $|+\rangle$ and $|-\rangle$ in terms of $|1\rangle$ and $|2\rangle$.

At time $t = 0$ the system is in state $|1\rangle$. Show that the probability that it will be found to be in state $|2\rangle$ at a later time t is

$$\sin^2 2\theta \sin^2(at).$$

Discuss how a formalism of this kind can be used in the context of neutrino oscillations. How might the existence of neutrino oscillations explain the solar neutrino problem? (This will be discussed in chapter 21 of volume 2.)

1.7 In an interesting speculation, it has been suggested (Arkani-Hamad *et al.* 1998, 1999, Antoniadis *et al.* 1998) that the weakness of gravity as observed in our (apparently) three-dimensional world could be due to the fact that gravity actually extends into additional 'compactified' dimensions (that is, dimensions which have the geometry of a circle, rather than of an infinite line). For the particles and forces of the Standard Model, however, such leakage into extra dimensions has to be confined to currently probed distances, which are of order M_{W}^{-1}.

(a) Consider Newtonian gravity in $(3 + d)$ spatial dimensions. Explain why you would expect that the gravitational potential will have the form

$$V_{\mathrm{N},3+d}(r) = -\frac{m_1 m_2 G_{\mathrm{N},3+d}}{r^{d+1}}. \tag{1.36}$$

[Think about how the '$1/r^2$' fall-off of the *force* is related to the *surface area* of a sphere in the case $d = 0$. Note that the formula works for $d = -2$! What happens in the case $d = -1$?]

(b) Show that $G_{\mathrm{N},3+d}$ has dimensions $(\mathrm{mass})^{-(2+d)}$. This allows us to introduce the 'true' Planck scale – i.e. the one for the underlying theory in $3 + d$ spatial dimensions – as $G_{\mathrm{N},3+d} = (M_{\mathrm{P},3+d})^{-(2+d)}$.

(c) Now suppose that the form (1.36) only holds when the distance r between the masses is much smaller R, the size of the compactified dimensions. If the masses are placed at distances $r \gg R$, their gravitational flux cannot continue to penetrate into the extra dimensions, and the potential (1.36) should reduce to the familiar three-dimensional one; so we must have

$$V_{\mathrm{N},3+d}(r \gg R) = -\frac{m_1 m_2 G_{\mathrm{N},3+d}}{R^d}\frac{1}{r}. \tag{1.37}$$

Show that this implies that

$$M_{\rm P}^2 = M_{{\rm P},3+d}^2 (R M_{{\rm P},3+d})^d. \tag{1.38}$$

(d) Suppose that $d = 2$ and $R \sim 1$ mm: what would $M_{{\rm P},3+d}$ be, in TeV? Suggest ways in which this theory might be tested experimentally. Taking $M_{{\rm P},3+d} \sim 1$ TeV, explore other possibilities for d and R.


Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

2

Electromagnetism as a Gauge Theory

2.1 Introduction

The previous chapter introduced the basic ideas of the Standard Model of particle physics, in which quarks and leptons interact via the exchange of gauge field quanta. We must now look more closely into what is the main concern of this book – namely, the particular nature of these '*gauge theories*'.

One of the relevant forces – electromagnetism – has been well understood in its classical guise for many years. Over a century ago, Faraday, Maxwell and others developed the theory of electromagnetic interactions, culminating in Maxwell's paper of 1864 (Maxwell 1864). Today Maxwell's theory still stands – unlike Newton's 'classical mechanics' which was shown by Einstein to require modifications at relativistic speeds, approaching the speed of light. Moreover, Maxwell's electromagnetism, when suitably married with quantum mechanics, gives us '*quantum electrodynamics*' or QED. We shall see in chapter 10 that this theory is in truly remarkable agreement with experiment. As we have already indicated, the theories of the weak and strong forces included in the Standard Model are generalizations of QED, and promise to be as successful as that theory. The simplest of the three, QED, is therefore our paradigmatic theory.

From today's perspective, the crucial thing about electromagnetism is that it is a theory in which the *dynamics* (i.e. the behaviour of the forces) is intimately related to a *symmetry* principle. In the everyday world, a symmetry operation is something that can be done to an object that leaves the object looking the same after the operation as before. By extension, we may consider mathematical operations – or 'transformations' – applied to the objects in our theory such that the physical laws look the same after the operations as they did before. Such transformations are usually called *invariances* of the laws. Familiar examples are, for instance, the translation and rotation invariance of all fundamental laws: Newton's laws of motion remain valid whether or not we translate or rotate a system of interacting particles. But of course – precisely because they do apply to all laws, classical or quantum – these two invariances have no special connection with any particular force law. Instead,

they constrain the form of the allowed laws to a considerable extent, but by no means uniquely determine them. Nevertheless, this line of argument leads one to speculate whether it might in fact be possible to impose further types of symmetry constraints so that the forms of the force laws *are* essentially determined. This would then be one possible answer to the question: why are the force laws the way they are? (Ultimately of course this only replaces one question by another!)

In this chapter we shall discuss electromagnetism from this point of view. This is not the historical route to the theory, but it is the one which generalizes to the other two interactions. This is why we believe it important to present the central ideas of this approach in the familiar context of electromagnetism at this early stage.

A distinction that is vital to the understanding of all these interactions is that between a *global* invariance and a *local* invariance. In a global invariance the same transformation is carried out at all space–time points: it has an 'everywhere simultaneously' character. In a local invariance different transformations are carried out at different individual space–time points. In general, as we shall see, a theory that is globally invariant will not be invariant under locally varying transformations. However, by introducing new force fields that interact with the original particles in the theory in a specific way, and which also transform in a particular way under the local transformations, a sort of local invariance can be restored. We will see all these things more clearly when we go into more detail, but the important conceptual point to be grasped is this: one may view these special force fields and their interactions as existing in order to permit certain local invariances to be true. The particular local invariance relevant to electromagnetism is the well-known *gauge invariance* of Maxwell's equations: in the quantum form of the theory this property is directly related to an invariance under *local phase transformations* of the quantum fields. A generalized form of this phase invariance also underlies the theories of the weak and strong interactions. For this reason they are all known as 'gauge theories'.

A full understanding of gauge invariance in electrodynamics can only be reached via the formalism of quantum field theory, which is not easy to master – and the theory of quantum gauge fields is particularly tricky, as we shall see in chapter 7. Nevertheless, many of the crucial ideas can be perfectly adequately discussed within the more familiar framework of ordinary quantum mechanics, rather than quantum field theory, treating electromagnetism as a purely classical field. This is the programme followed in the rest of part I of this volume. In the present chapter we shall discuss these ideas in the context of non-relativistic quantum mechanics; in the following two chapters, we shall explore the generalization to relativistic quantum mechanics, for particles of spin-0 (via the Klein–Gordon equation) and spin-$\frac{1}{2}$ (via the Dirac equation). While containing substantial physics in their own right, these chapters constitute essential groundwork for the quantum field treatment in parts II–IV.

2.2 The Maxwell equations: current conservation

> Question: Would you distinguish local conservation laws from global conservation laws.
> Feynman: If a cat were to disappear in Pasadena and at the same time appear in Erice, that would be an example of global conservation of cats. This is not the way cats are conserved. Cats or charge or baryons are conserved in a much more continuous way. If any of these quantities begin to disappear in a region, then they begin to appear in a neighbouring region. Consequently, we can identify the flow of charge out of a region with the disappearance of charge inside the region. This identification of the divergence of a flux with the time rate of change of a charge density is called a local conservation law. A local conservation law implies that the total charge is conserved globally, but the reverse does not hold. However, relativistically it is clear that non-local global conservation laws cannot exist, since to a moving observer the cat will appear in Erice before it disappears in Pasadena.
>
> —From the question-and-answer session following a lecture by R. P. Feynman at the 1964 International School of Physics 'Ettore Majorana' (Feynman 1965b).

We begin by considering the basic laws of classical electromagnetism, the Maxwell equations. We use a system of units (Heaviside–Lorentz) which is convenient in particle physics (see appendix C). Before Maxwell's work these laws were

$$\boldsymbol{\nabla} \cdot \boldsymbol{E} = \rho_{\mathrm{em}} \qquad \text{(Gauss' law)} \tag{2.1}$$

$$\boldsymbol{\nabla} \times \boldsymbol{E} = -\frac{\partial \boldsymbol{B}}{\partial t} \qquad \text{(Faraday–Lenz laws)} \tag{2.2}$$

$$\boldsymbol{\nabla} \cdot \boldsymbol{B} = 0 \qquad \text{(no magnetic charges)} \tag{2.3}$$

and, for steady currents,

$$\boldsymbol{\nabla} \times \boldsymbol{B} = \boldsymbol{j}_{\mathrm{em}} \qquad \text{(Ampère's law).} \tag{2.4}$$

Here ρ_{em} is the charge density and $\boldsymbol{j}_{\mathrm{em}}$ is the current density; these densities act as 'sources' for the $\boldsymbol{E}$ and $\boldsymbol{B}$ fields. Maxwell noticed that taking the divergence of this last equation leads to conflict with the continuity equation for electric charge

$$\frac{\partial \rho_{\mathrm{em}}}{\partial t} + \boldsymbol{\nabla} \cdot \boldsymbol{j}_{\mathrm{em}} = 0. \tag{2.5}$$

Since

$$\boldsymbol{\nabla} \cdot (\boldsymbol{\nabla} \times \boldsymbol{B}) = 0 \tag{2.6}$$

from (2.4) there follows the result

$$\boldsymbol{\nabla} \cdot \boldsymbol{j}_{\mathrm{em}} = 0. \tag{2.7}$$

This can only be true in situations where the charge density is constant in time. For the general case, Maxwell modified Ampère's law to read

$$\boldsymbol{\nabla} \times \boldsymbol{B} = \boldsymbol{j}_{\mathrm{em}} + \frac{\partial \boldsymbol{E}}{\partial t} \tag{2.8}$$

which is now consistent with (2.5). Equations (2.1)–(2.3), together with (2.8), constitute Maxwell's equations in free space (apart from the sources).

It is worth spending a moment on the vitally important continuity equation (2.5) – note the Feynman quotation at the start of this section. Let us integrate this equation over any arbitrary volume Ω, and write the result as

$$\frac{\partial}{\partial t} \int_{\Omega} \rho_{\mathrm{em}} \mathrm{d}V = - \int_{\Omega} \boldsymbol{\nabla} \cdot \boldsymbol{j}_{\mathrm{em}} \mathrm{d}V. \tag{2.9}$$

Equation (2.9) states that the rate of decrease of charge in any arbitrary volume Ω is due precisely and only to the flux of current out of its surface; that is, no net charge can be created or destroyed in Ω. Since Ω can be made as small as we please, this means that *electric charge must be locally conserved*: a process in which charge is created at one point and destroyed at a distant one is not allowed, despite the fact that it conserves the charge overall or 'globally'. The ultimate reason for this is that the global form of charge conservation would necessitate the instantaneous propagation of signals (such as 'now, create a positron over there'), and this conflicts with special relativity – a theory which, historically, flowered from the soil of electrodynamics. The extra term introduced by Maxwell – the 'electric displacement current' – owes its place in the dynamical equations to a local conservation requirement.

We remark at this point that we have just introduced another local/global distinction, similar to that discussed earlier in connection with invariances. In this case the distinction applies to a conservation law, but since invariances are related to conservation laws in both classical and quantum mechanics, we should perhaps not be too surprised by this. However, as with invariances, conservation laws – such as charge conservation in electromagnetism – play a central role in gauge theories in that they are closely related to the dynamics. The point is simply illustrated by asking how we could measure the charge of a newly created subatomic particle X. There are two conceptually different ways:

(i) We could arrange for X to be created in a reaction such as

$$\mathrm{A} + \mathrm{B} \rightarrow \mathrm{C} + \mathrm{D} + \mathrm{X}$$

where the charges of A, B, C and D are already known. In this case we can use *charge conservation* to determine the charge of X.

(ii) We could see how particle X responded to known electromagnetic fields. This uses *dynamics* to determine the charge of X.

Either way gives the same answer: it is the conserved charge which determines the particle's response to the field. By contrast, there are several other conservation laws that seem to hold in particle physics, such as lepton number and baryon number, that apparently have no dynamical counterpart (cf the remarks at the end of section 1.3.6). To determine the baryon number of a newly produced particle, we have to use B conservation and tot up the total baryon number on either side of the reaction. As far as we know there is no baryonic force field.

Thus gauge theories are characterized by a close interrelation between *three* conceptual elements: symmetries, conservation laws and dynamics. In fact, it is now widely believed that the *only* exact quantum number conservation laws are those which have an associated gauge theory force field – see comment (i) in section 2.6. Thus one might suspect that baryon number is not absolutely conserved – as is indeed the case in proposed unified gauge theories of the strong, weak and electromagnetic interactions. In this discussion we have briefly touched on the connection between two pairs of these three elements: symmetries $\leftrightarrow$ dynamics; and conservation laws $\leftrightarrow$ dynamics. The precise way in which the remaining link is made – between the symmetry of electromagnetic gauge invariance and the conservation law of charge – is more technical. We will discuss this connection with the help of simple ideas from quantum field theory in chapter 7, section 7.4. For the present we continue with our study of the Maxwell equations and, in particular, of the gauge invariance they exhibit.

2.3 The Maxwell equations: Lorentz covariance and gauge invariance

In classical electromagnetism, and especially in quantum mechanics, it is convenient to introduce the vector potential $A_\mu(x)$ in place of the fields $\boldsymbol{E}$ and $\boldsymbol{B}$. We write:

$$\boldsymbol{B} = \boldsymbol{\nabla} \times \boldsymbol{A} \tag{2.10}$$

$$\boldsymbol{E} = -\boldsymbol{\nabla} V - \frac{\partial \boldsymbol{A}}{\partial t} \tag{2.11}$$

which defines the 3-vector potential $\boldsymbol{A}$ and the scalar potential V. With these definitions, equations (2.2) and (2.3) are then automatically satisfied.

The origin of gauge invariance in classical electromagnetism lies in the fact that the potentials $\boldsymbol{A}$ and V are not unique for given physical fields $\boldsymbol{E}$ and $\boldsymbol{B}$. The transformations that $\boldsymbol{A}$ and V may undergo while preserving $\boldsymbol{E}$ and $\boldsymbol{B}$ (and hence the Maxwell equations) unchanged are called gauge transformations, and the associated invariance of the Maxwell equations is called gauge invariance.

What are these transformations? Clearly $\boldsymbol{A}$ can be changed by

$$\boldsymbol{A} \to \boldsymbol{A}' = \boldsymbol{A} + \boldsymbol{\nabla}\chi \tag{2.12}$$

where χ is an arbitrary function, with no change in $\boldsymbol{B}$ since $\boldsymbol{\nabla} \times \boldsymbol{\nabla} f = 0$, for any scalar function f. To preserve $\boldsymbol{E}$, V must then change simultaneously by

$$V \to V' = V - \frac{\partial \chi}{\partial t}. \tag{2.13}$$

These transformations can be combined into a single compact equation by introducing the 4-vector potential[1]:

$$A^\mu = (V, \boldsymbol{A}) \tag{2.14}$$

and noting (from problem 2.1) that the differential operators $(\partial/\partial t, -\boldsymbol{\nabla})$ form the components of a 4-vector operator ∂^μ. A gauge transformation is then specified by

$$\boxed{A^\mu \to A'^\mu = A^\mu - \partial^\mu \chi.} \tag{2.15}$$

The Maxwell equations can also be written in a manifestly *Lorentz covariant form* (see appendix D) using the 4-current $j^\mu_{\rm em}$ given by

$$j^\mu_{\rm em} = (\rho_{\rm em}, \boldsymbol{j}_{\rm em}) \tag{2.16}$$

in terms of which the continuity equation takes the form (problem 2.1):

$$\partial_\mu j^\mu_{\rm em} = 0. \tag{2.17}$$

The Maxwell equations (2.1) and (2.8) then become (problem 2.2):

$$\partial_\mu F^{\mu\nu} = j^\nu_{\rm em} \tag{2.18}$$

where we have defined the field strength tensor:

$$F^{\mu\nu} \equiv \partial^\mu A^\nu - \partial^\nu A^\mu. \tag{2.19}$$

Under the gauge transformation

$$A^\mu \to A'^\mu = A^\mu - \partial^\mu \chi \tag{2.20}$$

$F^{\mu\nu}$ remains unchanged:

$$F^{\mu\nu} \to F'^{\mu\nu} = F^{\mu\nu} \tag{2.21}$$

so $F^{\mu\nu}$ is gauge invariant and so, therefore, are the Maxwell equations in

[1] See appendix D for relativistic notation and for an explanation of the very important concept of *covariance*, which we are about to invoke in the context of Lorentz transformations, and will use again in the next section in the context of gauge transformations; we shall also use it in other contexts in later chapters.

the form (2.18). The 'Lorentz-covariant and gauge-invariant field equations' satisfied by A^μ then follow from equations (2.18) and (2.19):

$$\Box A^\nu - \partial^\nu(\partial_\mu A^\mu) = j^\nu_{\rm em}. \tag{2.22}$$

Since gauge transformations turn out to be of central importance in the quantum theory of electromagnetism, it would be nice to have some insight into why Maxwell's equations are gauge invariant. The all-important 'fourth' equation (2.8) was inferred by Maxwell from local charge conservation, as expressed by the continuity equation

$$\partial_\mu j^\mu_{\rm em} = 0. \tag{2.23}$$

The field equation

$$\partial_\mu F^{\mu\nu} = j^\nu_{\rm em} \tag{2.24}$$

then of course automatically embodies (2.23). The mathematical reason it does so is that $F^{\mu\nu}$ is a four-dimensional kind of 'curl'

$$F^{\mu\nu} \equiv \partial^\mu A^\nu - \partial^\nu A^\mu \tag{2.25}$$

which (as we have seen in (2.21)) is unchanged by a gauge transformation

$$A^\mu \to A'^\mu = A^\mu - \partial^\mu \chi. \tag{2.26}$$

Hence there is the suggestion that the gauge invariance is related in some way to charge conservation. However, the connection is not so simple. Wigner (1949) has given a simple argument to show that the principle that no physical quantity can depend on the absolute value of the electrostatic potential, when combined with energy conservation, implies the conservation of charge. Wigner's argument relates charge (and energy) conservation to an invariance under transformation of the electrostatic potential by a constant: charge conservation alone does not seem to require the more general space–time-dependent transformation of gauge invariance.

Changing the value of the electrostatic potential by a constant amount is an example of what we have called a *global* transformation (since the change in the potential is the same everywhere). Invariance under this global transformation is related to a conservation law: that of charge. But this global invariance is not sufficient to generate the full Maxwellian dynamics. However, as remarked by 't Hooft (1980), one can regard equations (2.12) and (2.13) as expressing the fact that the *local* change in the electrostatic potential V (the $\partial\chi/\partial t$ term in (2.13)) can be compensated – in the sense of leaving the Maxwell equations unchanged – by a corresponding local change in the magnetic vector potential $\boldsymbol{A}$. Thus by including magnetic effects, the global

invariance under a change of V by a constant can be extended to a local invariance (which is a much more restrictive condition to satisfy). Hence there is a beginning of a suggestion that one might almost 'derive' the complete Maxwell equations, which unify electricity and magnetism, from the requirement that the theory be expressed in terms of potentials in such a way as to be invariant under local (gauge) transformations on those potentials. Certainly special relativity must play a role too: this also links electricity and magnetism, via the magnetic effects of charges as seen by an observer moving relative to them. If a 4-vector potential A^μ is postulated, and it is then demanded that the theory involve it only in a way which is insensitive to local changes of the form (2.15), one is led naturally to the idea that the physical fields enter only via the quantity $F^{\mu\nu}$, which is invariant under (2.15). From this, one might conjecture the field equation on grounds of Lorentz covariance.

It goes without saying that this is certainly not a 'proof' or 'derivation' of the Maxwell equations. Nevertheless, the idea that *dynamics* (in this case, the complete interconnection of electric and magnetic effects) may be intimately related to a local invariance requirement (in this case, electromagnetic gauge invariance) turns out to be a fruitful one. As indicated in section 2.1, it is generally the case that, when a certain global invariance is generalized to a local one, the existence of a new 'compensating' field is entailed, interacting in a specified way. The first example of dynamical theory 'derived' from a local invariance requirement seems to be the theory of Yang and Mills (1954) (see also Shaw 1955). Their work was extended by Utiyama (1956), who developed a general formalism for such compensating fields. As we have said, these types of dynamical theories, based on local invariance principles, are called gauge theories.

It is a remarkable fact that the interactions in the Standard Model of particle physics are of precisely this type. We have briefly discussed the Maxwell equations in this light, and we will continue with (quantum) electrodynamics in the following two sections. The two other fundamental interactions – the strong interaction between quarks and the weak interaction between quarks and leptons – also seem to be described by gauge theories (of essentially the Yang–Mills type), as we shall see in detail in the second volume of this book. A fourth example, but one which we shall not pursue in this book, is that of general relativity (the theory of gravitational interactions). Utiyama (1956) showed that this theory could be arrived at by generalizing the global (space–time independent) coordinate transformations of special relativity to local ones; as with electromagnetism, the more restrictive local invariance requirements entailed the existence of a new field – the gravitational one – with an (almost) prescribed form of interaction. Unfortunately, despite this 'gauge' property, no consistent quantum field theory of general relativity is known.

In order to proceed further, we must now discuss how such (gauge) ideas are incorporated into quantum mechanics.

2.4 Gauge invariance (and covariance) in quantum mechanics

The Lorentz force law for a non-relativistic particle of charge q moving with velocity $\boldsymbol{v}$ under the influence of both electric and magnetic fields is

$$\boldsymbol{F} = q\boldsymbol{E} + q\boldsymbol{v} \times \boldsymbol{B}. \tag{2.27}$$

It may be derived, via Hamilton's equations, from the classical Hamiltonian[2]

$$H = \frac{1}{2m}(\boldsymbol{p} - q\boldsymbol{A})^2 + qV. \tag{2.28}$$

The Schrödinger equation for such a particle in an electromagnetic field is

$$\left(\frac{1}{2m}(-\mathrm{i}\boldsymbol{\nabla} - q\boldsymbol{A})^2 + qV\right)\psi(\boldsymbol{x}, t) = \mathrm{i}\frac{\partial\psi(\boldsymbol{x}, t)}{\partial t} \tag{2.29}$$

which is obtained from the classical Hamiltonian by the usual prescription, $\boldsymbol{p} \to -\mathrm{i}\boldsymbol{\nabla}$, for Schrödinger's wave mechanics ($\hbar = 1$). Note the appearance of the operator combinations

$$\boxed{\begin{aligned} &\boldsymbol{D} \equiv \boldsymbol{\nabla} - \mathrm{i}q\boldsymbol{A} \\ &D^0 \equiv \partial/\partial t + \mathrm{i}qV \end{aligned}} \tag{2.30}$$

in place of $\boldsymbol{\nabla}$ and $\partial/\partial t$, in going from the free-particle Schrödinger equation to the electromagnetic field case.

The solution $\psi(\boldsymbol{x}, t)$ of the Schrödinger equation (2.29) describes completely the state of the particle moving under the influence of the potentials V, $\boldsymbol{A}$. However, these potentials are not unique, as we have already seen: they can be changed by a gauge transformation

$$\boldsymbol{A} \to \boldsymbol{A}' = \boldsymbol{A} + \boldsymbol{\nabla}\chi \tag{2.31}$$

$$V \to V' = V - \partial\chi/\partial t \tag{2.32}$$

and the Maxwell equations for the fields $\boldsymbol{E}$ and $\boldsymbol{B}$ will remain the same. This immediately raises a serious question: if we carry out such a change of potentials in equation (2.29), will the solution of the resulting equation describe the same physics as the solution of equation (2.29)? If it does, we shall be able to assume the validity of Maxwell's theory for the quantum world; if not, some modification will be necessary, since the gauge symmetry possessed by the Maxwell equations will be violated in the quantum theory.

[2] We set $\hbar = c = 1$ throughout (see appendix B).

The answer to the question just posed is evidently negative, since it is clear that the same 'ψ' cannot possibly satisfy both (2.29) and the analogous equation with $(V, \boldsymbol{A})$ replaced by $(V', \boldsymbol{A}')$. Unlike Maxwell's equations, the Schrödinger equation is not gauge invariant. But we must remember that the wavefunction ψ is not a directly observable quantity, as the electromagnetic fields $\boldsymbol{E}$ and $\boldsymbol{B}$ are. Perhaps ψ does not need to remain unchanged (invariant) when the potentials are changed by a gauge transformation. In fact, in order to have any chance of 'describing the same physics' in terms of the gauge-transformed potentials, *we will have to allow ψ to change as well.* This is a crucial point: for quantum mechanics to be consistent with Maxwell's equations it is necessary for the gauge transformations (2.31) and (2.32) of the Maxwell potentials to be accompanied also by a transformation of the quantum-mechanical wavefunction, $\psi \to \psi'$, where ψ' satisfies the equation

$$\left(\frac{1}{2m}(-\mathrm{i}\boldsymbol{\nabla} - q\boldsymbol{A}')^2 + qV'\right)\psi'(\boldsymbol{x}, t) = \mathrm{i}\frac{\partial\psi'(\boldsymbol{x}, t)}{\partial t}. \tag{2.33}$$

Note that the *form* of (2.33) is exactly the same as the form of (2.29) – it is this that will effectively ensure that both 'describe the same physics'. Readers of appendix D will expect to be told that – if we can find such a ψ' – we may then assert that (2.29) is *gauge covariant*, meaning that it maintains the same form under a gauge transformation. (The transformations relevant to this use of 'covariance' are gauge transformations.)

Since we know the relations (2.31) and (2.32) between $\boldsymbol{A}$, V and $\boldsymbol{A}'$, V', we can actually find what $\psi'(\boldsymbol{x}, t)$ must be in order that equation (2.33) be consistent with (2.29). We shall state the answer and then verify it; then we shall discuss the physical interpretation. The required $\psi'(\boldsymbol{x}, t)$ is

$$\psi'(\boldsymbol{x}, t) = \exp[\mathrm{i}q\chi(\boldsymbol{x}, t)]\psi(\boldsymbol{x}, t) \tag{2.34}$$

where χ is the same space–time-dependent function as appears in equations (2.31) and (2.32). To verify this we consider

$$\begin{aligned}(-\mathrm{i}\boldsymbol{\nabla} - q\boldsymbol{A}')\psi' &= [-\mathrm{i}\boldsymbol{\nabla} - q\boldsymbol{A} - q(\boldsymbol{\nabla}\chi)][\exp(\mathrm{i}q\chi)\psi] \\ &= q(\boldsymbol{\nabla}\chi)\exp(\mathrm{i}q\chi)\psi + \exp(\mathrm{i}q\chi)\cdot(-\mathrm{i}\boldsymbol{\nabla}\psi) \\ &\quad + \exp(\mathrm{i}q\chi)\cdot(-q\boldsymbol{A}\psi) - q(\boldsymbol{\nabla}\chi)\exp(\mathrm{i}q\chi)\psi. \end{aligned} \tag{2.35}$$

The first and the last terms cancel leaving the result:

$$(-\mathrm{i}\boldsymbol{\nabla} - q\boldsymbol{A}')\psi' = \exp(\mathrm{i}q\chi)\cdot(-\mathrm{i}\boldsymbol{\nabla} - q\boldsymbol{A})\psi \tag{2.36}$$

which may be written using equation (2.30) as:

$$(-\mathrm{i}\boldsymbol{D}'\psi') = \exp(\mathrm{i}q\chi)\cdot(-\mathrm{i}\boldsymbol{D}\psi). \tag{2.37}$$

Thus, although the space–time-dependent phase factor feels the action of the gradient operator $\boldsymbol{\nabla}$, it 'passes through' the combined operator $\boldsymbol{D}'$ and converts it into $\boldsymbol{D}$: in fact comparing the equations (2.34) and (2.37), we see that

$\boldsymbol{D}'\psi'$ bears to $\boldsymbol{D}\psi$ exactly the same relation as ψ' bears to ψ. In just the same way we find (cf equation (2.30))

$$(\mathrm{i}D^{0\prime}\psi') = \exp(\mathrm{i}q\chi)\cdot(\mathrm{i}D^0\psi) \tag{2.38}$$

where we have used equation (2.32) for V'. Once again, $D^{0\prime}\psi'$ is simply related to $D^0\psi$. Repeating the operation which led to equation (2.37) we find

$$\begin{aligned}\frac{1}{2m}(-\mathrm{i}\boldsymbol{D}')^2\psi' &= \exp(\mathrm{i}q\chi)\cdot\frac{1}{2m}(-\mathrm{i}\boldsymbol{D})^2\psi \\ &= \exp(\mathrm{i}q\chi)\cdot\mathrm{i}D^0\psi \qquad \text{(using equation (2.29))} \\ &= \mathrm{i}D^{0\prime}\psi' \qquad \text{(using equation (2.30))}.\end{aligned} \tag{2.39}$$

Equation (2.39) is just (2.33) written in the D notation of equation (2.30), so we have verified that (2.34) is the correct relationship between ψ' and ψ to ensure consistency between equations (2.29) and (2.33). Precisely this consistency is summarized by the statement that (2.29) is gauge covariant.

Do ψ and ψ' describe the same physics, in fact? The answer is yes, but it is not quite trivial. It is certainly obvious that the probability densities $|\psi|^2$ and $|\psi'|^2$ are equal, since in fact ψ and ψ' in equation (2.34) are related by a *phase* transformation. However, we can be interested in other observables involving the derivative operators $\boldsymbol{\nabla}$ or $\partial/\partial t$ – for example, the current, which is essentially $\psi^*(\boldsymbol{\nabla}\psi) - (\boldsymbol{\nabla}\psi)^*\psi$. It is easy to check that this current is *not* invariant under (2.34), because the phase $\chi(\boldsymbol{x},t)$ is $\boldsymbol{x}$-dependent. But equations (2.37) and (2.38) show us what we must do to construct *gauge-invariant currents*: namely, we must replace $\boldsymbol{\nabla}$ by $\boldsymbol{D}$ (and in general also $\partial/\partial t$ by D^0) since then:

$$\psi^{*\prime}(\boldsymbol{D}'\psi') = \psi^*\exp(-\mathrm{i}q\chi)\cdot\exp(\mathrm{i}q\chi)\cdot(\boldsymbol{D}\psi) = \psi^*\boldsymbol{D}\psi \tag{2.40}$$

for example. Thus the identity of the physics described by ψ and ψ' is indeed ensured. Note, incidentally, that the *equality* between the first and last terms in (2.40) is indeed a statement of *(gauge) invariance*.

We summarize these important considerations by the statement that the gauge invariance of Maxwell equations re-emerges as a covariance in quantum mechanics provided we make the combined transformation

$$\boxed{\begin{aligned}&\boldsymbol{A}\to\boldsymbol{A}' = \boldsymbol{A} + \boldsymbol{\nabla}\chi \\ &V\to V' = V - \partial\chi/\partial t \\ &\psi\to\psi' = \exp(\mathrm{i}q\chi)\psi\end{aligned}} \tag{2.41}$$

on the potential and on the wavefunction.

The Schrödinger equation is non-relativistic, but the Maxwell equations are of course fully relativistic. One might therefore suspect that the prescriptions discovered here are actually true relativistically as well, and this is indeed

the case. We shall introduce the spin-0 and spin-$\frac{1}{2}$ relativistic equations in chapter 3. For the present we note that (2.30) can be written in manifestly Lorentz covariant form as

$$\boxed{D^\mu \equiv \partial^\mu + \mathrm{i}qA^\mu} \tag{2.42}$$

in terms of which (2.37) and (2.38) become

$$-\mathrm{i}D'^\mu\psi' = \exp(\mathrm{i}q\chi)\cdot(-\mathrm{i}D^\mu\psi). \tag{2.43}$$

It follows that any equation involving the operator ∂^μ can be made gauge invariant under the combined transformation

$$\begin{aligned} A^\mu &\rightarrow A'^\mu = A^\mu - \partial^\mu\chi \\ \psi &\rightarrow \psi' = \exp(\mathrm{i}q\chi)\psi \end{aligned}$$

if ∂^μ is replaced by D^μ. In fact, we seem to have a very simple prescription for obtaining the wave equation for a particle in the presence of an electromagnetic field from the corresponding *free particle* wave equation: make the replacement

$$\boxed{\partial^\mu \rightarrow D^\mu \equiv \partial^\mu + \mathrm{i}qA^\mu.} \tag{2.44}$$

In the following section this will be seen to be the basis of the so-called 'gauge principle' whereby, in accordance with the idea advanced in the previous sections, the form of the *interaction* is determined by the insistence on (local) gauge invariance.

One final remark: this new kind of derivative

$$D^\mu \equiv \partial^\mu + \mathrm{i}qA^\mu \tag{2.45}$$

turns out to be of fundamental importance – it will be the operator which generalizes from the (Abelian) phase symmetry of QED (see comment (iii) of section 2.6) to the (non-Abelian) phase symmetry of our weak and strong interaction theories. It is called the '*gauge covariant derivative*', the term being usually shortened to 'covariant derivative' in the present context. The geometrical significance of this term will be explained in volume 2.

2.5 The argument reversed: the gauge principle

In the preceding section, we took it as *known* that the Schrödinger equation, for example, for a charged particle in an electromagnetic field, has the form

$$\left[\frac{1}{2m}(-\mathrm{i}\boldsymbol{\nabla} - q\boldsymbol{A})^2 + qV\right]\psi = \mathrm{i}\partial\psi/\partial t. \tag{2.46}$$

We then checked its gauge invariance under the combined transformation

$$\begin{aligned} \boldsymbol{A} \to \boldsymbol{A}' &= \boldsymbol{A} + \boldsymbol{\nabla}\chi \\ V \to V' &= V - \partial\chi/\partial t \\ \psi \to \psi' &= \exp(\mathrm{i}q\chi)\psi. \end{aligned} \tag{2.47}$$

We now want to reverse the argument: we shall start by demanding that our theory is invariant under the *space–time-dependent phase transformation*

$$\psi(\boldsymbol{x}, t) \to \psi'(\boldsymbol{x}, t) = \exp[\mathrm{i}q\chi(\boldsymbol{x}, t)]\psi(\boldsymbol{x}, t). \tag{2.48}$$

We shall demonstrate that such a phase invariance is not possible for a free theory, but rather requires an *interacting* theory involving a (4-vector) field whose interactions with the charged particle are precisely determined, and which undergoes the transformation

$$\boldsymbol{A} \to \boldsymbol{A}' = \boldsymbol{A} + \boldsymbol{\nabla}\chi \tag{2.49}$$

$$V \to V' = V - \partial\chi/\partial t \tag{2.50}$$

when $\psi \to \psi'$. The demand of this type of phase invariance will have then dictated the form of the interaction – this is the basis of the *gauge principle*.

Before proceeding we note that the resulting equation – which will of course turn out to be (2.29) – will not strictly speaking be invariant under (2.48), but rather covariant (in the gauge sense), as we saw in the preceding section. Nevertheless, we shall in this section sometimes continue (slightly loosely) to speak of 'local phase invariance'. When we come to implement these ideas in quantum field theory in chapter 7 (section 7.4), using the Lagrangian formalism, we shall see that the relevant Lagrangians are indeed invariant under (2.48).

We therefore focus attention on the phase of the wavefunction. The absolute phase of a wavefunction in quantum mechanics cannot be measured; only relative phases are measurable, via some sort of interference experiment. A simple example is provided by the diffraction of particles by a two-slit system. Downstream from the slits, the wavefunction is a coherent superposition of two components, one originating from each slit: symbolically,

$$\psi = \psi_1 + \psi_2. \tag{2.51}$$

The probability distribution $|\psi|^2$ will then involve, in addition to the separate intensities $|\psi_1|^2$ and $|\psi_2|^2$, the *interference* term

$$2\ \mathrm{Re}(\psi_1^*\psi_2) = 2|\psi_1||\psi_2|\cos\delta \tag{2.52}$$

where δ $(= \delta_1 - \delta_2)$ is the *phase difference* between components ψ_1 and ψ_2. The familiar pattern of alternating intensity maxima and minima is then attributed to variation in the phase difference δ. Where the components are in phase, the interference is constructive and $|\psi|^2$ has a maximum; where they are out

of phase, it is destructive and $|\psi|^2$ has a minimum. It is clear that if the individual phases δ_1 and δ_2 are each shifted by the same amount, there will be no observable consequences, since only the phase difference δ enters.

The situation in which the wavefunction can be changed in a certain way without leading to any observable effects is precisely what is entailed by a symmetry or invariance principle in quantum mechanics. In the case under discussion, the invariance is that of a constant overall change in phase. In performing calculations it is necessary to make some definite choice of phase; that is, to adopt a 'phase convention'. The invariance principle guarantees that any such choice, or convention, is equivalent to any other.

Invariance under a constant change in phase is an example of a *global* invariance, according to the terminology introduced in the previous section. We make this point quite explicit by writing out the transformation as

$$\boxed{\begin{array}{c}\psi \to \psi' = \mathrm{e}^{\mathrm{i}\alpha}\psi \\ \alpha = \text{constant}\end{array}} \qquad \text{global phase invariance.} \tag{2.53}$$

That α in (2.53) is a constant, the same for all space–time points, expresses the fact that once a phase convention (choice of α) has been made at one space–time point, the same must be adopted at all other points. Thus in the two-slit experiment we are not free to make a *local* chance of phase: for example, as discussed by 't Hooft (1980), inserting a half-wave plate behind just one of the slits will certainly have observable consequences.

There is a sense in which this may seem an unnatural state of affairs. Once a phase convention has been adopted at one space–time point, the same convention must be adopted at all other ones: the half-wave plate must extend instantaneously across all of space, or not at all. Following this line of thought, one might then be led to 'explore the possibility' of requiring invariance under *local* phase transformations: that is, independent choices of phase convention at each space–time point. By itself, the foregoing is not a compelling motivation for such a step. However, as we pointed out in section 2.3, such a move from a global to a local invariance is apparently of crucial significance in classical electromagnetism and general relativity, and seems now to provide the key to an understanding of the other interactions in the Standard Model. Let us see, then, where the demand of '*local* phase invariance'

$$\boxed{\psi(\boldsymbol{x},t) \to \psi'(\boldsymbol{x},t) = \exp[\mathrm{i}\alpha(\boldsymbol{x},t)]\psi(\boldsymbol{x},t)} \qquad \text{local phase invariance} \tag{2.54}$$

leads us.

There is immediately a problem: this is *not* an invariance of the free-particle Schrödinger equation or of any free-particle relativistic wave equation! For example, if the original wavefunction $\psi(\boldsymbol{x},t)$ satisfied the free-particle Schrödinger equation

$$\frac{1}{2m}(-\mathrm{i}\boldsymbol{\nabla}^2)\psi(\boldsymbol{x},t) = \mathrm{i}\partial\psi(\boldsymbol{x},t)/\partial t \tag{2.55}$$

then the wavefunction ψ', given by the local phase transformation, will not, since both $\boldsymbol{\nabla}$ and $\partial/\partial t$ now act on $\alpha(\boldsymbol{x}, t)$ in the phase factor. Thus local phase invariance is not an invariance of the free-particle wave equation. If we wish to satisfy the demands of local phase invariance, we are obliged to modify the free-particle Schrödinger equation into something for which there is a local phase invariance – or rather, more accurately, a corresponding covariance. But this modified equation will no longer describe a free particle: in other words, the freedom to alter the phase of a charged particle's wavefunction locally is only possible if some kind of force field is introduced in which the particle moves. In more physical terms, the covariance will now be manifested in the inability to distinguish observationally between the effect of making a local change in phase convention and the effect of some new field in which the particle moves.

What kind of field will this be? In fact, we know immediately what the answer is, since the local phase transformation

$$\psi \to \psi' = \exp[\mathrm{i}\alpha(\boldsymbol{x}, t)]\psi \tag{2.56}$$

with $\alpha = q\chi$ is just the phase transformation associated with electromagnetic gauge invariance! Thus we must modify the Schrödinger equation

$$\frac{1}{2m}(-\mathrm{i}\boldsymbol{\nabla})^2\psi = \mathrm{i}\partial/\partial t \tag{2.57}$$

to

$$\frac{1}{2m}(-\mathrm{i}\boldsymbol{\nabla} - q\boldsymbol{A})^2\psi = (\mathrm{i}\partial/\partial t - qV)\psi \tag{2.58}$$

and satisfy the local phase invariance

$$\psi \to \psi' = \exp[\mathrm{i}\alpha(\boldsymbol{x}, t)]\psi \tag{2.59}$$

by demanding that $\boldsymbol{A}$ and V transform by

$$\begin{aligned} \boldsymbol{A} &\to \boldsymbol{A}' = \boldsymbol{A} + q^{-1}\boldsymbol{\nabla}\alpha \\ V &\to V' = V - q^{-1}\partial\alpha/\partial t \end{aligned} \tag{2.60}$$

when $\psi \to \psi'$. The modified wave equation is of course precisely the Schrödinger equation describing the interaction of the charged particle with the electromagnetic field described by $\boldsymbol{A}$ and V.

In a Lorentz covariant treatment, $\boldsymbol{A}$ and V will be regarded as parts of a 4-vector A^μ, just as $-\boldsymbol{\nabla}$ and $\partial/\partial t$ are parts of ∂^μ (see problem 2.1). Thus the presence of the vector field A^μ, interacting in a 'universal' prescribed way with any particle of charge q, is dictated by local phase invariance. A vector field such as A^μ, introduced to guarantee local phase invariance, is called a 'gauge field'. The principle that the interaction should be so dictated by the phase (or gauge) invariance is called the *gauge principle*: it allows us to write down the wave equation for the interaction directly from the free particle equation

via the replacement (2.44)[3]. As before, the method clearly generalizes to the four-dimensional case.

2.6 Comments on the gauge principle in electromagnetism

Comment (i)

A properly sceptical reader may have detected an important sleight of hand in the previous discussion. Where exactly did the electromagnetic charge appear from? The trouble with our argument as so far presented is that we could have defined fields $\boldsymbol{A}$ and V so that they coupled equally to all particles – instead we smuggled in a factor q.

Actually we can do a bit better than this. We can use the fact that the electromagnetic charge is absolutely conserved to claim that there can be no quantum mechanical interference between states of different charge q. Hence different phase changes are allowed within each 'sector' of definite q:

$$\psi' = \exp(\mathrm{i}q\chi)\psi \tag{2.61}$$

let us say. When this becomes a local transformation, $\chi \to \chi(\boldsymbol{x}, t)$, we shall need to cancel a term $q\boldsymbol{\nabla}\chi$, which will imply the presence of a '$-q\boldsymbol{A}$' term, as required. Note that such an argument is only possible for an *absolutely* conserved quantum number q – otherwise we cannot split up the states of the system into non-communicating sectors specified by different values of q. Reversing this line of reasoning, a conservation law such as baryon number conservation, with no related gauge field, would therefore now be suspected of not being absolutely conserved.

We still have not tied down why q is the electromagnetic charge and not some other absolutely conserved quantum number. A proper discussion of the reasons for identifying A^μ with the electromagnetic potential and q with the particle's charge will be given in chapter 7 with the help of quantum field theory.

Comment (ii)

Accepting these identifications, we note that the form of the interaction contains but one parameter, the electromagnetic charge q of the particle in question. It is the *same* whatever the type of particle with charge q, whether it be lepton, hadron, nucleus, ion, atom, etc. Precisely this type of 'universality' is present in the weak couplings of quarks and leptons, as we shall see in volume 2. This strongly suggests that some form of gauge principle must be

[3] Actually the electromagnetic interaction is uniquely specified by this procedure only for particles of spin-0 or $\frac{1}{2}$. The spin-1 case will be discussed in volume 2.

at work in generating weak interactions as well. The associated symmetry or conservation law is, however, of a very subtle kind. Incidentally, although all particles of a given charge q interact electromagnetically in a universal way, there is nothing at all in the preceding argument to indicate why, in nature, the charges of observed particles are all integer multiples of one basic charge.

Comment (iii)

Returning to comment (i), we may wish that we had not had to introduce the absolute conservation of charge as a separate axiom. As remarked earlier, at the end of section 2.2, we should like to relate that conservation law to the symmetry involved, namely invariance under (2.54). It is worth looking at the nature of this symmetry in a little more detail. It is not a symmetry which – as in the case of translation and rotation invariances for instance – involves changes in the space–time coordinates $\boldsymbol{x}$ and t. Instead, it operates on the *real and imaginary parts of the wavefunction.* Let us write

$$\psi = \psi_{\mathrm{R}} + \mathrm{i}\psi_{\mathrm{I}}. \tag{2.62}$$

Then

$$\psi' = \mathrm{e}^{\mathrm{i}\alpha}\psi = \psi'_{\mathrm{R}} + \mathrm{i}\psi'_{\mathrm{I}} \tag{2.63}$$

can be written as

$$\begin{aligned} \psi'_{\mathrm{R}} &= (\cos\alpha)\psi_{\mathrm{R}} - (\sin\alpha)\psi_{\mathrm{I}} \\ \psi'_{\mathrm{I}} &= (\sin\alpha)\psi_{\mathrm{R}} + \cos\alpha)\psi_{\mathrm{I}} \end{aligned} \tag{2.64}$$

from which we can see that it is indeed a kind of 'rotation', but in the ψ_{R}–ψ_{I} plane, whose 'coordinates' are the real and imaginary parts of the wavefunction. We call this plane an *internal* space and the associated symmetry an *internal symmetry.* Thus our phase invariance can be looked upon as a kind of internal space rotational invariance.

We can imagine doing two successive such transformations

$$\psi \to \psi' \to \psi'' \tag{2.65}$$

where

$$\psi'' = \mathrm{e}^{\mathrm{i}\beta}\psi' \tag{2.66}$$

and so

$$\psi'' = \mathrm{e}^{\mathrm{i}(\alpha+\beta)}\psi = \mathrm{e}^{\mathrm{i}\delta}\psi \tag{2.67}$$

with $\delta = \alpha + \beta$. This is a transformation of the same form as the original one. The set of all such transformations forms what mathematicians call a *group*, in this case U(1), meaning the group of all unitary one-dimensional matrices. A unitary matrix $\mathbf{U}$ is one such that

$$\mathbf{U}\mathbf{U}^{\dagger} = \mathbf{U}^{\dagger}\mathbf{U} = \mathbf{1} \tag{2.68}$$

where $\mathbf{1}$ is the identity matrix and † denotes the Hermitian conjugate. A

one-dimensional matrix is of course a single number – in this case a complex number. Condition (2.68) limits this to being a simple phase: the set of phase factors of the form $e^{i\alpha}$, where α is any real number, form the elements of a U(1) group. These are just the factors that enter into our gauge (or phase) transformations for wavefunctions. Thus we say that the electromagnetic gauge group is U(1). We must remember, however, that it is a *local* U(1), meaning (cf (2.54)) that the phase parameters $\alpha, \beta, \ldots$ depend on the space–time point x.

The transformations of the U(1) group have the simple property that it does not matter in what order they are performed: referring to (2.65)–(2.67), we would have got the same final answer if we had done the β 'rotation' first and then the α one, instead of the other way around; this is because, of course,

$$\exp(i\alpha) \cdot \exp(i\beta) = \exp[i(\alpha + \beta)] = \exp(i\beta) \cdot \exp(i\alpha). \tag{2.69}$$

This property remains true even in the 'local' case when α and β depend on x. Mathematicians call U(1) an *Abelian* group: different transformations commute. We shall see later (in volume 2) that the 'internal' symmetry spaces relevant to the strong and weak gauge invariances are not so simple. The 'rotations' in these cases are more like full three-dimensional rotations of real space, rather than the two-dimensional rotation of (2.64). We know that, in general, such real-space rotations do *not* commute, and the same will be true of the strong and weak rotations. Their gauge groups are called *non-Abelian*.

Once again, we shall have to wait until chapter 7 before understanding how the symmetry represented by (2.63) is really related to the conservation law of charge.

Comment (iv)

The attentive reader may have picked up one further loose end. The vector potential $\boldsymbol{A}$ is related to the magnetic field $\boldsymbol{B}$ by

$$\boldsymbol{B} = \boldsymbol{\nabla} \times \boldsymbol{A}. \tag{2.70}$$

Thus if $\boldsymbol{A}$ has the special form

$$\boldsymbol{A} = \boldsymbol{\nabla} f \tag{2.71}$$

$\boldsymbol{B}$ will vanish. The question we must answer, therefore, is: how do we know that the $\boldsymbol{A}$ field introduced by our gauge principle is not of the form (2.71), leading to a trivial theory ($\boldsymbol{B} = \boldsymbol{0}$)? The answer to this question will lead us on a very worthwhile detour.

The Schrödinger equation with $\boldsymbol{\nabla} f$ as the vector potential is

$$\frac{1}{2m}(-i\boldsymbol{\nabla} - q\boldsymbol{\nabla} f)^2 \psi = E\psi. \tag{2.72}$$

We can write the formal solution to this equation as

$$\psi = \exp\left(iq \int_{-\infty}^{\boldsymbol{x}} \boldsymbol{\nabla} f \cdot d\boldsymbol{l}\right) \cdot \psi(f = 0) \tag{2.73}$$

which may be checked by using the fact that

$$\frac{\partial}{\partial a}\int^{a} f(t)\,\mathrm{d}t = f(a). \tag{2.74}$$

The notation $\psi(f=0)$ means just the free-particle solution with $f=0$; the line integral is taken along an arbitrary path ending in the point $\boldsymbol{x}$. But we have

$$\mathrm{d}f = \frac{\partial f}{\partial x}\,\mathrm{d}x + \frac{\partial f}{\partial y}\,\mathrm{d}y + \frac{\partial f}{\partial z}\,\mathrm{d}z \equiv \boldsymbol{\nabla} f\cdot\mathrm{d}\boldsymbol{l}. \tag{2.75}$$

Hence the integral can be done trivially and the solution becomes

$$\psi = \exp[\mathrm{i}q(f(\boldsymbol{x}) - f(-\infty))]\cdot\psi(f=0). \tag{2.76}$$

We say that the phase factor introduced by the (in reality, field-free) vector potential $\boldsymbol{A} = \boldsymbol{\nabla} f$ is *integrable*: the effect of this particular $\boldsymbol{A}$ is merely to multiply the free-particle solution by an $\boldsymbol{x}$-dependent phase (apart from a trivial constant phase). Since this $\boldsymbol{A}$ should give no real electromagnetic effect, we must hope that such a change in the wavefunction is also somehow harmless. Indeed Dirac showed (Dirac 1981, pp 92–3) that such a phase factor corresponds merely to a redefinition of the momentum operator $\hat{\boldsymbol{p}}$. The essential point is that (in one dimension, say) $\hat{p}$ is defined ultimately by the commutator ($\hbar = 1$)

$$[\hat{x}, \hat{p}] = \mathrm{i}. \tag{2.77}$$

Certainly the familiar choice

$$\hat{p} = -\mathrm{i}\frac{\partial}{\partial x} \tag{2.78}$$

satisfies this commutation relation. But we can also add any function of x to $\hat{p}$, and this modified $\hat{p}$ will be still satisfactory since x commutes with any function of x. More detailed considerations by Dirac showed that this arbitrary function must actually have the form $\partial F/\partial x$, where F is arbitrary. Thus

$$\hat{p}' = -\mathrm{i}\frac{\partial}{\partial x} + \frac{\partial F}{\partial x} \tag{2.79}$$

is an acceptable momentum operator. Consider then the quantum mechanics defined by the wavefunction $\psi(f=0)$ and the momentum operator $\hat{p} = -\mathrm{i}\partial/\partial x$. Under the unitary transformation (cf (2.76))

$$\psi(f=0) \to \mathrm{e}^{\mathrm{i}qf(x)}\psi(f=0) \tag{2.80}$$

$\hat{p}$ will be transformed to

$$\hat{p} \to \mathrm{e}^{\mathrm{i}qf(x)}\hat{p}\mathrm{e}^{-\mathrm{i}qf(x)}. \tag{2.81}$$

But the right-hand side of this equation is just $\hat{p} - q\partial f/\partial x$ (problem 2.3), which is an equally acceptable momentum operator, identifying qf with the F of Dirac. Thus the case $\boldsymbol{A} = \boldsymbol{\nabla} f$ is indeed equivalent to the field-free case.

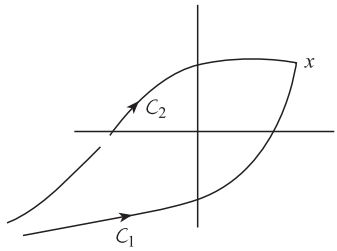


FIGURE 2.1
Two paths $\mathcal{C}_1$ and $\mathcal{C}_2$ (in two dimensions for simplicity) from $-\infty$ to the point $\boldsymbol{x}$.

What of the physically interesting case in which $\boldsymbol{A}$ is *not* of the form $\boldsymbol{\nabla} f$? The equation is now

$$\frac{1}{2m}(-\mathrm{i}\boldsymbol{\nabla} - q\boldsymbol{A})^2\psi = E\psi \tag{2.82}$$

to which the solution is

$$\psi = \exp\left(\mathrm{i}q\int_{-\infty}^{\boldsymbol{x}} \boldsymbol{A}\cdot\mathrm{d}\boldsymbol{l}\right)\cdot\psi(\boldsymbol{A}=0). \tag{2.83}$$

The line integral can now not be done so trivially: one says that the $\boldsymbol{A}$-field has produced a *non-integrable phase factor*. There is more to this terminology than the mere question of whether the integral is easy to do. The crucial point is that the integral now depends on the *path followed* in reaching the point $\boldsymbol{x}$, whereas the integrable phase factor in (2.73) depends only on the end-points of the integral, not on the path joining them.

Consider two paths $\mathcal{C}_1$ and $\mathcal{C}_2$ (figure 2.1) from $-\infty$ to the point $\boldsymbol{x}$. The difference in the two line integrals is the integral over a *closed* curve $\mathcal{C}$, which can be evaluated by Stokes' theorem:

$$\int_{\mathcal{C}_1}^{\boldsymbol{x}} \boldsymbol{A}\cdot\mathrm{d}\boldsymbol{l} - \int_{\mathcal{C}_2}^{\boldsymbol{x}} \boldsymbol{A}\cdot\mathrm{d}\boldsymbol{l} = \oint_{\mathcal{C}} \boldsymbol{A}\cdot\mathrm{d}\boldsymbol{l} = \iint_S \boldsymbol{\nabla}\times\boldsymbol{A}\cdot\mathrm{d}\boldsymbol{S} = \iint_S \boldsymbol{B}\cdot\mathrm{d}\boldsymbol{S} \tag{2.84}$$

where S is any surface spanning the curve $\mathcal{C}$. In this form we see that if $\boldsymbol{A} = \boldsymbol{\nabla} f$, then indeed the line integrals over $\mathcal{C}_1$ and $\mathcal{C}_2$ are equal since $\boldsymbol{\nabla}\times\boldsymbol{\nabla} f = 0$, but if $\boldsymbol{B} = \boldsymbol{\nabla}\times\boldsymbol{A}$ is not zero, the difference between the integrals is determined by the enclosed flux of $\boldsymbol{B}$.

This analysis turns out to imply the existence of a remarkable phenomenon – the Aharonov–Bohm effect, named after its discoverers (Aharonov and Bohm 1959). Suppose we go back to our two-slit experiment of section 2.5, only this time we imagine that a long thin solenoid is inserted between the slits, so that the components ψ_1 and ψ_2 of the split beam pass one on each side of the solenoid (figure 2.2). After passing round the solenoid, the beams are

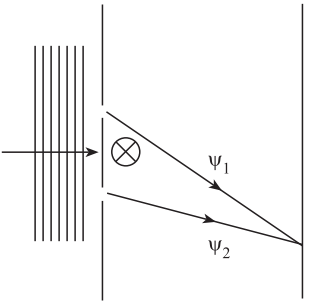


FIGURE 2.2
The Aharonov–Bohm effect.

recombined, and the resulting interference pattern is observed downstream. At any point $\boldsymbol{x}$ of the pattern, the phase of the ψ_1 and ψ_2 components will be modified – relative to the $\boldsymbol{B} = \boldsymbol{0}$ case – by factors of the form (2.83). These factors depend on the respective paths, which are different for the two components ψ_1 and ψ_2. The phase difference between these components, which determines the interference pattern, will therefore involve the $\boldsymbol{B}$-dependent factor (2.84). Thus, even though the field $\boldsymbol{B}$ is essentially totally contained within the solenoid, and the beams themselves have passed through $\boldsymbol{B} = \boldsymbol{0}$ regions only, there is nevertheless an observable effect on the pattern provided $\boldsymbol{B} \neq \boldsymbol{0}$! This effect – a shift in the pattern as $\boldsymbol{B}$ varies – was first confirmed experimentally by Chambers (1960), soon after its prediction by Aharonov and Bohm. It was anticipated in work by Ehrenburg and Siday (1949); further references and discussion are contained in Berry (1984).

Comment (v)

In conclusion, we must emphasize that there is ultimately no compelling logic for the vital leap to a local phase invariance from a global one. The latter is, by itself, both necessary and sufficient in quantum field theory to guarantee local charge conservation. Nevertheless, the gauge principle – deriving interactions from the requirement of local phase invariance – provides a satisfying conceptual unification of the interactions present in the Standard Model. In volume 2 of this book we shall consider generalizations of the electromagnetic gauge principle. It will be important always to bear in mind that any attempt to base theories of non-electromagnetic interactions on some kind of gauge principle can only make sense if there is an exact symmetry involved. The reason for this will only become clear when we consider the *renormalizability* of QED in chapter 11.

Problems

2.1

(a) A Lorentz transformation in the x^1 direction is given by

$$\begin{aligned} t' &= \gamma(t - vx^1) \\ x^{1\prime} &= \gamma(-vt + x^1) \\ x^{2\prime} &= x^2, \qquad x^{3\prime} = x^3 \end{aligned}$$

where $\gamma = (1 - v^2)^{-1/2}$ and $c = 1$. Write down the inverse of this transformation (i.e. express (t, x^1) in terms of $(t', x^{1\prime})$), and use the 'chain rule' of partial differentiation to show that, under the Lorentz transformation, the two quantities $(\partial/\partial t, -\partial/\partial x^1)$ transform in the same way as (t, x^1).

[The general result is that the four-component quantity $(\partial/\partial t, -\partial/\partial x^1, -\partial/\partial x^2, -\partial/\partial x^3) \equiv (\partial/\partial t, -\boldsymbol{\nabla})$ transforms in the same way as (t, x^1, x^2, x^3). Four-component quantities transforming this way are said to be 'contravariant 4-vectors', and are written with an upper 4-vector index; thus $(\partial/\partial t, -\boldsymbol{\nabla}) \equiv \partial^\mu$. Upper indices can be lowered by using the metric tensor $g_{\mu\nu}$, see appendix D, which reverses the sign of the spatial components. Thus $\partial^\mu = (\partial/\partial t, \partial/\partial x_1, \partial/\partial x_2, \partial/\partial x_3)$. Similarly the four quantities $(\partial/\partial t, \boldsymbol{\nabla}) = (\partial/\partial t, \partial/\partial x^1, \partial/\partial x^2, \partial/\partial x^3)$ transform as $(t, -x^1, -x^2, -x^3)$ and are a 'covariant 4-vector', denoted by ∂_μ.]

(b) Check that equation (2.5) can be written as (2.17).

2.2 How many independent components does the field strength $F^{\mu\nu}$ have? Express each component in terms of electric and magnetic field components. Hence verify that equation (2.18) correctly reproduces both equations (2.1) and (2.8).

2.3 Verify the result

$$\mathrm{e}^{\mathrm{i}qf(x)}\hat{p}\mathrm{e}^{-\mathrm{i}qf(x)} = \hat{p} - q\frac{\partial f}{\partial x}.$$

3

Relativistic Quantum Mechanics

It is clear that the non-relativistic Schrödinger equation is quite inadequate to analyse the results of experiments at energies far higher than the rest mass energies of the particles involved. Besides, the quarks and leptons have spin-$\frac{1}{2}$, a degree of freedom absent from the Schrödinger wavefunction. We therefore need two generalizations – from non-relativistic to relativistic for spin-0 particles, and from spin-0 to spin-$\frac{1}{2}$. The first step is to the Klein–Gordon equation (section 3.1), the second to the Dirac equation (section 3.2). Then after some further work on solutions of the Dirac equation (sections 3.3–3.4), we shall consider (section 3.5) some simple consequences of including the electromagnetic interaction via the gauge principle replacement (2.44).

3.1 The Klein–Gordon equation

The non-relativistic Schrödinger equation may be put into correspondence with the non-relativistic energy–momentum relation

$$E = \boldsymbol{p}^2/2m \tag{3.1}$$

by means of the operator replacements[1]

$$E \rightarrow \mathrm{i}\partial/\partial t \tag{3.2}$$

$$\boldsymbol{p} \rightarrow -\mathrm{i}\boldsymbol{\nabla}, \tag{3.3}$$

these differential operators being understood to act on the Schrödinger wavefunction.

For a relativistic wave equation we must start with the correct relativistic energy–momentum relation. Energy and momentum appear as the 'time' and 'space' components of the momentum 4-vector

$$p^\mu = (E, \boldsymbol{p}) \tag{3.4}$$

which satisfy the mass-shell condition

$$p^2 = p_\mu p^\mu = E^2 - \boldsymbol{p}^2 = m^2. \tag{3.5}$$

[1] Recall $\hbar = c = 1$ throughout (see appendix B).

Since energy and momentum are merely different components of a 4-vector, an attempt to base a relativistic theory on the relation

$$E = +(\boldsymbol{p}^2 + m^2)^{1/2} \tag{3.6}$$

is unattractive, as well as having obvious difficulties in interpretation for the square root operator. Schrödinger, before settling for the less ambitious non-relativistic Schrödinger equation, and later Klein and Gordon, attempted to build relativistic quantum mechanics (RQM) from the squared relation

$$E^2 = \boldsymbol{p}^2 + m^2. \tag{3.7}$$

Using the operator replacements for E and $\boldsymbol{p}$ we are led to

$$-\partial^2\phi/\partial t^2 = (-\boldsymbol{\nabla}^2 + m^2)\phi \tag{3.8}$$

which is the Klein–Gordon equation (KG equation). We consider the case of a one-component scalar wavefunction $\phi(\boldsymbol{x}, t)$: one expects this to be appropriate for the description of spin-0 bosons.

3.1.1 Solutions in coordinate space

In terms of the D'Alembertian operator

$$\Box \equiv \partial_\mu \partial^\mu = \frac{\partial^2}{\partial t^2} - \boldsymbol{\nabla}^2 \tag{3.9}$$

the KG equation reads:

$$(\Box + m^2)\phi(\boldsymbol{x}, t) = 0. \tag{3.10}$$

Let us look for a plane-wave solution of the form

$$\phi(\boldsymbol{x}, t) = N\mathrm{e}^{-\mathrm{i}Et + \mathrm{i}\boldsymbol{p}\cdot\boldsymbol{x}} = N\mathrm{e}^{-\mathrm{i}p\cdot x} \tag{3.11}$$

where we have written the exponent in suggestive 4-vector scalar product notation

$$p \cdot x = p_\mu x^\mu = Et - \boldsymbol{p} \cdot \boldsymbol{x} \tag{3.12}$$

and N is a normalization factor which need not be decided upon here (see section 8.1.1). In order that this wavefunction be a solution of the KG equation, we find by direct substitution that E must be related to $\boldsymbol{p}$ by the condition

$$E^2 = \boldsymbol{p}^2 + m^2. \tag{3.13}$$

This looks harmless enough, but it actually implies that for a given 3-momentum $\boldsymbol{p}$ there are in fact *two* possible solutions for the energy, namely

$$E = \pm(\boldsymbol{p}^2 + m^2)^{1/2}. \tag{3.14}$$

As Schrödinger and others quickly found, it is not possible to ignore the negative solutions without obtaining inconsistencies. What then do these negative-energy solutions mean?

3.1.2 Probability current for the KG equation

In exactly the same way as for the non-relativistic Schrödinger equation, it is possible to derive a conservation law for a 'probability current' of the KG equation. We have

$$\frac{\partial^2\phi}{\partial t^2} - \boldsymbol{\nabla}^2\phi + m^2\phi = 0 \tag{3.15}$$

and by multiplying this equation by ϕ^*, and subtracting ϕ times the complex conjugate of equation (3.15), one obtains, after some manipulation (see problem 3.1), the result

$$\frac{\partial\rho}{\partial t} + \boldsymbol{\nabla}\cdot\boldsymbol{j} = 0 \tag{3.16}$$

where

$$\rho = \mathrm{i}\left[\phi^*\frac{\partial\phi}{\partial t} - \left(\frac{\partial\phi^*}{\partial t}\right)\phi\right] \tag{3.17}$$

and

$$\boldsymbol{j} = \mathrm{i}^{-1}[\phi^*\boldsymbol{\nabla}\phi - (\boldsymbol{\nabla}\phi^*)\phi] \tag{3.18}$$

(the derivatives $(\partial_\mu\phi^*)$ act only within the bracket). In explicit 4-vector notation this conservation condition reads (cf problem 2.1 and equation (D.4) in appendix D)

$$\partial_\mu j^\mu = 0 \tag{3.19}$$

with

$$j^\mu \equiv (\rho, \boldsymbol{j}) = \mathrm{i}[\phi^*\partial^\mu\phi - (\partial^\mu\phi^*)\phi]. \tag{3.20}$$

Since ϕ of (3.11) is Lorentz invariant and ∂^μ is a contravariant 4-vector, equation (3.20) shows explicitly that j^μ is a contravariant 4-vector, as anticipated in the notation.

The spatial current $\boldsymbol{j}$ is identical in form to the Schrödinger current, but for the KG case the 'probability density' now contains time derivatives since the KG equation is second order in $\partial/\partial t$. This means that ρ is not constrained to be positive definite – so how can ρ represent a probability density? We can see this problem explicitly for the plane-wave solutions

$$\phi = N\mathrm{e}^{-\mathrm{i}Et + \mathrm{i}\boldsymbol{p}\cdot\boldsymbol{x}} \tag{3.21}$$

which give (problem 3.1)

$$\rho = 2|N|^2E \tag{3.22}$$

and E can be positive or negative: that is, the sign of ρ is the sign of energy.

Historically, this problem of negative probabilities coupled with that of negative energies led to the abandonment of the KG equation. For the moment we will follow history, and turn to the Dirac equation. We shall see in section 3.4, however, how the negative-energy solutions of the KG equation do after all have a role to play, following Feynman's interpretation, in processes involving antiparticles. Later, in chapters 5–7, we shall see how this interpretation arises naturally within the formalism of quantum field theory.

3.2 The Dirac equation

In the case of the KG equation it is clear why the problem arose:

(i) In constructing a wave equation in close correspondence with the squared energy–momentum relation

$$E^2 = \boldsymbol{p}^2 + m^2$$

we immediately allowed negative-energy solutions.

(ii) The KG equation has a $\partial^2/\partial t^2$ term: this leads to a continuity equation with a 'probability density' containing $\partial/\partial t$, and hence to negative probabilities.

Dirac approached these problems in his characteristically direct way. In order to obtain a positive-definite probability density $\rho \geq 0$, he required an equation linear in $\partial/\partial t$. Then, for relativistic covariance (see chapter 4), the equation must also be linear in $\boldsymbol{\nabla}$. He postulated the equation (Dirac 1928)

$$\begin{aligned} \mathrm{i}\frac{\partial\psi(\boldsymbol{x},t)}{\partial t} &= \left[-\mathrm{i}\left(\alpha_1\frac{\partial}{\partial x^1} + \alpha_2\frac{\partial}{\partial x^2} + \alpha_3\frac{\partial}{\partial x^3}\right) + \beta m\right]\psi(\boldsymbol{x},t) \\ &= (-\mathrm{i}\boldsymbol{\alpha}\cdot\boldsymbol{\nabla} + \beta m)\psi(\boldsymbol{x},t). \end{aligned} \tag{3.23}$$

What are the α's and β? To find the conditions on the α's and β, consider what we require of a relativistic wave equation:

(i) the correct relativistic relation between E and $\boldsymbol{p}$, namely

$$E = +(\boldsymbol{p}^2 + m^2)^{1/2}$$

(ii) the equation should be covariant under Lorentz transformations.

We shall postpone discussion of (ii) until the following chapter. To solve requirement (i), Dirac in fact demanded that his wavefunction ψ satisfy, in addition, a KG-type condition

$$-\partial^2\psi/\partial t^2 = (-\boldsymbol{\nabla}^2 + m^2)\psi. \tag{3.24}$$

We note with hindsight that we have once more opened the door to negative-energy solutions: Dirac's remarkable achievement was to turn this apparent defect into one of the triumphs of theoretical physics!

We can now derive conditions on $\boldsymbol{\alpha}$ and β. We have

$$\mathrm{i}\partial\psi/\partial t = (-\mathrm{i}\boldsymbol{\alpha}\cdot\boldsymbol{\nabla} + \beta m)\psi \tag{3.25}$$

and so, squaring the operator on both sides,

$$\begin{aligned}\left(\mathrm{i}\frac{\partial}{\partial t}\right)^2\psi &= (-\mathrm{i}\boldsymbol{\alpha}\cdot\boldsymbol{\nabla}+\beta m)(-\mathrm{i}\boldsymbol{\alpha}\cdot\boldsymbol{\nabla}+\beta m)\psi \\ &= -\sum_{i=1}^{3}\alpha_i^2\frac{\partial^2\psi}{(\partial x^i)^2}-\sum_{\substack{i,j=1\\ i>j}}^{3}(\alpha_i\alpha_j+\alpha_j\alpha_i)\frac{\partial^2\psi}{\partial x^i\partial x^j} \\ &\quad -\mathrm{i}m\sum_{i=1}^{3}(\alpha_i\beta+\beta\alpha_i)\frac{\partial\psi}{\partial x^i}+\beta^2m^2\psi. \qquad (3.26)\end{aligned}$$

But by our assumption that ψ also satisfies the KG condition, we must have

$$\left(\mathrm{i}\frac{\partial}{\partial t}\right)^2\psi=-\sum_{i=1}^{3}\frac{\partial^2\psi}{(\partial x^i)^2}+m^2\psi. \qquad (3.27)$$

It is thus evident that the α's and β cannot be ordinary, classical, commuting quantities. Instead they must satisfy the following *anticommutation relations* in order to eliminate the unwanted terms on the right-hand side of equation (3.26):

$$\alpha_i\beta+\beta\alpha_i = 0 \qquad i=1,2,3 \qquad (3.28)$$

$$\alpha_i\alpha_j+\alpha_j\alpha_i = 0 \qquad i,j=1,2,3;\ i\neq j. \qquad (3.29)$$

In addition we require

$$\alpha_i^2=\beta^2=1. \qquad (3.30)$$

Dirac proposed that the α's and β should be interpreted as matrices, acting on a wavefunction which had several components arranged as a column vector. Anticipating somewhat the results of the next section, we would expect that, since each such component obeys the same wave equation, the physical states which they represent would have the same energy. This would mean that the different components represent some *degeneracy*, associated with a new degree of freedom.

The degree of freedom is, of course, *spin* – an entirely quantum mechanical angular momentum, analogous to (but not equivalent to) orbital angular momentum. Consider, for example, the wavefunctions for the 2p state in the simple non-relativistic theory of the hydrogen atom. There are three of them, all degenerate with energy given by the $n=2$ Bohr energy. The three corresponding states all have orbital angular momentum quantum number l equal to 1; they differ in their values of the 'magnetic' quantum number m (i.e. the eigenvalue of the z-component of the orbital angular momentum operator $\hat{L}_z$). Specifically, these three wavefunctions have the form (omitting normalization constants) $(r\sin\theta\mathrm{e}^{\mathrm{i}\phi}, r\sin\theta\mathrm{e}^{-\mathrm{i}\phi}, r\cos\theta)\mathrm{e}^{-r/2r_\mathrm{B}}$, where r_B is the Bohr radius. Remembering the expressions for the Cartesian coordinates x, y and z in terms of the spherical polar coordinates r, θ and ϕ, we see that by a suitable

linear combination (always allowed for degenerate states) we can write these wavefunctions as $(x, y, z)f(r)$, where again a normalization factor has been omitted. In this form it is plain that the multiplicity of the p-state wavefunctions can be interpreted in simple geometrical terms: they are effectively the components of a *vector* (multiplication by the scalar function $f(r)$ does not affect this).

The several components of the Dirac wavefunction together make up a similar, but quite distinct, object called a *spinor*. We shall have more to say about this in chapter 4. For the moment we continue with the problem of finding the matrices α_i and β to satisfy (3.28)–(3.30).

As problem 3.2 shows, the smallest possible dimension of the matrices for which the Dirac conditions can be satisfied is 4×4. One conventional choice of the α's and β is

$$\alpha_i = \begin{pmatrix} \mathbf{0} & \sigma_i \\ \sigma_i & \mathbf{0} \end{pmatrix} \qquad \beta = \begin{pmatrix} \mathbf{1} & \mathbf{0} \\ \mathbf{0} & -\mathbf{1} \end{pmatrix} \tag{3.31}$$

where we have written these 4×4 matrices in 2×2 'block diagonal' form, the σ_i's are the 2×2 *Pauli matrices*, $\mathbf{1}$ is the 2×2 unit matrix, and $\mathbf{0}$ is the 2×2 null matrix. The Pauli matrices (see appendix A) are defined by

$$\sigma_x = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \qquad \sigma_y = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix} \qquad \sigma_z = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}. \tag{3.32}$$

Readers unfamiliar with the labour-saving 'block' form of (3.31) should verify, both by using the corresponding explicit 4×4 matrices, such as

$$\alpha_1 = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix} \tag{3.33}$$

and so on, and by the block diagonal form, that this choice does indeed satisfy the required conditions. These are

$$\{\alpha_i, \beta\} = 0 \tag{3.34}$$

$$\{\alpha_i, \alpha_j\} = 2\delta_{ij}\mathbf{1} \tag{3.35}$$

$$\beta^2 = \mathbf{1} \tag{3.36}$$

where $\{\mathbf{A}, \mathbf{B}\}$ is the anticommutator of two matrices, $\mathbf{AB} + \mathbf{BA}$, and $\mathbf{1}$ is here the 4×4 unit matrix.

At this point we can already begin to see that the extra multiplicity is very likely to have something to do with an angular momentum-like degree of freedom. In fact, if we define the spin matrices $\boldsymbol{S}$ by $\boldsymbol{S} = \frac{1}{2}\boldsymbol{\sigma}$ ($\hbar = 1$), we find from (3.32) that

$$[S_x, S_y] = iS_z \tag{3.37}$$

(with obvious cyclic permutations), which are precisely the commutation relations satisfied by the components $\hat{J}_x$, $\hat{J}_y$ and $\hat{J}_z$ of the angular momentum operator $\hat{\boldsymbol{J}}$ in quantum mechanics (see appendix A). Furthermore, the eigenvalues of S_z are $\pm\frac{1}{2}$, and of $\boldsymbol{S}^2$ are $s(s+1)$ with $s=\frac{1}{2}$. So these matrices undoubtedly represent quantum mechanical angular momentum operators, appropriate to a state with angular momentum quantum number $j=\frac{1}{2}$. This is precisely what 'spin' is. We will discuss this in more detail in section 3.3.

It is important to note that the choice (3.31) of α and β is not unique. In fact, all matrices related to these by any unitary 4×4 matrix $\mathbf{U}$ (which thus preserves the anticommutation relations) are allowed:

$$\alpha_i' = \mathbf{U}\alpha_i\mathbf{U}^{-1} \tag{3.38}$$

$$\beta' = \mathbf{U}\beta\mathbf{U}^{-1}. \tag{3.39}$$

Another commonly used representation is provided by the matrices

$$\boldsymbol{\alpha} = \begin{pmatrix} \boldsymbol{\sigma} & \mathbf{0} \\ \mathbf{0} & -\boldsymbol{\sigma} \end{pmatrix} \qquad \beta = \begin{pmatrix} \mathbf{0} & \mathbf{1} \\ \mathbf{1} & \mathbf{0} \end{pmatrix}. \tag{3.40}$$

The reader may check (problem 3.2) that these matrices also satisfy (3.34)–(3.36).

Unless otherwise stated, we shall use the standard representation (3.31). This is generally convenient for 'low energy' applications – that is, when the momentum $|\boldsymbol{p}|$ is significantly smaller than the mass m. In that case, βm will be the largest term in the Dirac Hamiltonian (see (3.23)), and it is sensible to have it in diagonal form. The choice (3.40), by contrast, is more natural when the mass is small compared with the energy or momentum.

3.2.1 Free-particle solutions

Since the Dirac Hamiltonian now involves 4×4 matrices, it is clear that we must interpret the Dirac wavefunction ψ as a four-component column vector – the so-called Dirac spinor. Let us look at the explicit form of the free-particle solutions. As in the KG case, we look for solutions in which the space–time behaviour is of plane-wave form and put

$$\psi = \omega \mathrm{e}^{-\mathrm{i}p\cdot x} \tag{3.41}$$

where ω is a four-component spinor independent of x, and $\mathrm{e}^{-\mathrm{i}p\cdot x}$, with $p^\mu = (E,\boldsymbol{p})$, is the plane-wave solution corresponding to 4-momentum p^μ. We substitute this into the Dirac equation

$$\mathrm{i}\partial\psi/\partial t = (-\mathrm{i}\boldsymbol{\alpha}\cdot\boldsymbol{\nabla} + \beta m)\psi \tag{3.42}$$

using the explicit $\boldsymbol{\alpha}$ and β matrices. In order to use the 2×2 block form, it is conventional (and convenient) to split the spinor ω into two two-component spinors ϕ and χ:

$$\omega = \begin{pmatrix} \phi \\ \chi \end{pmatrix}. \tag{3.43}$$

We obtain the matrix equation (see problem 3.3)

$$E\begin{pmatrix}\phi\\ \chi\end{pmatrix}=\begin{pmatrix}m\mathbf{1} & \boldsymbol{\sigma}\cdot\boldsymbol{p}\\ \boldsymbol{\sigma}\cdot\boldsymbol{p} & -m\mathbf{1}\end{pmatrix}\begin{pmatrix}\phi\\ \chi\end{pmatrix} \tag{3.44}$$

representing two coupled equations for ϕ and χ:

$$(E-m)\phi=\boldsymbol{\sigma}\cdot\boldsymbol{p}\chi \tag{3.45}$$

and

$$(E+m)\chi=\boldsymbol{\sigma}\cdot\boldsymbol{p}\phi. \tag{3.46}$$

Solving for χ from (3.46), the general four-component spinor may be written (without worrying about normalization for the moment)

$$\omega=\begin{pmatrix}\phi\\ \dfrac{\boldsymbol{\sigma}\cdot\boldsymbol{p}}{E+m}\phi\end{pmatrix}. \tag{3.47}$$

What is the relation between E and $\boldsymbol{p}$ for this to be a solution of the Dirac equation? If we substitute χ from (3.46) into (3.45) and remember that (problem 3.4)

$$(\boldsymbol{\sigma}\cdot\boldsymbol{p})^2=\boldsymbol{p}^2\mathbf{1} \tag{3.48}$$

we find that

$$(E-m)(E+m)\phi=\boldsymbol{p}^2\phi \tag{3.49}$$

for any ϕ. Hence we arrive at the same result as for the KG equation in that for a given value of $\boldsymbol{p}$, two values of E are allowed:

$$E=\pm(\boldsymbol{p}^2+m^2)^{1/2} \tag{3.50}$$

i.e. positive *and* negative solutions are still admitted.

The Dirac equation does not therefore solve this problem. What about the probability current?

3.2.2 Probability current for the Dirac equation

Consider the following quantity which we denote (suggestively) by ρ:

$$\rho=\psi^\dagger(x)\psi(x). \tag{3.51}$$

Here $\psi^\dagger$ is the Hermitian conjugate row vector of the column vector ψ. In terms of components

$$\rho=(\psi_1^*,\psi_2^*,\psi_3^*,\psi_4^*)\begin{pmatrix}\psi_1\\ \psi_2\\ \psi_3\\ \psi_4\end{pmatrix} \tag{3.52}$$

so

$$\rho = \sum_{a=1}^{4} |\psi_a|^2 > 0 \tag{3.53}$$

and we see that ρ is a scalar density which is explicitly positive-definite. This is one property we require of a probability density: in addition, we require a conservation law, coming from the Dirac equation, and a corresponding probability current density. In fact (see problem 3.5) we can demonstrate, using the Dirac equation,

$$\mathrm{i}\partial\psi/\partial t = (-\mathrm{i}\boldsymbol{\alpha}\cdot\boldsymbol{\nabla} + \beta m)\psi \tag{3.54}$$

and its Hermitian conjugate

$$-\mathrm{i}\partial\psi^\dagger = \psi^\dagger(+\mathrm{i}\boldsymbol{\alpha}\cdot\overleftarrow{\boldsymbol{\nabla}} + \beta m) \tag{3.55}$$

that there is a conservation law of the required form

$$\partial\rho/\partial t + \boldsymbol{\nabla}\cdot\boldsymbol{j} = 0. \tag{3.56}$$

The notation $\psi^\dagger\overleftarrow{\boldsymbol{\nabla}}$ requires some comment: it is shorthand for three row matrices

$$\psi^\dagger\overleftarrow{\boldsymbol{\nabla}}_x \equiv \partial\psi^\dagger/\partial x \qquad \text{etc.}$$

(recall that $\psi^\dagger$ is a row matrix).

In equation (3.56), with ρ being given by (3.51), the probability current density $\boldsymbol{j}$ is

$$\boldsymbol{j}(x) = \psi^\dagger(x)\boldsymbol{\alpha}\psi(x) \tag{3.57}$$

representing a 3-vector with components

$$(\psi^\dagger\alpha_1\psi, \psi^\dagger\alpha_2\psi, \psi^\dagger\alpha_3\psi). \tag{3.58}$$

We therefore have a positive-definite ρ and an associated $\boldsymbol{j}$ satisfying the required conservation law (3.56), which, as usual, we can write in invariant form as $\partial_\mu j^\mu = 0$, where

$$j^\mu = (\rho, \boldsymbol{j}). \tag{3.59}$$

Thus j^μ is an acceptable probability current, unlike the current for the KG equation – as we might have anticipated.

The form of equation (3.56) implies that j^μ of (3.59) is a contravariant 4-vector (cf equation (D.4)), as we verified explicitly in the KG case. The corresponding verification is more difficult in the Dirac case, since the Dirac spinor ψ transforms non-trivially under Lorentz transformations, unlike the KG wavefunction ϕ. We shall come back to this problem in chapter 4.

We now turn to further discussion of the spin degree of freedom, postponing consideration of the negative-energy solutions until section 3.4.

3.3 Spin

Four-momentum is not the only physical property of a particle obeying the Dirac equation. We must now interpret the column vector (Dirac spinor) part, ω, of the solution (3.41). The particular properties of the $\boldsymbol{\sigma}$-matrices, appearing in the $\boldsymbol{\alpha}$-matrices, have already led us to think in terms of spin. A further indication that this is correct comes when we consider the explicit form of ω given in (3.47). In this equation the two-component spinor ϕ is completely arbitrary. It may be chosen in just two linearly independent ways, for example

$$\phi_\uparrow = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \qquad \phi_\downarrow = \begin{pmatrix} 0 \\ 1 \end{pmatrix} \tag{3.60}$$

which (as the notation of course indicates) are in fact eigenvectors of $S_z = \frac{1}{2}\sigma_z$ with eigenvalues $\pm\frac{1}{2}$ ('up' and 'down' along the z-axis). Remember that, in quantum mechanics, linear combinations of wavefunctions can be formed using complex numbers as superposition coefficients, in general; so the most general ϕ can always be written as

$$\phi = \begin{pmatrix} a \\ b \end{pmatrix} = a\phi_\uparrow + b\phi_\downarrow \tag{3.61}$$

where a and b are complex numbers. Hence, there are precisely *two* linearly independent solutions, for a given 4-momentum, just as we would expect for a quantum system with $j = \frac{1}{2}$ (the multiplicity is $2j + 1$, in general).

In the rest frame of the particle ($\boldsymbol{p} = \mathbf{0}$) this interpretation is straightforward. In this case choosing (3.60) for the two independent ϕ's, the solutions (3.61) for $E = m$ reduce to

$$\underset{\text{(a)}}{\begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix} \mathrm{e}^{-\mathrm{i}mt}} \qquad \text{and} \qquad \underset{\text{(b)}}{\begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \end{pmatrix} \mathrm{e}^{-\mathrm{i}mt}}. \tag{3.62}$$

Since we have degeneracy between these two solutions (both have $E = m$) there must be some operator which commutes with the energy operator, and whose eigenvalues would distinguish the solutions (3.62). In this case the energy operator is just βm (from (3.54) setting $-\mathrm{i}\boldsymbol{\nabla}$ to zero, since $\boldsymbol{p} = \mathbf{0}$) and the required operator commuting with β is

$$\Sigma_z = \begin{pmatrix} \sigma_z & 0 \\ 0 & \sigma_z \end{pmatrix} \tag{3.63}$$

which has eigenvalues 1 (twice) and -1 (twice). Our rest-frame spinors appearing in (3.62) are indeed eigenstates of Σ_z, with eigenvalues ± 1 as can be easily verified.

Generalizing (3.63), we introduce the three matrices $\mathbf{\Sigma}$ where

$$\mathbf{\Sigma} = \begin{pmatrix} \boldsymbol{\sigma} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\sigma} \end{pmatrix}. \tag{3.64}$$

Then the operators $\frac{1}{2}\mathbf{\Sigma}$ are such that

$$[\tfrac{1}{2}\Sigma_x, \tfrac{1}{2}\Sigma_y] = \mathrm{i}\tfrac{1}{2}\Sigma_z \tag{3.65}$$

and $(\frac{1}{2}\mathbf{\Sigma})^2 = \frac{3}{4}\mathbf{I}$ where $\mathbf{I}$ is now the unit 4×4 matrix. These are just the properties expected of quantum-mechanical angular momentum operators (see appendix A) belonging to magnitude $j = \frac{1}{2}$ (we already know that the eigenvalues of $\frac{1}{2}\Sigma_z$ are $\pm\frac{1}{2}$). So we can interpret $\frac{1}{2}\mathbf{\Sigma}$ as spin-$\frac{1}{2}$ operators appropriate to our rest-frame solutions; and – at least in the rest frame – we may say that the Dirac equation describes a particle of spin-$\frac{1}{2}$.

It seems reasonable to suppose that the magnitude of a spin of a particle could not be changed by doing a Lorentz transformation, as would be required in order to discuss the spin in a general frame with $\boldsymbol{p} \neq \mathbf{0}$. But $\frac{1}{2}\mathbf{\Sigma}$ is then no longer a suitable spin operator, since it fails to commute with the energy operator, which is now $(\boldsymbol{\alpha} \cdot \boldsymbol{p} + \beta m)$ from (3.54), for a plane-wave solution with momentum $\boldsymbol{p}$. Yet there are still just two independent states for a given 4-momentum as our explicit solution (3.47) shows: ϕ can still be chosen in only two linearly independent ways. Hence there must be some operator which does commute with $\boldsymbol{\alpha} \cdot \boldsymbol{p} + \beta m$, and whose eigenvalues can be used to distinguish the two states. Actually this condition is not enough to specify such an operator uniquely, and several choices are common. One of the most useful is the *helicity* operator $h(\boldsymbol{p})$ defined by

$$h(\boldsymbol{p}) = \begin{pmatrix} \dfrac{\boldsymbol{\sigma} \cdot \boldsymbol{p}}{|\boldsymbol{p}|} & \mathbf{0} \\ \mathbf{0} & \dfrac{\boldsymbol{\sigma} \cdot \boldsymbol{p}}{|\boldsymbol{p}|} \end{pmatrix} \tag{3.66}$$

which (see problem 3.6) does commute with $\boldsymbol{\alpha} \cdot \boldsymbol{p} + \beta m$. We can therefore choose our general $\boldsymbol{p} \neq \mathbf{0}$ states to be eigenstates of $h(\boldsymbol{p})$. These will be called 'helicity states': physically they are eigenstates of $\mathbf{\Sigma}$ resolved along the direction of $\boldsymbol{p}$.

Using (3.48) it is easy to see that the eigenvalues of $h(\boldsymbol{p})$ are $+1$ (twice) and -1 (twice). Our general four-component spinor (3.47) is therefore an eigenstate of $h(\boldsymbol{p})$ if

$$\begin{pmatrix} \dfrac{\boldsymbol{\sigma} \cdot \boldsymbol{p}}{|\boldsymbol{p}|} & \mathbf{0} \\ \mathbf{0} & \dfrac{\boldsymbol{\sigma} \cdot \boldsymbol{p}}{|\boldsymbol{p}|} \end{pmatrix} \begin{pmatrix} \phi \\ \dfrac{\boldsymbol{\sigma} \cdot \boldsymbol{p}}{E+m}\phi \end{pmatrix} = \pm \begin{pmatrix} \phi \\ \dfrac{\boldsymbol{\sigma} \cdot \boldsymbol{p}}{E+m}\phi \end{pmatrix}. \tag{3.67}$$

Taking the + sign first, this will hold if

$$\frac{\boldsymbol{\sigma} \cdot \boldsymbol{p}}{|\boldsymbol{p}|}\phi_+ = \phi_+ \tag{3.68}$$

where the + subscript has been added to indicate that this ϕ is a solution of (3.68). Such a ϕ_+ is called a two-component helicity spinor. The explicit form of ϕ_+ can be found by solving (3.68) – see problem 3.7. Similarly, the four-component spinor will be an eigenstate of $h(\boldsymbol{p})$ belonging to the eigenvalue -1 if it contains ϕ_- where

$$\frac{\boldsymbol{\sigma}\cdot\boldsymbol{p}}{|\boldsymbol{p}|}\phi_- = -\phi_-. \tag{3.69}$$

Again, these two choices ϕ_+ and ϕ_- are linearly independent.

3.4 The negative-energy solutions

In this section we shall first look more closely at the form of both the positive- and negative-energy solutions of the Dirac equation, and we shall then concentrate on the physical interpretation of the negative-energy solutions of both the Dirac and the KG equations.

It will be convenient, from now on, to reserve the symbol 'E' for the *positive* square root in (3.50): $E = +(\boldsymbol{p}^2 + m^2)$. The general 4-momentum in the plane-wave solution (3.41) will be denoted by $p^\mu = (p^0, \boldsymbol{p})$ where p^0 may be either positive or negative. With this notation equation (3.44) becomes

$$p^0 \begin{pmatrix} \phi \\ \chi \end{pmatrix} = \begin{pmatrix} m\mathbf{1} & \boldsymbol{\sigma}\cdot\boldsymbol{p} \\ \boldsymbol{\sigma}\cdot\boldsymbol{p} & -m\mathbf{1} \end{pmatrix} \begin{pmatrix} \phi \\ \chi \end{pmatrix} \tag{3.70}$$

in our original representation for $\boldsymbol{\alpha}$ and β.

3.4.1 Positive-energy spinors

For these

$$p^0 = +(\boldsymbol{p}^2 + m^2)^{1/2} \equiv E > 0. \tag{3.71}$$

We eliminate χ and obtain positive-energy spinors in the form

$$\omega^{1,2} = N \begin{pmatrix} \phi^{1,2} \\ \dfrac{\boldsymbol{\sigma}\cdot\boldsymbol{p}}{E+m}\phi^{1,2} \end{pmatrix}, \tag{3.72}$$

with $\phi^{1\dagger}\phi^1 = \phi^{2\dagger}\phi^2 = 1$. We shall now choose N so that for these positive-energy solutions $\omega^\dagger\omega = 2E$. In this case the spinors will be denoted by $u(p,s)$, where (problem 3.8)

$$u(p,s) = (E+m)^{1/2} \begin{pmatrix} \phi^s \\ \dfrac{\boldsymbol{\sigma}\cdot\boldsymbol{p}}{E+m}\phi^s \end{pmatrix} \qquad s = 1,2 \tag{3.73}$$

and s labels the spin degree of freedom in some suitable way (e.g. the helicity eigenvalues). The complete plane-wave solution ψ for such a positive 4-momentum state is then

$$\psi = u(p,s)\mathrm{e}^{-\mathrm{i}p_+\cdot x} \tag{3.74}$$

with $p_+^\mu = (E, \boldsymbol{p})$.

3.4.2 Negative-energy spinors

Now we look for spinors appropriate to the solution

$$p^0 = -(\boldsymbol{p}^2 + m^2)^{1/2} \equiv -E < 0 \tag{3.75}$$

(E is always defined to be positive). Consider first what are appropriate solutions at rest. We have now

$$p^0 = -m \qquad \boldsymbol{p} = \boldsymbol{0} \tag{3.76}$$

and

$$-m\begin{pmatrix}\phi\\ \chi\end{pmatrix} = \begin{pmatrix} m\mathbf{1} & \mathbf{0}\\ \mathbf{0} & -\,m\mathbf{1}\end{pmatrix}\begin{pmatrix}\phi\\ \chi\end{pmatrix} \tag{3.77}$$

leading to

$$\phi = 0. \tag{3.78}$$

Thus the two independent negative-energy solutions at rest are just

$$\omega(p^0 = -m, s) = \begin{pmatrix} 0\\ \chi^s\end{pmatrix}. \tag{3.79}$$

The solution for finite momentum $+\boldsymbol{p}$, i.e. for 4-momentum $(-E, \boldsymbol{p})$, is then

$$\omega(p^0 = -E, \boldsymbol{p}, s) = \begin{pmatrix} \dfrac{-\boldsymbol{\sigma}\cdot\boldsymbol{p}}{E+m}\chi^s\\ \chi^s\end{pmatrix} \tag{3.80}$$

with ${\chi^s}^\dagger\chi^s = 1$. However, it is clearly much more in keeping with relativity if, in addition to changing the sign of E, we also change the sign of $\boldsymbol{p}$ and consider solutions corresponding to negative 4-momentum $(-E, -\boldsymbol{p}) = -p_+^\mu$. We therefore define

$$\omega(p^0 = -E, -\boldsymbol{p}, s) \equiv \omega^{3,4} = N\begin{pmatrix} \dfrac{\boldsymbol{\sigma}\cdot\boldsymbol{p}}{E+m}\chi^{1,2}\\ \chi^{1,2}\end{pmatrix}. \tag{3.81}$$

Adopting the same N as in (3.73) implies the same normalization ($\omega^\dagger\omega = 2E$) for (3.81) as in (3.73); in this case the spinors are called $v(p,s)$ where (problem 3.8)

$$v(p,s) = (E+m)^{1/2}\begin{pmatrix} \dfrac{\boldsymbol{\sigma}\cdot\mathbf{p}}{E+m}\chi^s\\ \chi^s\end{pmatrix} \qquad s = 1, 2. \tag{3.82}$$

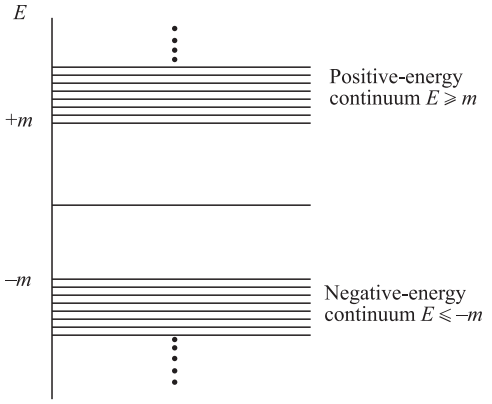


FIGURE 3.1
Energy levels for Dirac particle.

(There is a small subtlety in the choice of χ^1 and χ^2 which we will come to shortly.) The solution ψ for such negative 4-momentum states is then

$$\psi = v(p,s)\mathrm{e}^{-\mathrm{i}(-p_+)\cdot x} = v(p,s)\mathrm{e}^{\mathrm{i}p_+\cdot x}. \tag{3.83}$$

3.4.3 Dirac's interpretation of the negative-energy solutions of the Dirac equation

The physical interpretation of the positive-energy solution (3.74) is straightforward, in terms of the ρ and $\boldsymbol{j}$ given in section 3.2.2. They describe spin-$\frac{1}{2}$ particles with 4-momentum $(E, \boldsymbol{p})$ and spin appropriate to the choice of ϕ^s; ρ and the energy p^0 are both positive.

Unfortunately ρ is also positive for the *negative*-energy solutions (3.83), so we cannot eliminate them on that account. This means that for a free Dirac particle (e.g. an electron) the available positive- and negative-energy levels are as shown in figure 3.1. This, in turn, implies that a particle with initially positive energy can 'cascade down' through the negative-energy levels, without limit; in this case no stable positive-energy state would exist!

In order to prevent positive-energy electrons making transitions to the lower, negative-energy states, Dirac postulated that the normal 'empty', or 'vacuum', state – that with no positive-energy electrons present – is such that all the negative-energy states are filled with electrons. The Pauli exclusion principle then forbids any positive-energy electrons from falling into these lower energy levels. The 'vacuum' now has infinite negative charge and energy, but since all observations represent *finite* fluctuations in energy and charge with respect to this vacuum, this leads to an acceptable theory. For example, if one negative-energy electron is absent from the Dirac sea, we have a 'hole'

relative to the normal vacuum:

$$\begin{aligned} \text{energy of 'hole'} &= -(E_{\text{neg}}) \rightarrow \text{positive energy} \\ \text{charge of 'hole'} &= -(q_{\text{e}}) \rightarrow \text{positive charge.} \end{aligned}$$

Thus the *absence* of a negative-energy electron is equivalent to the *presence* of a positive-energy positively charged version of the electron, that is a positron. In the same way, the absence of a 'spin-up' negative-energy electron is equivalent to the presence of a 'spin-down' positive-energy positron. This last point is the reason for the subtlety in the choice of χ^s mentioned after (3.82): we choose

$$\chi^1 = \begin{pmatrix} 0 \\ 1 \end{pmatrix} \qquad \chi^2 = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \tag{3.84}$$

the opposite way round from the choice for the positive-energy spinors (3.73).

Dirac's brilliant re-interpretation of (unfilled) negative-energy solutions in terms of antiparticles is one of the triumphs of theoretical physics[2]: Carl Anderson received the Nobel Prize for his discovery of the positron in 1932 (Anderson 1932).

In this way it proved possible to obtain sensible results from the Dirac equation and its negative-energy solutions. It is clear, however, that the theory is no longer really a 'single-particle' theory, since we can excite electrons from the infinite 'sea' of filled negative-energy states that constitute the normal 'empty state'. For example, if we excite one negative-energy electron to a positive-energy state, we have in the final state a positive-energy electron plus a positive-energy positron 'hole' in the vacuum: this corresponds physically to the process of $\mathrm{e}^+\mathrm{e}^-$ pair creation. Thus this way of dealing with the negative-energy problem for fermions leads us directly to the need for a quantum field theory. The appropriate formalism will be presented later, in section 7.2.

3.4.4 Feynman's interpretation of the negative-energy solutions of the KG and Dirac equations

It is clear that despite its brilliant success for spin-$\frac{1}{2}$ particles, Dirac's interpretation cannot be applied to spin-0 particles, since bosons are not subject to the exclusion principle. Besides, spin-0 particles also have their corresponding antiparticles (e.g. π^+ and π^-), and so do spin-1 particles (W^+ and W^-, for instance). A consistent picture for both bosons and fermions does emerge from quantum field theory, as we shall see in chapters 5–7, which is perhaps one of the strongest reasons for mastering it. Nevertheless, it is useful to have an alternative, non-field-theoretic, interpretation of the negative-energy solutions which works for both bosons and fermions. Such an interpretation is due

[2] At that time, this was not universally recognized. For example, Pauli (1933) wrote: 'Dirac has tried to identify holes with antielectrons... we do not believe that this explanation can be seriously considered.'

to Feynman: in essence, the idea is that the negative 4-momentum solutions will be used to describe antiparticles, for both bosons and fermions.

We begin with bosons – for example pions, which for the present purposes we take to be simple spin-0 particles whose wavefunctions obey the KG equation. We decide by convention that the π^+ is the 'particle'. We will then have

$$\text{positive 4-momentum } \pi^+ \text{ solutions: } N\mathrm{e}^{-\mathrm{i}p\cdot x} \tag{3.85}$$

$$\text{negative 4-momentum } \pi^+ \text{ solutions: } N\mathrm{e}^{\mathrm{i}p\cdot x} \tag{3.86}$$

where $p^\mu = [(m^2 + \boldsymbol{p}^2)^{1/2}, \boldsymbol{p}]$. The electromagnetic current for a free physical (positive-energy) π^+ is given by the probability current for a positive-energy solution multiplied by the charge $Q(= +e)$:

$$j^\mu_{\rm em}(\pi^+) = (+e) \times (\text{probability current for positive energy } \pi^+) \tag{3.87}$$

$$= (+e)2|N|^2[(m^2 + \boldsymbol{p}^2)^{1/2}, \boldsymbol{p}] \tag{3.88}$$

using (3.20) and (3.85) (see problem 3.1). What about the current for the π^-? For free physical π^- particles of positive energy $(m^2 + \boldsymbol{p}^2)^{1/2}$ and momentum $\boldsymbol{p}$ we expect

$$j^\mu_{\rm em}(\pi^-) = (-e)2|N|^2[(m^2 + \boldsymbol{p}^2)^{1/2}, \boldsymbol{p}] \tag{3.89}$$

by simply changing the sign of the charge in (3.88). But it is evident that (3.89) may be written as

$$j^\mu_{\rm em}(\pi^-) = (+e)2|N|^2[-(m^2 + \boldsymbol{p}^2)^{1/2}, -\boldsymbol{p}] \tag{3.90}$$

which is just $j^\mu_{\rm em}(\pi^+)$ with *negative* 4-momentum. This suggests some equivalence between antiparticle solutions with positive 4-momentum and particle solutions with negative 4-momentum.

Can we push this equivalence further? Consider what happens when a system A absorbs a π^+ with positive 4-momentum p: its charge increases by $+e$, and its 4-momentum increases by p. Now suppose that A emits a physical π^- with 4-momentum k, where the energy k^0 is positive. Then the charge of A will increase by $+e$, and its 4-momentum will decrease by k. Now this increase in the charge of A could equally well be caused by the absorption of a π^+ – and indeed we can make the effect (as far as A is concerned) of the π^- emission process fully equivalent to a π^+ absorption process if we say that the equivalent absorbed π^+ has negative 4-momentum, $-k$; in particular the equivalent absorbed π^+ has negative energy $-k^0$. In this way, we view the emission of a physical 'antiparticle' π^- with positive 4-momentum k as equivalent to the absorption of a 'particle' π^+ with (unphysical) negative 4-momentum $-k$. Similar reasoning will apply to the absorption of a π^- of positive 4-momentum, which is equivalent to the emission of a π^+ of negative 4-momentum. Thus we are led to the following hypothesis (due to Feynman):

The emission (absorption) of an antiparticle of 4-momentum p^μ is physically equivalent to the absorption (emission) of a particle of 4-momentum $-p^\mu$.

FIGURE 3.2
Coulomb scattering of a π^- by a static charge Ze illustrating the Feynman interpretation of negative 4-momentum states.

In other words the unphysical negative 4-momentum solutions of the 'particle' equation do have a role to play: they can be used to describe physical processes involving positive 4-momentum antiparticles, if we reverse the role of 'entry' and 'exit' states.

The idea is illustrated in figure 3.2, for the case of Coulomb scattering of a π^- particle by a static charge Ze, which will be discussed later in section 8.1.3. By convention we are taking π^- to be the antiparticle. In the physical process of figure 3.2(a) the incoming physical antiparticle π^- has 4-momentum p_i, and the final π^- has 4-momentum p_f: both E_i and E_f are, of course, positive. Figure 3.2(b) shows how the amplitude for the process can be calculated using π^+ solutions with negative 4-momentum. The initial state π^- of 4-momentum p_i becomes a final state π^+ with 4-momentum $-p_i$, and similarly the final state π^- of 4-momentum p_f becomes an initial state π^+ of 4-momentum $-p_f$. Note that in this and similar figures, the sense of the arrows always indicates the 'flow' of 4-momentum, positive 4-momentum corresponding to forward flow.

It is clear that the basic physical idea here is not limited to bosons. But there is a difference between the KG and Dirac cases in that the Dirac equation was explicitly designed to yield a probability density (and probability current density) which was independent of the sign of the energy:

$$\rho = \psi^\dagger\psi \qquad \boldsymbol{j} = \psi^\dagger\boldsymbol{\alpha}\psi. \tag{3.91}$$

Thus for any solutions of the form

$$\psi = \omega\phi(\boldsymbol{x}, t) \tag{3.92}$$

we have

$$\rho = \omega^\dagger\omega|\phi(\boldsymbol{x}, t)|^2 \tag{3.93}$$

and

$$\boldsymbol{j} = \omega^\dagger\boldsymbol{\alpha}\omega|\phi(\boldsymbol{x}, t)|^2 \tag{3.94}$$

and $\rho \geq 0$ always. We nevertheless want to set up a correspondence so that positive-energy solutions describe *electrons* (taken to be the 'particle', by convention, in this case) and negative-energy solutions describe *positrons*, if we reverse the sense of incoming and outgoing waves. For the KG case this was straightforward, since the probability current was proportional to the 4-momentum:

$$j^\mu(\mathrm{KG}) \sim p^\mu. \tag{3.95}$$

We were therefore able to set up the correspondence for the electromagnetic current of π^+ and π^-:

$$\pi^+ : \quad j^\mu_{\mathrm{em}} \sim ep^\mu \qquad \text{positive energy } \pi^+ \tag{3.96}$$

$$\pi^- : \quad j^\mu_{\mathrm{em}} \sim (-e)p^\mu \qquad \text{positive energy } \pi^- \tag{3.97}$$

$$\equiv (+e)(-p^\mu) \qquad \text{negative energy } \pi^+. \tag{3.98}$$

This simple connection does not hold for the Dirac case since $\rho \geq 0$ for both signs of the energy. It is still possible to set up the correspondence, but now an extra minus sign must be inserted 'by hand' whenever we have a negative-energy fermion in the final state. We shall make use of this rule in section 8.2.4. We therefore state the Feynman hypothesis for fermions:

The invariant amplitude for the emission (absorption) of an antifermion of 4-momentum p^μ and spin projection s_z in the rest frame is equal to the amplitude (minus the amplitude) for the absorption (emission) of a fermion of 4-momentum $-p^\mu$ and spin projection $-s_z$ in the rest frame.

As we shall see in chapters 5–7, the Feynman interpretation of the negative-energy solutions is naturally embodied in the field theory formalism.

3.5 Inclusion of electromagnetic interactions via the gauge principle: the Dirac prediction of $g = 2$ for the electron

Having set up the relativistic spin-0 and spin-$\frac{1}{2}$ free-particle wave equations, we are now in a position to use the machinery developed in chapter 2, in order to include electromagnetic interactions. All we have to do is make the replacement

$$\partial^\mu \to D^\mu \equiv \partial^\mu + \mathrm{i}qA^\mu \tag{3.99}$$

for a particle of charge q. For the spin-0 KG equation (3.10) we obtain, after some rearrangement (problem 3.9),

$$(\Box + m^2)\phi = -\mathrm{i}q(\partial_\mu A^\mu + A^\mu \partial_\mu)\phi + q^2 A^2 \phi \tag{3.100}$$

$$= -\hat{V}_{\mathrm{KG}}\phi. \tag{3.101}$$

Note that the potential $\hat{V}_{\rm KG}$ contains the differential operator ∂_μ; the sign of $\hat{V}_{\rm KG}$ is a convention chosen so as to maintain the same relative sign between $\boldsymbol{\nabla}^2$ and $\hat{V}$ as in the Schrödinger equation – for example that in (A.5).

For the Dirac equation the replacement (3.99) leads to

$$\mathrm{i}\frac{\partial\psi}{\partial t} = [\boldsymbol{\alpha}\cdot(-\mathrm{i}\boldsymbol{\nabla} - q\boldsymbol{A}) + \beta m + qA^0]\psi \tag{3.102}$$

where $A^\mu = (A^0, \boldsymbol{A})$. The potential due to A^μ is therefore $\hat{V}_{\rm D} = qA^0\mathbf{1} - q\boldsymbol{\alpha}\cdot\boldsymbol{A}$, which is a 4×4 matrix acting on the Dirac spinor.

The non-relativistic limit of (3.102) is of great importance, both physically and historically. It was, of course, first obtained by Dirac; and it provided, in 1928, a sensational explanation of why the g-factor of the electron had the value $g = 2$, which was then the empirical value, without any theoretical basis.

By way of background, recall from appendix A that the Schrödinger equation for a non-relativistic spinless particle of charge q in a magnetic field $\boldsymbol{B}$ described by a vector potential $\boldsymbol{A}$ such that $\boldsymbol{B} = \boldsymbol{\nabla}\times\boldsymbol{A}$ is

$$-\frac{1}{2m}\boldsymbol{\nabla}^2\psi - \frac{q}{2m}\boldsymbol{B}\cdot\hat{\boldsymbol{L}}\psi + \frac{q^2}{2m}\boldsymbol{A}^2\psi = \mathrm{i}\frac{\partial\psi}{\partial t}. \tag{3.103}$$

Taking $\boldsymbol{B}$ along the z-axis, the $\boldsymbol{B}\cdot\hat{\boldsymbol{L}}$ term will cause the usual splitting (into states of different magnetic quantum number) of the $(2l+1)$-fold degeneracy associated with a state of definite l. In particular, though, there should be no splitting of the hydrogen ground state which has $l = 0$. But experimentally splitting into two levels is observed, indicating a two-fold degeneracy and thus (see earlier) a $j = \frac{1}{2}$-like degree of freedom.

Uhlenbeck and Goudsmit (1925) suggested that the doubling of the hydrogen ground state could be explained if the electron were given an additional quantum number corresponding to an angular-momentum-like observable, having magnitude $j = \frac{1}{2}$. The operators $\boldsymbol{S} = \frac{1}{2}\boldsymbol{\sigma}$ which we have already met serve to represent such a *spin* angular momentum. If the contribution to the energy operator of the particle due to its spin $\boldsymbol{S}$ enters into the effective Schrödinger equation in exactly the same way as that due to its orbital angular momentum, then we would expect an additional term on the left-hand side of (3.103) of the form

$$-\frac{q}{2m}\boldsymbol{B}\cdot\boldsymbol{S}. \tag{3.104}$$

The corresponding wavefunction must now have two (spinor) components, acted on by the 2×2 matrices in $\boldsymbol{S}$.

The energy difference between the two levels with eigenvalues $S_z = \pm\frac{1}{2}$ would then be $qB/2m$ in magnitude. Experimentally the splitting was found to be just twice this value. Thus empirically the term (3.104) was modified to

$$-g\frac{q}{2m}\boldsymbol{B}\cdot\boldsymbol{S} \tag{3.105}$$

where g is the 'gyromagnetic ratio' of the particle, with $g \approx 2$. Let us now see

how Dirac deduced the term (3.105), with the precise value $g = 2$, from his equation.

To achieve a non-relativistic limit, we expect that we have somehow to reduce the four-component Dirac equation to one involving just two components, since the desired term (3.105) is only a 2×2 matrix. Looking at the explicit form (3.72) for the free-particle positive-energy solutions, we see that the lower two components are of order v (i.e. v/c with $c = 1$) times the upper two. This suggests that, to get a non-relativistic limit, we should regard the lower two components of ψ as being small (at least in the specific representation we are using for $\boldsymbol{\alpha}$ and β). However, since (3.102) includes the A^μ-field, this will have to be demonstrated (see (3.112)). Also, if we write the total energy operator as $m + \hat{H}_1$, we expect $\hat{H}_1$ to be the non-relativistic energy operator.

We let

$$\psi = \begin{pmatrix} \Psi \\ \Phi \end{pmatrix} \tag{3.106}$$

where Ψ and Φ are not free-particle solutions, and they carry the space–time dependence as well as the spinor character (each has two components). We set

$$\hat{H}_1 = \boldsymbol{\alpha} \cdot (-\mathrm{i}\boldsymbol{\nabla} - q\boldsymbol{A}) + \beta m + qA^0 - m \tag{3.107}$$

where a 4×4 unit matrix multiplying the last two terms is understood. Then

$$\begin{aligned} \hat{H}_1 \begin{pmatrix} \Psi \\ \Phi \end{pmatrix} &= \begin{pmatrix} \mathbf{0} & \boldsymbol{\sigma} \cdot (-\mathrm{i}\boldsymbol{\nabla} - q\boldsymbol{A}) \\ \boldsymbol{\sigma} \cdot (-\mathrm{i}\boldsymbol{\nabla} - q\boldsymbol{A}) & \mathbf{0} \end{pmatrix} \begin{pmatrix} \Psi \\ \Phi \end{pmatrix} \\ &\quad - 2m \begin{pmatrix} 0 \\ \Phi \end{pmatrix} + qA^0 \begin{pmatrix} \Psi \\ \Phi \end{pmatrix}. \end{aligned} \tag{3.108}$$

Multiplying out (3.108), we obtain

$$\hat{H}_1 \Psi = \boldsymbol{\sigma} \cdot (-\mathrm{i}\boldsymbol{\nabla} - q\boldsymbol{A})\Phi + qA^0 \Psi \tag{3.109}$$

$$\hat{H}_1 \Phi = \boldsymbol{\sigma} \cdot (-\mathrm{i}\boldsymbol{\nabla} - q\boldsymbol{A})\Psi + qA^0 \Phi - 2m\Phi. \tag{3.110}$$

From (3.110), we obtain

$$(\hat{H}_1 - qA^0 + 2m)\Phi = \boldsymbol{\sigma} \cdot (-\mathrm{i}\boldsymbol{\nabla} - q\boldsymbol{A})\psi. \tag{3.111}$$

So, if $\hat{H}_1$ (or rather any matrix element of it) is $\ll m$ and if A^0 is positive or, if negative, much less in magnitude than m/e, we can deduce

$$\Phi \sim (\text{velocity}) \times \Psi \tag{3.112}$$

as in the free case, provided that the magnetic energy $\sim \boldsymbol{\sigma} \cdot \boldsymbol{A}$ is not of order m. Further, if $\hat{H}_1 \ll m$ and the conditions on the fields are met, we can drop $\hat{H}_1$ and qA^0 on the left-hand side of (3.111), as a first approximation, so that

$$\Phi \approx \frac{\boldsymbol{\sigma} \cdot (-\mathrm{i}\boldsymbol{\nabla} - q\boldsymbol{A})}{2m} \Psi. \tag{3.113}$$

Hence, in (3.109),

$$\hat{H}_1\Psi \approx \frac{1}{2m}\{\boldsymbol{\sigma}\cdot(-\mathrm{i}\boldsymbol{\nabla}-q\boldsymbol{A})\}^2\Psi + qA^0\Psi. \tag{3.114}$$

The right-hand side of (3.114) should therefore be the non-relativistic energy operator for a spin-$\frac{1}{2}$ particle of charge q and mass m in a field A^μ.

Consider then the case $A^0 = 0$ which is sufficient for the discussion of g. We need to evaluate

$$\{\boldsymbol{\sigma}\cdot(-\mathrm{i}\boldsymbol{\nabla}-q\boldsymbol{A})\}^2\Psi. \tag{3.115}$$

This requires care, because although it is true that (for example) $(\boldsymbol{\sigma}\cdot\boldsymbol{p})^2 = \boldsymbol{p}^2$ if $\boldsymbol{p} = (p_x, p_y, p_z)$ are ordinary numbers which commute with each other, the components of '$-\mathrm{i}\boldsymbol{\nabla} - q\boldsymbol{A}$' do *not* commute due to the presence of the differential operator $\boldsymbol{\nabla}$, and the fact that $\boldsymbol{A}$ depends on $\boldsymbol{r}$. In problem 3.10 it is shown that

$$\{\boldsymbol{\sigma}\cdot(-\mathrm{i}\boldsymbol{\nabla}-q\boldsymbol{A})\}^2\Psi = (-\mathrm{i}\boldsymbol{\nabla}-q\boldsymbol{A})^2\Psi - q\boldsymbol{\sigma}\cdot\boldsymbol{B}\Psi. \tag{3.116}$$

The first term on the right-hand side of (3.116) when inserted into (3.114), gives precisely the spin-0 non-relativistic Hamiltonian appearing on the left-hand side of (3.103) (see appendix A), while the second term in (3.116) yields exactly (3.105) with $g = 2$, recalling that $\boldsymbol{S} = \frac{1}{2}\boldsymbol{\sigma}$. Thus the non-relativistic reduction of the Dirac equation leads to the prediction $g = 2$ for a spin-$\frac{1}{2}$ particle.

In actual fact, the measured g-factor of the electron (and muon) is slightly greater than this value: $g_{\mathrm{exp}} = 2(1 + a)$. The 'anomaly' a, which is of order 10^{-3} in size, is measured with quite extraordinary precision (see section 11.7) for both the e^- and e^+. This small correction can also be computed with equally extraordinary accuracy, using the full theory of QED, as we shall briefly explain in chapter 11. The agreement between theory and experiment is phenomenal and is one example of such agreement exhibited by our 'paradigm theory'.

It may be worth noting that spin-$\frac{1}{2}$ hadrons, such as the proton, have g-factors very different from the Dirac prediction. This is because they are, as we know, composite objects and are thus (in this respect) more like atoms in nuclei than 'elementary particles'.

Problems

3.1

(a) In natural units $\hbar = c = 1$ and with $2m = 1$, the Schrödinger equation may be written as

$$-\boldsymbol{\nabla}^2\psi + V\psi - \mathrm{i}\partial\psi/\partial t = 0.$$

Multiply this equation from the left by ψ^* and multiply the complex conjugate of this equation by ψ (assume V is real). Subtract the two equations and show that your answer may be written in the form of a continuity equation

$$\partial\rho/\partial t + \boldsymbol{\nabla}\cdot\boldsymbol{j} = 0$$

where $\rho = \psi^*\psi$ and $\boldsymbol{j} = \mathrm{i}^{-1}[\psi^*(\boldsymbol{\nabla}\psi) - (\boldsymbol{\nabla}\psi^*)\psi]$.

(b) Perform the same operations for the Klein–Gordon equation and derive the corresponding 'probability' density current. Show also that for a free-particle solution

$$\phi = N\mathrm{e}^{-\mathrm{i}p\cdot x}$$

with $p^\mu = (E, \boldsymbol{p})$, the probability current $j^\mu = (\rho, \boldsymbol{j})$ is proportional to p^μ.

3.2

(a) Prove the following properties of the matrices α_i and β:

(i) α_i and β $(i = 1, 2, 3)$ are all Hermitian [*Hint*: what is the Hamiltonian?].

(ii) $\mathrm{Tr}\alpha_i = \mathrm{Tr}\beta = 0$ where 'Tr' means the trace, i.e. the sum of the diagonal elements [*Hint*: use $\mathrm{Tr}(\boldsymbol{AB}) = \mathrm{Tr}(\boldsymbol{BA})$ for any matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ – and prove this too!].

(iii) The eigenvalues of α_i and β are ± 1 [*Hint*: square α_i and β].

(iv) The dimensionality of α_i and β is even [*Hint*: the trace of a matrix is equal to the sum of its eigenvalues].

(b) Verify explicitly that the matrices $\boldsymbol{\alpha}$ and β of (3.31), and of (3.40), satisfy the Dirac conditions (3.34) – (3.36).

3.3 For free-particle solutions of the Dirac equation

$$\psi = \omega\mathrm{e}^{-\mathrm{i}p\cdot x}$$

the four-component spinor ω may be written in terms of the two-component spinors

$$\omega = \begin{pmatrix} \phi \\ \chi \end{pmatrix}.$$

From the Dirac equation for ψ

$$\mathrm{i}\partial\psi/\partial t = (-\mathrm{i}\boldsymbol{\alpha}\cdot\boldsymbol{\nabla} + \beta m)\psi$$

using the explicit forms for the Dirac matrices

$$\boldsymbol{\alpha} = \begin{pmatrix} \mathbf{0} & \boldsymbol{\sigma} \\ \boldsymbol{\sigma} & \mathbf{0} \end{pmatrix} \qquad \beta = \begin{pmatrix} \mathbf{1} & \mathbf{0} \\ \mathbf{0} & -\mathbf{1} \end{pmatrix}$$

show that ϕ and χ satisfy the coupled equations

$$\begin{aligned}(E-m)\phi &= \boldsymbol{\sigma}\cdot\boldsymbol{p}\chi \\ (E+m)\chi &= \boldsymbol{\sigma}\cdot\boldsymbol{p}\phi\end{aligned}$$

where $p^\mu = (E, \boldsymbol{p})$.

3.4

(a) Using the explicit forms for the 2×2 Pauli matrices, verify the commutation (square brackets) and anticommutation (braces) relation [note the summation convention for repeated indices: $\epsilon_{ijk}\sigma_k \equiv \sum_{k=1}^{3}\epsilon_{ijk}\sigma_k$]:

$$[\sigma_i,\sigma_j] = 2\mathrm{i}\epsilon_{ijk}\sigma_k \qquad \{\sigma_i,\sigma_j\} = 2\delta_{ij}\mathbf{1}$$

where ϵ_{ijk} is the usual antisymmetric tensor

$$\epsilon_{ijk} = \begin{cases} +1 & \text{for an even permutation of 1, 2, 3} \\ -1 & \text{for an odd permutation of 1, 2, 3} \\ 0 & \text{if two or more indices are the same,}\end{cases}$$

δ_{ij} is the usual Kronecker delta, and $\mathbf{1}$ is the 2×2 matrix. Hence show that

$$\sigma_i\sigma_j = \delta_{ij}\mathbf{1} + \mathrm{i}\epsilon_{ijk}\sigma_k.$$

(b) Use this last identity to prove the result

$$(\boldsymbol{\sigma}\cdot\boldsymbol{a})(\boldsymbol{\sigma}\cdot\boldsymbol{b}) = \boldsymbol{a}\cdot\boldsymbol{b}\mathbf{1} + \mathrm{i}\boldsymbol{\sigma}\cdot\boldsymbol{a}\times\boldsymbol{b}.$$

Using the explicit 2×2 form for

$$\boldsymbol{\sigma}\cdot\boldsymbol{p} = \begin{pmatrix} p_z & p_x - ip_y \\ p_x + ip_y & -p_z\end{pmatrix}$$

show that

$$(\boldsymbol{\sigma}\cdot\boldsymbol{p})^2 = \boldsymbol{p}^2\mathbf{1}.$$

3.5 Verify the conservation equation (3.56).

3.6 Check that $h(\boldsymbol{p})$ as given by (3.66) does commute with $\boldsymbol{\alpha}\cdot\boldsymbol{p} + \beta m$, the momentum–space free Dirac Hamiltonian.

3.7 Let ϕ be an arbitrary two-component spinor, and let $\hat{\boldsymbol{u}}$ be a unit vector.

(a) Show that $\frac{1}{2}(1 + \boldsymbol{\sigma}\cdot\hat{\boldsymbol{u}})\phi$ is an eigenstate of $\boldsymbol{\sigma}\cdot\hat{\boldsymbol{u}}$ with eigenvalue $+1$. The operator $\frac{1}{2}(1 + \boldsymbol{\sigma}\cdot\hat{\boldsymbol{u}})$ is called a projector operator for the $\boldsymbol{\sigma}\cdot\hat{\boldsymbol{u}} = +1$ eigenstate since when acting on any ϕ this is what it 'projects out'. Write down a similar operator which projects out the $\boldsymbol{\sigma}\cdot\hat{\boldsymbol{u}} = -1$ eigenstate.

(b) Construct two two-component spinors ϕ_+ and ϕ_- which are eigenstates of $\boldsymbol{\sigma}\cdot\hat{\boldsymbol{u}}$ belonging to eigenvalues ± 1, and normalized to $\phi_r^\dagger\phi_s = \delta_{rs}$ for $(r,s) = (+,-)$, for the case $\hat{\boldsymbol{u}} = (\sin\theta\cos\phi, \sin\theta\sin\phi, \cos\theta)$ [*Hint*: take the arbitrary $\phi = \binom{1}{0}$].

3.8 Positive-energy spinors $u(p,s)$ are defined by

$$u(p,s) = (E+m)^{1/2}\begin{pmatrix} \phi^s \\ \dfrac{\boldsymbol{\sigma}\cdot\boldsymbol{p}}{E+m}\phi^s \end{pmatrix} \qquad s = 1,2$$

with $\phi^{s\dagger}\phi^s = 1$. Verify that these satisfy $u^\dagger u = 2E$.

In a similar way, negative-energy spinors $v(p,s)$ are defined by

$$v(p,s) = (E+m)^{1/2}\begin{pmatrix} \dfrac{\boldsymbol{\sigma}\cdot\boldsymbol{p}}{E+m}\chi^s \\ \chi^s \end{pmatrix} \qquad s = 1,2$$

with $\chi^{s\dagger}\chi^s = 1$. Verify that $v^\dagger v = 2E$.

3.9 Using the KG equation together with the replacement $\partial^\mu \to \partial^\mu + \mathrm{i}qA^\mu$, find the form of the potential $\hat{V}_{\mathrm{KG}}$ in the corresponding equation

$$(\Box + m^2)\phi = -\hat{V}_{\mathrm{KG}}\phi$$

in terms of A^μ.

3.10 Evaluate

$$\{\boldsymbol{\sigma}\cdot(-\mathrm{i}\boldsymbol{\nabla} - q\boldsymbol{A})\}^2\psi$$

by following the subsequent steps (or doing it your own way):

(a) Multiply the operator by itself to get

$$\{(\boldsymbol{\sigma}\cdot -\mathrm{i}\boldsymbol{\nabla})^2 + \mathrm{i}q(\boldsymbol{\sigma}\cdot\boldsymbol{\nabla})(\boldsymbol{\sigma}\cdot\boldsymbol{A}) + \mathrm{i}q(\boldsymbol{\sigma}\cdot\boldsymbol{A})(\boldsymbol{\sigma}\cdot\boldsymbol{\nabla}) + q^2(\boldsymbol{\sigma}\cdot\boldsymbol{A})^2\}\psi.$$

The first and last terms are, respectively, $-\boldsymbol{\nabla}^2$ and $q^2\boldsymbol{A}^2$ where the 2×2 unit matrix $\mathbf{1}$ is understood. The second and third terms are $\mathrm{i}q(\boldsymbol{\sigma}\cdot\boldsymbol{\nabla})(\boldsymbol{\sigma}\cdot\boldsymbol{A}\psi)$ and $\mathrm{i}q(\boldsymbol{\sigma}\cdot\boldsymbol{A})(\boldsymbol{\sigma}\cdot\boldsymbol{\nabla}\psi)$. These may be simplified using the identity of problem 4.4(b), but we must be careful to treat $\boldsymbol{\nabla}$ correctly as a differential operator.

(b) Show that $(\boldsymbol{\sigma}\cdot\boldsymbol{\nabla})(\boldsymbol{\sigma}\cdot\boldsymbol{A})\psi = \boldsymbol{\nabla}\cdot(\boldsymbol{A}\psi) + \mathrm{i}\boldsymbol{\sigma}\cdot\{\boldsymbol{\nabla}\times(\boldsymbol{A}\psi)\}$. Now use $\boldsymbol{\nabla}\times(\boldsymbol{A}\psi) = (\boldsymbol{\nabla}\times\boldsymbol{A})\psi - \boldsymbol{A}\times\boldsymbol{\nabla}\psi$ to simplify the last term.

(c) Similarly, show that $(\boldsymbol{\sigma}\cdot\boldsymbol{A})(\boldsymbol{\sigma}\cdot\boldsymbol{\nabla})\psi = \boldsymbol{A}\cdot\boldsymbol{\nabla}\psi + \mathrm{i}\boldsymbol{\sigma}\cdot(\boldsymbol{A}\times\boldsymbol{\nabla}\psi)$.

(d) Hence verify (3.116).

4

Lorentz Transformations and Discrete Symmetries

In this chapter we shall review various *covariances* (see appendix D) of the KG and Dirac equations, concentrating mainly on the latter. First, we consider Lorentz transformations (rotations and velocity transformations) and show how the scalar KG wavefunction and the 4-component Dirac spinor must transform in order that the respective equations be covariant under these transformations. Then we perform a similar task for the discrete transformations of parity, charge conjugation and time reversal. The results enable us to construct 'bilinear covariants' having well-defined behaviour (scalar, pseudoscalar, vector, etc.) under these transformations. This is essential for later work, for two reasons: first, we shall be able to do dynamical calculations in a way that is manifestly covariant under Lorentz transformations; and secondly we shall be ready to study physical problems in which the discrete transformations are, or are not, actual symmetries of the real world, a topic to which we shall return in the second volume.

4.1 Lorentz transformations

4.1.1 The KG equation

In order to ensure that the laws of physics are the same in all inertial frames, we require our relativistic wave equations to be *covariant* under Lorentz transformations – that is, they must have the same form in the two different frames (see appendix D). In the case of the KG equation

$$(\Box + m^2)\phi(x) = -\mathrm{i}q[\partial_\mu A^\mu(x) + A^\mu(x)\partial_\mu]\phi(x) + q^2A^2(x)\phi(x) \qquad (4.1)$$

for a particle of charge q in the field A^μ, this requirement is taken care of, almost automatically, by the notation. Consider a Lorentz transformation such that $x \to x'$. A^μ will transform by the usual 4-vector transformation law (i.e. like x^μ), which we write as $A^\mu(x) \to A'^\mu(x')$. Similarly we write the transform of ϕ as $\phi(x) \to \phi'(x')$. Then in the primed coordinate frame physics must be described by the equation

$$(\Box' + m^2)\phi'(x') = -\mathrm{i}q[\partial'_\mu A'^\mu(x') + A'^\mu(x')\partial'_\mu]\phi'(x') + q^2A'^2(x')\phi'(x'). \quad (4.2)$$

Now the 4-dimensional dot products appearing in (4.2) are all invariant under the Lorentz transformation, so that (4.2) can be written as

$$(\Box + m^2)\phi'(x') = -\mathrm{i}q[\partial_\mu A^\mu(x) + A^\mu(x)\partial_\mu]\phi'(x') + q^2A^2(x)\phi'(x'), \qquad (4.3)$$

and we see that the wavefunction in the primed frame may be identified (up to a phase) with that in the unprimed frame:

$$\phi'(x') = \phi(x). \qquad (4.4)$$

Equation (4.4) is the condition for the KG equation to be covariant under Lorentz transformations. Since x' is a known function of x, given by the angles and velocities parametrizing the transformation, equation (4.4) enables one to construct the correct function ϕ' which the primed observers must use, in order to be consistent with the unprimed observers.

By way of illustration, consider a rotation of the coordinate system by an angle α in a positive sense about the x-axis; then the position vector referred to the new system is $\boldsymbol{x}' = (x', y', z')$ where

$$\begin{pmatrix} x' \\ y' \\ z' \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\alpha & \sin\alpha \\ 0 & -\sin\alpha & \cos\alpha \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix}, \qquad (4.5)$$

which we shall write as

$$\boldsymbol{x}' = \mathbf{R}_x(\alpha)\,\boldsymbol{x}. \qquad (4.6)$$

Correspondingly, equation (4.4) is, in this case,

$$\phi'(\mathbf{R}_x(\alpha)\,\boldsymbol{x}) = \phi(\boldsymbol{x}), \qquad (4.7)$$

which can also be written as

$$\phi'(\boldsymbol{x}) = \phi(\mathbf{R}_x^{-1}(\alpha)\,\boldsymbol{x}). \qquad (4.8)$$

It is convenient to begin with an 'infinitesimal rotation', where the angle α in (4.5) is replaced by ϵ_x such that $\cos\epsilon_x \approx 1$ and $\sin\epsilon_x \approx \epsilon_x$. Then it is easy to verify that (4.5) becomes

$$\boldsymbol{x}' = \mathbf{R}_x(\epsilon_x)\,\boldsymbol{x} = \boldsymbol{x} - \boldsymbol{\epsilon} \times \boldsymbol{x} \qquad (4.9)$$

where $\boldsymbol{\epsilon} = (\epsilon_x, 0, 0)$. For a general infinitesimal rotation, we simply replace this $\boldsymbol{\epsilon}$ by a general one, $(\epsilon_x, \epsilon_y, \epsilon_z)$. For such a rotation, condition (4.8) becomes

$$\phi'(\boldsymbol{x}) = \phi(\boldsymbol{x} + \boldsymbol{\epsilon} \times \boldsymbol{x}). \qquad (4.10)$$

Expanding the right hand side to first order in $\boldsymbol{\epsilon}$ we obtain

$$\begin{aligned} \phi'(\boldsymbol{x}) &= \phi(\boldsymbol{x}) + (\boldsymbol{\epsilon} \times \boldsymbol{x}) \cdot \boldsymbol{\nabla}\phi = \phi(\boldsymbol{x}) + \boldsymbol{\epsilon} \cdot (\boldsymbol{x} \times \boldsymbol{\nabla})\phi \\ &= (1 + \mathrm{i}\boldsymbol{\epsilon} \cdot \hat{\boldsymbol{L}})\phi(\boldsymbol{x}) \end{aligned} \qquad (4.11)$$

where $\hat{\boldsymbol{L}}$ is the vector angular momentum operator $\boldsymbol{x} \times -\mathrm{i}\boldsymbol{\nabla}$.

The rule for finite rotations may be obtained from the infinitesimal form by using the result

$$\mathrm{e}^{A} = \lim_{n\to\infty}(1 + A/n)^n \tag{4.12}$$

generalized to differential operators (the exponential of a matrix being understood as the infinite series $\exp\boldsymbol{A} = \mathbf{1} + \boldsymbol{A} + \frac{1}{2}\boldsymbol{A}^2 + \dots$). Let $\boldsymbol{\epsilon} = \boldsymbol{\alpha}/n$, where $\boldsymbol{\alpha} = (\alpha_x, \alpha_y, \alpha_z)$ are three real finite parameters; we may think of the direction of $\boldsymbol{\alpha}$ as representing the axis of the rotation, and the magnitude of $\boldsymbol{\alpha}$ as representing the angle of rotation. Then applying the transformation (4.11) n times, and letting n tend to infinity, we obtain for the finite rotation

$$\phi'(\boldsymbol{x}) = \mathrm{e}^{\mathrm{i}\boldsymbol{\alpha}\cdot\hat{\boldsymbol{L}}}\phi(\boldsymbol{x}) \equiv \hat{U}_{\mathrm{R}}(\boldsymbol{\alpha})\phi(\boldsymbol{x}). \tag{4.13}$$

Note that $\hat{U}_{\mathrm{R}}(\boldsymbol{\alpha})$ is a unitary operator, since $\hat{U}_{\mathrm{R}}^{\dagger}$ is the inverse rotation.

Equation (4.13) is, of course, the familiar rule for rotations of scalar wavefunctions, exhibiting the intimate connection between rotations and angular momentum in quantum mechanics. We recall that if a Hamiltonian is invariant under rotations, then the operators $\hat{\boldsymbol{L}}$ commute with the Hamiltonian and angular momentum is conserved.

A similar calculation may be done for velocity transformations ('boosts'), leading to corresponding operators $\hat{\boldsymbol{K}}$ – see problem 4.1.

4.1.2 The Dirac equation

The case of the Dirac equation is more complicated, because (unlike the KG ϕ) the wavefunction has more than one component, corresponding to the fact that it describes a spin-1/2 particle. There is, however, a direct connection between the angular momentum associated with a wavefunction, and the way that the wavefunction transforms under rotations of the coordinate system. To take a simple case, the 2p wavefunctions mentioned in section 3.2 correspond to $l = 1$ on the one hand and, on the other, to the components of a vector – indeed the most basic vector of all, the position vector $\boldsymbol{x} = (x, y, z)$ itself. If we rotate the coordinate system in the way represented by (4.5), the components in the primed system transform into simple linear combinations of the components in the original system.

Very much the same thing happens in the case of spinor wavefunctions, except that they transform in a way different from – though closely related to – that of vectors. In the present section we shall discuss how this works for three-dimensional rotations of the spatial coordinate system, and explain how it generalizes to boosts, which include transformations of the time coordinate as well. It will be convenient to use the alternative representation (3.40) for the Dirac matrices. In this representation, the components ϕ, χ of the free-particle 4-spinor ω of (3.43) satisfy

$$E\phi = \boldsymbol{\sigma}\cdot\boldsymbol{p}\phi + m\chi \tag{4.14}$$

$$E\chi = -\boldsymbol{\sigma}\cdot\boldsymbol{p}\chi + m\phi \tag{4.15}$$

rather than (3.45) and (3.46).

As before, we start with the infinitesimal rotation (4.9). Since $\boldsymbol{p}$ is a vector, it transforms in the same way as $\boldsymbol{x}$, so that under an infinitesimal rotation $\boldsymbol{p}$ becomes $\boldsymbol{p}'$ where

$$\boldsymbol{p}' = \boldsymbol{p} - \boldsymbol{\epsilon} \times \boldsymbol{p}. \tag{4.16}$$

The question for us now is: how do the spinors ϕ and χ transform under this same rotation of the coordinate system?

The essential point is that in the new coordinate system the defining equations (4.14) and (4.15) should take exactly the same form, namely

$$E\phi' = \boldsymbol{\sigma} \cdot \boldsymbol{p}'\phi' + m\chi' \tag{4.17}$$

$$E\chi' = -\boldsymbol{\sigma} \cdot \boldsymbol{p}'\chi' + m\phi' \tag{4.18}$$

where ϕ' and χ' are the spinors in the new coordinate system, and we have used the fact that both E and m do not change under rotations. Our task is to find ϕ' and χ' in terms of ϕ and χ.

Since both ϕ and χ are 2-component spinors, we might guess from (4.11) that the answer is

$$\phi' = (1 + \mathrm{i}\boldsymbol{\sigma} \cdot \boldsymbol{\epsilon}/2)\phi, \qquad \chi' = (1 + \mathrm{i}\boldsymbol{\sigma} \cdot \boldsymbol{\epsilon}/2)\chi, \tag{4.19}$$

since the $\boldsymbol{\sigma}/2$ are the spin-1/2 matrices, taking the place of $\hat{\boldsymbol{L}}$. To check that this is, in fact, the correct transformation law, we proceed as follows.[1] First, multiply (4.14) from the left by the matrix $(1 + \mathrm{i}\boldsymbol{\sigma} \cdot \boldsymbol{\epsilon}/2)$: then, since E and m commute with all matrices, the result is

$$E\phi' = (1 + \mathrm{i}\boldsymbol{\sigma} \cdot \boldsymbol{\epsilon}/2)\boldsymbol{\sigma} \cdot \boldsymbol{p}\phi + m\chi' \tag{4.20}$$

$$= (1 + \mathrm{i}\boldsymbol{\sigma} \cdot \boldsymbol{\epsilon}/2)\boldsymbol{\sigma} \cdot \boldsymbol{p}(1 - \mathrm{i}\boldsymbol{\sigma} \cdot \boldsymbol{\epsilon}/2)\phi' + m\chi' \tag{4.21}$$

where we have used

$$(1 + \mathrm{i}\boldsymbol{\sigma} \cdot \boldsymbol{\epsilon}/2)^{-1} \approx (1 - \mathrm{i}\boldsymbol{\sigma} \cdot \boldsymbol{\epsilon}/2) \tag{4.22}$$

to first order in $\boldsymbol{\epsilon}$. Keeping only first order terms in $\boldsymbol{\epsilon}$, the first term on the right hand side of (4.21) is

$$(\boldsymbol{\sigma} \cdot \boldsymbol{p} + \frac{1}{2}\mathrm{i}\boldsymbol{\sigma} \cdot \boldsymbol{\epsilon}\ \boldsymbol{\sigma} \cdot \boldsymbol{p} - \frac{1}{2}\mathrm{i}\boldsymbol{\sigma} \cdot \boldsymbol{p}\ \boldsymbol{\sigma} \cdot \boldsymbol{\epsilon})\phi'. \tag{4.23}$$

This can be simplified using the result from problem 3.4(b):

$$\boldsymbol{\sigma} \cdot \boldsymbol{a}\ \boldsymbol{\sigma} \cdot \boldsymbol{b} = \boldsymbol{a} \cdot \boldsymbol{b} + \mathrm{i}\boldsymbol{\sigma} \cdot \boldsymbol{a} \times \boldsymbol{b}, \tag{4.24}$$

provided all the components of $\boldsymbol{a}$ and $\boldsymbol{b}$ commute. Applying (4.24), (4.23) becomes

$$[\boldsymbol{\sigma} \cdot \boldsymbol{p} + \frac{\mathrm{i}}{2}(\boldsymbol{\epsilon} \cdot \boldsymbol{p} + \mathrm{i}\boldsymbol{\sigma} \cdot \boldsymbol{\epsilon} \times \boldsymbol{p}) - \frac{\mathrm{i}}{2}(\boldsymbol{\epsilon} \cdot \boldsymbol{p} + \mathrm{i}\boldsymbol{\sigma} \cdot \boldsymbol{p} \times \boldsymbol{\epsilon})]\phi' \tag{4.25}$$

$$= (\boldsymbol{\sigma} \cdot \boldsymbol{p} - \boldsymbol{\sigma} \cdot \boldsymbol{\epsilon} \times \boldsymbol{p})\phi' = \boldsymbol{\sigma} \cdot \boldsymbol{p}'\phi'. \tag{4.26}$$

[1] We shall derive (4.19), and the corresponding rule for velocity transformations, equation (4.42) below, in appendix M of volume 2 using group theory.

Hence (4.21) is just

$$E\phi' = \boldsymbol{\sigma}\cdot\boldsymbol{p}'\phi' + m\chi' \tag{4.27}$$

as required in (4.17). We can similarly check the correctness of the transformation law (4.19) for χ.

The transformation rule for a finite rotation may be obtained from the infinitesimal form by using the result (4.12) applied to matrices. Then for a finite rotation we obtain the result

$$\phi' = \exp(\mathrm{i}\boldsymbol{\sigma}\cdot\boldsymbol{\alpha}/2)\ \phi, \qquad \chi' = \exp(\mathrm{i}\boldsymbol{\sigma}\cdot\boldsymbol{\alpha}/2)\ \chi. \tag{4.28}$$

We note that the behaviour of ϕ and χ under rotations is the same: equation (4.28) is the way all 2-component spinors transform under rotations.

By way of an illustration, consider the case of the finite rotation (4.5). Here $\boldsymbol{\alpha} = (\alpha, 0, 0)$, and the transformation matrix is

$$\exp(\mathrm{i}\sigma_x\alpha/2) = 1 + \mathrm{i}\sigma_x\alpha/2 + \frac{1}{2}(\mathrm{i}\sigma_x\alpha/2)^2 + \ldots. \tag{4.29}$$

Multiplying out the terms in (4.29) and remembering that $\sigma_x^2 = 1$, we see that the transformation matrix is

$$\cos\alpha/2 + \mathrm{i}\sigma_x\sin\alpha/2 = \begin{pmatrix} \cos\alpha/2 & \mathrm{i}\sin\alpha/2 \\ \mathrm{i}\sin\alpha/2 & \cos\alpha/2 \end{pmatrix}. \tag{4.30}$$

This means that the components ϕ_1, ϕ_2 of the spinor ϕ transform according to the rule

$$\phi_1' = \cos\alpha/2\ \phi_1 + \mathrm{i}\sin\alpha/2\ \phi_2 \tag{4.31}$$

$$\phi_2' = \mathrm{i}\sin\alpha/2\ \phi_1 + \cos\alpha/2\ \phi_2, \tag{4.32}$$

for this particular rotation. The transformed components are linear combinations of the original components, but it is the half-angle $\alpha/2$ that enters, not α.

Let us denote the finite transformation matrix by $\boldsymbol{U}$, so that

$$\boldsymbol{U} = \exp(\mathrm{i}\boldsymbol{\sigma}\cdot\boldsymbol{\alpha}/2) \quad \text{and} \quad \boldsymbol{U}^\dagger = \exp(-\mathrm{i}\boldsymbol{\sigma}\cdot\boldsymbol{\alpha}/2). \tag{4.33}$$

It follows that

$$\boldsymbol{U}\boldsymbol{U}^\dagger = \boldsymbol{U}^\dagger\boldsymbol{U} = \mathbf{1}, \tag{4.34}$$

since the rotation parametrized by $-\boldsymbol{\alpha}$ clearly undoes the rotation parametrized by $\boldsymbol{\alpha}$. So $\boldsymbol{U}$ is a 2×2 unitary matrix. It follows that the normalization of ϕ and χ is preserved under rotations: $\phi'^\dagger\phi' = \phi^\dagger\phi$, and $\chi'^\dagger\chi' = \chi^\dagger\chi$. The free-particle Dirac probability density $\rho = \psi^\dagger\psi = \phi^\dagger\phi + \chi^\dagger\chi$ is therefore also (as we expect) invariant under rotations.

More interestingly, we can examine the way the free-particle current density

$$\boldsymbol{j} = \psi^\dagger\boldsymbol{\alpha}\psi = \phi^\dagger\boldsymbol{\sigma}\phi - \chi^\dagger\boldsymbol{\sigma}\chi \tag{4.35}$$

transforms under rotations. Of course, it should behave as a 3-vector, and this is checked in problem 4.2(a).

We now turn to the behaviour of the spinors ϕ and χ under boosts, which mix $\boldsymbol{x}$ and t, or equivalently $\boldsymbol{p}$ and E. For example, consider a Lorentz velocity transformation (boost) from a frame S to a frame S$'$ which is moving with speed u with respect to S along the common x-axis. Then the energy E and momentum p_x of a particle in S are transformed to E' and p'_x in S$'$ where (cf (D.1))

$$E' = \cosh\vartheta\ E - \sinh\vartheta\ p_x \tag{4.36}$$

$$p'_x = \cosh\vartheta\ p_x - \sinh\vartheta\ E, \tag{4.37}$$

where $\cosh\vartheta = (1-u^2)^{-1/2} \equiv \gamma(u)$, and $\sinh\vartheta = \gamma(u)u$. As before, we start with an infinitesimal transformation, where ϑ is replaced by η_x such that $\cosh\eta_x \approx 1$ and $\sinh\eta_x \approx \eta_x$. Then (4.36) and (4.37) become $E' = E - \eta_x p_x$, $p'_x = p_x - \eta_x E$. For the general infinitesimal boost parametrized by $\boldsymbol{\eta} = (\eta_x, \eta_y, \eta_z)$, the transformation law for $(E, \boldsymbol{p})$ is

$$E' = E - \boldsymbol{\eta}\cdot\boldsymbol{p} \tag{4.38}$$

$$\boldsymbol{p}' = \boldsymbol{p} - \boldsymbol{\eta}E. \tag{4.39}$$

Once again, we have to determine ϕ' and χ' such that the transformed versions of (4.14) and (4.15) are

$$(E' - \boldsymbol{\sigma}\cdot\boldsymbol{p}')\phi' = m\chi' \tag{4.40}$$

$$(E' + \boldsymbol{\sigma}\cdot\boldsymbol{p}')\chi' = m\phi'. \tag{4.41}$$

Note that this time E does transform, according to (4.38).

The required ϕ' and χ' are

$$\phi' = (1 - \boldsymbol{\sigma}\cdot\boldsymbol{\eta}/2)\phi, \qquad \chi' = (1 + \boldsymbol{\sigma}\cdot\boldsymbol{\eta}/2)\chi. \tag{4.42}$$

The spinors ϕ and χ behaved the same under rotations, but they transform differently under boosts. There are two kinds of 2-component spinors, ϕ-type and χ-type, in the representation (3.40), which are distinguished by their behaviour under boosts. The group theory behind this will be explained in appendix M of volume 2.

To verify the rule (4.42), take equation (4.14) in the form (4.40) and multiply from the left by the matrix $(1 + \boldsymbol{\sigma}\cdot\boldsymbol{\eta}/2)$, to obtain

$$(1 + \boldsymbol{\sigma}\cdot\boldsymbol{\eta}/2)(E - \boldsymbol{\sigma}\cdot\boldsymbol{p})\phi = m\chi', \tag{4.43}$$

or equivalently

$$(1 + \boldsymbol{\sigma}\cdot\boldsymbol{\eta}/2)(E - \boldsymbol{\sigma}\cdot\boldsymbol{p})(1 + \boldsymbol{\sigma}\cdot\boldsymbol{\eta}/2)\phi' = m\chi', \tag{4.44}$$

where we have used $(1 - \boldsymbol{\sigma}\cdot\boldsymbol{\eta}/2)^{-1} \approx (1 + \boldsymbol{\sigma}\cdot\boldsymbol{\eta}/2)$. For (4.44) to be consistent with (4.40) we require

$$(1 + \boldsymbol{\sigma}\cdot\boldsymbol{\eta}/2)(E - \boldsymbol{\sigma}\cdot\boldsymbol{p})(1 + \boldsymbol{\sigma}\cdot\boldsymbol{\eta}/2) = E' - \boldsymbol{\sigma}\cdot\boldsymbol{p}'. \tag{4.45}$$

Keeping only first order terms in $\boldsymbol{\eta}$, the left hand side of (4.45) is

$$E - \boldsymbol{\sigma}\cdot\boldsymbol{p} + E\boldsymbol{\sigma}\cdot\boldsymbol{\eta} - \frac{1}{2}(\boldsymbol{\sigma}\cdot\boldsymbol{p}\;\boldsymbol{\sigma}\cdot\boldsymbol{\eta} + \boldsymbol{\sigma}\cdot\boldsymbol{\eta}\;\boldsymbol{\sigma}\cdot\boldsymbol{p}) \tag{4.46}$$

$$= E - \boldsymbol{\eta}\cdot\boldsymbol{p} - \boldsymbol{\sigma}\cdot(\boldsymbol{p} - \boldsymbol{\eta}E) \tag{4.47}$$

$$= E' - \boldsymbol{\sigma}\cdot\boldsymbol{p}' \tag{4.48}$$

as required for the right hand side of (4.45).

For a finite boost ϕ and χ transform by the 'exponentiation' of (4.42), namely

$$\phi' = \exp(-\boldsymbol{\sigma}\cdot\boldsymbol{\vartheta}/2)\;\phi, \qquad \chi' = \exp(\boldsymbol{\sigma}\cdot\boldsymbol{\vartheta}/2)\;\chi \tag{4.49}$$

where the three real parameters $\boldsymbol{\vartheta} = (\vartheta_x, \vartheta_y, \vartheta_z)$ specify the direction and magnitude of the boost. In contrast to (4.28), the transformations (4.49) are *not* unitary. If we denote the matrix $\exp(-\boldsymbol{\sigma}\cdot\boldsymbol{\vartheta}/2)$ by $\boldsymbol{B}$, we have $\boldsymbol{B} = \boldsymbol{B}^\dagger$ rather than $\boldsymbol{B}^{-1} = \boldsymbol{B}^\dagger$. So $\boldsymbol{B}$ does not leave $\phi^\dagger\phi$ and $\chi^\dagger\chi$ invariant. Actually this is no surprise. We already know from section 4.1.2 that the density $\phi^\dagger\phi + \chi^\dagger\chi$ ought to transform as the fourth component ρ of the 4-vector $j^\mu = (\rho, \boldsymbol{j})$. Let us check this for our infinitesimal boost:

$$\begin{aligned}
\rho' &= \phi'^\dagger\phi' + \chi'^\dagger\chi' \\
&= \phi^\dagger(1 - \boldsymbol{\sigma}\cdot\boldsymbol{\eta}/2)(1 - \boldsymbol{\sigma}\cdot\boldsymbol{\eta}/2)\phi + \chi^\dagger(1 + \boldsymbol{\sigma}\cdot\boldsymbol{\eta}/2)(1 + \boldsymbol{\sigma}\cdot\boldsymbol{\eta}/2)\;\chi \\
&= \phi^\dagger\phi + \chi^\dagger\chi - \phi^\dagger\boldsymbol{\sigma}\phi\cdot\boldsymbol{\eta} + \chi^\dagger\boldsymbol{\sigma}\chi\cdot\boldsymbol{\eta} \\
&= \rho - \boldsymbol{\eta}\cdot\boldsymbol{j}
\end{aligned} \tag{4.50}$$

as required by (4.38). Similarly, it may be verified (problem 4.2(b)) that $\boldsymbol{j}$ transforms as the 3-vector part of the 4-vector j^μ, under this infinitesimal boost.

On the other hand, the products $\phi^\dagger\chi$ and $\chi^\dagger\phi$ are clearly invariant under the transformation (4.49), since the exponential factors cancel. This means that the quantity $\omega^\dagger\beta\omega$ is a Lorentz invariant.

At this point it is beginning to be clear that a more 'covariant-looking' notation would be very desirable. In the case of the KG probability current, the 4-vector index μ was clearly visible in the expression on the right-hand side of (3.20), but there is nothing similar in the Dirac case so far. In problem 4.3 the four 'γ matrices' are introduced, defined by $\gamma^\mu = (\gamma^0, \boldsymbol{\gamma})$ with $\gamma^0 = \beta$ and $\boldsymbol{\gamma} = \beta\boldsymbol{\alpha}$, together with the quantity $\bar{\psi} \equiv \psi^\dagger\gamma^0$, in terms of which the Dirac ρ of (3.51) and $\boldsymbol{j}$ of (3.57) can be written as $\bar{\psi}(x)\gamma^0\psi(x)$ and $\bar{\psi}(x)\boldsymbol{\gamma}\psi(x)$ respectively. The complete Dirac 4-current is then

$$j^\mu = \bar{\psi}(x)\gamma^\mu\psi(x). \tag{4.51}$$

For free particle solutions, we (and problem 4.2) have established that j^μ of (4.51) indeed transforms as a 4-vector under infinitesimal rotations and boosts. We have also just seen that the quantity $\bar{\psi}\psi$ is an invariant.

We end this section by illustrating the use of the finite boost transformations (4.49). Consider two frames S and S', such that in S a particle is at rest

with $E = m, \boldsymbol{p} = \mathbf{0}$, and with spin up along the z-axis; in S', the particle has energy E', momentum $\boldsymbol{p}' = (0, 0, p')$, and spin up along the z-axis. If we apply a boost such that S' has velocity $(0, 0, -v')$ relative to S, where $v' = p'/E'$, then E and $\boldsymbol{p}$ become

$$E' = \cosh\vartheta' E = m\gamma(v') \tag{4.52}$$

$$p' = \sinh\vartheta' E = mv'\gamma(v') \tag{4.53}$$

as required. Now consider the forms of the 4-spinors in S and S'. In S, from (4.14) and (4.15) we have simply $\phi = \chi$, and if we normalize such that $\bar{u}u = 2m$ we may take

$$u_S = \sqrt{m}\begin{pmatrix} \phi_+ \\ \phi_+ \end{pmatrix}, \qquad \phi_+ = \begin{pmatrix} 1 \\ 0 \end{pmatrix}. \tag{4.54}$$

In S' the spinor is

$$u_{S'} = N\begin{pmatrix} \phi_+ \\ \left(\frac{E'-\sigma_z p'}{m}\right)\phi_+ \end{pmatrix} = N\begin{pmatrix} \phi_+ \\ \left(\frac{E'-p'}{m}\right)\phi_+ \end{pmatrix} \tag{4.55}$$

where the normalization N is determined (since $\bar{u}u$ is invariant) from the condition $\bar{u}_{S'}u_{S'} = 2m$ to be $N = (E' + p')^{1/2}$, giving

$$u_{S'} = \begin{pmatrix} (E'+p')^{1/2}\ \phi_+ \\ (E'-p')^{1/2}\ \phi_+ \end{pmatrix}. \tag{4.56}$$

But we can also calculate $u_{S'}$ by applying the transformation (4.49) with $\tanh\vartheta' = -v'$ to u_S. Then the upper two components become

$$\phi' = \sqrt{m}\ \mathrm{e}^{\sigma_z\vartheta'/2}\phi_+ = \sqrt{m}\ \mathrm{e}^{\vartheta'/2}\phi_+, \tag{4.57}$$

while the lower two components become

$$\chi' = \sqrt{m}\ \mathrm{e}^{-\vartheta'/2}\phi_+. \tag{4.58}$$

Now we can write

$$\mathrm{e}^{\vartheta'/2} = (\mathrm{e}^{\vartheta'})^{1/2} = (\cosh\vartheta' + \sinh\vartheta')^{1/2} = \left(\frac{E'+p'}{m}\right)^{1/2} \tag{4.59}$$

and

$$\mathrm{e}^{-\vartheta'/2} = \left(\frac{E'-p'}{m}\right)^{1/2}; \tag{4.60}$$

and so we recover (4.56).

4.2 Discrete transformations: P, C and T

The transformations we considered in section 4.1 are known as 'continuous', because the parameters involved (angles, speeds) vary continuously. This is essentially the reason we were able to build up finite transformations from infinitesimal ones, which differ only slightly from the identity transformation: finite transformations could be reached continuously from the identity. But there is another class of transformations, called 'discrete', which cannot be reached continuously from the identity. Examples of discrete transformations are parity (or space inversion), charge conjugation, and time reversal, and their combinations. Although these discrete transformations are important primarily in weak interactions, which we shall not cover until the second volume, it is useful to discuss the behaviour of Dirac wavefunctions under discrete transformations at this stage. Among other things, more light will be cast on antiparticles.

4.2.1 Parity

The parity (or space inversion) transformation **P** is defined by

$$\mathbf{P} : \boldsymbol{x} \to \boldsymbol{x}' = -\boldsymbol{x}, \quad t \to t; \tag{4.61}$$

that is, **P** inverts the spatial coordinates. It follows that **P** also inverts momenta ($\boldsymbol{p} \to -\boldsymbol{p}$) but does not change angular momenta ($\boldsymbol{x} \times \boldsymbol{p} \to \boldsymbol{x} \times \boldsymbol{p}$) or spin ($\boldsymbol{\sigma} \to \boldsymbol{\sigma}$). We already see that there are two kinds of 3-vectors: polar 3-vectors which change sign under **P** and axial vectors which do not. For example, the electric field $\boldsymbol{E}$ and the vector potential $\boldsymbol{A}$ are polar vectors, while the magnetic field $\boldsymbol{B}$ is an axial vector. There are also scalar quantities (such as $\boldsymbol{x} \cdot \boldsymbol{p}$) which do not change sign under **P**, and pseudoscalar quantities (such as $\boldsymbol{\sigma} \cdot \boldsymbol{p}$) which do.

Consider first the KG equation (4.1). Since $\boldsymbol{A}$ is a polar vector, it changes sign under parity, as does $\boldsymbol{\nabla}$, while both $\partial/\partial t$ and A^0 remain the same. The scalar products $\partial_\mu A^\mu$ and $A^\mu \partial_\mu$ are therefore invariant under parity, as are $\Box$ and A^2. Hence we may identify $\phi_{\mathbf{P}}(x') = \phi(x)$, or equivalently

$$\phi_{\mathbf{P}}(\boldsymbol{x}) = \phi(-\boldsymbol{x}) \equiv \hat{\mathbf{P}}_0 \phi(\boldsymbol{x}), \tag{4.62}$$

where $\hat{\mathbf{P}}_0$ is the coordinate inversion operator. Note that we are calling the transformed wavefunction $\phi_{\mathbf{P}}$ rather than yet another ϕ' since we need to keep track of what transformation we are considering. If we take $\phi(\boldsymbol{x})$ to be a positive-energy free particle solution with energy E and momentum $\boldsymbol{p}$, $\phi_{\mathbf{P}}$ will describe a positive energy particle with momentum $-\boldsymbol{p}$, as we expect.

Now let us study the covariance of the free particle Dirac equation

$$\mathrm{i}\frac{\partial \psi(\boldsymbol{x}, t)}{\partial t} = -\mathrm{i}\boldsymbol{\alpha} \cdot \boldsymbol{\nabla}\psi(\boldsymbol{x}, t) + \beta m \psi(\boldsymbol{x}, t) \tag{4.63}$$

under $\mathbf{P}$. Equation (4.63) will be covariant under (4.61) if we can find a wavefunction $\psi_{\mathbf{P}}(\boldsymbol{x}', t)$ for observers using the transformed coordinate system such that their Dirac equation has exactly the same form in their system as (4.63):

$$\mathrm{i}\frac{\partial \psi_{\mathbf{P}}}{\partial t}(\boldsymbol{x}', t) = -\mathrm{i}\boldsymbol{\alpha} \cdot \boldsymbol{\nabla}' \psi_{\mathbf{P}}(x', t) + \beta m \psi_{\mathbf{P}}(\boldsymbol{x}', t). \tag{4.64}$$

Now we know that $\boldsymbol{\nabla}' = -\boldsymbol{\nabla}$, since $\boldsymbol{x}' = -\boldsymbol{x}$. Hence (4.64) becomes

$$\mathrm{i}\frac{\partial \psi_{\mathbf{P}}}{\partial t}(\boldsymbol{x}', t) = \mathrm{i}\boldsymbol{\alpha} \cdot \boldsymbol{\nabla} \psi_{\mathbf{P}}(\boldsymbol{x}', t) + \beta m \psi_{\mathbf{P}}(\boldsymbol{x}', t). \tag{4.65}$$

Multiplying this equation from the left by β and using $\beta\boldsymbol{\alpha} = -\boldsymbol{\alpha}\beta$ we find

$$\frac{\mathrm{i}\partial}{\partial t}[\beta \psi_{\mathbf{P}}(\boldsymbol{x}', t)] = -\mathrm{i}\boldsymbol{\alpha} \cdot \boldsymbol{\nabla}[\beta \psi_{\mathbf{P}}(\boldsymbol{x}', t)] + \beta m [\beta \psi_{\mathbf{P}}(\boldsymbol{x}', t)]. \tag{4.66}$$

Comparing (4.66) and (4.63), it follows that we may consistently translate between ψ and $\psi_{\mathbf{P}}$ using the relation

$$\psi(\boldsymbol{x}, t) = \beta \psi_{\mathbf{P}}(-\boldsymbol{x}, t), \tag{4.67}$$

or equivalently

$$\psi_{\mathbf{P}}(\boldsymbol{x}, t) = \beta \psi(-\boldsymbol{x}, t) \equiv \beta \hat{\mathbf{P}}_0 \psi(\boldsymbol{x}, t). \tag{4.68}$$

Equation (4.68) is the required relation between the wavefunctions in the two systems; it may be compared to (4.4) and (4.62).

In principle we could include an arbitrary phase factor $\eta_{\mathbf{P}}$ on the right hand of (4.68) and (4.62); such a phase leaves the normalization of ϕ and ψ, and all bilinears of the form $\bar{\psi}$ (gamma matrix) ψ unaltered. The possibility of such a phase factor did not arise in the case of Lorentz transformations, since for infinitesimal ones the transformed ψ' and the original ψ differ only infinitesimally (not by a finite phase factor). But the parity transformation cannot be built up out of infinitesimal steps – the coordinate system is either reflected or it is not. We will choose $\eta_{\mathbf{P}} = 1$.

As an example of (4.68), consider the free particle solutions in the standard form (3.41), (3.72):

$$\psi(\boldsymbol{x}, t) = N \begin{pmatrix} \phi \\ \frac{\boldsymbol{\sigma} \cdot \boldsymbol{p}}{E+m} \phi \end{pmatrix} \exp(-\mathrm{i}Et + \mathrm{i}\boldsymbol{p} \cdot \boldsymbol{x}). \tag{4.69}$$

Then

$$\psi_{\mathbf{P}}(\boldsymbol{x}, t) = \beta \psi(-\boldsymbol{x}, t) = N \begin{pmatrix} \phi \\ \frac{-\boldsymbol{\sigma} \cdot \boldsymbol{p}}{E+m} \phi \end{pmatrix} \exp(-\mathrm{i}Et - \mathrm{i}\boldsymbol{p} \cdot \boldsymbol{x}) \tag{4.70}$$

which can be conveniently summarized by the simple statement that the three-momentum $\boldsymbol{p}$ as seen in the parity transformed system is minus that in the original one, as expected. Note that $\boldsymbol{\sigma}$ does not change sign.

It is also interesting to look at the behaviour of the spinors ϕ and χ in the representation (3.40), where they satisfy the equations (4.14) and (4.15). Under parity $\boldsymbol{p} \to -\boldsymbol{p}$, so we can immediately see that $\phi_{\mathbf{P}} = \chi$ and $\chi_{\mathbf{P}} = \phi$. Thus the 2-component spinors ϕ and χ are (in this representation) interchanged under parity.

The analysis leading to (4.68) may be extended to the case of the Dirac equation (3.102) for a particle of charge q in the field A^μ. As already noted, $\boldsymbol{A}$ is a polar vector, transforming under like $\boldsymbol{x}$ or $\boldsymbol{\nabla}$; the scalar potential A^0 is invariant under parity. The combination $(-\mathrm{i}\boldsymbol{\nabla} - q\boldsymbol{A})$ therefore changes sign under parity, and the manipulations following (4.65) proceed as before.

We may introduce a corresponding parity operator $\hat{\mathbf{P}}$, which is unitary and acts on wavefunctions so as to change ψ into $\psi_{\mathbf{P}}$; then

$$\hat{\mathbf{P}}\psi(\boldsymbol{x}, t) = \beta\psi(-\boldsymbol{x}, t) = \beta\hat{\mathbf{P}}_0\psi(\boldsymbol{x}, t), \tag{4.71}$$

so that

$$\hat{\mathbf{P}} = \beta\hat{\mathbf{P}}_0. \tag{4.72}$$

Applying $\hat{\mathbf{P}}$ twice, we find

$$\hat{\mathbf{P}}^2\psi(\boldsymbol{x}, t) = \psi(\boldsymbol{x}, t) \tag{4.73}$$

which implies that the eigenvalues of $\hat{\mathbf{P}}$ are ± 1.

For example, the positive energy rest-frame spinors ((3.73) with $\boldsymbol{p} = \mathbf{0}$)) are eigenstates of $\hat{\mathbf{P}}$ with eigenvalue $+1$, and the negative energy rest-frame spinors are eigenstates of $\hat{\mathbf{P}}$ with eigenvalue -1. Such rest-frame eigenvalues of $\hat{\mathbf{P}}$ are called intrinsic parities. The correspondence between negative energy solutions and antiparticles, discussed in the preceding section, then suggests that a fermion and its antiparticle have opposite intrinsic parity (note that the parity eigenvalue is multiplicative). We shall be able to derive this result after quantization of the Dirac field, in chapter 7.

As usual in quantum mechanics, we may consider the action of $\hat{\mathbf{P}}$ on operators as well as wavefunctions. In particular, the parity transform of a Dirac Hamiltonian $\hat{H}(\boldsymbol{x})$ will be

$$\hat{\mathbf{P}}\hat{H}(\boldsymbol{x})\hat{\mathbf{P}}^\dagger = \beta\hat{\mathbf{P}}_0\hat{H}(\boldsymbol{x})\hat{\mathbf{P}}_0^\dagger\beta. \tag{4.74}$$

If the Hamiltonian is invariant under parity, the right hand side of (4.74) will equal $\hat{H}$ and the operator $\hat{\mathbf{P}}$ will commute with $\hat{H}$; the eigenvalue of $\hat{\mathbf{P}}$ will then be conserved. The reader may easily check that the Hamiltonian for the charged particle in a field A^μ is parity invariant, using $\hat{\mathbf{P}}_0\boldsymbol{A}\hat{\mathbf{P}}_0^\dagger = -\boldsymbol{A}$.

With the rule (4.68) in hand, we can examine how various *bilinear covariants*, such as $\bar{\psi}\psi$ or $\bar{\psi}\gamma^\mu\psi$, transform under parity. For example,

$$\bar{\psi}_{\mathbf{P}}(\boldsymbol{x}', t)\psi_{\mathbf{P}}(\boldsymbol{x}', t) = \psi^\dagger(\boldsymbol{x}, t)\beta\beta\beta\psi(\boldsymbol{x}, t) = \bar{\psi}(\boldsymbol{x}, t)\psi(\boldsymbol{x}, t), \tag{4.75}$$

showing that $\bar{\psi}\psi$ is a scalar. Similarly, for a 4-vector

$$v^\mu(\boldsymbol{x}, t) = (v^0(\boldsymbol{x}, t), \boldsymbol{v}(\boldsymbol{x}, t)) = \bar{\psi}(\boldsymbol{x}, t)\gamma^\mu\psi(\boldsymbol{x}, t), \tag{4.76}$$

the reader may check in problem 4.4(a) that v^0 is a scalar and $\boldsymbol{v}$ is a polar vector.

More interesting possibilities emerge when we introduce a new γ-matrix, γ_5, defined by

$$\gamma_5 = \mathrm{i}\gamma^0\gamma^1\gamma^2\gamma^3. \tag{4.77}$$

This matrix has the defining property that it anticommutes with the γ^μ matrices:

$$\{\gamma_5, \gamma^\mu\} = 0. \tag{4.78}$$

Consider now the quantity $p(\boldsymbol{x}, t) \equiv \bar{\psi}(\boldsymbol{x}, t)\gamma_5\psi(\boldsymbol{x}, t)$. We find

$$\bar{\psi}_{\mathbf{P}}(\boldsymbol{x}', t)\gamma_5\psi_{\mathbf{P}}(\boldsymbol{x}', t) = \psi^\dagger(\boldsymbol{x}, t)\beta\gamma_5\beta\psi(\boldsymbol{x}, t) = -\bar{\psi}(\boldsymbol{x}, t)\psi(\boldsymbol{x}, t), \tag{4.79}$$

so that $p(\boldsymbol{x}, t)$ is a pseudoscalar. Similarly, the reader may verify in problem 4.4(b) that the quantity $a^\mu(\boldsymbol{x}, t) \equiv \bar{\psi}(\boldsymbol{x}, t)\gamma_5\gamma^\mu\psi(\boldsymbol{x}, t)$ transforms under (infinitesimal) rotations and boosts as a 4-vector, but that under parity $a^0(\boldsymbol{x}, t)$ is a pseudoscalar and $\boldsymbol{a}(\boldsymbol{x}, t)$ is an axial vector.

Matrix elements formed from v^μ and a^μ would have to be Lorentz invariant, of the form $v_\mu v^\mu$, $a_\mu a^\mu$, or $v_\mu a^\mu$. For the first of these, we find (shortening the notation)

$$v_{\mathbf{P}\mu}v^\mu_{\mathbf{P}} = v^0v^0 - (-\boldsymbol{v})\cdot(-\boldsymbol{v}) = v_\mu v^\mu, \tag{4.80}$$

and similarly $a_{\mathbf{P}\mu}a^\mu_{\mathbf{P}} = a_\mu a^\mu$. Thus both of these matrix elements are scalars, taking the same form in both systems. However, this is not true of $v_\mu a^\mu$:

$$v_{\mathbf{P}\mu}a^\mu_{\mathbf{P}} = v^0(-a^0) - (-\boldsymbol{v})\cdot(\boldsymbol{a}) = -v_\mu a^\mu, \tag{4.81}$$

showing that this quantity is a pseudoscalar, changing sign when we change systems. By itself, such a sign change would be irrelevant, since observables will depend on the modulus squared of the matrix element. If, however, the matrix element for a process has the form $(v_\mu - a_\mu)(v^\mu - a^\mu)$, for example, where both scalar and pseudoscalar parts are present, then the physics in one coordinate system and in the parity-transformed system will not be the same. One says 'parity is violated': only one of the systems can represent the real world; parity is conserved if physics in the two coordinate systems is the same.

Lee and Yang (1956) were the first to point out that, while there was strong evidence for parity conservation in strong and electromagnetic interactions, its status in weak interactions was at that time untested. They proposed that a clear signal of parity violation could be found in weak decays from initially polarized states (i.e. $< \boldsymbol{s} > \neq \boldsymbol{0}$): if the distribution of final state particles depends on odd powers of the cosine of the angle between the initial spin direction and the final momentum, then parity is violated (note that $< \boldsymbol{s} > \cdot \boldsymbol{p}$ is a pseudoscalar). The first experiment to demonstrate parity violation was performed by Wu *et al.* (1957), using the β-decay of polarized ^{60}Co. Lee and Yang (1956) also remarked that parity violation in the decay

$$\pi^+ \to \mu^+ + \nu_\mu \tag{4.82}$$

implies that the spin of the muon will be polarized along the direction of its momentum, and furthermore that the angular distribution of positrons in the subsequent decay

$$\mu^+ \to \mathrm{e}^+ + \bar{\nu}_\mu + \nu_\mathrm{e} \tag{4.83}$$

would (as in the ^{60}Co experiment) serve as an analyser. This suggestion was quickly confirmed by Garwin *et al.* (1957) and by Friedman and Telegdi (1957); in the rest frame of the pion, the μ^+ spin is aligned opposite to its momentum, a situation that would be reversed in the parity transformed frame.

The end result of many years of research was to establish that the currents responsible for weak interactions of quarks and leptons have precisely the '$v^\mu - a^\mu$' structure, leading to the observed parity violation (see volume 2).

4.2.2 Charge conjugation

Dirac's hole theory led him to the remarkable prediction of the positron, and suggested a new kind of symmetry: to each charged spin-1/2 particle there must correspond an antiparticle with the opposite charge and the same mass. Feynman's interpretation of the negative energy solutions of the KG and Dirac equations assumes that this symmetry holds for both bosons and fermions. We now explore the idea of particle-antiparticle symmetry more formally.

We begin with the KG equation for a spin-0 particle of mass m and charge q in an electromagnetic field A^μ, namely equation (4.1). Inspection of this equation shows at once that the wave function $\phi_\mathbf{C}$ of a particle with the same mass and charge $-q$ is related to the original wavefunction ϕ by

$$\phi_\mathbf{C} = \eta_\mathbf{C}\phi^* \tag{4.84}$$

where $\eta_\mathbf{C}$ is an arbitrary phase factor which we shall take to be unity. Equation (4.84) tells us how to connect the solutions of the particle (charge q) and antiparticle (charge $-q$) equations. When applied to free-particle solutions of the KG equation, the transformation (4.84) relates positive and negative 4-momentum solutions, as expected in the Feynman interpretation of the latter.

We may extend the transformation (4.84) to a symmetry operation for the KG equation (4.1) if we introduce an operation which changes the sign of A^μ. Then the combined operation 'take the complex conjugate of ϕ and change A^μ to $-A^\mu$' is a formal symmetry of (4.84), in the sense that the wavefunction ϕ^* in the field $-A^\mu$ satisfies exactly the same equation as does the wavefunction ϕ in the field A^μ. Of course, we have just seen that ϕ^* is the antiparticle wavefunction, so it is no surprise that the dynamics of the antiparticle in a field $-A^\mu$ is the same as that of the particle in a field A^μ. Still, this is symmetry of the KG equation, which we will call charge conjugation, denoted by **C**:

$$\mathbf{C} : \phi \to \phi_\mathbf{C} = \phi^*, \qquad A^\mu \to A^\mu_\mathbf{C} = -A^\mu. \tag{4.85}$$

We can ask: how does the electromagnetic current behave under this transformation? The expression for the KG current is found by multiplying the free-particle probability current by the charge q, and by replacing ∂^μ by the gauge-invariant operator $D^\mu = \partial^\mu + \mathrm{i}qA^\mu$. This leads to

$$\begin{aligned} j^\mu_{\mathrm{KG\ em}}(\phi, A^\mu) &= \mathrm{i}q\{\phi^*(\partial^\mu + \mathrm{i}qA^\mu)\phi - [(\partial^\mu + \mathrm{i}qA^\mu)\phi]^*\phi\} \\ &= \mathrm{i}q[\phi^*\partial^\mu\phi - (\partial^\mu\phi^*)\phi] - 2q^2A^\mu\phi^*\phi. \end{aligned} \tag{4.86}$$

The current for $\phi_{\mathbf{C}}, A^\mu_{\mathbf{C}}$ is then

$$\begin{aligned} j^\mu_{\mathrm{KG\ em}}(\phi_{\mathbf{C}}, A^\mu_{\mathbf{C}}) &= \mathrm{i}q[\phi^*_{\mathbf{C}}\partial^\mu\phi_{\mathbf{C}} - (\partial^\mu\phi^*_{\mathbf{C}})\phi_{\mathbf{C}}] - 2q^2A^\mu_{\mathbf{C}}\phi^*_{\mathbf{C}}\phi_{\mathbf{C}} \\ &= \mathrm{i}q[\phi\ \partial^\mu\phi^* - (\partial^\mu\phi)\phi^*] + 2q^2A^\mu\phi\ \phi^* \\ &= -j^\mu_{\mathrm{KG\ em}}(\phi, A^\mu). \end{aligned} \tag{4.87}$$

As we would hope, the KG current changes sign under $\mathbf{C}$.

Now consider the Dirac equation for a particle of mass m and charge q in a field A^μ, which we write in the form

$$\frac{\partial\psi}{\partial t} = (-\boldsymbol{\alpha}\cdot\boldsymbol{\nabla} + \mathrm{i}q\boldsymbol{\alpha}\cdot\boldsymbol{A} - \mathrm{i}\beta m - \mathrm{i}qA^0)\psi. \tag{4.88}$$

We want to relate solutions of this equation to the solution $\psi_{\mathbf{C}}$ of the same equation with q replaced by $-q$. As in the KG case, we begin by writing down the complex conjugate equation,

$$\begin{aligned} \frac{\partial\psi^*}{\partial t} &= (-\alpha_1\partial^1 + \alpha_2\partial^2 - \alpha_3\partial^3 \\ &\quad - \mathrm{i}q\alpha_1\partial^1 + \mathrm{i}q\alpha_2\partial^2 - \mathrm{i}q\alpha_3\partial^3 + \mathrm{i}\beta m + \mathrm{i}qA^0)\psi^* \end{aligned} \tag{4.89}$$

where we have used the fact that α_1, α_3 and β are real and α_2 is pure imaginary, which is the case in both the standard representation of the Dirac matrices, and the representation (3.40). Now imagine multiplying (4.89) from the left by a matrix $\mathbf{c}$, with the properties that it commutes with α_1 and α_3, but anticommutes with α_2 and β. Then (4.89) will become

$$\mathbf{c}\frac{\partial\psi^*}{\partial t} = (-\boldsymbol{\alpha}\cdot\boldsymbol{\nabla} - \mathrm{i}q\boldsymbol{\alpha}\cdot\boldsymbol{A} - \mathrm{i}\beta m + \mathrm{i}qA^0)\ \mathbf{c}\psi^* \tag{4.90}$$

which is just (4.88) with q replaced by $-q$. So we may identify the charge-conjugate Dirac wavefunction as

$$\psi_{\mathbf{C}} = \eta_{\mathbf{C}}\ \mathbf{c}\psi^* \tag{4.91}$$

where $\eta_{\mathbf{C}}$ is the usual arbitrary phase factor. The required $\mathbf{c}$ is

$$\mathbf{c} = \beta\alpha_2 = \gamma^2 \tag{4.92}$$

as the reader may easily verify. It is customary to choose $\eta_{\mathbf{C}} = \mathrm{i}$, and so finally the connection between $\psi_{\mathbf{C}}$ and ψ is

$$\psi_{\mathbf{C}}(x) = \mathbf{C}_0\psi^*(x), \quad \text{where} \quad \mathbf{C}_0 = \mathrm{i}\gamma^2. \tag{4.93}$$

Let us look at the effect of the transformation (4.93) on free-particle solutions of the Dirac equation. Referring to (3.73) we find that a positive energy spinor is transformed to

$$\begin{aligned} u_{\mathbf{C}}(p,s) &= (E+m)^{1/2}\ \mathrm{i}\gamma^2 \begin{pmatrix} \phi^{s*} \\ \frac{\boldsymbol{\sigma}^*\cdot\boldsymbol{p}}{E+m}\ \phi^{s*} \end{pmatrix} \\ &= (E+m)^{1/2} \begin{pmatrix} \frac{\boldsymbol{\sigma}\cdot\boldsymbol{p}}{E+m}(-\mathrm{i}\sigma_2\phi^{s*}) \\ -\mathrm{i}\sigma_2\phi^{s*} \end{pmatrix}, \end{aligned} \tag{4.94}$$

where we have used $\sigma_2^* = -\sigma_2$, $\sigma_2\sigma_1 = -\sigma_1\sigma_2$ and $\sigma_2\sigma_3 = -\sigma_3\sigma_2$. The 4-spinor (4.94) is a *negative* energy solution $v(p,s)$ as in (3.82), identifying $-\mathrm{i}\sigma_2\phi^{s*}$ with χ^s. Accordingly we have shown that

$$u_{\mathbf{C}}(p,s) = v(p,s). \tag{4.95}$$

Similarly, as the reader may check,

$$v_{\mathbf{C}}(p,s) = \mathrm{i}\gamma^2 v^*(p,s) = u(p,s). \tag{4.96}$$

So from a positive energy free-particle spinor associated with 4-momentum p and spin s the transformation (4.93) produces a negative energy free-particle spinor associated with the same 4-momentum and spin, and vice versa: that is, u and v are charge-conjugate spinors.

At this point we may wonder if it is possible to construct a *self-conjugate* 4-spinor. Such a spinor would be appropriate for a fermionic particle which is the same as its antiparticle – that is, for a *Majorana fermion*, so named after Ettore Majorana who first raised this possibility (Majorana 1937). To pursue this idea, it is convenient to use the representation (3.40) for the Dirac matrices again, in order to keep track of the Lorentz transformation property of the Majorana spinor. Consider the 4-spinor

$$\omega_{\mathrm{M}} = \begin{pmatrix} \phi \\ \mathrm{i}\sigma_2\phi^* \end{pmatrix}. \tag{4.97}$$

Then

$$\omega_{\mathrm{M}\mathbf{C}} = \mathrm{i}\gamma^2\omega_{\mathrm{M}}^* = \begin{pmatrix} 0 & -\mathrm{i}\sigma_2 \\ \mathrm{i}\sigma_2 & 0 \end{pmatrix} \begin{pmatrix} \phi^* \\ \mathrm{i}\sigma_2\phi \end{pmatrix} = \begin{pmatrix} \phi \\ \mathrm{i}\sigma_2\phi^* \end{pmatrix} = \omega_{\mathrm{M}}, \tag{4.98}$$

so that indeed ω_{M} is self-conjugate. The Lorentz transformation property of ω_{M} is consistent, since we may easily show (problem 4.4(c)) that the 2-spinor $\sigma_2\phi^*$ transforms as a χ-type spinor. The reader can construct a similar self-conjugate 4-spinor using χ rather than ϕ.

A self-conjugate fermion has to carry no distinguishing quantum number, such as electromagnetic charge. The only known neutral fermions are the neutrinos, and until quite recently it was assumed that they are Dirac fermions, with distinct antiparticles (the relevant distinguishing quantum number being

lepton number). However, as we shall see in volume 2, owing to their very small mass, it is hard to discriminate between the two possibilities (Majorana and Dirac) for neutrinos, and a definitive answer will have to await the result of a crucial experiment, the search for neutrinoless double beta decay, which is only possible for Majorana neutrinos.

Returning to more conventional matters, we extend (as in the KG case) the transformation (4.93) to a formal symmetry of the Dirac equation by including the sign change of A^μ, so that $\mathbf{C}$ for the Dirac equation is

$$\mathbf{C} : \psi \to \psi_{\mathbf{C}} = \mathrm{i}\gamma^2\psi^*, \qquad A^\mu \to -A^\mu. \tag{4.99}$$

We now examine how the electromagnetic current behaves under $\mathbf{C}$ in the Dirac case. The Dirac charge density is the probability density $\psi^\dagger\psi$ multiplied by the charge q, and the electromagnetic 3-current is the probability current $\psi^\dagger\boldsymbol{\alpha}\psi$ multiplied by q:

$$j^\mu_{\mathrm{D\ em}} = (q\psi^\dagger\psi, q\psi^\dagger\boldsymbol{\alpha}\psi) = q\bar{\psi}\gamma^\mu\psi. \tag{4.100}$$

Consider the charge density: under the transformation (4.93) this becomes

$$q\psi^\dagger_{\mathbf{C}}\psi_{\mathbf{C}} = q\psi^{\mathrm{T}}\gamma^{2\dagger}\gamma^2\psi^* = q\psi^{\mathrm{T}}\alpha_2\beta\beta\alpha_2\psi^* = q\psi^{\mathrm{T}}\psi^*. \tag{4.101}$$

In terms of the four components of ψ, the product $\psi^{\mathrm{T}}\psi^*$ is $\psi_1\psi_1^* + \psi_2\psi_2^* + \psi_3\psi_3^* + \psi_4\psi_4^*$. These components are ordinary functions which commute with each other, so $\psi^{\mathrm{T}}\psi^* = \psi^{*\mathrm{T}}\psi = \psi^\dagger\psi$; hence

$$q\psi^\dagger_{\mathbf{C}}\psi_{\mathbf{C}} = q\psi^\dagger\psi \tag{4.102}$$

and the charge density does not change sign under $\mathbf{C}$. Similarly, one finds that the electromagnetic 3-current does not change sign either.

These results can be interpreted in the hole theory picture: the current due to a physical positive energy antiparticle of charge q and momentum $\boldsymbol{p}$ is regarded as the same as that of a missing negative energy particle of charge $-q$ and momentum $\boldsymbol{p}$. Our charge conjugation operation explicitly constructs the positive energy antiparticle wavefunction from the negative energy particle one.

Yet this is not really what we want a true charge conjugation operator to do: which is, rather, to change a positive energy particle into a positive energy antiparticle. The same inadequacy was true in the KG case also. There is no way of representing such an operation in a single particle wavefunction formalism. The appropriate formalism is quantum field theory, in which $\psi(x)$ becomes a quantum field operator (as do bosonic fields), and there is a unitary quantum field operator $\hat{\mathbf{C}}$ with the required property. We shall see in chapter 7 that fermionic operators anticommute with each other, and that this is just what is needed to ensure that the current changes sign under $\hat{\mathbf{C}}$. Bosonic fields, on the other hand, obey commutation rather than anticommutation relations, and this safeguards the change in sign of the bosonic current.

We have approached charge conjugation following the historical route, which is to say via the electromagnetic interaction. But we can ask whether (true) **C** is a good symmetry of other interactions, for example the weak interaction. Consider applying **C** to the reaction (4.82), so that it becomes

$$\pi^- \to \mu^- + \bar{\nu}_\mu. \tag{4.103}$$

If **C** was a good symmetry, the (parity-violating) longitudinal polarization of the μ^- in (4.103) should be the same as that of the μ^+ in (4.82). But in fact it is the opposite, the μ^- spin being aligned along the direction of its momentum. So **C**, like **P**, is violated in weak interactions. It is a good symmetry in electromagnetic and strong interactions.

4.2.3 CP

It has probably occurred to the reader that, although **C** and **P** are each violated in the decays (4.82) and (4.103), the combined transformation **CP** might be a good symmetry: particles are changed to antiparticles, the sense of longitudinal polarization is reversed, and the corresponding decays occur. Indeed, the rates for these two decays are the same, and **CP** is conserved. For a while, after 1956, it was hoped that **CP** would prove to be always conserved, so as to avoid a 'lopsided' distinction between right and left, and between matter and antimatter. But before long Christenson *et al.* (1964) reported evidence for **CP** violation in the decays of neutral K-mesons, a result soon confirmed by other experiments.

As we mentioned in section 1.2.2, it was the difficulty of incorporating **CP** violation into the 2-generation electroweak theory that led Kobayashi and Maskawa (1973) to propose a third generation of quarks, which allowed a **CP** violating parameter to be included quite naturally. **CP** violation in K-decays is a small effect (of order one part in 10^3), but in 1980 Carter and Sanda (1980) showed that considerably larger effects, up to 20%, could be expected in rare decays of neutral B mesons, according to the framework of Kobayashi and Maskawa (KM). Some 20 years later, the 'B factories' at the asymmetric e^-e^+ colliders PEPII and KEKB began producing B mesons by the many millions, and intensive study of **CP** violation in the $B^0(d\bar{b}) - \bar{B}^0(\bar{d}b)$ systems followed at the BaBar and Belle detectors. Remarkably, all observations to date are consistent with the original KM parametrization. We shall return to this topic when we discuss weak interactions in volume 2, specifically in chapter 21. Meanwhile we refer to Bettini (2008), chapter 8, for an introductory overview.

It is worth pausing here to note the significance of **CP** violation. First of all, it implies that there is an absolute distinction between matter and antimatter and, as a consequence, between left and right: these are not merely a matter of convention. For example, the rate for the process

$$B^0 \to K^+\pi^- \tag{4.104}$$

is some 20% greater (Nakamura *et al.* 2010) than the rate for the **CP**-conjugate process

$$\bar{\mathrm{B}}^0 \to \mathrm{K}^-\pi^+. \tag{4.105}$$

(Note that the $\bar{\mathrm{B}}^0$ state is conventionally defined as the **CP** transform of the B^0 state). So the pion distinguished by being emitted in the higher-yielding reaction (4.104) defines 'negatively charged', and the polarization of the muon in its decay (4.103) defines what is a right-handed screw sense.

Secondly, **CP** (and **C**) violation is one of the three conditions[2] established by Sakharov (1967) that would enable a universe containing initially equal amounts of matter and antimatter, when created in the Big Bang, to evolve into the matter-dominated universe we see today – rather than simply having the required imbalance as an initial condition. Within the Standard Model, all known **CP** violating effects are attributable to the KM mechanism. But calculations show (Huet and Sather 1995) that the matter-antimatter asymmetry generated from this source is very many orders of magnitude too small. This is, therefore, one area of physics where the Standard Model fails.

Thirdly, **CP** violation is directly connected to the violation of another discrete symmetry, namely time reversal **T**, because very general principles of quantum field theory imply that the product **CPT** (in any order) is conserved – the **CPT** theorem. This theorem states (Lüders 1954, 1957, Pauli 1957) that **CPT** must be an exact symmetry for any Lorentz invariant quantum field theory constructed out of local fields, with a Hermitian Hamiltonian, and quantized according to the usual spin-statistics rule (integer spin particles are bosons, half-odd integer spin particles are fermions). Thus any violation of **CP** implies a violation of **T** if **CPT** is to be conserved.

We shall return to **CPT** presently, but first let us deal with **T**.

4.2.4 Time reversal

The time reversal transformation **T** is defined by

$$\mathbf{T} : \boldsymbol{x} \to \boldsymbol{x}' = \boldsymbol{x}, \quad t \to t' = -t; \tag{4.106}$$

that is, **T** reverses the direction of time. It follows that **T** reverses momenta ($\boldsymbol{p} \to -\boldsymbol{p}$) and angular momenta ($\boldsymbol{x} \times \boldsymbol{p} \to -\boldsymbol{x} \times \boldsymbol{p}$). Let us also note how the electromagnetic potentials transform under **T**: A^0 does not change, being generated by static charges, while **A** changes sign, since it is produced by currents; that is,

$$A^0_{\mathbf{T}}(t') = A^0(t) \qquad \boldsymbol{A}_{\mathbf{T}}(t') = -\boldsymbol{A}(t). \tag{4.107}$$

It follows that the electric field $\boldsymbol{E}$ does not change sign under **T**, but the magnetic field $\boldsymbol{B}$ does. It is easily checked that these prescriptions ensure that the Maxwell equations are covariant under **T**.

[2]The other two are (a) the existence of baryon number violating transitions and (b) a time when the **C**, **CP** and baryon number violating transitions proceeded out of thermal equilibrium.

Consider first the behaviour of the KG equation for a particle of charge q in the field A^μ:

$$(\Box + m^2)\phi(t) = -\mathrm{i}q[\partial_\mu A^\mu(t) + A^\mu(t)\partial_\mu]\phi(t) + q^2A^2(t)\phi(t). \tag{4.108}$$

The equation in the time-reversed system is

$$(\Box + m^2)\phi_{\mathbf{T}}(t') = -\mathrm{i}q[\partial'_\mu A^\mu_{\mathbf{T}}(t') + A^\mu_{\mathbf{T}}(t')\partial'_\mu]\phi_{\mathbf{T}}(t') + q^2A^2_{\mathbf{T}}\phi_{\mathbf{T}}(t'). \tag{4.109}$$

Using (4.107) we obtain

$$\partial'_\mu A^\mu_{\mathbf{T}}(t') = -\partial_\mu A^\mu(t), \quad A^\mu_{\mathbf{T}}(t')\partial'_\mu = -A^\mu(t)\partial_\mu, \quad A^2_{\mathbf{T}}(t') = A^2(t). \tag{4.110}$$

It follows that we can identify

$$\phi_{\mathbf{T}}(t') = \phi^*(t) \tag{4.111}$$

up to an arbitrary phase factor, here chosen to be unity. If ϕ is a positive-energy free particle solution, ϕ^* represents a particle of positive energy in the time-reversed system, with momentum $-\boldsymbol{p}$ as expected.

Now consider the behaviour under **T** of the Dirac equation for a particle of charge q in a field A^μ,

$$\mathrm{i}\frac{\partial\psi(t)}{\partial t} = \{\boldsymbol{\alpha}\cdot[-\mathrm{i}\boldsymbol{\nabla} - q\boldsymbol{A}(t)] + \beta m + qA^0(t)\}\psi(t) \tag{4.112}$$

where we have suppressed the spatial coordinate arguments. In the time-reversed system, the corresponding equation is

$$\mathrm{i}\frac{\partial\psi_{\mathbf{T}}(t')}{\partial t'} = \{\boldsymbol{\alpha}\cdot[-\mathrm{i}\boldsymbol{\nabla} - q\boldsymbol{A}_{\mathbf{T}}(t')] + \beta m + qA^0_{\mathbf{T}}(t')\}\psi_{\mathbf{T}}(t'). \tag{4.113}$$

To relate $\psi_{\mathbf{T}}$ to ψ we start by taking the complex conjugate of (4.112) so as to obtain

$$-\mathrm{i}\frac{\partial\psi^*(t)}{\partial t} = \{\boldsymbol{\alpha}^*\cdot[\mathrm{i}\boldsymbol{\nabla} - q\boldsymbol{A}(t)] + \beta^* m + qA^0(t)\}\psi^*(t) \tag{4.114}$$

which we may rewrite as

$$\mathrm{i}\frac{\partial\psi^*(t)}{\partial t'} = \{\boldsymbol{\alpha}^*\cdot[\mathrm{i}\boldsymbol{\nabla} + q\boldsymbol{A}_{\mathbf{T}}(t')] + \beta^* m + qA^0_{\mathbf{T}}(t')\}\psi^*(t). \tag{4.115}$$

Now suppose a unitary matrix $U_{\mathbf{T}}$ exists such that

$$U_{\mathbf{T}}\boldsymbol{\alpha}^*U^\dagger_{\mathbf{T}} = -\boldsymbol{\alpha}, \qquad U_{\mathbf{T}}\beta^*U^\dagger_{\mathbf{T}} = \beta; \tag{4.116}$$

then it is clear that the Dirac equation will be covariant under **T** with the identification

$$\psi_{\mathbf{T}}(t') = U_{\mathbf{T}}\psi^*(t). \tag{4.117}$$

In either of the two representations of the Dirac matrices which we have been using, α_1, α_3 and β are real, while α_2 is pure imaginary; it follows that $U_{\mathbf{T}}$ must commute with α_2 and β, and anticommute with α_1 and α_2. A suitable $U_{\mathbf{T}}$ is

$$U_{\mathbf{T}} = \mathrm{i}\alpha_1\alpha_3 \tag{4.118}$$

where the phase is a conventional choice.

Let us check what is the effect of the transformation (4.117) on a positive-energy plane wave solution (3.74). In the representation (3.31) $U_{\mathbf{T}}$ is given by

$$U_{\mathbf{T}} = \begin{pmatrix} \sigma_2 & 0 \\ 0 & \sigma_2 \end{pmatrix} \tag{4.119}$$

and so

$$\begin{aligned} \psi_{\mathbf{T}}(\boldsymbol{x}, t') &= (E+m)^{1/2} \begin{pmatrix} \sigma_2 & 0 \\ 0 & \sigma_2 \end{pmatrix} \begin{pmatrix} \phi^* \\ \frac{\boldsymbol{\sigma}^* \cdot \boldsymbol{p}}{E+m}\phi^* \end{pmatrix} \exp(\mathrm{i}Et - \mathrm{i}\boldsymbol{p}\cdot\boldsymbol{x}) \\ &= (E+m)^{1/2} \begin{pmatrix} \sigma_2\phi^* \\ \frac{\boldsymbol{\sigma}\cdot\boldsymbol{p}'}{E+m}\sigma_2\phi^* \end{pmatrix} \exp(-\mathrm{i}Et' + \mathrm{i}\boldsymbol{p}'\cdot\boldsymbol{x}), \end{aligned} \tag{4.120}$$

which is a positive-energy solution with the expected momentum $\boldsymbol{p}' = -\boldsymbol{p}$, and with the transformed spinor wavefunction $\sigma_2\phi^*$. If we take ϕ to be a helicity eigenstate

$$\frac{\boldsymbol{\sigma}\cdot\boldsymbol{p}}{|\boldsymbol{p}|}\phi_\lambda = \lambda\phi_\lambda \tag{4.121}$$

where $\lambda = \pm 1$, then it follows that

$$\frac{\boldsymbol{\sigma}\cdot\boldsymbol{p}'}{|\boldsymbol{p}'|}\sigma_2\phi^*_\lambda = \lambda\sigma_2\phi^*_\lambda, \tag{4.122}$$

and the helicity is unchanged.

As in the case of parity, we may introduce an operator $\hat{\mathbf{T}}$ which changes ϕ to $\phi_{\mathbf{T}}$ for the KG equation, and ψ to $\psi_{\mathbf{T}}$ for the Dirac equation. Then

$$\hat{\mathbf{T}}(\mathrm{KG}) = \mathbf{K}\hat{\mathbf{T}}_0 \tag{4.123}$$

and

$$\hat{\mathbf{T}}(\mathrm{Dirac}) = U_{\mathbf{T}}\mathbf{K}\hat{\mathbf{T}}_0 \tag{4.124}$$

where $\mathbf{K}$ is the complex conjugation operator, and $\hat{\mathbf{T}}_0$ is the time coordinate reversal operator. The appearance of $\mathbf{K}$ is a general feature of time-reversal in quantum mechanics (Wigner 1964), and has important consequences.[3] Because the transformations involve complex conjugation, the scalar product of

[3]Complex conjugation also appeared in our discussion of $\mathbf{C}$ in section 4.2.2, but as indicated there the true operator $\hat{\mathbf{C}}$ of quantum field is unitary. Even in quantum field theory, however, the time-reversal operator involves complex conjugation, as we shall see in section 7.5.3.

two wavefunctions $< \psi_2|\psi_1 >$ is not equal to the corresponding quantity $< \psi_{2\mathbf{T}}|\psi_{1\mathbf{T}} >$, as it would be in the case of parity, for example, or for any other transformation represented by a unitary operator. Instead, we have

$$< \psi_2|\psi_1 >=< \psi_{2\mathbf{T}}|\psi_{1\mathbf{T}} >^* . \tag{4.125}$$

Note, however, that the probability $| < \psi_2|\psi_1 > |^2$ is still preserved.

If we consider the matrix element of any operator $\hat{O}$, then since $\hat{O}\psi_1$ is itself a wavefunction, we must have

$$< \psi_2|\hat{O}|\psi_1 >=< \psi_2|\hat{O}\psi_1 >=< \psi_{2\mathbf{T}}|\hat{\mathbf{T}}\hat{O}\psi_1 >^*=< \psi_{2\mathbf{T}}|\hat{\mathbf{T}}\hat{O}\hat{\mathbf{T}}^{-1}|\psi_{1\mathbf{T}} >^* \tag{4.126}$$

where $\hat{\mathbf{T}}\hat{O}\hat{\mathbf{T}}^{-1}$ is the operator in the time-reversed system. In particular, if we take $\hat{O}$ to be a Hermitian interaction potential $\hat{V}$, which is time-reversal invariant, then time-reversal invariance implies the relation

$$< \psi_2|\hat{V}|\psi_1 >=< \psi_{2\mathbf{T}}|\hat{V}|\psi_{1\mathbf{T}} >^*=< \psi_{1\mathbf{T}}|\hat{V}|\psi_{2P\mathbf{T}} > . \tag{4.127}$$

Now $< \psi_2|\hat{V}|\psi_1 >$ is the amplitude for the state represented by ψ_1 to make a transition to the state represented by ψ_2 to first order in the potential $\hat{V}$ (see section M.3 of appendix M). Equation (4.127) therefore relates this amplitude to one for the inverse transition, involving time-reversed states. The relation in fact holds for the complete (all orders) transition operator $\hat{T}$ (see for example Lee 1981, section 13.5), and enables one to relate rates and cross sections for reactions and their inverses.

For strong interactions, these relations are straightforward to test, and confirm that strong interactions are **T**-invariant. So are electromagnetic interactions. In weak interactions, where the violation of **CP** and the conservation of **CPT** implies that **T** is violated, it is generally very difficult if not impossible to set up the conditions for an inverse reaction to occur (consider the inverse of neutron decay, $\mathrm{n} \to \mathrm{pe}^-\bar{\nu}_\mathrm{e}$, for example). However, one such test is possible in neutral K-decays (Kabir 1970). We can check whether the rate for a particle tagged at its production as a K^0 to decay in a way that identifies it as a $\bar{\mathrm{K}}^0$ is equal to the rate for a particle tagged as $\bar{\mathrm{K}}^0$ at its production to decay in a way that identifies it as a K^0. The experiment (Angelopoulos *et al.* 1998) showed a **T**-violating difference in these rates. The parameters determining these reactions had actually been well determined by other measurements; still, this was an independent and direct demonstration of **T** violation. Evidence for **T** violation in B-meson transitions has been reported by Alvarez and Szynkman (2008), developing a test suggested by Banuls and Bernabeu (1999, 2000).

We can also examine the behaviour of various bilinears under **T**. For example, the reader may easily check the results

$$\bar{\psi}_\mathbf{T}(x')\psi_\mathbf{T}(x') = \bar{\psi}_(x)\psi(x), \qquad \bar{\psi}_\mathbf{T}(x')\gamma_5\psi_\mathbf{T}(x') = -\bar{\psi}(x)\gamma_5\psi(x). \tag{4.128}$$

Time reversal symmetry will be violated if the theory contains both even and

odd amplitudes under **T**. An interesting example is provided by the amplitude

$$-\mathrm{i}d_{\mathrm{e}}\bar{\psi}(x)\sigma^{\mu\nu}\gamma_5\psi(x)F_{\mu\nu}, \tag{4.129}$$

where

$$\sigma^{\mu\nu} = \frac{\mathrm{i}}{2}(\gamma^\mu\gamma^\nu - \gamma^\nu\gamma^\mu) \tag{4.130}$$

and where $F_{\mu\nu}$ is an external electric field with non-vanishing components $F_{0i} = E^i$. In the representation (3.31),

$$\sigma^{0i}\gamma_5 = \mathrm{i}\begin{pmatrix} \sigma_i & 0 \\ 0 & \sigma_i \end{pmatrix} \equiv \mathrm{i}\Sigma_i, \tag{4.131}$$

and (4.129) reduces to

$$d_{\mathrm{e}}\bar{\psi}(x)\boldsymbol{\Sigma}\psi(x)\cdot\boldsymbol{E}. \tag{4.132}$$

Problem 4.5 shows that the quantity (4.132) is odd under **T**, and it is easy to check that it is also odd under **P**. A non-zero value of such a term would correspond to an electric dipole moment for a spin-1/2 particle (compare the analogous quantity $d_{\mathrm{m}}\bar{\psi}(x)\boldsymbol{\Sigma}\psi(x)\cdot\boldsymbol{B}$ for the magnetic dipole moment, which is even under **P** and **T**). Experiment places very strong limits on possible electric dipole moments (Nakamura *et al.* 2010) for the neutron, proton and electron:

$$d_{\mathrm{n}} < 0.29\times10^{-25}\ e\ \mathrm{cm} \tag{4.133}$$

$$d_{\mathrm{p}} < 0.54\times10^{-23}\ e\ \mathrm{cm} \tag{4.134}$$

$$d_{\mathrm{e}} = (0.069\pm0.074)\times10^{-26}\ e\ \mathrm{cm} \tag{4.135}$$

Although these numbers seem tiny, calculations of the d_{n} in the Standard Model produce a result some 6 or 7 orders of magnitude smaller than (4.133). However, these experimental limits impose strong constraints on theories which go beyond the Standard Model, and which may typically contain the possibility of larger **T** and **CP** violating effects.

4.2.5 CPT

We denote the product **CPT** by $\boldsymbol{\theta}$, and the corresponding operator by $\hat{\boldsymbol{\theta}}$. As already mentioned, for any conventional quantum field theory, and certainly for the Standard Model, the transformation $\boldsymbol{\theta}$ is an invariance of the theory. One immediate consequence of this invariance is the equality of particle and antiparticle masses. This is easily demonstrated. Let $|\mathrm{X}, s_z>$ be the state of a particle X at rest with z-component of spin equal to s_z. The mass of X is given by the expectation value

$$M_{\mathrm{X}} = <\mathrm{X}, s_z|\hat{H}|\mathrm{X}, s_z>, \tag{4.136}$$

where $\hat{H}$ is the total Hamiltonian. Clearly M_{X} is real, and independent of s_z. Now the operator $\hat{\boldsymbol{\theta}}$ involves $\hat{\mathbf{T}}$, and therefore we must be careful to use

(4.126) rather than the usual rule for unitary operators. So from (4.126) we have

$$M_{\rm X} =< {\rm X}, s_z|\hat{H}|{\rm X}, s_z >^* =< {\rm X}, s_z|\hat{\boldsymbol{\theta}}^{-1}\hat{\boldsymbol{\theta}}\hat{H}\hat{\boldsymbol{\theta}}^{-1}\hat{\boldsymbol{\theta}}|{\rm X}, s_z >. \tag{4.137}$$

If the Hamiltonian is **CPT** invariant, then $\hat{\boldsymbol{\theta}}\hat{H}\hat{\boldsymbol{\theta}}^{-1} = \hat{H}$. Also, we know the action of $\hat{\mathbf{P}}, \hat{\mathbf{C}}$ and $\hat{\mathbf{T}}$ on the states, from the previous results. Equation (4.137) then becomes

$$M_{\rm X} =< \bar{\rm X}, -s_z|\hat{H}|\bar{\rm X}, -s_z >= M_{\bar{\rm X}}, \tag{4.138}$$

stating the equality of particle and antiparticle masses. The most sensitive test of (4.138) is provided by the ${\rm K}^0 - \bar{\rm K}^0$ system, where the currently quoted limit for the mass difference is (Nakamura *et al.* 2010)

$$\frac{|M^0_{\rm K} - M^0_{\bar{\rm K}}|}{M_{\rm average}} < 8 \times 10^{-19} \quad \text{at} \quad 90\% \quad \text{C.L.} \tag{4.139}$$

$\boldsymbol{\theta}$-invariance also implies that the charges of a charged particle and its antiparticle are equal in magnitude but opposite in sign, as are their magnetic moments; and in the case of unstable particles it implies that their lifetimes are equal, to first order in the interaction responsible for the decay (Lee 1981). All current data support these equalities (Nakamura *et al.* 2010). Other tests involve analysis of the implications of $\boldsymbol{\theta}$-invariance as applied to transition amplitudes. As an example, we refer to a recent analysis of K-decays by Abouziad *et al.* (2011), both with and without the assumption of $\boldsymbol{\theta}$-invariance. The results were consistent with $\boldsymbol{\theta}$-invariance.

Problems

4.1 Consider an infinitesimal boost along the x-axis,

$$t' = t - \eta x \tag{4.140}$$

$$x' = x - \eta t. \tag{4.141}$$

Show that the KG wavefunction transforms according to

$$\phi'(x,t) = (1 + {\rm i}\eta\hat{K}_x)\phi, \tag{4.142}$$

where

$$\hat{K}_x = -{\rm i}\; x\; \partial/\partial t - {\rm i}\; t\; \partial/\partial x. \tag{4.143}$$

Defining similar operators $\hat{K}_y, \hat{K}_z$ for boosts in the y and z directions, show that

$$[\hat{K}_x, \hat{K}_y] = -{\rm i}\hat{L}_z. \tag{4.144}$$

4.2 In this problem, use the representation (3.40) for the Dirac matrices, as in section 4.1.2.

(a) Using the rule (4.19) for the transformation of the spinor ϕ under an infinitesimal rotation of the coordinate system, verify that $\phi^\dagger \boldsymbol{\sigma} \phi$ transforms as a 3-vector. [*Hint*: you need to show that $\phi'^\dagger \boldsymbol{\sigma} \phi' = \phi^\dagger \boldsymbol{\sigma} \phi - \boldsymbol{\epsilon} \times \phi^\dagger \boldsymbol{\sigma} \phi$; use the results of problem 3.4(a).] Show also that the free-particle Dirac probability current density is a 3-vector.

(b) Using the rule (4.42) for the transformation of ϕ and χ under an infinitesimal boost, verify that $\boldsymbol{j} = \phi^\dagger \boldsymbol{\sigma} \phi - \chi^\dagger \boldsymbol{\sigma} \chi$ transforms as the 3-vector part of the 4-vector $(\rho, \boldsymbol{j})$. [*Hint*: you need to show that $\boldsymbol{j}' = \boldsymbol{j} - \boldsymbol{\eta} \rho$.]

4.3

(a) Defining the four 'γ matrices'

$$\gamma^\mu = (\gamma^0, \boldsymbol{\gamma})$$

where $\gamma^0 = \beta$ and $\boldsymbol{\gamma} = \beta \boldsymbol{\alpha}$, show that the Dirac equation can be written in the form $(\mathrm{i}\gamma^\mu \partial_\mu - m)\psi = 0$. Find the anticommutation relations of the γ matrices. Show that the positive energy spinors $u(p, s)$ satisfy $(\not{p} - m)u(p, s) = 0$, and that the negative energy spinors $v(p, s)$ satisfy $(\not{p} + m)v(p, s) = 0$, where $\not{p} = \gamma^\mu p_\mu$ (pronounced 'p-slash').

(b) Define the conjugate spinor

$$\bar{\psi}(x) = \psi^\dagger(x)\gamma^0$$

and use the previous result to find the equation satisfied by $\bar{\psi}$ in γ matrix notation.

(c) The Dirac probability current may be written as

$$j^\mu = \bar{\psi}(x)\gamma^\mu \psi(x).$$

Show that it satisfies the conservation law

$$\partial_\mu j^\mu = 0.$$

4.4

(a) Verify that, under $\mathbf{P}$, $\bar{\psi}(\boldsymbol{x}, t)\gamma^0 \psi(\boldsymbol{x}, t)$ is a scalar, and that $\bar{\psi}(\boldsymbol{x}, t)\boldsymbol{\gamma}\psi(\boldsymbol{x}, t)$ is a polar vector.

(b) Verify that $a^\mu(\boldsymbol{x}, t) = \bar{\psi}(\boldsymbol{x}, t)\gamma_5 \gamma^\mu \psi(\boldsymbol{x}, t)$ transforms under infinitesimal rotations and boosts as a 4-vector; and that under $\mathbf{P}$ $a^0(\boldsymbol{x})$ is a pseudoscalar, and $\boldsymbol{a}(\boldsymbol{x}, t)$ is an axial vector.

(c) Show that $\sigma_2 \phi^*$ transforms under rotations and boosts as a χ-type spinor, and that $\sigma_2 \chi^*$ transforms as a ϕ-type spinor.

4.5 Verify that $\bar{\psi}(\boldsymbol{x}, t)\boldsymbol{\Sigma}\psi(\boldsymbol{x}, t) \cdot \boldsymbol{E}$ of (4.132) is odd under $\mathbf{T}$.

4.6 The Galilean transformation (non-relativistic boost) is defined by

$$\boldsymbol{x}' = \boldsymbol{x} - \boldsymbol{v}t, \quad t' = t.$$

Show that the free-particle time-dependent Schrödinger equation is covariant under this transformation if the wavefunction transforms according to the rule $\psi'(\boldsymbol{x}', t') = \exp[\mathrm{i}f(\boldsymbol{x}, t)]\psi(\boldsymbol{x}, t)$, where $f(\boldsymbol{x}, t)$ satisfies the condition

$$-\frac{\partial f}{\partial t} - \boldsymbol{v} \cdot \boldsymbol{\nabla} f + \mathrm{i}\boldsymbol{v} \cdot \boldsymbol{\nabla} = \frac{1}{2m}(\boldsymbol{\nabla} f)^2 - \frac{\mathrm{i}}{2m}\boldsymbol{\nabla}^2 f - \frac{\mathrm{i}}{m}\boldsymbol{\nabla} f \cdot \boldsymbol{\nabla}.$$

Find constants a and $\boldsymbol{b}$ such that the function $f = at + \boldsymbol{b} \cdot \boldsymbol{x}$ satisfies this condition. Show that the resulting transformation rule is consistent with the way you expect a plane wave solution to transform.


Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

Part II

Introduction to Quantum Field Theory

It was a wonderful world my father told me about.

You might wonder what he got out of it all. I went to MIT. I went to Princeton. I went home and he said, 'Now you've got a science education. I have always wanted to know something that I have never understood; and so, my son, I want you to explain it to me.' I said yes.

He said, 'I understand that they say that light is emitted from an atom when it goes from one state to another, from an excited state to a state of lower energy.'

I said 'That's right.'

'And light is a kind of particle, a photon I think they call it.'

'Yes.'

'So if the photon comes out of the atom when it goes from the excited to the lower state, the photon must have been in the atom in the excited state.'

I said, 'Well, no.'

He said, 'Well, how do you look at it so you can think of a particle photon coming out without it having been in there in the excited state?'

I thought a few minutes, and I said, 'I'm sorry; I don't know. I can't explain it to you.'

He was very disappointed after all these years and years trying to teach me something, that it came out with such poor results.

—R. P. Feynman, *The Physics Teacher*, vol 7, No 6, September 1969

All the fifty years of conscious brooding have brought me no closer to the answer to the question, 'What are light quanta?' Of course today every rascal thinks he knows the answer, but he is deluding himself.

—A. Einstein (1951)

Quoted in 'Einstein's research on the nature of light'
E. Wolf (1979), *Optic News*, vol 5, No 1, page 39.

I never satisfy myself until I can make a mechanical model of a thing. If I can make a mechanical model I can understand it. As long as I cannot make a mechanical model all the way through I cannot understand; and that is why I cannot get the electromagnetic theory.

—Sir William Thomson, Lord Kelvin, 1884 *Notes of Lectures on Molecular Dynamics and the Wave Theory of Light delivered at the Johns Hopkins University, Baltimore*, stenographic report by A. S. Hathaway (Baltimore: Johns Hopkins University) Lecture XX, pp 270–1.

5

Quantum Field Theory I: The Free Scalar Field

In this chapter we shall give an elementary introduction to quantum field theory, which is the established 'language' of the Standard Model of particle physics. Even so long after Maxwell's theory of the (classical) electromagnetic field, the concept of a 'disembodied' field is not an easy one; and we are going to have to add the complications of quantum mechanics to it. In such a situation, it is helpful to have some physical model in mind. For most of us, as for Lord Kelvin, this still means a mechanical model. Thus in the following two sections we begin by considering a mechanical model for a quantum field. At the end, we shall – like Maxwell – throw away the 'mechanism' and have simply quantum field theory. Section 5.1 describes this programme qualitatively; section 5.2 presents a more complete formalism, for the simple case of a field whose quanta are massless, and move in only one spatial dimension. The appropriate generalizations for massive quanta in three dimensions are given in section 5.3.

5.1 The quantum field: (i) descriptive

Mechanical systems are usefully characterized by the number of *degrees of freedom* they possess: thus a one-dimensional pendulum has one degree of freedom, two coupled one-dimensional pendulums have two degrees of freedom – which may be taken to be their angular displacements, for example. A scalar field $\phi(x,t)$ corresponds to a system with an infinite number of degrees of freedom, since at each continuously varying point x an independent 'displacement' $\phi(x,t)$, which also varies with time, has to be determined. Thus quantum field theory involves two major mathematical steps: the description of continuous systems (fields) which have infinitely many degrees of freedom, and the application of *quantum* theory to such systems. These two aspects are clearly separable. It is certainly easier to begin by considering systems with a discrete – but possibly very large – number of degrees of freedom, for example a solid. We shall treat such systems first classically and then quantum mechanically. Then, returning to the classical case, we shall allow the number

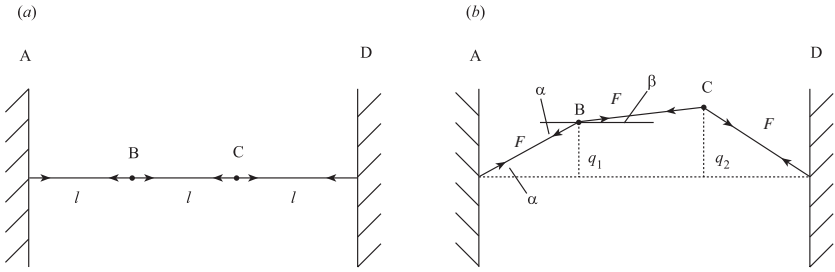


FIGURE 5.1
A vibrating system with two degrees of freedom: (a) two mass points at rest, with the strings under tension; (b) a small transverse displacement.

of degrees of freedom to become infinite, so that the system corresponds to a classical field. Finally, we shall apply quantum mechanics directly to fields.

We begin by considering a rather small solid – one that has only two atoms free to move. The atoms, each of mass m, are connected by a string, and each is connected to a fixed support by a similar string (figure 5.1(a)); all the strings are under tension F. We consider *small transverse vibrations* of the atoms (figure 5.1(b)), and we call $q_r(t)$ $(r = 1, 2)$ the transverse displacements. We are interested in the total energy E of the system. According to classical mechanics, this is equal to the sum of the kinetic energies $\frac{1}{2}m\dot{q}_r^2$ of each atom, together with a potential energy V which can be calculated as follows. Referring to figure 5.1(b), when atom 1 is displaced by q_1, it experiences a restoring force

$$F_1 = F\sin\alpha - F\sin\beta \tag{5.1}$$

assuming a constant tension F along the string. For small displacements q_1 and q_2 (i.e. $q_{1,2} \ll l$) we have

$$\begin{aligned} \sin\alpha &= q_1/(l^2 + q_1^2)^{1/2} \approx q_1/l \\ \sin\beta &= (q_2 - q_1)/[l^2 + (q_2 - q_1)^2]^{1/2} \approx (q_2 - q_1)/l \end{aligned} \tag{5.2}$$

where terms of order $(q_{1,2}/l)^3$ and higher have been neglected. Thus the restoring force on particle 1 is, in this approximation,

$$F_1 = k(2q_1 - q_2) \tag{5.3}$$

with $k = F/l$. Similarly, the restoring force on particle 2 is

$$F_2 = k(2q_2 - q_1) \tag{5.4}$$

and the equations of motion are

$$m\ddot{q}_1 = -k(2q_1 - q_2) \tag{5.5}$$

$$m\ddot{q}_2 = -k(2q_2 - q_1). \tag{5.6}$$

The potential energy is then determined (up to an irrelevant constant) by the requirement that (5.5) and (5.6) are of the form

$$m\ddot{q}_1 = -\partial V/\partial q_1 \tag{5.7}$$

$$m\ddot{q}_2 = -\partial V/\partial q_2. \tag{5.8}$$

Thus we deduce that

$$V = k(q_1^2 + q_2^2 - q_1 q_2). \tag{5.9}$$

Equations (5.5) and (5.6) form a pair of *linear*, *coupled* differential equations. Each of the italicized words is important. By 'linear', is meant that only the first power of q_1 and q_2 and their time derivatives appear in the equations of motion; terms such as q_1^2, $q_1 q_2$, $\dot{q}_1^2$, q_1^3 and so on would render the equations of motion 'nonlinear'. This linear/nonlinear distinction is a crucial one in dynamics. Most importantly, the solutions of linear differential equations may be added together with constant coefficients ('linearly superposed') to make new valid solutions of the equations. In contrast, solutions of nonlinear differential equations – besides being very hard to find! – cannot be linearly superposed to get new solutions. In addition, nonlinear dynamical equations may typically lead to *chaotic motion*.

The notion of linearity/nonlinearity carries over also into the equations of motion for fields. In this context, an equation for a field $\phi(x,t)$ is said to be linear if ϕ and its space – or time – derivatives appear only to the first power. As we shall see, this is true for Maxwell's equations for the electromagnetic field and it is, of course, the mathematical reason behind all the physics of such things as interference and diffraction, which may be understood precisely in terms of superposition of solutions of these equations. Likewise the equations of quantum mechanics (e.g. Schrödinger's equation) are all linear in this sense, consistent with the principle of superposition in quantum mechanics.

It is clear, then, that in looking at simple mechanical models as a guide to the field systems in which we will ultimately be interested, we should consider ones in which the equations of motion are linear. In the present case, this is true, but only because we have made the approximation that q_1 and q_2 are small (compared to l). Referring to equation (5.2), we can immediately see that if we had kept the full expression for $\sin\alpha$ and $\sin\beta$, the resulting equations of motion would have been highly nonlinear. A similar 'small displacement' approximation has to be made in determining the familiar wave equation, describing waves on continuous strings, for example (see (5.29) later). Most significantly, however, quantum mechanics is believed to be a linear theory *without* any approximation.

The appearance of only linear terms in q_1 and q_2 in the equations of motion implies, via (5.7) and (5.8), that the potential energy can only involve quadratic powers of the q's, i.e. q_1^2, q_2^2 and $q_1 q_2$, as in (5.9). Once again, had we used the general expression for the potential energy in a stretched string as 'tension×extension' we would have obtained an expression containing all powers of the q's via such terms as $\{[l^2 + q_1^2]^{1/2} - l\}$.

We turn now to the *coupled* aspect of (5.5) and (5.6). By this we mean that the right-hand side of the q_1 equation depends on q_2 as well as q_1, and similarly for the q_2 equation. This 'mathematical' coupling has its origin in the term $-kq_1q_2$ in V, which corresponds to the 'physical' coupling of the string BC connecting the two atoms. If this coupling were absent, equations (5.5) and (5.6) would describe two independent (uncoupled) harmonic oscillators, each of frequency $(2k/m)^{1/2}$. When we consider the addition of more and more particles (see later) we certainly do not want them to vibrate independently, otherwise we would not be able to get wave-like displacements propagating through the system. So we need to retain at least this minimal kind of 'quadratic' coupling.

With the coupling, the solutions of (5.5) and (5.6) are not quite so obvious. However, a simple step makes the equations much easier. Suppose we add the two equations so as to obtain

$$m(\ddot{q}_1 + \ddot{q}_2) = -k(q_1 + q_2) \tag{5.10}$$

and subtract them to obtain

$$m(\ddot{q}_1 - \ddot{q}_2) = -3k(q_1 - q_2). \tag{5.11}$$

A remarkable thing has happened: the two *combinations* $q_1 + q_2$ and $q_1 - q_2$ of the original coordinates satisfy *uncoupled* equations – which are of course very easy to solve. The combination $q_1 + q_2$ oscillates with frequency $\omega_1 = (k/m)^{1/2}$, while $q_1 - q_2$ oscillates with frequency $\omega_2 = (3k/m)^{1/2}$.

Let us introduce

$$Q_1 = (q_1 + q_2)/\sqrt{2} \qquad Q_2 = (q_1 - q_2)/\sqrt{2} \tag{5.12}$$

(the $\sqrt{2}$'s are for later convenience). Then the solutions of (5.10) and (5.11) are:

$$Q_1(t) = A\cos\omega_1 t + B\sin\omega_1 t \tag{5.13}$$

$$Q_2(t) = C\cos\omega_2 t + D\sin\omega_2 t. \tag{5.14}$$

Suppose that the initial conditions are such that

$$q_1(0) = q_2(0) = a \qquad \dot{q}_1(0) = \dot{q}_2(0) = 0 \tag{5.15}$$

i.e. the atoms are released from rest, at equal transverse displacements a. In terms of the Q_r's, the conditions (5.15) are

$$\begin{aligned} Q_2(0) = \dot{Q}_2(0) &= 0 \\ Q_1(0) = \sqrt{2}a \qquad \dot{Q}_1(0) &= 0. \end{aligned} \tag{5.16}$$

Thus from (5.13) and (5.14) we find that the complete solution, for these initial conditions, is

$$Q_1(t) = \sqrt{2}a\cos\omega_1 t \tag{5.17}$$

$$Q_2(t) = 0. \tag{5.18}$$

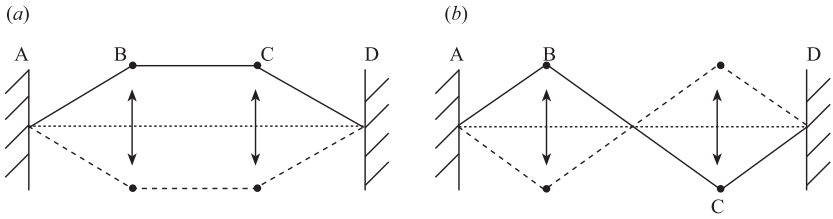


FIGURE 5.2
Motion in the two normal modes: (a) frequency ω_1; (b) frequency ω_2.

We see from (5.18) that the motion is such that $q_1 = q_2$ throughout, and from (5.17) that the system vibrates with a single definite frequency ω_1. A form of motion in which the system as a whole moves with a definite frequency is called a '*normal mode*' or simply a 'mode' for short. Figure 5.2(a) shows two 'snapshot' configurations of our two-atom system when it is oscillating in the mode characterized by $q_1 = q_2$. In this mode, only $Q_1(t)$ changes; $Q_2(t)$ is always zero. Another mode also exists in which $q_1 = -q_2$ at all times: here $Q_1(t)$ is zero and $Q_2(t)$ oscillates with frequency ω_2. Figure 5.2(b) shows two snapshots of the atoms when they are vibrating in this second mode. The coordinate combinations Q_1, Q_2, in terms of which this 'single frequency motion' occurs, are called 'normal mode coordinates' or '*normal coordinates*' for short.

In general, the initial conditions will not be such that the motion is a pure mode; both $Q_1(t)$ and $Q_2(t)$ will be non-zero. From (5.12) we have

$$q_1(t) = [Q_1(t) + Q_2(t)]/\sqrt{2} \tag{5.19}$$

and

$$q_2(t) = [Q_1(t) - Q_2(t)]/\sqrt{2} \tag{5.20}$$

so that q_1 and q_2 are expressed as a sum of two terms oscillating with frequencies ω_1 and ω_2. We say the system is in 'a superposition of modes'. Nevertheless, the mode idea is still very important as regards the total energy of the system, as we shall now see. The kinetic energy can be written in terms of the mode coordinates Q_r as

$$T = \tfrac{1}{2}m\dot{Q}_1^2 + \tfrac{1}{2}m\dot{Q}_2^2 \tag{5.21}$$

while the potential energy V of (5.9) becomes

$$V = \tfrac{1}{2}m\omega_1^2 Q_1^2 + \tfrac{1}{2}m\omega_2^2 Q_2^2 \equiv V(Q_1, Q_2). \tag{5.22}$$

The total energy is therefore

$$E = [\tfrac{1}{2}m\dot{Q}_1^2 + \tfrac{1}{2}m\dot{Q}_2^2] + [\tfrac{1}{2}m\omega_1^2 Q_1^2 + \tfrac{1}{2}m\omega_2^2 Q_2^2]. \tag{5.23}$$

This equation shows that, when written in terms of the normal coordinates, the total energy contains *no* couplings terms of the form $Q_1 Q_2$; indeed, the energy has the remarkable form of a simple sum of two independent uncoupled oscillators, one with characteristic frequency ω_1, the other with frequency ω_2. The energy (5.23) has exactly the form appropriate to a system of two *non-interacting* 'things', each executing simple harmonic motion: the 'things' are actually the two modes. Modes do not interact, whereas the original atoms do! Of course, this decoupling in the expression for the total energy is reflected in the decoupling of the equations of motion for the Q variables:

$$m\ddot{Q}_r = -\frac{\partial V(Q_1, Q_2)}{\partial Q_r} \qquad r = 1, 2. \tag{5.24}$$

It is most important to realize that the modes are non-interacting by virtue of the fact that we ignored higher than quadratic terms in $V(q_1, q_2)$. Although the simple change of variables $(q_1, q_2) \to (Q_1, Q_2)$ of (5.12) does remove the $q_1 q_2$ coupling, this would not be the case if, say, cubic terms in V were to be considered. Such higher order 'anharmonic' corrections would produce couplings between the modes – indeed, this will be the basis of the quantum field theory description of particle interactions (see the following chapter)!

The system under discussion had just two degrees of freedom. We began by describing it in terms of the obvious degree of freedom, the physical displacements of the two atoms q_1 and q_2. But we have learned that it is very illuminating to describe it in terms of the normal coordinate combinations Q_1 and Q_2. The normal coordinates are really the relevant degrees of freedom. Of course, for just two particles, the choice between the q_r's and the Q_r's may seem rather academic; but the important point – and the reason for going through these simple manipulations in detail – is that the basic idea of the normal mode, and of normal coordinates, generalizes immediately to the much less trivial N-atom problem (and also to the field problem). For N atoms there are (for one-dimensional displacements) N degrees of freedom, and if we take them to be the actual atomic displacements, the total energy will be

$$E = \sum_{r=1}^{N} \tfrac{1}{2} m \dot{q}_r^2 + V(q_1, \ldots, q_r) \tag{5.25}$$

which includes all the couplings between atoms. We assume, as before, that the q_r's are small enough so that only quadratic terms need to be kept in V (a constant is as usual irrelevant, and the linear terms vanish if the q_r's are the displacements from equilibrium). In this case, the equations of motion will be linear. By a linear transformation of the form (generalizing (5.12))

$$Q_r = \sum_{s=1}^{N} a_{rs} q_s \tag{5.26}$$

it is possible to write E as a sum of N separate terms, just as in (5.23):

$$E = \sum_{r=1}^{N} [\tfrac{1}{2} m\dot{Q}_r^2 + \tfrac{1}{2} m\omega_r^2 Q_r^2]. \tag{5.27}$$

The Q_r's are the normal coordinates and the ω_r's are the normal frequencies, and there are N of them. If only one of the Q_r's is non-zero, the N atoms are moving in a single mode. The fact that the total energy in (5.27) is a *sum* of N single-mode energies allows us to say that our N-atom solid *behaves as if it consisted of N separate and free harmonic oscillators* – which, however, are *not* to be identified with the coordinates of the original atoms. Once again, and now much more crucially, it is the *mode coordinates* that are the relevant degrees of freedom rather than those of the original particles.

The second stage in our programme is to treat such systems quantum mechanically, as we should certainly have to for a real solid. It is still true that – if the potential energy is a quadratic function of the displacements – the transformation (5.26) allows us to write the total energy as a sum of N mode energies, each of which has the form of a harmonic oscillator. Now, however, these oscillators obey the laws of quantum mechanics, so that each mode oscillator exists only in certain definite states, whose energy eigenvalues are quantized. For each mode of frequency ω_r, the allowed energy values are

$$\epsilon_r = (n_r + \tfrac{1}{2})\hbar\omega_r \tag{5.28}$$

where n_r is a positive integer or zero. This is in sharp contrast to the classical case, of course, in which arbitrary values are allowed for the oscillator energies. The total energy eigenvalue then has the form

$$E = \sum_{r=1}^{N} (n_r + \tfrac{1}{2})\hbar\omega_r. \tag{5.29}$$

The frequencies ω_r are determined by the interatomic forces and are common to both the classical and quantum descriptions; in quantum theory, though, *the states of definite energy of the vibrating N-body system are characterized by the values of a set of integers* $(n_1, n_2, \ldots, n_N)$*, which determine the energies of each mode oscillator.*

For each mode oscillator, $\hbar\omega_r$ measures the quantum of vibrational energy; the energy of an allowed mode state is determined uniquely by the number n_r of such quanta of energy in the state. We now make a profound reinterpretation of this result (first given, almost *en passant* by Born, Heisenberg and Jordan (Born *et al.* 1926) in one of the earliest papers on quantum mechanics). We forget about the original N degrees of freedom $q_1, q_2, \ldots, q_N$ and the original N 'atoms', which indeed are only remembered in (5.29) via the fact that there are N different mode frequencies ω_r. Instead we concentrate on the *quanta* and treat *them* as 'things' which really determine the behaviour of our quantum system. We say that 'in a state with energy $(n_r + \frac{1}{2})\hbar\omega_r$ there

are n_r quanta present'. For the state characterized by $(n_1, n_2, \ldots, n_N)$ there are n_1 quanta of mode 1 (frequency ω_1), n_2 of mode 2, ... and n_N of mode N. Note particularly that although the number of modes N is fixed, the values of the n_r's are unrestricted, except insofar as the total energy is fixed. Thus we are moving from a 'fixed number' picture (N degrees of freedom) to a 'variable number' picture (the n_r's restricted only by the total energy constraint (5.29)). In the case of a real solid, these quanta of vibrational energy are called *phonons*. We summarize the point we have reached by the important statement that *a phonon is an elementary quantum of vibrational excitation.*

Now we take one step backward in order, afterwards, to take two steps forward. We return to the classical mechanical model with N harmonically interacting degrees of freedom. It is possible to imagine increasing the number N to infinity, and decreasing the interatomic spacing a to zero, in such a way that the product Na stays finite, say $Na = \ell$. We then have a classical *continuous* system – for example a string of length ℓ. (We stay in one dimension for simplicity.) The transverse vibrations of this string are now described by a *field* $\phi(x,t)$, where at each point x of the string $\phi(x,t)$ measures the displacement from equilibrium, at the time t, of a small element of string around the point x. Thus we have passed from a system described by a discrete number of degrees of freedom, $q_r(t)$ or $Q_r(t)$, to one described by a continuous degree of freedom, the displacement field $\phi(x,t)$. The discrete suffix r has become the continuous argument x – and to prepare for later abstraction, we have denoted the displacement by $\phi(x,t)$ rather than, say, $q(x,t)$.

In the continuous problem the analogue of the small-displacement assumption, which limited the potential energy in the discrete case to quadratic powers, implies that $\phi(x,t)$ obeys the wave equation

$$\frac{1}{c^2}\frac{\partial^2\phi(x,t)}{\partial t^2} = \frac{\partial^2\phi(x,t)}{\partial x^2} \tag{5.30}$$

where c is the wave propagation velocity. Note that (5.30) is linear, but only by virtue of having made the small-displacement assumption. Again, we consider first the classical treatment of this system. Our aim is to find, for this continuous field problem, the analogue of the normal coordinates – or in physical terms, the *modes* of vibration – which were so helpful in the discrete case. Fortunately, the string's modes are very familiar. By imposing suitable boundary conditions at each end of the string, we determine the allowed wavelengths of waves travelling along the string. Suppose, for simplicity, that the string is stretched between $x = 0$ and $x = \ell$. This constrains $\phi(x,t)$ to vanish at these end points. A suitable form for $\phi(x,t)$ which does this is

$$\phi_r(x,t) = A_r(t)\sin\left(\frac{r\pi x}{\ell}\right) \tag{5.31}$$

where $r = 1, 2, 3, \ldots$, which expresses the fact that an exact number of half-wavelengths must fit onto the interval $(0, \ell)$. Inserting (5.31) into (5.30), we find

$$\ddot{A}_r = -\omega_r^2 A_r \tag{5.32}$$

FIGURE 5.3
String motion in two normal modes: (*a*) $r = 1$ in equation (5.31); (*b*) $r = 2$.

where

$$\omega_r^2 = r^2\pi^2c^2/\ell^2. \tag{5.33}$$

Thus the amplitude $A_r(t)$ of the particular waveform (5.31) executes simple harmonic motion with frequency ω_r. Each motion of the string which has a definite wavelength also has a definite frequency; it is therefore precisely a mode. Figure 5.3(*a*) shows two snapshots of the string when it is oscillating in the mode for which $r = 1$, and figure 5.3(*b*) shows the same for the mode $r = 2$; these may be compared with figures 5.2(*a*) and (*b*). Just as in the discrete case, the general motion of the string is a superposition of modes

$$\phi(x,t) = \sum_{r=1}^{\infty} A_r(t)\sin\left(\frac{r\pi x}{\ell}\right); \tag{5.34}$$

in short, a Fourier series!

We must now examine the total energy of the vibrating string, which we expect to be greatly simplified by the use of the mode concept. The total energy is the continuous analogue of the discrete summation in (5.25), namely the integral

$$E = \int_0^\ell \left[\frac{1}{2}\rho\left(\frac{\partial\phi}{\partial t}\right)^2 + \frac{1}{2}\rho c^2\left(\frac{\partial\phi}{\partial x}\right)^2\right] \mathrm{d}x \tag{5.35}$$

where the first term is the kinetic energy and the second is the potential energy (ρ is the mass per unit length of the string, assumed constant). As noted earlier, the potential energy term arises from an approximation which limits it to the quadratic power. To relate this to the earlier discrete case, note that the derivative may be regarded as $[\phi(x+\delta x)-\phi(x)]/\delta x$ as $\delta x \to 0$, so that the square of the derivative involves the 'nearest neighbour coupling' $\phi(x+\delta x)\phi(x)$, analogous to the q_1q_2 term in (5.9).

Inserting (5.34) into (5.35), and using the orthonormality of the sine functions on the interval $(0,\ell)$, one obtains (problem 5.1) the crucial result

$$E = (\ell/2)\sum_{r=1}^{\infty}[\tfrac{1}{2}\rho\dot{A}_r^2 + \tfrac{1}{2}\rho\omega_r^2A_r^2]. \tag{5.36}$$

Indeed, just as in the discrete case, the total energy of the string can be

written as a sum of individual mode energies. We note that *the Fourier amplitude A_r acts as a normal coordinate*. Comparing (5.36) with (5.27), we see that the string behaves exactly like a system of independent uncoupled oscillators, the only difference being that now there are an infinite number of them, corresponding to the infinite number of degrees of freedom in the continuous field $\phi(x,t)$. The normal coordinates $A_r(t)$ are, for many purposes, a much more relevant set of degrees of freedom than the original displacements $\phi(x,t)$.

The final step is to apply quantum mechanics to this classical field system. Once again, the total energy is equivalent to that of a sum of (infinitely many) mode oscillators, each of which has to be quantized. The total energy eigenvalue has the form (5.29), except that now the sum extends to infinity:

$$E = \sum_{r=1}^{\infty} (n_r + \tfrac{1}{2})\hbar\omega_r. \tag{5.37}$$

The excited states of the quantized field $\hat{\phi}(x,t)$ are characterized by saying how many phonons of each frequency are present; the ground state has no phonons at all. We remark that as $\ell \to \infty$, the mode sum in (5.36) or (5.37) will be replaced by an integral over a continuous frequency variable.

We have now completed, in outline, the programme introduced earlier, ending up with the quantization of a 'mechanical' system. All of the foregoing, it must be clearly emphasized, is absolutely basic to modern solid state physics. The essential idea – quantizing independent modes – can be applied to an enormous variety of 'oscillations'. In all cases the crucial concept is the elementary excitation – the mode quantum. Thus we have plasmons (quanta of plasma oscillations), magnons (magnetic oscillations), ..., as well as phonons (vibrational oscillations). All this is securely anchored in the physics of many-body systems.

Now we come to the use of these ideas as an *analogy*, to help us understand the (presumably non-mechanical) quantum fields with which we shall actually be concerned in this book – for example the electromagnetic field. Consider a region of space containing electromagnetic fields. These fields obey (a three-dimensional version of) the wave equation (5.30), with c now standing for the speed of light. By imposing suitable boundary conditions, the total electromagnetic energy in any region of space can be written as a sum of mode energies. Each mode has the form of an oscillator, whose amplitude is (see (5.31)) the Fourier component of the wave, for a given wavelength. These oscillators are all quantized. Their quanta are called photons. Thus, *a photon is an elementary quantum of excitation of the electromagnetic field*.

So far the only kind of 'particle' we have in our relativistic quantum field theoretic world is the photon. What about the electron, say? Well, recalling Feynman again, 'There is one lucky break, however – electrons behave just like light'. In other words, we shall also regard an electron as an elementary quantum of excitation of an 'electron field'. What is 'waving' to supply the

vibrations for this electron field? We do not answer this question just as we did not for the photon. We *postulate* a relativistic quantum field for the electron which obeys some suitable wave equation – in this case, for non-interacting electrons, the Dirac equation. The field is expanded as a sum of Fourier components, as with the electromagnetic field. Each component behaves as an independent oscillator degree of freedom (and there are, of course, an infinite number of them); the quanta of these oscillators are electrons.

Actually this, though correctly expressing the basic idea, omits one crucial factor, which makes it almost fraudulently oversimplified. There is of course one very big difference between photons and electrons. The former are *bosons* and the latter are *fermions*; photons have spin angular momentum of one (in unit of $\hbar$), electrons of one-half. It is very difficult, if not downright impossible, to construct any mechanical model at all which has fermionic excitations. Phonons have spin-1, in fact, corresponding to the three states of polarization of the corresponding vibrational waves. But 'phonons' carrying spin-$\frac{1}{2}$ are hard to come by. No matter, you may say, Maxwell has weaned us away from jelly, so we shall be grown up and boldly postulate the electron field as a basic *thing*.

Certainly this is what we do. But we also know that fermionic particles, like electrons, have to obey an exclusion principle: no two identical fermions can have the same quantum numbers. In chapter 7, we shall learn how the idea sketched here must be modified for fields whose quanta are fermions.

5.2 The quantum field: (ii) Lagrange–Hamilton formulation

5.2.1 The action principle: Lagrangian particle mechanics

We must now make the foregoing qualitative picture more mathematically precise. It is clear that we would like a formalism capable of treating, within a single overall framework, the mechanics of both fields and particles, in both classical and quantum aspects. Remarkably enough, such a framework does exist (and was developed long before quantum field theory): Hamilton's principle of *least action*, with the action defined in terms of a *Lagrangian*. We strongly recommend the reader with no prior acquaintance with this profound approach to physical laws read chapter 19 of volume 2 of Feynman's *Lectures on Physics* (Feynman 1964).

The least action approach differs radically from the more familiar one which can conveniently be called 'Newtonian'. Consider the simplest case, that of classical particle mechanics. In the Newtonian approach, equations of motion are postulated which involve forces as the essential physical input; from these, the trajectories of the particle can be calculated. In the least

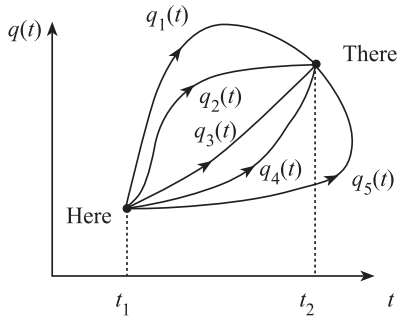


FIGURE 5.4
Possible space–time trajectories from 'Here' ($q(t_1)$) to 'There' ($q(t_2)$).

action approach, equations of motion are *not* postulated as basic, and the primacy of forces yields to that of *potentials*. The path by which a particle actually travels is determined by the postulate (or principle) that it has to follow that particular path, out of infinitely many possible ones, for which a certain quantity – the *action* – is minimized. The action S is defined by

$$S = \int_{t_1}^{t_2} L(q(t), \dot{q}(t))\,\mathrm{d}t \tag{5.38}$$

where $q(t)$ is the position of the particle as a function of time, $\dot{q}(t)$ is its velocity and the all-important function L is the Lagrangian. Given L as an explicit function of the variables $q(t)$ and $\dot{q}(t)$, we can imagine evaluating S for all sorts of possible $q(t)$'s starting at time t_1 and ending at time t_2. We can draw these different *possible trajectories* on a q versus t diagram as in figure 5.4. For each path we evaluate S: the *actual* path is the one for which S is smallest, by hypothesis.

But what is L? In simple cases (as we shall verify later) L is just $T - V$, the difference of kinetic and potential energies. Thus for a single particle in a potential V

$$L = \tfrac{1}{2}m\dot{x}^2 - V(x). \tag{5.39}$$

Knowing $V(x)$, we can try and put the 'action principle' into action. However, how can we set about finding which trajectory minimizes S? It is quite interesting to play with some simple specific examples and actually calculate S for several 'fictitious' trajectories – i.e. ones that we know from the Newtonian approach are *not* followed by the particle – and try and get a feeling for what the actual trajectory that minimizes S might be like (of course it is the Newtonian one – see problem 5.2). But clearly this is not a practical answer to the general problem of finding the $q(t)$ that minimizes S. Actually, we can solve this problem by calculus.

Our problem is something like the familiar one of finding the point t_0 at which a certain function $f(t)$ has a stationary value. In the present case, however, the function S is not a simple function of t – rather it is a function of the entire set of points $q(t)$. It is a *function of the function* $q(t)$, or a '*functional*' of $q(t)$. We want to know what particular '$q_c(t)$' minimizes S.

By analogy with the single-variable case, we consider a small variation $\delta q(t)$ in the path from $q(t_1)$ to $q(t_2)$. At the minimum, the change δS corresponding to the change δq must vanish. This change in the action is given by

$$\delta S = \int_{t_1}^{t_2} \left(\frac{\partial L}{\partial q(t)} \delta q(t) + \frac{\partial L}{\partial \dot{q}(t)} \delta \dot{q}(t) \right) \mathrm{d}t. \tag{5.40}$$

Using $\delta \dot{q}(t) = \mathrm{d}(\delta q(t))/\mathrm{d}t$ and integrating the second term by parts yields

$$\delta S = \int_{t_1}^{t_2} \delta q(t) \left[\frac{\partial L}{\partial q(t)} - \frac{\mathrm{d}}{\mathrm{d}t} \frac{\partial L}{\partial \dot{q}(t)} \right] \mathrm{d}t + \left[\frac{\partial L}{\partial \dot{q}(t)} \delta q(t) \right]_{t_1}^{t_2}. \tag{5.41}$$

Since we are considering variations of path in which all trajectories start at t_1 and end at t_2, $\delta q(t_1) = \delta q(t_2) = 0$. So the condition that S be stationary is

$$\delta S = \int_{t_1}^{t_2} \delta q(t) \left[\frac{\partial L}{\partial q(t)} - \frac{\mathrm{d}}{\mathrm{d}t} \frac{\partial L}{\partial \dot{q}(t)} \right] \mathrm{d}t = 0. \tag{5.42}$$

Since this must be true for *arbitrary* $\delta q(t)$, we must have

$$\boxed{\frac{\partial L}{\partial q(t)} - \frac{\mathrm{d}}{\mathrm{d}t} \frac{\partial L}{\partial \dot{q}(t)} = 0.} \tag{5.43}$$

This is the celebrated Euler–Lagrange equation of motion. Its solution gives the '$q_c(t)$' which the particle actually follows.

We can see how this works for the simple case (5.39) where q is the coordinate x. We have immediately

$$\partial L/\partial \dot{x} = m\dot{x} = p \tag{5.44}$$

and

$$\partial L/\partial x = -\partial V/\partial x = F \tag{5.45}$$

where p and F are, respectively, the momentum and the force of the Newtonian approach. The Euler–Lagrange equation then reads

$$F = \mathrm{d}p/\mathrm{d}t \tag{5.46}$$

precisely the Newtonian equation of motion. For the special case of a harmonic oscillator (obviously fundamental for the quantum field idea, as section 5.1 should have made clear), we have

$$L = \tfrac{1}{2} m\dot{x}^2 - \tfrac{1}{2} m\omega^2 x^2 \tag{5.47}$$

which can be immediately generalized to N independent oscillators (see section 5.1) via

$$L = \sum_{r=1}^{N} (\tfrac{1}{2} m \dot{Q}_r^2 - \tfrac{1}{2} m \omega_r^2 Q_r^2). \tag{5.48}$$

For many dynamical systems, the Lagrangian has the form '$T - V$' indicated in (5.47) and (5.48).

Our next step will be to replace classical particle mechanics by quantum particle mechanics. The standard way to do this is via the Hamiltonian formulation of classical mechanics, which we will now briefly review for the simple system with Lagrangian (5.39). In Hamiltonian dynamics, the variables used are not the Lagrangian ones of position x and velocity $\dot{x}$, but rather the position x and the canonical momentum p, where p is defined by

$$p = \frac{\partial L}{\partial \dot{x}}. \tag{5.49}$$

The place of the Lagrangian is taken by the Hamiltonian $H(x, p)$ which is defined by

$$H(x, p) = p\dot{x} - L. \tag{5.50}$$

Using (5.39) for L we find $p = m\dot{x}$, and placing this result in (5.50) we obtain

$$H(x, p) = \frac{p^2}{2m} + V(x) \tag{5.51}$$

which in this case is just the total energy, expressed in terms of x and p. Instead of the Euler-Lagrange equation we have the Hamiltonian equations of motion, which are

$$\frac{\partial H}{\partial p} = \dot{x} \tag{5.52}$$

and

$$\frac{\partial H}{\partial x} = -\dot{p}. \tag{5.53}$$

For the case (5.51) these equations yield

$$p/m = \dot{x} \tag{5.54}$$

and

$$\dot{p} = -\partial V / \partial x. \tag{5.55}$$

Equation (5.54) is just the familiar relation of p to $\dot{x}$, and (5.55) is the Newtonian equation of motion. In the same way, the reader may check that the Hamiltonian for the assembly of oscillators described by the Lagrangian (5.48) is

$$H = \sum_{r=1}^{N} (\frac{P_r^2}{2m} + \frac{1}{2} m \omega_r^2 Q_r^2) \tag{5.56}$$

where $P_r = m\dot{Q}_r$.

With this in hand, we turn to quantum particle mechanics.

5.2.2 Quantum particle mechanics à la Heisenberg–Lagrange–Hamilton

It seems likely that a particularly direct correspondence between the quantum and the classical cases will be obtained if we use the Heisenberg formulation (or 'picture') of quantum mechanics (see appendix I). In the Schrödinger picture, the dynamical variables such as position x are independent of time, and the time dependence is carried by the wavefunction. Thus we seem to have nothing like the $q(t)$'s. However, one can always do a unitary transformation to the Heisenberg picture, in which the wavefunction is fixed and the dynamical variables change with time. This is what we want in order to parallel the classical quantities $q(t)$. But of course there is one fundamental difference between quantum mechanics and classical mechanics: in the former, the dynamical variables are operators which in general do not commute. In particular, the fundamental commutator states that ($\hbar = 1$)

$$[\hat{q}(t), \hat{p}(t)] = \mathrm{i} \tag{5.57}$$

where $\hat{}$ indicates the operator character of the quantity. Here $\hat{p}$ is defined by the generalization of (5.44):

$$\hat{p} = \partial \hat{L}/\partial \dot{\hat{q}}. \tag{5.58}$$

In this formulation of quantum mechanics we do not have the Schrödinger-type equation of motion. Instead we have the Heisenberg equation of motion

$$\dot{\hat{A}} = -\mathrm{i}[\hat{A}, \hat{H}] \tag{5.59}$$

where the Hamiltonian operator $\hat{H}$ is defined in terms of the Lagrangian operator $\hat{L}$ by (cf (5.50))

$$\hat{H} = \hat{p}\dot{\hat{q}} - \hat{L} \tag{5.60}$$

and $\hat{A}$ is any dynamical observable. For example, in the oscillator case

$$\hat{L} = \tfrac{1}{2}m\dot{\hat{q}}^2 - \tfrac{1}{2}m\omega^2\hat{q}^2 \tag{5.61}$$

$$\hat{p} = m\dot{\hat{q}} \tag{5.62}$$

and

$$\hat{H} = \frac{1}{2m}\hat{p}^2 + \frac{1}{2}m\omega^2\hat{q}^2 \tag{5.63}$$

which is the total energy operator. Note that $\hat{p}$, obtained from the Lagrangian using (5.58), had better be consistent with the Heisenberg equation of motion for the operator $\hat{A} = \hat{q}$. The Heisenberg equation of motion for $\hat{A} = \hat{p}$ leads to

$$\dot{\hat{p}} = -m\omega^2\hat{q} \tag{5.64}$$

which is an operator form of Newton's law for the harmonic oscillator. Using the expression for $\hat{p}$ (5.62), we find

$$\ddot{\hat{q}} = -\omega^2\hat{q}. \tag{5.65}$$

Now, although this looks like the familiar classical equation of motion for the position of the oscillator – and recovering it from the Lagrangian formalism is encouraging – we must be very careful to appreciate that this is an equation stating how an *operator* evolves with time. Where the quantum particle will actually be *found* is an entirely different matter. By sandwiching (5.65) between wavefunctions, we can at once see that the *average* position of the particle will follow the classical trajectory (remember that wavefunctions are independent of time in the Heisenberg formulation). But *fluctuations* about this trajectory will certainly occur: a quantum particle does not follow a ray-like classical trajectory. Come to think of it, neither does a photon!

In the original formulations of quantum theory, such fluctuations were generally taken to imply that the very notion of a 'path' was no longer a useful one. However, just as the differential equations satisfied by operators in the Heisenberg picture are quantum generalizations of Newtonian mechanics, so there is an analogous quantum generalization of the 'path-contribution to the action' approach to classical mechanics. The idea was first hinted at by Dirac (1933, 1981, section 32), but it was Feynman who worked it out completely. The book by Feynman and Hibbs (1965) presents a characteristically fascinating discussion – here we only wish to indicate the central idea. We ask: how does a particle get from the point $q(t_1)$ at time t_1 to the point $q(t_2)$ at t_2? Referring back to figure 5.4, in the classical case we imagined (infinitely) many possible paths $q_i(t)$, of which, however, only *one* was the actual path followed, namely the one we called $q_{\rm c}(t)$ which minimized the action integral (5.38) as a functional of $q(t)$. In the quantum case, however, we previously noted that a particle will no longer follow any definite path, because of quantum fluctuations. But rather than, as a consequence, throwing away the whole idea of a path, Feynman's insight was to appreciate that the 'opposite' viewpoint is also possible: since unique paths are forbidden in quantum theory, we should in principle include *all possible* paths! In other words, we take all the trajectories on figure 5.4 as physically possible (together with all the other infinitely many ways of accomplishing the trip).

However, surely not all paths are equally *likely*: after all, we must presumably recover the classical trajectory as $\hbar \to 0$, in some sense. Thus we must find an appropriate weighting for the paths. Feynman's recipe is beautifully simple: weight each path by the factor

$$\mathrm{e}^{\mathrm{i}S/\hbar} \tag{5.66}$$

where S is the action for that particular path. At first sight this is a rather strange proposal, since all paths – even the classical one – are weighted by a quantity which is of unit modulus. But of course contributions of the form (5.66) from all the paths have to be added coherently – just as we superposed the amplitudes in the 'two-slit' discussion in section 2.5. What distinguishes the classical path $q_{\rm c}(t)$ is that it makes S stationary under small changes of path: thus in its vicinity paths have a strong tendency to add up constructively, while far from it the phase factors will tend to produce cancellations.

The amount a quantum particle can 'stray' from the classical path depends on the magnitude of the corresponding action relative to $\hbar$, the quantum of action: the scale of coherence is set by $\hbar$.

In summary, then, the quantum mechanical amplitude to go from $q(t_1)$ to $q(t_2)$ is proportional to

$$\sum_{\text{all paths } q(t)} \exp\left(\frac{\mathrm{i}}{\hbar}\int_{t_1}^{t_2} L(q(t),\dot{q}(t))\,\mathrm{d}t\right). \tag{5.67}$$

There is an evident generalization to quantum field theory. We shall not, however, make use of the 'path integral' approach to quantum field theory in this volume. Its use was, in fact, decisive in obtaining the Feynman rules for non-Abelian gauge theories; and it is the only approach suitable for numerical studies of quantum field theories (how can operators be simulated numerically?). Nevertheless, for a first introduction to quantum field theory, there is still much to be said for the traditional approach based on 'quantizing the modes', and this is the path we shall follow in the rest of this volume. Not the least of its advantages is that it contains the intuitively powerful 'calculus' of creation and annihilation operators, as we now describe. We shall return to the path integral formalism in chapter 16 of volume 2.

5.2.3 Interlude: the quantum oscillator

As we saw in section 5.1, we need to know the energy spectrum and associated states of a quantum harmonic oscillator. This is a standard problem, but there is one particular way of solving it – the 'operator' approach due to Dirac (1981, chapter 6) – that is so crucial to all subsequent development that we include a discussion here in the body of the text.

For the oscillator Hamiltonian

$$\hat{H} = \frac{1}{2m}\hat{p}^2 + \frac{1}{2}m\omega^2\hat{q}^2 \tag{5.68}$$

if $\hat{p}$ and $\hat{q}$ were not operators, we could attempt to factorize the Hamiltonian in the form '$(q+\mathrm{i}p)(q-\mathrm{i}p)$' (apart from the factors of $2m$ and ω). In the quantum case, in which $\hat{p}$ and $\hat{q}$ do not commute, it still turns out to be very helpful to introduce such combinations. If we define the operator

$$\hat{a} = \frac{1}{\sqrt{2}}\left(\sqrt{m\omega}\hat{q} + \frac{\mathrm{i}}{\sqrt{m\omega}}\hat{p}\right) \tag{5.69}$$

and its Hermitian conjugate

$$\hat{a}^\dagger = \frac{1}{\sqrt{2}}\left(\sqrt{m\omega}\hat{q} - \frac{\mathrm{i}}{\sqrt{m\omega}}\hat{p}\right) \tag{5.70}$$

the Hamiltonian may be written as (see problem 5.4)

$$\hat{H} = \tfrac{1}{2}(\hat{a}^\dagger\hat{a} + \hat{a}\hat{a}^\dagger)\omega = (\hat{a}^\dagger\hat{a} + \tfrac{1}{2})\omega. \tag{5.71}$$

The second form for $\hat{H}$ may be obtained from the first using the commutation relation between $\hat{a}$ and $\hat{a}^\dagger$

$$[\hat{a}, \hat{a}^\dagger] = 1 \tag{5.72}$$

derived using the fundamental commutator between $\hat{p}$ and $\hat{q}$. Using this basic commutator (5.72) and our expression for $\hat{H}$, (5.71), one can prove the relations (see problem 5.4)

$$\begin{aligned} [\hat{H}, \hat{a}] &= -\omega \hat{a} \\ [\hat{H}, \hat{a}^\dagger] &= \omega \hat{a}^\dagger. \end{aligned} \tag{5.73}$$

Consider now a state $|n\rangle$ which is an eigenstate of $\hat{H}$ with energy E_n:

$$\hat{H}|n\rangle = E_n|n\rangle. \tag{5.74}$$

Using this definition and the commutators (5.73), we can calculate the energy of the states $(\hat{a}^\dagger|n\rangle)$ and $(\hat{a}|n\rangle)$. We find

$$\hat{H}(\hat{a}^\dagger|n\rangle) = (E_n + \omega)(\hat{a}^\dagger|n\rangle) \tag{5.75}$$

$$\hat{H}(\hat{a}|n\rangle) = (E_n - \omega)(\hat{a}|n\rangle). \tag{5.76}$$

Thus the operators $\hat{a}^\dagger$ and $\hat{a}$ respectively raise and lower the energy of $|n\rangle$ by one unit of ω ($\hbar = 1$). Now since $\hat{H} \sim \hat{p}^2 + \hat{q}^2$ with $\hat{p}$ and $\hat{q}$ Hermitian, we can prove that $\langle\psi|\hat{H}|\psi\rangle$ is positive-definite for any state $|\psi\rangle$. Thus the operator $\hat{a}$ cannot lower the energy indefinitely: there must exist a lowest state $|0\rangle$ such that

$$\boxed{\hat{a}|0\rangle = 0.} \tag{5.77}$$

This defines the lowest-energy state of the system; its energy is

$$\hat{H}|0\rangle = \tfrac{1}{2}\omega|0\rangle \tag{5.78}$$

the 'zero-point energy' of the quantum oscillator. The first excited state is

$$|1\rangle = \hat{a}^\dagger|0\rangle \tag{5.79}$$

with energy $(1 + \frac{1}{2})\omega$. The nth state has energy $(n + \frac{1}{2})\omega$ and is proportional to $(\hat{a}^\dagger)^n|0\rangle$. To obtain a normalization

$$\langle n|n\rangle = 1 \tag{5.80}$$

the correct normalization factor can be shown to be (problem 5.4)

$$|n\rangle = \frac{1}{\sqrt{n!}}(\hat{a}^\dagger)^n|0\rangle. \tag{5.81}$$

Returning to the eigenvalue equation for $\hat{H}$, we have arrived at the result

$$\hat{H}|n\rangle = (\hat{a}^\dagger\hat{a} + \tfrac{1}{2})\omega|n\rangle = (n + \tfrac{1}{2})\omega|n\rangle \tag{5.82}$$

so that the state $|n\rangle$ defined by (5.81) is an eigenstate of the *number operator* $\hat{n} = \hat{a}^\dagger \hat{a}$, with integer eigenvalue n:

$$\hat{n}|n\rangle = n|n\rangle. \tag{5.83}$$

It is straightforward to generalize all the foregoing to a system whose Lagrangian is a sum of N independent oscillators, as in (5.48):

$$\hat{L} = \sum_{r=1}^{N} (\tfrac{1}{2} m \dot{\hat{q}}_r^2 - \tfrac{1}{2} m \omega_r^2 \hat{q}_r^2). \tag{5.84}$$

The required generalization of the basic commutation relations (5.57) is

$$\begin{aligned} [\hat{q}_r, \hat{p}_s] &= \mathrm{i}\delta_{rs} \\ [\hat{q}_r, \hat{q}_s] &= [\hat{p}_r, \hat{p}_s] = 0 \end{aligned} \tag{5.85}$$

since the different oscillators labelled by the index r or s are all independent. The Hamiltonian is (cf (5.56))

$$\hat{H} = \sum_{r=1}^{N} [(1/2m)\hat{p}_r^2 + \tfrac{1}{2} m \omega_r^2 \hat{q}_r^2] \tag{5.86}$$

$$= \sum_{r=1}^{N} (\hat{a}_r^\dagger \hat{a}_r + \tfrac{1}{2}) \omega_r \tag{5.87}$$

with $\hat{a}_r$ and $\hat{a}_r^\dagger$ defined via the analogues of (5.69) and (5.70). Since the eigenvalues of *each* number operator $\hat{n}_r = \hat{a}_r^\dagger \hat{a}_r$ are n_r, by the previous results, the eigenvalues of $\hat{H}$ indeed have the form (5.29),

$$E = \sum_{r=1}^{N} (n_r + \tfrac{1}{2}) \omega_r. \tag{5.88}$$

The corresponding eigenstates are products $|n_1\rangle|n_2\rangle \cdots |n_N\rangle$ of N individual oscillator eigenstates, where $|n_r\rangle$ contains n_r quanta of excitation, of frequency ω_r; the product state is usually abbreviated to $|n_1, n_2, \ldots, n_N\rangle$. In the ground state of the system, each individual oscillator is unexcited: this state is $|0, 0, \ldots, 0\rangle$, which is abbreviated to $|0\rangle$, where it is understood that

$$\hat{a}_r|0\rangle = 0 \qquad \text{for all } r. \tag{5.89}$$

The operators $\hat{a}_r^\dagger$ *create oscillator quanta*; the operators $\hat{a}_r$ *destroy oscillator quanta*.

5.2.4 Lagrange–Hamilton classical field mechanics

We now consider how to use the Lagrange–Hamilton approach for a *field*, starting again with the classical case and limiting ourselves to one dimension to start with.

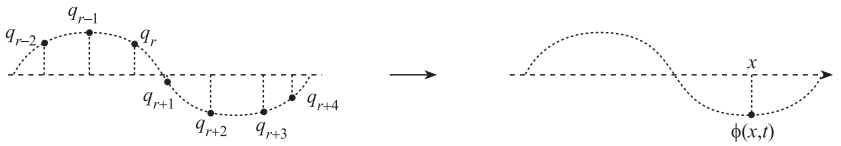


FIGURE 5.5
The passage from a large number of discrete degrees of freedom (mass points) to a continuous degree of freedom (field).

As explained in the previous section, we shall have in mind the $N \to \infty$ limit of the N degrees of freedom case

$$\{q_r(t); r = 1, 2, \ldots, N\} \underset{N\to\infty}{\longrightarrow} \phi(x, t) \tag{5.90}$$

where x is now a continuous variable labelling the displacement of the 'string' (to picture a concrete system, see figure 5.5). At each point x we have an independent degree of freedom $\phi(x, t)$ – thus the *field* system has a 'continuous infinity' of degrees of freedom. We now formulate everything in terms of a Lagrangian density $\mathcal{L}$:

$$S = \int \mathrm{d}t\, L \tag{5.91}$$

where (in one dimension)

$$L = \int \mathrm{d}x\, \mathcal{L}. \tag{5.92}$$

Equation (5.90) suggests that ϕ has dimension of [length], and since in the discrete case $L = T - V$, $\mathcal{L}$ has dimension [energy/length]. (In general $\mathcal{L}$ has dimension [energy/volume].)

A new feature arises because ϕ is now a continuous function of x, so that $\mathcal{L}$ can depend on $\partial\phi/\partial x$ as well as on ϕ and $\dot{\phi} = \partial\phi/\partial t$: $\mathcal{L} = \mathcal{L}(\phi, \partial\phi/\partial x, \dot{\phi})$.

As before, we postulate the same fundamental principle

$$\delta S = 0 \tag{5.93}$$

meaning that the dynamics of the field ϕ is governed by minimizing S. This time the total variation is given by

$$\delta S = \int \mathrm{d}t \int \left[\frac{\partial \mathcal{L}}{\partial \phi}\delta\phi + \frac{\partial \mathcal{L}}{\partial(\partial\phi/\partial x)}\delta\left(\frac{\partial\phi}{\partial x}\right) + \frac{\partial \mathcal{L}}{\partial \dot{\phi}}\delta\dot{\phi} \right] \mathrm{d}x. \tag{5.94}$$

Integrating the $\delta\dot{\phi}$ by parts in t, and the $\delta(\partial\phi/\partial x)$ by parts in x, and discarding the resulting 'surface' terms, we obtain

$$\delta S = \int \mathrm{d}t \int \mathrm{d}x\, \delta\phi \left[\frac{\partial \mathcal{L}}{\partial \phi} - \frac{\partial}{\partial x}\left(\frac{\partial \mathcal{L}}{\partial(\partial\phi/\partial x)}\right) - \frac{\partial}{\partial t}\left(\frac{\partial \mathcal{L}}{\partial \dot{\phi}}\right) \right]. \tag{5.95}$$

Since $\delta\phi$ is an *arbitrary* function, the requirement $\delta S = 0$ yelds the Euler–Lagrange *field* equation

$$\frac{\partial \mathcal{L}}{\partial \phi} - \frac{\partial}{\partial x}\left(\frac{\partial \mathcal{L}}{\partial(\partial\phi/\partial x)}\right) - \frac{\partial}{\partial t}\left(\frac{\partial \mathcal{L}}{\partial \dot{\phi}}\right) = 0. \tag{5.96}$$

The generalization to three dimensions is

$$\boxed{\frac{\partial \mathcal{L}}{\partial \phi} - \boldsymbol{\nabla}\cdot\left(\frac{\partial \mathcal{L}}{\partial(\boldsymbol{\nabla}\phi)}\right) - \frac{\partial}{\partial t}\left(\frac{\partial \mathcal{L}}{\partial \dot{\phi}}\right) = 0.} \tag{5.97}$$

As an example, consider

$$\mathcal{L}_\rho = \frac{1}{2}\rho\left(\frac{\partial\phi}{\partial t}\right)^2 - \frac{1}{2}\rho c^2\left(\frac{\partial\phi}{\partial x}\right)^2 \tag{5.98}$$

where the factor ρ (mass density) and c (a velocity) have been introduced to get the dimension of $\mathcal{L}$ right. Inserting this into the Euler–Lagrangian field equation (5.96), we obtain

$$\frac{\partial^2\phi}{\partial x^2} - \frac{1}{c^2}\frac{\partial^2\phi}{\partial t^2} = 0 \tag{5.99}$$

which is precisely the wave equation (5.30) for the one-dimensional string, now obtained via the Euler–Lagrange field equations. Note that the Lagrange density $\mathcal{L}$ has the expected form (cf (5.48)) of 'kinetic energy density minus potential energy density'.

For the final step – the passage to quantum mechanics for a field system – we shall be interested in the Hamiltonian (total energy) of the system, just as we were for the discrete case. Though we shall not actually *use* the Hamiltonian in the classical field case, we shall introduce it here, generalizing it to the quantum theory in the following section. We recall that Hamiltonian mechanics is formulated in terms of coordinate variables ('q') and momentum variables ('p'), rather than the q and $\dot{q}$ of Lagrangian mechanics. In the continuum (field) case, the Hamiltonian H is written as the integral of a density $\mathcal{H}$ (we remain in one dimension)

$$H = \int \mathrm{d}x\, \mathcal{H} \tag{5.100}$$

while the coordinates $q_r(t)$ become the 'coordinate field' $\phi(x,t)$. The question is what is the corresponding 'momentum field'?

The answer to this is provided by a continuum version of the generalized momentum derived from the Lagrangian approach (cf equation (5.44))

$$p = \partial L/\partial\dot{q}. \tag{5.101}$$

We define a 'momentum field' $\pi(x,t)$ – technically called the 'momentum canonically conjugate to ϕ' – by

$$\pi(x,t) = \partial\mathcal{L}/\partial\dot{\phi}(x,t) \tag{5.102}$$

where $\mathcal{L}$ is now the Lagrangian density. Note that π has dimensions of a momentum density. In the classical particle mechanics case we define the Hamiltonian by

$$H(p,q) = p\dot{q} - L. \tag{5.103}$$

Here we define a Hamiltonian density $\mathcal{H}$ by

$$\mathcal{H}(\phi,\pi) = \pi(x,t)\dot{\phi}(x,t) - \mathcal{L}. \tag{5.104}$$

Let us see how all this works for the one-dimensional string with $\mathcal{L}$ given by

$$\mathcal{L}_\rho = \frac{1}{2}\rho\left(\frac{\partial\phi}{\partial t}\right)^2 - \frac{1}{2}\rho c^2\left(\frac{\partial\phi}{\partial x}\right)^2. \tag{5.105}$$

We have

$$\pi(x,t) = \rho\partial\phi/\partial t \tag{5.106}$$

and

$$\begin{aligned}\mathcal{H}_\rho &= \frac{1}{\rho}\pi^2 - \frac{1}{2}\left[\frac{1}{\rho}\pi^2 - \rho c^2\left(\frac{\partial\phi}{\partial x}\right)^2\right] \\ &= \frac{1}{2}\left[\frac{1}{\rho}\pi^2 + \rho c^2\left(\frac{\partial\phi}{\partial x}\right)^2\right]\end{aligned} \tag{5.107}$$

so that

$$H_\rho = \int_0^\ell \left[\frac{1}{2\rho}\pi^2(x,t) + \frac{1}{2}\rho c^2\left(\frac{\partial\phi(x,t)}{\partial x}\right)^2\right] \mathrm{d}x. \tag{5.108}$$

This has exactly the form we expect (see (5.35)), thus verifying the plausibility of the above prescription.

Inserting the mode expansion (5.34) into (5.92) and (5.105) we obtain the result (just as in (5.36) and problem 5.1)

$$L_\rho = \int_0^\ell \mathrm{d}x\ \mathcal{L}_\rho = \frac{\ell}{2}\sum_{r=1}^{\infty}\left[\frac{1}{2}\rho\dot{A}_r^2 - \frac{1}{2}\rho\omega_r^2 A_r^2\right], \tag{5.109}$$

confirming that the system is equivalent to an infinite number of oscillators. The momentum canonically conjugate to A_r is

$$p_r = \frac{\partial L_\rho}{\partial\dot{A}_r} = \frac{\ell}{2}\rho\dot{A}_r \tag{5.110}$$

and the Hamiltonian is

$$H_\rho = \sum_{r=1}^{\infty} \frac{p_r^2}{\ell\rho} + \frac{\ell}{4}\rho\omega_r^2 A_r^2. \tag{5.111}$$

We may cast (5.111) into nicer form by the change of variables

$$P_r = \sqrt{2/\ell}\; p_r, \quad Q_r = \sqrt{\ell/2}\; A_r, \tag{5.112}$$

in terms of which

$$H_\rho = \sum_{r=1}^{\infty} \frac{P_r^2}{2\rho} + \frac{1}{2}\rho\omega_r^2 Q_r^2 \tag{5.113}$$

just as in (5.56), with $N \to \infty$.

5.2.5 Heisenberg–Lagrange–Hamilton quantum field mechanics

Finally, we are ready to quantize classical field formalism, and arrive at a quantum field mechanics – at least for the scalar field $\phi(x,t)$. If we were dealing with the case in which $\phi(x,t)$ represented the displacement of a one-dimensional stretched string, quantization would be straightforward. We would take the classical Hamiltonian (5.113) and promote the mode coordinates Q_r and their conjugate momenta P_r to operators satisfying commutation relations of the form (5.85). The rest of the analysis would be exactly as in equations (5.86) to (5.89), except that the number of modes N is infinite. But in the case of the general scalar field, we do not want to impose the boundary conditions $\phi(0,t) = \phi(\ell,t) = 0$, which led to the mode expansion (5.34). It is then not so clear how to proceed.

Fortunately, the Lagrange-Hamilton field formalism does indicate the way forward, which is one good reason for developing it in the first place. (Another is that it is very well suited to the analysis of symmetries, a crucial aspect of gauge theories – see chapter 7.) In the previous section we introduced the 'coordinate-like' field $\phi(x,t)$ and (via the Lagrangian) the 'momentum-like' field $\pi(x,t)$. To pass to the quantized version of the field theory, we mimic the procedure followed in the discrete case and promote both the quantities ϕ and π to operators $\hat{\phi}$ and $\hat{\pi}$, in the Heisenberg picture. As usual, the distinctive feature of quantum theory is the non-commutativity of certain basic quantities in the theory – for example, the fundamental commutator ($\hbar = 1$)

$$[\hat{q}_r(t), \hat{p}_s(t)] = \mathrm{i}\delta_{rs} \tag{5.114}$$

of the discrete case. Thus we expect that the operators $\hat{\phi}$ and $\hat{\pi}$ will obey some commutation relation which is a continuum generalization of (5.114). The commutator will be of the form $[\hat{\phi}(x,t), \hat{\pi}(y,t)]$, since – recalling figure 5.5 – the discrete index r or s becomes the continuous variable x or y; we

also note that (5.114) is between operators at equal times. The continuum generalization of the δ_{rs} symbol is the Dirac δ function, $\delta(x-y)$, with the properties

$$\int_{-\infty}^{\infty} \delta(x)\,\mathrm{d}x = 1 \tag{5.115}$$

$$\int_{-\infty}^{\infty} \delta(x-y)f(x)\,\mathrm{d}x = f(y) \tag{5.116}$$

for all reasonable functions f (see appendix E). Thus the fundamental commutator of quantum field theory is taken to be

$$\boxed{[\hat{\phi}(x,t), \hat{\pi}(y,t)] = \mathrm{i}\delta(x-y)} \tag{5.117}$$

in the one-dimensional case, with obvious generalization to the three-dimensional case via the symbol $\delta^3(\boldsymbol{x}-\boldsymbol{y})$. Remembering that we have set $\hbar = 1$, it is straightforward to check that the dimensions are consistent on both sides. Variables $\hat{\phi}$ and $\hat{\pi}$ obeying such a commutation relation are said to be 'conjugate' to each other.

What about the commutator of two $\hat{\phi}$'s or two $\hat{\pi}$'s? In the discrete case, two different $\hat{q}$'s (in the Heisenberg picture) will commute at equal times, $[\hat{q}_r(t), \hat{q}_s(t)] = 0$, and so will two different $\hat{p}$'s. We therefore expect to supplement (5.117) with

$$[\hat{\phi}(x,t), \hat{\phi}(y,t)] = [\hat{\pi}(x,t), \hat{\pi}(y,t)] = 0. \tag{5.118}$$

Let us now proceed to explore the effect of these fundamental commutator assumptions, for the case of the Lagrangian density which yielded the wave equation via the Euler–Lagrange equations, namely

$$\hat{\mathcal{L}}_\rho = \frac{1}{2}\rho\left(\frac{\partial\hat{\phi}}{\partial t}\right)^2 - \frac{1}{2}\rho c^2\left(\frac{\partial\hat{\phi}}{\partial x}\right)^2. \tag{5.119}$$

If we remove ρ, and set $c = 1$, we obtain

$$\hat{\mathcal{L}} = \frac{1}{2}\left(\frac{\partial\hat{\phi}}{\partial t}\right)^2 - \frac{1}{2}\left(\frac{\partial\hat{\phi}}{\partial x}\right)^2 \tag{5.120}$$

for which the Euler–Lagrangian equation yields the field equation

$$\frac{\partial^2\hat{\phi}}{\partial t^2} - \frac{\partial^2\hat{\phi}}{\partial x^2} = 0. \tag{5.121}$$

We can think of (5.121) as a highly simplified (spin-0, one-dimensional) version of the wave equation satisfied by the electromagnetic potentials. We may guess, then, that the associated quanta are massless, as we shall soon confirm.

The Lagrangian density (5.120) is our prototype quantum field Lagrangian (one often slips into leaving out the word 'density'). Applying the quantized version of (5.95) we then have

$$\hat{\pi}(x,t) = \frac{\partial \hat{\mathcal{L}}}{\partial \dot{\hat{\phi}}(x,t)} = \dot{\hat{\phi}}(x,t) \tag{5.122}$$

and the Hamiltonian density is

$$\hat{\mathcal{H}} = \hat{\pi}\dot{\hat{\phi}} - \hat{\mathcal{L}} = \frac{1}{2}\hat{\pi}^2 + \frac{1}{2}\left(\frac{\partial \hat{\phi}}{\partial x}\right)^2. \tag{5.123}$$

The total Hamiltonian is

$$\hat{H} = \int \hat{\mathcal{H}}\,\mathrm{d}x = \int \frac{1}{2}\left[\hat{\pi}^2 + \left(\frac{\partial \hat{\phi}}{\partial x}\right)^2\right]\,\mathrm{d}x. \tag{5.124}$$

It is not immediately clear how to find the eigenvalues and eigenstates of the operator $\hat{H}$. However, it is exactly at this point that all our preliminary work on *normal modes* comes into its own. If we can write the Hamiltonian as some kind of sum over independent oscillators – i.e. modes – we shall know how to proceed. For the classical string with fixed end points which was considered in section 5.1, the mode expansion was simply a Fourier expansion. In the present case, we want to allow the field to extend throughout all of space, without the periodicity imposed by fixed-end boundary conditions. In that case, the Fourier series is replaced by a Fourier integral, and standing waves are replaced by travelling waves. For the classical field obeying the wave equation (5.30) there are plane-wave solutions

$$\phi(x,t) \propto \mathrm{e}^{\mathrm{i}kx - \mathrm{i}\omega t} \tag{5.125}$$

where ($c = 1$)

$$\omega = k \tag{5.126}$$

which is just the dispersion relation of light *in vacuo*. The general field may be Fourier expanded in terms of these solutions:

$$\phi(x,t) = \int_{-\infty}^{\infty} \frac{\mathrm{d}k}{2\pi\sqrt{2\omega}}[a(k)\mathrm{e}^{\mathrm{i}kx - \mathrm{i}\omega t} + a^*(k)\mathrm{e}^{-\mathrm{i}kx + \mathrm{i}\omega t}] \tag{5.127}$$

where we have required ϕ to be real. (The rather fussy factors $(2\pi\sqrt{2\omega})^{-1}$ are purely conventional, and determine the normalization of the expansion coefficients a, a^* and $\hat{a}$, $\hat{a}^\dagger$ later; in turn, the latter enter into the definition, and normalization, of the states – see (5.143)). Similarly, the 'momentum field' $\pi = \dot{\phi}$ is expanded as

$$\pi = \int_{-\infty}^{\infty} \frac{\mathrm{d}k}{2\pi\sqrt{2\omega}}(-\mathrm{i}\omega)[a(k)\mathrm{e}^{\mathrm{i}kx - \mathrm{i}\omega t} - a^*(k)\mathrm{e}^{-\mathrm{i}kx + \mathrm{i}\omega t}]. \tag{5.128}$$

We quantize these mode expressions by promoting $\phi \to \hat{\phi}$, $\pi \to \hat{\pi}$ and assuming the commutator (5.117). Thus we write

$$\hat{\phi} = \int_{-\infty}^{\infty} \frac{\mathrm{d}k}{2\pi\sqrt{2\omega}} [\hat{a}(k)\mathrm{e}^{\mathrm{i}kx - \mathrm{i}\omega t} + \hat{a}^\dagger(k)\mathrm{e}^{-\mathrm{i}kx + \mathrm{i}\omega t}] \tag{5.129}$$

and similarly for $\hat{\pi}$. The commutator (5.117) now *determines* the commutators of the *mode operators* $\hat{a}$ and $\hat{a}^\dagger$:

$$\begin{aligned} [\hat{a}(k), \hat{a}^\dagger(k')] &= 2\pi\delta(k - k') \\ [\hat{a}(k), \hat{a}(k')] &= [\hat{a}^\dagger(k), \hat{a}^\dagger(k')] = 0 \end{aligned} \tag{5.130}$$

as shown in problem 5.6. These are the desired continuum analogues of the *discrete* oscillator commutation relations

$$\begin{aligned} [\hat{a}_r, \hat{a}_s^\dagger] &= \delta_{rs} \\ [\hat{a}_r, \hat{a}_s] &= [\hat{a}_r^\dagger, \hat{a}_s^\dagger] = 0. \end{aligned} \tag{5.131}$$

The precise factor in front of the δ-function in (5.130) depends on the normalization choice made in the expansion of $\hat{\phi}$, (5.129). Problem 5.6 also shows that the commutation relations (5.130) lead to (5.118) as expected.

The form of the $\hat{a}$, $\hat{a}^\dagger$ commutation relations (5.130) already suggests that the $\hat{a}(k)$ and $\hat{a}^\dagger(k)$ operators are precisely the single-quantum destruction and creation operators for the continuum problem. To verify this interpretation and find the eigenvalues of $\hat{H}$, we now insert the expansion for $\hat{\phi}$ and $\hat{\pi}$ into $\hat{H}$ of (5.124). One finds the remarkable result (problem 5.7)

$$\hat{H} = \int_{-\infty}^{\infty} \frac{\mathrm{d}k}{2\pi} \left\{ \frac{1}{2} [\hat{a}^\dagger(k)\hat{a}(k) + \hat{a}(k)\hat{a}^\dagger(k)]\omega \right\}. \tag{5.132}$$

Comparing this with the single-oscillator result

$$\hat{H} = \tfrac{1}{2}(\hat{a}^\dagger\hat{a} + \hat{a}\hat{a}^\dagger)\omega \tag{5.133}$$

shows that, as anticipated in section 5.1, each classical mode of the field can be quantized, and behaves like a separate oscillator coordinate, with its own frequency $\omega = k$. The operator $\hat{a}^\dagger(k)$ creates, and $\hat{a}(k)$ destroys, a quantum of the k mode. The factor $(2\pi)^{-1}$ in $\hat{H}$ arises from our normalization choice.

We note that in the field operator $\hat{\phi}$ of (5.129), those terms which destroy quanta go with the factor $\mathrm{e}^{-\mathrm{i}\omega t}$, while those which create quanta go with $\mathrm{e}^{+\mathrm{i}\omega t}$. This choice is deliberate and is consistent with the 'absorption' and 'emission' factors $\mathrm{e}^{\pm\mathrm{i}\omega t}$ of ordinary time-dependent perturbation theory in quantum mechanics (cf equation (A.33) of appendix A).

What is the mass of these quanta? We know that their frequency ω is related to their wavenumber k by (5.126), which – restoring $\hbar$'s and c's – can be regarded as equivalent to $\hbar\omega = \hbar ck$, or $E = cp$, where we use the Einstein

and de Broglie relations. This is precisely the E–p relation appropriate to a *massless* particle, as expected.

What is the energy spectrum? We expect the ground state to be determined by the continuum analogue of

$$\hat{a}_r|0\rangle = 0 \qquad \text{for all } r; \tag{5.134}$$

namely

$$\hat{a}(k)|0\rangle = 0 \qquad \text{for all } k. \tag{5.135}$$

However, there is a problem with this. If we allow the Hamiltonian of (5.132) to act on $|0\rangle$ the result is not (as we would expect) zero, because of the $\hat{a}(k)\hat{a}^\dagger(k)$ term (the other term does give zero by (5.135)). In the *single* oscillator case, we rewrote $\hat{a}\hat{a}^\dagger$ in terms of $\hat{a}^\dagger\hat{a}$ by using the commutation relation (5.72), and this led to the 'zero-point energy', $\frac{1}{2}\omega$, of the oscillator ground state. Adopting the same strategy here, we write $\hat{H}$ of (5.132) as

$$\hat{H} = \int \frac{\mathrm{d}k}{2\pi}\hat{a}^\dagger(k)\hat{a}(k)\omega + \int \frac{\mathrm{d}k}{2\pi}\frac{1}{2}[\hat{a}(k), \hat{a}^\dagger(k)]\omega. \tag{5.136}$$

Now consider $\hat{H}|0\rangle$: we see from the definition of the vacuum (5.135) that the first term will give zero as expected – but the second term is infinite, since the commutation relation (5.130) produces the infinite quantity '$\delta(0)$' as $k \to k'$; moreover, the k integral diverges.

This term is obviously the continuum analogue of the zero-point energy $\frac{1}{2}\omega$ – but because there are infinitely many oscillators, it is infinite. The conventional ploy is to argue that only energy *differences*, relative to a conveniently defined ground state, really matter – so that we may discard the infinite constant in (5.136). Then the ground state $|0\rangle$ has energy zero, by definition, and the eigenvalues of $\hat{H}$ are of the form

$$\int \frac{\mathrm{d}k}{2\pi} n(k)\omega \tag{5.137}$$

where $n(k)$ is the number of quanta (counted by the number operator $\hat{a}^\dagger(k)\hat{a}(k)$) of energy $\omega = k$. For each definite k, and hence ω, the spectrum is like that of the simple harmonic oscillator. The process of going from (5.132) to (5.136) *without* the second term is called 'normally ordering' the $\hat{a}$ and $\hat{a}^\dagger$ operators: in a 'normally ordered' expression, all $\hat{a}^\dagger$'s are to the left of all $\hat{a}$'s, with the result that the vacuum value of such expressions is by definition zero.

It has to be admitted that the argument that only energy differences matter is false as far as gravity is concerned, which couples to all sources of energy. It would ultimately be desirable to have theories in which the vacuum energy came out finite from the start (as actually happens in 'supersymmetric' field theories – see for example Weinberg (1995), p 325); see also comment (3).

We proceed on to the excited states. Any desired state in which excitation quanta are present can be formed by the appropriate application of $\hat{a}^\dagger(k)$ operators to the ground state $|0\rangle$. For example, a two-quantum state containing

one quantum of momentum k_1 and another of momentum k_2 may be written (cf (5.81))

$$|k_1, k_2\rangle \propto \hat{a}^\dagger(k_1)\hat{a}^\dagger(k_2)|0\rangle. \tag{5.138}$$

A general state will contain an arbitrary number of quanta.

Once again, and this time more formally, we have completed the programme outlined in section 5.1, ending up with the 'quantization' of a classical field $\phi(x,t)$, as exemplified in the basic expression (5.129), together with the interpretation of the operators $\hat{a}(k)$ and $\hat{a}^\dagger(k)$ as destruction and creation operators for mode quanta. We have, at least implicitly, still retained up to this point the 'mechanical model' of some material object oscillating – some kind of infinitely extended 'jelly'. We now throw away the mechanical props and embrace the unadorned quantum field theory! We do not ask *what* is waving, we simply postulate a field – such as ϕ – and quantize it. *Its quanta of excitation are what we call particles* – for example, photons in the electromagnetic case.

We end this long section with some further remarks about the formalism, and the physical interpretation of our quantum field $\hat{\phi}$.

Comment (1)

The alert reader, who has studied appendix I, may be worried about the following (possible) consistency problem. The fields $\hat{\phi}$ and $\hat{\pi}$ are Heisenberg picture operators, and obey the equations of motion

$$\dot{\hat{\phi}}(x,t) = -\mathrm{i}[\hat{\phi}(x,t), \hat{H}] \tag{5.139}$$

$$\dot{\hat{\pi}}(x,t) = -\mathrm{i}[\hat{\pi}(x,t), \hat{H}] \tag{5.140}$$

where $\hat{H}$ is given by (5.132). It is a good exercise to check (problem 5.8(a)) that (5.139) yields just the expected relation $\dot{\hat{\phi}}(x,t) = \hat{\pi}(x,t)$ (cf (5.122)). Thus (5.140) becomes

$$\ddot{\hat{\phi}}(x,t) = -\mathrm{i}[\hat{\pi}(x,t), \hat{H}]. \tag{5.141}$$

However, we have assumed in our work here that $\hat{\phi}$ obeyed the wave equation (cf.(5.121))

$$\ddot{\hat{\phi}} = \frac{\partial^2}{\partial x^2}\hat{\phi}(x,t) \tag{5.142}$$

as a consequence of the quantized version of the Euler–Lagrange equation (5.96). Thus the right-hand sides of (5.141) and (5.142) need to be the same, for consistency – and they are: see problem 5.8(b). Thus – at least in this case – the Heisenberg operator equations of motion are consistent with the Euler–Lagrange equations.

Comment (2)

Following on from this, we may note that this formalism encompasses both the wave and the particle aspects of matter and radiation. The former is evi-

dent from the plane-wave expansion functions in the expansion of $\hat{\phi}$, (5.129), which in turn originate from the fact that $\hat{\phi}$ obeys the wave equation (5.121). The latter follows from the discrete nature of the energy spectrum and the associated operators $\hat{a}$, $\hat{a}^\dagger$ which refer to individual quanta i.e. *particles*.

Comment (3)

Next, we may ask: what is the meaning of the ground state $|0\rangle$ for a quantum field? It is undoubtedly the state with $n(k) = 0$ for all k, i.e. the state with no quanta in it – and hence no *particles* in it, on our new interpretation. It is therefore the vacuum! As we shall see later, this understanding of the vacuum as the ground state of a field system is fundamental to much of modern particle physics – for example, to quark confinement and to the generation of mass for the weak vector bosons. Note that although we discarded the overall (infinite) constant in $\hat{H}$, differences in zero-point energies *can* be detected; for example, in the Casimir effect (Casimir 1948, Kitchener and Prosser 1957, Sparnaay 1958, Lamoreaux 1997, 1998). These and other aspects of the quantum field theory vacuum are discussed in Aitchison (1985).

Comment (4)

Consider the two-particle state (5.138): $|k_1, k_2\rangle \propto \hat{a}^\dagger(k_1)\hat{a}^\dagger(k_2)|0\rangle$. Since the $\hat{a}^\dagger$ operators commute, (5.130), this state is symmetric under the interchange $k_1 \leftrightarrow k_2$. This is an inevitable feature of the formalism as so far developed – there is no possible way of distinguishing one quantum of energy from another, and we expect the two-quantum state to be indifferent to the order in which the quanta are put in it. However, this has an important implication for the *particle* interpretation: since the state is symmetric under interchange of the particle labels k_1 and k_2, it must describe identical *bosons*. How the formalism is modified in order to describe the antisymmetric states required for two fermionic quanta will be discussed in section 7.2.

Comment (5)

Finally, the reader may well wonder how to connect the quantum field theory formalism to ordinary 'wavefunction' quantum mechanics. The ability to see this connection will be important in subsequent chapters and it is indeed quite simple. Suppose we form a state containing one quantum of the $\hat{\phi}$ field, with momentum k':

$$|k'\rangle = N\hat{a}^\dagger(k')|0\rangle \tag{5.143}$$

where N is a normalization constant. Now consider the amplitude $\langle 0|\hat{\phi}(x,t)|k'\rangle$. We expand this out as

$$\langle 0|\hat{\phi}(x,t)|k'\rangle = \langle 0| \int \frac{\mathrm{d}k}{2\pi\sqrt{2\omega}} [\hat{a}(k)\mathrm{e}^{\mathrm{i}kx - \mathrm{i}\omega t} + \hat{a}^\dagger(k)\mathrm{e}^{-\mathrm{i}kx + \mathrm{i}\omega t}] N\hat{a}^\dagger(k')|0\rangle. \tag{5.144}$$

The '$\hat{a}^\dagger\hat{a}^\dagger$' term will give zero since $\langle 0|\hat{a}^\dagger = 0$. For the other term we use the commutation relation (5.130) to write it as

$$\langle 0| \int \frac{N\mathrm{d}k}{2\pi\sqrt{2\omega}} [\hat{a}^\dagger(k')\hat{a}(k) + 2\pi\delta(k-k')]\mathrm{e}^{\mathrm{i}kx-\mathrm{i}\omega t}|0\rangle = N\frac{\mathrm{e}^{\mathrm{i}k'x-\mathrm{i}\omega' t}}{\sqrt{2\omega'}} \qquad (5.145)$$

using the vacuum condition once again, and integrating over the δ function using the property (5.116) which sets $k = k'$ and hence $\omega = \omega'$. The vacuum is normalized to unity, $\langle 0|0\rangle = 1$. The normalization constant N can be adjusted according to the desired convention for the normalization of the states and wavefunctions. The result is just the plane-wave *wavefunction* for a particle in the state $|k'\rangle$! Thus we discover that the vacuum to one-particle matrix elements of the field operators are just the familiar wavefunctions of single-particle quantum mechanics. In this connection we can explain some common terminology. The path to quantum field theory that we have followed is sometimes called 'second quantization' – ordinary single-particle quantum mechanics being the first-quantized version of the theory.

5.3 Generalizations: four dimensions, relativity and mass

In the previous section we have shown how quantum mechanics may be married to field theory, but we have considered only one spatial dimension, for simplicity. Now we must generalize to three and incorporate the demands of relativity. This is very easy to do in the Lagrangian approach, for the scalar field $\phi(\boldsymbol{x}, t)$. 'Scalar' means that the field has only one independent component at each point $(\boldsymbol{x}, t)$ – unlike the electromagnetic field, for instance, for which the analogous quantity has four components, making up a 4-vector field $A^\mu(\boldsymbol{x}, t) = (A_0(\boldsymbol{x}, t), \boldsymbol{A}(\boldsymbol{x}, t))$ (see chapter 7). In the quantum case, a one-component field (or wavefunction) is appropriate for spin-0 particles.

As we saw in (5.97), the three-dimensional Euler–Lagrange equations are

$$\frac{\partial\mathcal{L}}{\partial\phi} - \boldsymbol{\nabla}\cdot\frac{\partial\mathcal{L}}{\partial(\boldsymbol{\nabla}\phi)} - \frac{\partial}{\partial t}\left(\frac{\partial\mathcal{L}}{\partial\dot{\phi}}\right) = 0 \qquad (5.146)$$

which may immediately be rewritten in relativistically invariant form

$$\frac{\partial\mathcal{L}}{\partial\phi} - \partial_\mu\left(\frac{\partial\mathcal{L}}{\partial(\partial_\mu\phi)}\right) = 0 \qquad (5.147)$$

where $\partial_\mu = \partial/\partial x^\mu$. Similarly, the action

$$S = \int \mathrm{d}t \int \mathrm{d}^3\boldsymbol{x}\,\mathcal{L} = \int \mathrm{d}^4x\,\mathcal{L} \qquad (5.148)$$

will be relativistically invariant if $\mathcal{L}$ is, since the volume element d^4x is invariant. Thus, to construct a relativistic field theory, we have to construct an invariant density $\mathcal{L}$ and use the already given covariant Euler–Lagrange equation. Thus our previous string Lagrangian

$$\mathcal{L}_\rho = \frac{1}{2}\rho\left(\frac{\partial\phi}{\partial t}\right)^2 - \frac{1}{2}\rho c^2\left(\frac{\partial\phi}{\partial x}\right)^2 \tag{5.149}$$

with $\rho = c = 1$ generalizes to

$$\mathcal{L} = \tfrac{1}{2}\partial_\mu\phi\partial^\mu\phi \tag{5.150}$$

and produces the invariant wave equation

$$\partial_\mu\partial^\mu\phi = \left(\frac{\partial^2}{\partial t^2} - \boldsymbol{\nabla}^2\right)\phi = 0. \tag{5.151}$$

All of this goes through just the same when the fields are quantized.

This invariant Lagrangian describes a field whose quanta are massless. To find the Lagrangian for the case of massive quanta, we need to find the Lagrangian that gives us the Klein–Gordon equation (see section 3.1)

$$(\Box + m^2)\phi(\boldsymbol{x}, t) = 0 \tag{5.152}$$

via the Euler–Lagrangian equations.

The answer is a simple generalization of (5.150):

$$\mathcal{L}_{\mathrm{KG}} = \tfrac{1}{2}\partial_\mu\phi\partial^\mu\phi - \tfrac{1}{2}m^2\phi^2. \tag{5.153}$$

The plane-wave solutions of the field equation – now the KG equation – have frequencies (or energies) given by

$$\omega^2 = \boldsymbol{k}^2 + m^2 \tag{5.154}$$

which is the correct energy–momentum relation for a massive particle.

How do we quantize this field theory? The four-dimensional analogue of the Fourier expansion of the field ϕ takes the form

$$\hat{\phi}(x) = \int_{-\infty}^{\infty} \frac{\mathrm{d}^3\boldsymbol{k}}{(2\pi)^3\sqrt{2\omega}}[\hat{a}(k)\mathrm{e}^{-\mathrm{i}k\cdot x} + \hat{a}^\dagger(k)\mathrm{e}^{\mathrm{i}k\cdot x}] \tag{5.155}$$

with a similar expansion for the 'conjugate momentum' $\hat{\pi} = \dot{\hat{\phi}}$:

$$\hat{\pi}(x) = \int_{-\infty}^{\infty} \frac{\mathrm{d}^3\boldsymbol{k}}{(2\pi)^3\sqrt{2\omega}}(-\mathrm{i}\omega)[\hat{a}(k)\mathrm{e}^{-\mathrm{i}k\cdot x} - \hat{a}^\dagger(k)\mathrm{e}^{\mathrm{i}k\cdot x}]. \tag{5.156}$$

Here $k\cdot x$ is the four-dimensional dot product $k\cdot x = \omega t - \boldsymbol{k}\cdot\boldsymbol{x}$, and $\omega = +(\boldsymbol{k}^2 + m^2)^{1/2}$. The Hamiltonian is found to be

$$\hat{H}_{\mathrm{KG}} = \int \mathrm{d}^3\boldsymbol{x}\hat{\mathcal{H}}_{\mathrm{KG}} = \int_{-\infty}^{\infty} \mathrm{d}^3\boldsymbol{x}\tfrac{1}{2}[\hat{\pi}^2 + \boldsymbol{\nabla}\hat{\phi}\cdot\boldsymbol{\nabla}\hat{\phi} + m^2\hat{\phi}^2] \tag{5.157}$$

and this can be expressed in terms of the $\hat{a}$'s and the $\hat{a}^\dagger$'s using the expansion for $\hat{\phi}$ and $\hat{\pi}$ and the commutator

$$[\hat{a}(k), \hat{a}^\dagger(k')] = (2\pi)^3 \delta^3(\boldsymbol{k} - \boldsymbol{k}') \tag{5.158}$$

with all others vanishing. The result is, as expected,

$$\hat{H}_{\mathrm{KG}} = \frac{1}{2} \int \frac{\mathrm{d}^3\boldsymbol{k}}{(2\pi)^3} [\hat{a}^\dagger(k)\hat{a}(k) + \hat{a}(k)\hat{a}^\dagger(k)]\omega \tag{5.159}$$

and, normally ordering as usual, we arrive at

$$\hat{H}_{\mathrm{KG}} = \int \frac{\mathrm{d}^3\boldsymbol{k}}{(2\pi)^3} \hat{a}^\dagger(k)\hat{a}(k)\omega. \tag{5.160}$$

This supports the physical interpretation of the mode operators $\hat{a}^\dagger$ and $\hat{a}$ as creation and destruction operators for quanta of the field $\hat{\phi}$ as before, except that now the energy–momentum relation for these particles is the relativistic one, for particles of mass m.

Since $\hat{\phi}$ is real ($\hat{\phi} = \hat{\phi}^\dagger$) and has no spin degrees of freedom, it is called a real scalar field. Only field quanta of one type enter – those created by $\hat{a}^\dagger$ and destroyed by $\hat{a}$. Thus $\hat{\phi}$ would correspond physically to a case where there was a unique particle state of a given mass m – for example the π^0 field. Actually, of course, we would not want to describe the π^0 in any fundamental sense in terms of such a field, since we know it is not a point-like object ('ϕ' is defined only at the single space–time point $(\boldsymbol{x}, t)$). The question of whether true 'elementary' scalar fields exist in nature is an interesting one: in the Standard Model, as we shall eventually see in volume 2, the Higgs field is a scalar field (though it contains several components with different charge). It remains to be seen if this field – and the associated quantum, the Higgs boson – is a scalar, and if so whether it is elementary or composite.

We have learned how to describe free relativistic spinless particles of finite mass as the quanta of a relativistic quantum field. We now need to understand *interactions* in quantum field theory.

Problems

5.1 Verify equation (5.36).

5.2 Consider one-dimensional motion under gravity so that $V(x) = -mgx$ in (5.39). Evaluate S of (5.38) for $t_1 = 0$, $t_2 = t_0$, for three possible trajectories:

(a) $x(t) = at$,

(b) $x(t) = \frac{1}{2}gt^2$ (the Newtonian result) and

(c) $x(t) = bt^3$

where the constants a and b are to be chosen so that all the trajectories end at the same point $x(t_0)$.

5.3

(a) Use (5.57) and (5.63) to verify that

$$\hat{p} = m\dot{\hat{q}}$$

is consistent with the Heisenberg equation of motion for $\hat{A} = \hat{q}$.

(b) By similar methods verify that

$$\dot{\hat{p}} = -m\omega^2\hat{q}.$$

5.4

(a) Rewrite the Hamiltonian $\hat{H}$ of (5.63) in terms of the operators $\hat{a}$ and $\hat{a}^\dagger$.

(b) Evaluate the commutator between $\hat{a}$ and $\hat{a}^\dagger$ and use this result together with your expression for $\hat{H}$ from part (a) to verify equation (5.73).

(c) Verify that for $|n\rangle$ given by equation (5.81) the normalization condition

$$\langle n|n\rangle = 1$$

is satisfied.

(d) Verify (5.83) directly using the commutation relation (5.72).

5.5 Treating ψ and ψ^* as independent classical fields, show that the Lagrangian density

$$\mathcal{L} = \mathrm{i}\psi^*\dot{\psi} - (1/2m)\boldsymbol{\nabla}\psi^* \cdot \boldsymbol{\nabla}\psi$$

gives the Schrödinger equation for ψ and ψ^* correctly.

5.6

(a) Verify that the commutation relations for $\hat{a}(k)$ and $\hat{a}^\dagger(k)$ (equations (5.130)) are consistent with the equal time commutation relation between $\hat{\phi}$ and $\hat{\pi}$ (equation (5.117)), and with (5.118).

(b) Consider the *unequal time* commutator $D(x_1, x_2) \equiv [\hat{\phi}(\boldsymbol{x}_1, t_1), \hat{\phi}(\boldsymbol{x}_2, t_2)]$, where $\hat{\phi}$ is a massive KG field in three dimensions. Show that

$$D(x_1, x_2) = \int \frac{\mathrm{d}^3\boldsymbol{k}}{(2\pi)^3 2E}[\mathrm{e}^{-\mathrm{i}k\cdot(x_1-x_2)} - \mathrm{e}^{\mathrm{i}k\cdot(x_1-x_2)}] \qquad (5.161)$$

where $k \cdot (x_1 - x_2) = E(t_1 - t_2) - \boldsymbol{k} \cdot (\boldsymbol{x}_1 - \boldsymbol{x}_2)$, and $E = (\boldsymbol{k}^2 + m^2)^{1/2}$. Note that D is not an operator, and that it depends only

on the difference of coordinates $x_1 - x_2$, consistent with translation invariance. Show that $D(x_1, x_2)$ vanishes for $t_1 = t_2$. Explain why the right-hand side of (5.161) is Lorentz invariant (see the exercise in appendix E), and use this fact to show that $D(x_1, x_2)$ vanishes for all *space-like* separations $(x_1 - x_2)^2 < 0$. Discuss the significance of this result – or see the discussion in section 6.3.2!

5.7 Insert the plane-wave expansions for the operators $\hat{\phi}$ and $\hat{\pi}$ into the equation for $\hat{H}$, (5.124), and verify equation (5.132). [*Hint*: note that ω is defined to be always positive, so that (5.126) should strictly be written $\omega = |k|$.]

5.8

(a) Use (5.117) and (5.124) to verify that $\hat{\pi}(x,t) = \dot{\hat{\phi}}(x,t)$ is consistent with the Heisenberg equation of motion for $\hat{\phi}(x,t)$. [*Hint*: write the integral in (5.124) as over y, not x!]

(b) Similarly, verify the consistency of (5.141) and (5.121).

6

Quantum Field Theory II: Interacting Scalar Fields

6.1 Interactions in quantum field theory: qualitative introduction

In the previous chapter we considered only free – i.e. non-interacting – quantum fields. The fact that they are non-interacting is evident in a number of ways. The mode expansions (5.129) and (5.155) are written in terms of the (free) plane-wave solutions of the associated wave equations. Also the Hamiltonians turned out to be just the sum of individual oscillator Hamiltonians for each mode frequency, as in (5.132) or (5.159). The energies of the quanta add up – they are non-interacting quanta. Finally, since the Hamiltonians are just sums of number operators

$$\hat{n}(k) = \hat{a}^\dagger(k)\hat{a}(k) \tag{6.1}$$

it is obvious that each such operator commutes with the Hamiltonian and is therefore a constant of the motion. Thus two waves, each with one excitation quantum, travelling towards each other will pass smoothly through each other and emerge unscathed on the other side – they will not interact at all.

How can we get the mode quanta to interact? If we return to our discussion of classical mechanical systems in section 5.1, we see that the crucial step in arriving at the 'sum over oscillators' form for the energy was the assumption that the potential energy was quadratic in the small displacements q_r. We expect that 'modes will interact' when we go *beyond this harmonic approximation*. The same is true in the continuous (wave or field) case. In the derivation of the appropriate wave equation you will find that somewhere an approximation like $\tan\phi \approx \phi$ or $\sin\phi \approx \phi$ is made. This linearizes the equation, and solutions to linear equations can be linearly superposed to make new solutions. If we retain higher powers of ϕ, such as ϕ^3, the resulting nonlinear equation has solutions that cannot be obtained by superposing two independent solutions. Thus two waves travelling towards each other will not just pass smoothly through each other: various forms of interaction and distortion of the original waveforms will occur.

What happens when we quantize such anharmonic systems? To gain some idea of the new features that emerge, consider just one 'anharmonic oscillator'

with Hamiltonian

$$\hat{H} = (1/2m)\hat{p}^2 + \tfrac{1}{2}m\omega^2\hat{q}^2 + \lambda\hat{q}^3. \tag{6.2}$$

In terms of the $\hat{a}$ and $\hat{a}^\dagger$ combinations this becomes

$$\hat{H} = \frac{1}{2}(\hat{a}^\dagger\hat{a} + \hat{a}\hat{a}^\dagger)\omega + \frac{\lambda}{(2m\omega)^{3/2}}(\hat{a} + \hat{a}^\dagger)^3 \tag{6.3}$$

$$\equiv \hat{H}_0 + \lambda\hat{H}' \tag{6.4}$$

where $\hat{H}_0$ is our previous free oscillator Hamiltonian. The algebraic tricks we used to find the spectrum of $\hat{H}_0$ do *not* work for this new $\hat{H}$ because of the addition of the $\hat{H}'$ interaction term. In particular, although $\hat{H}_0$ commutes with the number operator $\hat{a}^\dagger\hat{a}$, $\hat{H}'$ does not. Therefore, whatever the eigenstates of $\hat{H}$ are, they will not in general have a definite number of '$\hat{H}_0$ quanta'. In fact, we cannot find an exact algebraic solution to this new eigenvalue problem, and we must resort to *perturbation theory* or to numerical methods.

The perturbative solution to this problem treats $\lambda\hat{H}'$ as a perturbation and expands the true eigenstates of $\hat{H}$ in terms of the eigenstates of $\hat{H}_0$:

$$|\bar{r}\rangle = \sum_n c_{rn}|n\rangle. \tag{6.5}$$

From this expansion we see that, as expected, the true eigenstates $|\bar{r}\rangle$ will 'contain different numbers of $\hat{H}_0$ quanta': $|c_{rn}|^2$ is the probability of finding n '$\hat{H}_0$ quanta' in the state $|\bar{r}\rangle$. Perturbation theory now proceeds by expanding the coefficients c_{rn} and exact energy eigenvalues $\bar{E}_r$ as power series in the strength λ of the perturbation. For example, the exact energy eigenvalue has the expansion

$$\bar{E}_r = E_r^{(0)} + \lambda E_r^{(1)} + \lambda^2 E_r^{(2)} + \cdots \tag{6.6}$$

where

$$\hat{H}_0|r\rangle = E_r^{(0)}|r\rangle \tag{6.7}$$

and

$$E_r^{(1)} = \langle r|\hat{H}'|r\rangle \tag{6.8}$$

$$E_r^{(2)} = \sum_{s\neq r} \frac{\langle r|\hat{H}'|s\rangle\langle s|\hat{H}'|r\rangle}{E_r^{(0)} - E_s^{(0)}}. \tag{6.9}$$

To evaluate the second-order shift in energy, we therefore need to consider matrix elements of the form

$$\langle s|(\hat{a} + \hat{a}^\dagger)^3|r\rangle. \tag{6.10}$$

Keeping careful track of the order of the $\hat{a}$ and $\hat{a}^\dagger$ operators, we can evaluate these matrix elements and find, in this case, that there are non-zero matrix elements for states $\langle s| = \langle r+3|$, $\langle r+1|$, $\langle r-1|$ and $\langle r-3|$.

What about the quantum mechanics of two coupled nonlinear oscillators? In the same way, the general state is assumed to be a superposition

$$|\bar{r}\rangle = \sum_{n_1,n_2} c_{r,n_1n_2}|n_1\rangle|n_2\rangle \tag{6.11}$$

of states of arbitrary numbers of quanta of the unperturbed oscillator Hamiltonians $\hat{H}_{0(1)}$ and $\hat{H}_{0(2)}$. States of the unperturbed system contain definite numbers n_1 and n_2, say, of the '1' and '2' quanta. Perturbation calculations of the interacting system will involve matrix elements connecting such $|n_1\rangle|n_2\rangle$ states to states $|n_1'\rangle|n_2'\rangle$ with different numbers of these quanta.

All this can be summarized by the remark that the typical feature of quantized interacting modes is that we need to consider processes in which the numbers of the different mode quanta are not constants of the motion. This is, of course, exactly what happens when we have collisions between high-energy particles. When far apart the particles, definite in number, are indeed free and are just the mode quanta of some quantized fields. But, when they interact, we must expect to see changes in the numbers of quanta, and can envisage processes in which the number of quanta which emerge finally as free particles is different from the number that originally collided. From the quantum mechanical examples we have discussed, we expect that these interactions will be produced by terms like $\hat{\phi}^3$ or $\hat{\phi}^4$, since the free – 'harmonic' – case has $\hat{\phi}^2$, analogous to $\hat{q}^2$ in the quantum mechanics example. Such terms arise in the solid state phonon application precisely from anharmonic corrections involving the atomic displacements. These terms lead to non-trivial phonon–phonon scattering, the treatment of which forms the basis of the quantum theory of thermal resistivity of insulators. In the quantum field theory case, when we have generalized the formalism to fermions and photons, the nonlinear interaction terms will produce e^+e^- scattering, $q\bar{q}$ annihilation and so on. As in the quantum mechanical case, the basic calculational method will be perturbation theory.

As remarked earlier, the trouble with all these 'real-life' cases is that they involve significant complications due to spin; the corresponding fields then have several components, with attendant complexity in the solutions of the associated free-particle wave equations (Maxwell, Dirac). So in this chapter we shall seek to explain the essence of *the perturbative approach to quantum field dynamics* – which we take to be essentially the Feynman graph version of Yukawa's exchange mechanism – in the context of simple models involving only scalar fields; Maxwell (vector) and Dirac (spinor) fields will be introduced in the following chapter. The route we follow to the 'Feynman rules' is the one first given (with remarkable clarity) by Dyson (1949a), which rapidly became the standard formulation.

Before proceeding it may be worth emphasizing that in introducing a 'non-harmonic' term such as $\hat{\phi}^3$ and thus departing from linearity in that sense, we are in no way affecting the basic linearity of state vector superposition in quantum mechanics (cf (6.11)), which continues to hold.

6.2 Perturbation theory for interacting fields: the Dyson expansion of the *S*-matrix

> On the third day of the journey a remarkable thing happened; going into a sort of semi-stupor as one does after 48 hours of bus-riding, I began to think very hard about physics, and particularly about the rival radiation theories of Schwinger and Feynman. Gradually my thoughts grew more coherent, and before I knew where I was, I had solved the problem that had been in the back of my mind all this year, which was to prove the equivalence of the two theories.
>
> —From a letter from F. J. Dyson to his parents, 18 September 1948, as quoted in Schweber (1994), p 505.

For definiteness, let us consider the Lagrangian

$$\hat{\mathcal{L}} = \tfrac{1}{2}\partial_\mu\hat{\phi}\partial^\mu\hat{\phi} - \tfrac{1}{2}m^2\hat{\phi}^2 - \lambda\hat{\phi}^3 \equiv \hat{\mathcal{L}}_{\mathrm{KG}} - \lambda\hat{\phi}^3 \tag{6.12}$$

with $\lambda > 0$. Equation (6.12) is like '$\hat{\mathcal{L}} = \hat{T} - \hat{V}$' where $\hat{V} = \frac{1}{2}(\boldsymbol{\nabla}\hat{\phi})^2 + \frac{1}{2}m^2\hat{\phi}^2 + \lambda\hat{\phi}^3$ is the 'potential'. Though simple, this Lagrangian is unfortunately not physically sensible. The classical particle analogue potential would have the form $V(q) = \frac{1}{2}\omega q^2 + \lambda q^3$. If we sketch $V(q)$ as a function of q we see that, for small λ, it retains the shape of an oscillator well near $q = 0$, but for q sufficiently large and negative it will 'turn over', tending ultimately to $-\infty$ as $q \to -\infty$. Classically we expect to be able to set up a successful perturbation theory for oscillations about the equilibrium position $q = 0$, provided that the amplitude of the oscillations is not so large as to carry the particle over the 'lip' of the potential; in the latter case, the particle will escape to $q = -\infty$, invalidating a perturbative approach. In the quantum mechanical case the same potential $V(q)$ is more problematical, since the particle can *tunnel* through the barrier separating it from the region where $V \to -\infty$. This means that the ground state will not be stable. An analogous disease affects the quantum field case – the supposed vacuum state will be unstable, and indeed the energy will not be positive-definite.

Nevertheless, as the reader may already have surmised, and we shall confirm later in this chapter, the 'ϕ-cubed' interaction is precisely of the form relevant to Yukawa's exchange mechanism. As we have seen in the previous section, such an interaction will typically give rise to matrix elements between one-quantum and two-quantum states, for example, exactly like the basic Yukawa emission and absorption process. In fact, all that is necessary to make the $\hat{\phi}^3$-type interaction physical is to let it describe, not the 'self-coupling' of a single field, but the 'interactive coupling' of at least two different fields. For example, we may have two scalar fields with quanta 'A' and 'B', and an interaction between them of the form $\lambda\hat{\phi}_{\mathrm{A}}^2\hat{\phi}_{\mathrm{B}}$. This will allow

processes such as $\mathrm{A} \leftrightarrow \mathrm{A} + \mathrm{B}$. Or we may have three such fields, and an interaction $\lambda\hat{\phi}_\mathrm{A}\hat{\phi}_\mathrm{B}\hat{\phi}_\mathrm{C}$, allowing $\mathrm{A} \leftrightarrow \mathrm{B} + \mathrm{C}$ and similar transitions. In these cases the problems with the $\hat{\phi}^3$ self-interaction do not arise. (Incidentally those problems can be eliminated by the addition of a suitable higher-power term, for instance $g\hat{\phi}^4$.) In later sections we shall be considering the 'ABC' model specifically, but for the present it will be simpler to continue with the single field $\hat{\phi}$ and the self-interaction $\lambda\hat{\phi}^3$, as described by the Lagrangian (6.12). The associated Hamiltonian is

$$\hat{H} = \hat{H}_\mathrm{KG} + \hat{H}' \tag{6.13}$$

where (as is usual in perturbation theory) we have separated the Hamiltonian into a part we can handle exactly, which is the free Klein–Gordon Hamiltonian

$$\hat{H}_\mathrm{KG} = \int \mathrm{d}^3\boldsymbol{x}\, \hat{\mathcal{H}}_\mathrm{KG} = \tfrac{1}{2}\int \mathrm{d}^3\boldsymbol{x}\, [\hat{\pi}^2 + (\boldsymbol{\nabla}\hat{\phi})^2 + m^2\hat{\phi}^2] \tag{6.14}$$

and the part we shall treat perturbatively

$$\hat{H}' = \int \mathrm{d}^3\boldsymbol{x}\, \hat{\mathcal{H}}' = \lambda \int \mathrm{d}^3\boldsymbol{x}\, \hat{\phi}^3. \tag{6.15}$$

6.2.1 The interaction picture

We begin with a crucial formal step. In our introduction to quantum field theory in the previous chapter, we worked in the Heisenberg picture (HP). There, however, we only dealt with free (non-interacting) fields. The time dependence of the operators as given by the mode expansion (5.155) is that generated by the free KG Hamiltonian (6.14) via the Heisenberg equations of motion (see problem 5.8). But as soon as we include the interaction term $\hat{H}'$, we cannot make progress in the HP, since we do not then know the time dependence of the operators – which is generated by the full Hamiltonian $\hat{H} = \hat{H}_\mathrm{KG} + \hat{H}'$.

Instead, we might consider using the Schrödinger picture (SP) in which the states change with time according to

$$\hat{H}|\psi(t)\rangle = \mathrm{i}\frac{\mathrm{d}}{\mathrm{d}t}|\psi(t)\rangle \tag{6.16}$$

and the operators are time-independent (see appendix I). Note that although (6.16) is a 'Schrödinger picture' equation, there is nothing non-relativistic about it: on the contrary, $\hat{H}$ is the relevant relativistic Hamiltonian. In this approach, the field operators appearing in the density $\hat{\mathcal{H}}$ are all evaluated at a fixed time, say $t = 0$ by convention, which is the time at which the Schrödinger and Heisenberg pictures coincide. At this fixed time, mode expansions of the form (5.155) with $t = 0$ are certainly possible, since the basis functions form a complete set.

One problem with this formulation, however, is that it is not going to be manifestly 'Lorentz invariant' (or covariant), because a particular time ($t = 0$)

has been singled out. In the end, physical quantities should come out correct, but it is much more convenient to have everything looking nice and consistent with relativity as we go along. This is one of the reasons for choosing to work in yet a third 'picture', an ingenious kind of half-way-house between the other two, called the 'interaction picture' (IP). We shall see other good reasons shortly.

In the HP, all the time dependence is carried by the operators and none by the state, while in the SP it is exactly the other way around. In the IP, both states and operators are time-dependent but in a way that is well adapted to perturbation theory, especially in quantum field theory. The operators have a time dependence generated by the free Hamiltonian $\hat{H}_0$, say, and so a 'free-particle' mode expansion like (5.155) survives intact (here $\hat{H}_0 = \hat{H}_{\rm KG}$). The states have a time dependence generated by the interaction $\hat{H}'$. Thus as $\hat{H}' \to 0$ we return to the free-particle HP.

The way this works formally is as follows. In terms of the time-independent SP operator $\hat{A}$ (cf appendix I), we define the corresponding IP operator $\hat{A}_{\rm I}(t)$ by

$$\hat{A}_{\rm I}(t) = {\rm e}^{{\rm i}\hat{H}_0 t}\hat{A}{\rm e}^{-{\rm i}\hat{H}_0 t}. \tag{6.17}$$

This is just like the definition of the HP operator $\hat{A}(t)$ in appendix I, except that $\hat{H}_0$ appears instead of the full $\hat{H}$. It follows that the time dependence of $\hat{A}_{\rm I}(t)$ is given by (I.8) with $\hat{H} \to \hat{H}_0$:

$$\frac{{\rm d}\hat{A}_{\rm I}(t)}{{\rm d}t} = -{\rm i}[\hat{A}_{\rm I}(t), \hat{H}_0]. \tag{6.18}$$

Equation (6.18) can also, of course, be derived by carefully differentiating (6.17). Thus – as mentioned already – the time dependence of $\hat{A}_{\rm I}(t)$ is generated by the free part of the Hamiltonian, by construction.

As applied to our model theory (6.12), then, our field $\hat{\phi}$ will now be specified as being in the IP, $\hat{\phi}_{\rm I}(\boldsymbol{x}, t)$. What about the field canonically conjugate to $\hat{\phi}_{\rm I}(t)$, in the case when the interaction is included? In the HP, as long as the interaction does not contain time derivatives, as is the case here, the field canonically conjugate to the interacting field remains the same as the free-field case:

$$\hat{\pi}(\boldsymbol{x}, t) = \frac{\partial\hat{\mathcal{L}}}{\partial\dot{\hat{\phi}}(\boldsymbol{x}, t)} = \frac{\partial\hat{\mathcal{L}}_{\rm KG}}{\partial\dot{\hat{\phi}}(\boldsymbol{x}, t)} = \dot{\hat{\phi}}(\boldsymbol{x}, t) \tag{6.19}$$

so that we continue to adopt the equal-time commutation relation

$$[\hat{\phi}(\boldsymbol{x}, t), \hat{\pi}(\boldsymbol{y}, t)] = {\rm i}\delta^3(\boldsymbol{x} - \boldsymbol{y}) \tag{6.20}$$

for the Heisenberg fields. But the IP fields are related to the HP fields by a unitary transformation $\hat{U}$, as we can see by combining (6.17) with (I.7):

$$\begin{aligned} \hat{A}_{\rm I}(t) &= {\rm e}^{{\rm i}\hat{H}_0 t}{\rm e}^{-{\rm i}\hat{H}t}\hat{A}(t){\rm e}^{{\rm i}\hat{H}t}{\rm e}^{-{\rm i}\hat{H}_0 t} \\ &= \hat{U}\hat{A}(t)\hat{U}^{-1} \end{aligned} \tag{6.21}$$

where $\hat{U} = \mathrm{e}^{\mathrm{i}\hat{H}_0 t}\mathrm{e}^{-\mathrm{i}\hat{H}t}$, and it is easy to check that $\hat{U}\hat{U}^\dagger = \hat{U}^\dagger\hat{U} = \hat{I}$. So taking equation (6.20) and pre-multiplying by $\hat{U}$ and post-multiplying by $\hat{U}^{-1}$ on both sides, we obtain

$$[\hat{\phi}_{\mathrm{I}}(\boldsymbol{x}, t), \hat{\pi}_{\mathrm{I}}(\boldsymbol{y}, t)] = \mathrm{i}\delta^3(\boldsymbol{x} - \boldsymbol{y}) \tag{6.22}$$

showing that, *in the interacting case*, the IP fields $\hat{\phi}_{\mathrm{I}}$ and $\hat{\pi}_{\mathrm{I}}$ obey the free field commutation relation. Thus in the IP case the interacting fields obey the same equations of motion and the same commutation relations as the free-field operators. It follows that the mode expansion (5.155), and the commutation relations (5.158) for the mode creation and annihilation operators, can be taken straight over for the IP operators.

We now turn to the states in the IP. To preserve consistency between the matrix elements in the Schrödinger and interaction pictures (cf the step from (I.6) to (I.7)) we define the corresponding IP state vector by

$$|\psi(t)\rangle_{\mathrm{I}} = \mathrm{e}^{\mathrm{i}\hat{H}_0 t}|\psi(t)\rangle \tag{6.23}$$

in terms of the SP state $|\psi(t)\rangle$. We now use (6.23) to find the equation of motion of $|\psi(t)\rangle_{\mathrm{I}}$. We have

$$\begin{aligned}
\mathrm{i}\frac{\mathrm{d}}{\mathrm{d}t}|\psi(t)\rangle_{\mathrm{I}} &= \mathrm{e}^{\mathrm{i}\hat{H}_0 t}\left\{-\hat{H}_0|\psi(t)\rangle + \mathrm{i}\frac{\mathrm{d}}{\mathrm{d}t}|\psi(t)\rangle\right\} \\
&= \mathrm{e}^{\mathrm{i}\hat{H}_0 t}\{-\hat{H}_0|\psi(t)\rangle + (\hat{H}_0 + \hat{H}')|\psi(t)\rangle\} \\
&= \mathrm{e}^{\mathrm{i}\hat{H}_0 t}\hat{H}'|\psi(t)\rangle \\
&= \mathrm{e}^{\mathrm{i}\hat{H}_0 t}\hat{H}'\mathrm{e}^{-\mathrm{i}\hat{H}_0 t}|\psi(t)\rangle_{\mathrm{I}}
\end{aligned} \tag{6.24}$$

or

$$\boxed{\mathrm{i}\frac{\mathrm{d}}{\mathrm{d}t}|\psi(t)\rangle_{\mathrm{I}} = \hat{H}'_{\mathrm{I}}(t)|\psi(t)\rangle_{\mathrm{I}}} \tag{6.25}$$

where

$$\hat{H}'_{\mathrm{I}} = \mathrm{e}^{\mathrm{i}\hat{H}_0 t}\hat{H}'\mathrm{e}^{-\mathrm{i}\hat{H}_0 t} \tag{6.26}$$

is the interaction Hamiltonian *in the interaction picture*. The italicised words are important: they mean that all operators in $\hat{H}'_{\mathrm{I}}$ have the (known) free-field time dependence, which would not be the case for $\hat{H}'$ in the HP. Thus, as mentioned earlier, the states in the IP have a time dependence generated by the interaction Hamiltonian, and this derivation has shown us that it is, in fact, the interaction Hamiltonian in the IP which is the appropriate generator of time change in this picture.

Equation (6.25) is a slightly simplified form of the Tomonaga–Schwinger equation, which formed the starting point of the approach to QED followed by Schwinger (Schwinger 1948b, 1949a, b) and independently by Tomonaga and his group (Tomonaga 1946, Koba, Tati and Tomonaga 1947a, b, Kanesawa and Tomonaga 1948a, b, Koba and Tomonaga 1948, Koba and Takeda 1948, 1949).

6.2.2 The S-matrix and the Dyson expansion

We now start the job of applying the IP formalism to scattering and decay processes in quantum field theory, treated in perturbation theory; for this, following Dyson (1949a, b), the crucial quantity is the *scattering matrix*, or S-matrix for short, which we now introduce. A scattering process may plausibly be described in the following terms. At a time $t \to -\infty$, long before any interaction has occurred, we expect the effect of $\hat{H}'_{\rm I}$ to be negligible so that, from (6.25), $|\psi(-\infty)\rangle_{\rm I}$ will be a constant state vector $|{\rm i}\rangle$, which is in fact an eigenstate of $\hat{H}_0$. Thus $|{\rm i}\rangle$ will contain a certain number of non-interacting particles with definite momenta, and $|\psi(-\infty)\rangle_{\rm I} = |{\rm i}\rangle$. As time evolves, the particles approach each other and may scatter, leading in the distant future (at $t \to \infty$) to another constant state $|\psi(\infty)\rangle_{\rm I}$ containing non-interacting particles. Note that $|\psi(\infty)\rangle_{\rm I}$ will in general contain many different components, each with (in principle) different numbers and types of particle; these different components in $|\psi(\infty)\rangle_{\rm I}$ will be denoted by $|{\rm f}\rangle$. The $\hat{S}$-*operator* is now defined via

$$|\psi(\infty)\rangle_{\rm I} = \hat{S}|\psi(-\infty)\rangle_{\rm I} = \hat{S}|{\rm i}\rangle. \tag{6.27}$$

A particular *S-matrix element* is then the amplitude for finding a particular final state $|{\rm f}\rangle$ in $|\psi(\infty)\rangle_{\rm I}$:

$$\langle {\rm f}|\psi(\infty)\rangle_{\rm I} = \langle {\rm f}|\hat{S}|{\rm i}\rangle \equiv S_{\rm fi}. \tag{6.28}$$

Thus we may write

$$|\psi(\infty)\rangle_{\rm I} = \sum_{\rm f} |{\rm f}\rangle\langle {\rm f}|\psi(\infty)\rangle_{\rm I} = \sum_{\rm f} S_{\rm fi}|{\rm f}\rangle. \tag{6.29}$$

It is clear that it is these S-matrix elements $S_{\rm fi}$ that we need to calculate, and the associated probabilities $|S_{\rm fi}|^2$.

Before proceeding we note an important property of $\hat{S}$. Assuming that $|\psi(\infty)\rangle_{\rm I}$ and $|{\rm i}\rangle$ are both normalized, we have

$$1 = {}_{\rm I}\langle\psi(\infty)|\psi(\infty)\rangle_{\rm I} = \langle {\rm i}|\hat{S}^\dagger\hat{S}|{\rm i}\rangle = \langle {\rm i}|{\rm i}\rangle \tag{6.30}$$

implying that $\hat{S}$ is *unitary*: $\hat{S}^\dagger\hat{S} = \hat{I}$. Taking matrix elements of this gives us the result

$$\sum_{\rm k} S^*_{\rm kf} S_{\rm ki} = \delta_{\rm fi}. \tag{6.31}$$

Putting ${\rm i} = {\rm f}$ in (6.31) yields $\sum_{\rm k} |S_{\rm ki}|^2 = 1$, which confirms that the expansion coefficients in (6.29) must obey the usual condition that the sum of all the partial probabilities must add up to 1. Note, however, that in the present case the states involved may contain different numbers of particles.

We set up a perturbation-theory approach to calculating $\hat{S}$ as follows. Integrating (6.25) subject to the condition at $t \to -\infty$ yields

$$|\psi(t)\rangle_{\rm I} = |{\rm i}\rangle - {\rm i}\int_{-\infty}^{t} \hat{H}'_{\rm I}(t')|\psi(t')\rangle_{\rm I}\,{\rm d}t'. \tag{6.32}$$

This is an integral equation in which the unknown $|\psi(t)\rangle_{\rm I}$ is buried under the integral on the right-hand side, rather similar to the one we encounter in non-relativistic scattering theory (equation (H.12) of appendix H). As in that case, we solve it iteratively. If $\hat{H}'_{\rm I}$ is neglected altogether, then the solution is

$$|\psi(t)\rangle_{\rm I}^{(0)} = |{\rm i}\rangle. \tag{6.33}$$

To get the first order in $\hat{H}'_{\rm I}$ correction to this, insert (6.33) in place of $|\psi(t')\rangle_{\rm I}$ on the right-hand side of (6.32) to obtain

$$|\psi(t)\rangle_{\rm I}^{(1)} = |{\rm i}\rangle + \int_{-\infty}^{t} (-{\rm i}\hat{H}'_{\rm I}(t_1)){\rm d}t_1|{\rm i}\rangle \tag{6.34}$$

recalling that $|{\rm i}\rangle$ is a constant state vector. Putting this back into (6.32) yields $|\psi(t)\rangle$ correct to second order in $\hat{H}'_{\rm I}$:

$$\begin{aligned} |\psi(t)\rangle_{\rm I}^{(2)} \quad &= \quad \Big\{1 + \int_{-\infty}^{t} (-{\rm i}\hat{H}'_{\rm I}(t_1))\,{\rm d}t_1 \\ &\quad + \int_{-\infty}^{t} {\rm d}t_1 \int_{-\infty}^{t_1} {\rm d}t_2\,(-{\rm i}\hat{H}'_{\rm I}(t_1))(-{\rm i}\hat{H}'_{\rm I}(t_2))\Big\}|{\rm i}\rangle \end{aligned} \tag{6.35}$$

which is as far as we intend to go. Letting $t \to \infty$ then gives us our *perturbative series for the $\hat{S}$-operator*:

$$\hat{S} = 1 + \int_{-\infty}^{\infty} (-{\rm i}\hat{H}'_{\rm I}(t_1))\,{\rm d}t_1 + \int_{-\infty}^{\infty} {\rm d}t_1 \int_{-\infty}^{t_1} {\rm d}t_2\,(-{\rm i}\hat{H}'_{\rm I}(t_1))(-{\rm i}\hat{H}'_{\rm I}(t_2)) + \cdots \tag{6.36}$$

with the dots indicating the higher-order terms, which are in fact summarized by the full formula

$$\hat{S} = \sum_{n=0}^{\infty} (-{\rm i})^n \int_{-\infty}^{\infty} {\rm d}t_1 \int_{-\infty}^{t_1} {\rm d}t_2 \cdots \int_{-\infty}^{t_{n-1}} {\rm d}t_n\, \hat{H}'_{\rm I}(t_1)\hat{H}'_{\rm I}(t_2)\ldots\hat{H}'_{\rm I}(t_n). \tag{6.37}$$

We could immediately start getting to work with (6.37), but there is one more useful technical adjustment to make. Remembering that

$$\hat{H}'_{\rm I}(t) = \int \hat{\mathcal{H}}'_{\rm I}(\boldsymbol{x}, t)\,{\rm d}^3\boldsymbol{x} \tag{6.38}$$

we can write the second term of (6.36) as

$$\iint_{t_1 > t_2} {\rm d}^4x_1\,{\rm d}^4x_2\,(-{\rm i}\hat{\mathcal{H}}'_{\rm I}(x_1))(-{\rm i}\hat{\mathcal{H}}'_{\rm I}(x_2)) \tag{6.39}$$

which looks much more symmetrical in $\boldsymbol{x} - t$. However, there is still an awkward asymmetry between the $\boldsymbol{x}$-integrals and the t-integrals because of the $t_1 > t_2$ condition. The t-integrals can be converted to run from $-\infty$ to ∞

without constraint, like the $\boldsymbol{x}$ ones, by a clever trick. Note that the *ordering* of the operators $\hat{\mathcal{H}}'_{\rm I}$ is significant (since they will contain non-commuting bits), and that it is actually given by the order of their time arguments, 'earlier' operators appearing to the right of 'later' ones. This feature must be preserved, obviously, when we let the t-integrals run over the full infinite domain. We can arrange for this by introducing the time-ordering symbol T, which is defined by

$$\begin{aligned} T(\hat{\mathcal{H}}'_{\rm I}(x_1)\hat{\mathcal{H}}'_{\rm I}(x_2)) &= \hat{\mathcal{H}}'_{\rm I}(x_1)\hat{\mathcal{H}}'_{\rm I}(x_2) \qquad \text{for } t_1 > t_2 \\ &= \hat{\mathcal{H}}'_{\rm I}(x_2)\hat{\mathcal{H}}'_{\rm I}(x_1) \qquad \text{for } t_1 < t_2 \end{aligned} \tag{6.40}$$

and similarly for more products, and for arbitrary operators. Then (see problem 6.1) (6.39) can be written as

$$\tfrac{1}{2}\iint \mathrm{d}^4x_1\,\mathrm{d}^4x_2\,T[(-\mathrm{i}\hat{\mathcal{H}}'_{\rm I}(x_1))(-\mathrm{i}\hat{\mathcal{H}}'_{\rm I}(x_2))] \tag{6.41}$$

where the integrals are now unrestricted. Applying a similar analysis to the general term gives us the *Dyson expansion of the $\hat{S}$ operator*:

$$\boxed{\hat{S} = \sum_{n=0}^{\infty} \frac{(-\mathrm{i})^n}{n!} \int \ldots \int \mathrm{d}^4x_1\,\mathrm{d}^4x_2 \ldots \mathrm{d}^4x_n\, T\{\hat{\mathcal{H}}'_{\rm I}(x_1)\hat{\mathcal{H}}'_{\rm I}(x_2)\cdots\hat{\mathcal{H}}'_{\rm I}(x_n)\}.} \tag{6.42}$$

This fundamental formula provides the bridge leading from the Tomonaga–Schwinger equation (6.25) to the Feynman amplitudes (Feynman 1949a, b), as we shall see in detail in section 7.3.2 for the 'ABC' case.

6.3 Applications to the 'ABC' theory

As previously explained, the simple self-interacting $\hat{\phi}^3$ theory is not respectable. Following Griffiths (2008) we shall instead apply the foregoing covariant perturbation theory to a hypothetical world consisting of three distinct types of scalar particles A, B and C, with masses $m_{\rm A}$, $m_{\rm B}$, $m_{\rm C}$. Each is described by a real scalar field which, if free, would obey the appropriate KG equation; the interaction term is $g\hat{\phi}_{\rm A}\hat{\phi}_{\rm B}\hat{\phi}_{\rm C}$. We shall from now on omit the IP subscript 'I', since all operators are taken to be in the IP. Thus the Hamiltonian is

$$\hat{H} = \hat{H}_0 + \hat{H}' \tag{6.43}$$

where

$$\hat{H}_0 = \tfrac{1}{2}\sum_{i=\mathrm{A,B,C}} \int [\hat{\pi}_i{}^2 + (\boldsymbol{\nabla}\hat{\phi}_i)^2 + m_i^2\hat{\phi}_i^2]\,\mathrm{d}^3\boldsymbol{x} \tag{6.44}$$

and

$$\hat{H}' = g \int \mathrm{d}^3\boldsymbol{x}\, \hat{\phi}_{\mathrm{A}}\hat{\phi}_{\mathrm{B}}\hat{\phi}_{\mathrm{C}} \equiv \int \mathrm{d}^3\boldsymbol{x}\, \hat{\mathcal{H}}'. \tag{6.45}$$

Each field $\hat{\phi}_i$, $(i = \mathrm{A}, \mathrm{B}, \mathrm{C})$ has a mode expansion of the form (5.143), and associated creation and annihilation operators $\hat{a}_i^\dagger$ and $\hat{a}_i$ which obey the commutation relations

$$[\hat{a}_i(k), \hat{a}_j^\dagger(k')] = (2\pi)^3\delta^3(\boldsymbol{k} - \boldsymbol{k}')\delta_{ij} \qquad i, j = \mathrm{A}, \mathrm{B}, \mathrm{C}. \tag{6.46}$$

The new feature in (6.46) is that operators associated with distinct particles commute. In a similar way, we also have $[\hat{a}_i, \hat{a}_j] = [\hat{a}_i^\dagger, \hat{a}_j^\dagger] = 0$.

6.3.1 The decay $\mathrm{C} \to \mathrm{A} + \mathrm{B}$

As our first application of (6.42), we shall calculate the decay rate (or resonance width) for the decay $\mathrm{C} \to \mathrm{A}+\mathrm{B}$, to lowest order in g. Admittedly this is not yet a realistic, physical, example; even so, the basic steps in the calculation are common to more complicated physical examples, such as $\mathrm{W}^- \to \mathrm{e}^- + \bar{\nu}_{\mathrm{e}}$.

We suppose that the initial state $|\mathrm{i}\rangle$ consists of one C particle with 4-momentum p_{C}, and that the final state in which we are interested is that with one A and one B particle present, with 4-momenta p_{A} and p_{B} respectively. We want to calculate the matrix element

$$S_{\mathrm{fi}} = \langle p_{\mathrm{A}}, p_{\mathrm{B}}|\hat{S}|p_{\mathrm{C}}\rangle \tag{6.47}$$

to lowest order in g. (Note that the '1' term in (6.36) cannot contribute here because the initial and final states are plainly orthogonal.) This means that we need to evaluate the amplitude

$$\mathcal{A}_{\mathrm{fi}}^{(1)} = -\mathrm{i}g\langle p_{\mathrm{A}}, p_{\mathrm{B}}| \int \mathrm{d}^4x\, \hat{\phi}_{\mathrm{A}}(x)\hat{\phi}_{\mathrm{B}}(x)\hat{\phi}_{\mathrm{C}}(x)|p_{\mathrm{C}}\rangle. \tag{6.48}$$

To proceed we need to decide on the normalization of our states $|p_i\rangle$. We will define (for $i = \mathrm{A}, \mathrm{B}, \mathrm{C}$)

$$|p_i\rangle = \sqrt{2E_i}\hat{a}_i^\dagger(p_i)|0\rangle \tag{6.49}$$

where $E_i = \sqrt{m_i^2 + \boldsymbol{p}_i^2}$, so that (using (6.46))

$$\langle p_i'|p_i\rangle = 2E_i(2\pi)^3\delta^3(\boldsymbol{p}_i' - \boldsymbol{p}_i). \tag{6.50}$$

The quantity $E_i\delta^3(\boldsymbol{p}_i' - \boldsymbol{p}_i)$ is Lorentz invariant. Note that the completeness relation for such states reads

$$\int \frac{\mathrm{d}^3\boldsymbol{p}_i}{(2\pi)^3}\frac{1}{2E_i}|p_i\rangle\langle p_i| = 1 \tag{6.51}$$

where the '1' on the right-hand side means the identity in the subspace of

such one-particle states, and zero for all other states. The normalization choice (6.49) corresponds (see comment (5) in section 5.2.5) to a wavefunction normalization of $2E_i$ particles per unit volume.

Consider now just the $\hat{\phi}_{\rm C}(x)|p_{\rm C}\rangle$ piece of (6.48). This is

$$\int \frac{{\rm d}^3\boldsymbol{k}}{(2\pi)^3}\frac{1}{\sqrt{2E_k}}[\hat{a}_{\rm C}(k){\rm e}^{-{\rm i}k\cdot x}+\hat{a}^\dagger_{\rm C}(k){\rm e}^{{\rm i}k\cdot x}]\sqrt{2E_{\rm C}}\hat{a}^\dagger_{\rm C}(p_{\rm C})|0\rangle \tag{6.52}$$

where $k=(E_k,\boldsymbol{k})$ and $E_k=\sqrt{\boldsymbol{k}^2+m^2_{\rm C}}$. The term with two $\hat{a}^\dagger_{\rm C}$'s will give zero when bracketed with a final state containing no C particles. In the other term, we use (6.46) together with $\hat{a}_{\rm C}(k)|0\rangle=0$ to reduce (6.52) to

$$\int \frac{{\rm d}^3\boldsymbol{k}}{(2\pi)^3}\frac{1}{\sqrt{2E_k}}(2\pi)^3\delta^3(\boldsymbol{p}_{\rm C}-\boldsymbol{k})\sqrt{2E_{\rm C}}{\rm e}^{-{\rm i}k\cdot x}|0\rangle={\rm e}^{-{\rm i}p_{\rm C}\cdot x}|0\rangle \tag{6.53}$$

where $p_{\rm C}=(\sqrt{\boldsymbol{p}^2_{\rm C}+m^2_{\rm C}},\boldsymbol{p}_{\rm C})$. In exactly the same way we find that, when bracketed with an initial state containing no A's or B's,

$$\langle p_{\rm A},p_{\rm B}|\hat{\phi}_{\rm A}(x)\hat{\phi}_{\rm B}(x)=\langle 0|e^{{\rm i}p_{\rm A}\cdot x}{\rm e}^{{\rm i}p_{\rm B}\cdot x}. \tag{6.54}$$

Hence the amplitude (6.48) becomes just

$$\mathcal{A}^{(1)}_{\rm fi}=-{\rm i}g\int {\rm d}^4x{\rm e}^{{\rm i}(p_{\rm A}+p_{\rm B}-p_{\rm C})\cdot x}=-{\rm i}g(2\pi)^4\delta^4(p_{\rm A}+p_{\rm B}-p_{\rm C}). \tag{6.55}$$

Unsurprisingly, but reassuringly, we have discovered that the amplitude vanishes unless the 4-momentum is conserved via the δ-function condition: $p_{\rm C}=p_{\rm A}+p_{\rm B}$.

It is clear that such a transition will not occur unless $m_{\rm C}>m_{\rm A}+m_{\rm B}$ (in the rest frame of the C, we need $m_{\rm C}=\sqrt{m^2_{\rm A}+\boldsymbol{p}^2}+\sqrt{m^2_{\rm B}+\boldsymbol{p}^2}$), so let us assume this to be the case. We would now like to calculate the rate for the decay ${\rm C}\to{\rm A}+{\rm B}$. To do this, we shall adopt a plausible generalization of the ordinary procedure followed in quantum mechanical time-dependent perturbation theory (the reader may wish to consult section H.3 of appendix H at this point, to see a non-relativistic analogue). The first problem is that the transition probability $|\mathcal{A}^{(1)}_{\rm fi}|^2$ apparently involves the square of the four-dimensional δ-function. This is bad news, since (to take a simple case, and using (E.53)) $\delta(x-a)\delta(x-a)=\delta(x-a)\delta(0)$ and $\delta(0)$ is infinite. In our case we have a four-fold infinity. This trouble has arisen because we have been using plane-wave solutions of our wave equation, and these notoriously lead to such problems. A proper procedure would set the whole thing up using wave packets, as is done, for instance, in Peskin and Schroeder (1995), section 4.5. An easier remedy is to adopt 'box normalization', in which we imagine that space has the finite volume V, and the interaction is turned on only for a time T. Then '$(2\pi)^4\delta^4(0)$' is effectively 'VT' (see Weinberg (1995, section 3.4)). Dividing this factor out, the transition rate per unit volume is then

$$\dot{P}_{\rm fi}=|\mathcal{A}^{(1)}_{\rm fi}|^2/VT=(2\pi)^4\delta^4(p_{\rm A}+p_{\rm B}-p_{\rm C})|\mathcal{M}_{\rm fi}|^2 \tag{6.56}$$

where (cf (6.55))

$$\mathcal{A}_{\rm fi}^{(1)} = (2\pi)^4\delta^4(p_{\rm A}+p_{\rm B}-p_{\rm C})\mathrm{i}\mathcal{M}_{\rm fi} \tag{6.57}$$

so that the *invariant amplitude* $\mathrm{i}\mathcal{M}_{\rm fi}$ is just $-\mathrm{i}g$, in this case.

Equation (6.56) is the probability per unit time for a transition to one specific final state $|\mathrm{f}\rangle$. But in the present case (and in all similar ones with at least two particles in the final state), the A + B final states form a continuum, and to get the total rate Γ we need to integrate $\dot{P}_{\rm fi}$ over all the continuum of final states, consistent with energy–momentum conservation. The corresponding differential decay rate $\mathrm{d}\Gamma$ is defined by $\mathrm{d}\Gamma = \dot{P}_{\rm fi}\mathrm{d}N_{\rm f}$ where $\mathrm{d}N_{\rm f}$ is the number of final states, per particle, lying in a momentum space volume $\mathrm{d}^3p_{\rm A}\mathrm{d}^3p_{\rm B}$ about $p_{\rm A}$ and $p_{\rm B}$. For the normalization (6.49), this number is

$$\mathrm{d}N_{\rm f} = \frac{\mathrm{d}^3\boldsymbol{p}_{\rm A}}{(2\pi)^3 2E_{\rm A}}\frac{\mathrm{d}^3\boldsymbol{p}_{\rm B}}{(2\pi)^3 2E_{\rm B}}. \tag{6.58}$$

Finally, to get a normalization-independent quantity we must divide by the number of decaying particles per unit volume, which is $2E_{\rm C}$. Thus our final formula for the decay rate is

$$\Gamma = \int \mathrm{d}\Gamma = \frac{1}{2E_{\rm C}}(2\pi)^4\int \delta^4(p_{\rm A}+p_{\rm B}-p_{\rm C})|\mathcal{M}_{\rm fi}|^2\frac{\mathrm{d}^3\boldsymbol{p}_{\rm A}}{(2\pi)^3 2E_{\rm A}}\frac{\mathrm{d}^3\boldsymbol{p}_{\rm B}}{(2\pi)^3 2E_{\rm B}}. \tag{6.59}$$

Note that the '$\mathrm{d}^3\boldsymbol{p}/2E$' factors are Lorentz invariant (see the exercise in appendix E) and so are all the other terms in (6.59) except $E_{\rm C}$, which contributes the correct Lorentz-transformation character for a rate (i.e. rate $\propto 1/\gamma$).

We now calculate the total rate Γ in the rest frame of the decaying C particle. In this case, the 3-momentum part of the δ^4 gives $\boldsymbol{p}_{\rm A}+\boldsymbol{p}_{\rm B}=\boldsymbol{0}$, so $\boldsymbol{p}_{\rm A}=\boldsymbol{p}=-\boldsymbol{p}_{\rm B}$, and the energy part becomes $\delta(E-m_{\rm C})$ where

$$E = \sqrt{m_{\rm A}^2+\boldsymbol{p}^2}+\sqrt{m_{\rm B}^2+\boldsymbol{p}^2} = E_{\rm A}+E_{\rm B}. \tag{6.60}$$

So the total rate is

$$\Gamma = \frac{1}{2m_{\rm C}}\frac{g^2}{(2\pi)^2}\int\frac{\mathrm{d}^3\boldsymbol{p}}{4E_{\rm A}E_{\rm B}}\delta(E-m_{\rm C}). \tag{6.61}$$

Differentiating (6.60) we find

$$\mathrm{d}E = \left(\frac{|\boldsymbol{p}|}{E_{\rm A}}+\frac{|\boldsymbol{p}|}{E_{\rm B}}\right)\mathrm{d}|\boldsymbol{p}| = \frac{|\boldsymbol{p}|E}{E_{\rm A}E_{\rm B}}\,\mathrm{d}|\boldsymbol{p}|. \tag{6.62}$$

Thus we may write

$$\mathrm{d}^3\boldsymbol{p} = 4\pi|\boldsymbol{p}|^2\,\mathrm{d}|\boldsymbol{p}| = 4\pi|\boldsymbol{p}|\frac{E_{\rm A}E_{\rm B}}{E}\,\mathrm{d}E \tag{6.63}$$

and use the energy δ-function in (6.61) to do the dE integral yielding finally

$$\Gamma = \frac{g^2}{8\pi}\frac{|\boldsymbol{p}|}{m_{\rm C}^2}. \tag{6.64}$$

The quantity $|\boldsymbol{p}|$ is actually determined from (6.60) now with $E = m_{\rm C}$; after some algebra, we find (problem 5.2)

$$|\boldsymbol{p}| = [m_{\rm A}^4 + m_{\rm B}^4 + m_{\rm C}^4 - 2m_{\rm A}^2 m_{\rm B}^2 - 2m_{\rm B}^2 m_{\rm C}^2 - 2m_{\rm C}^2 m_{\rm A}^2]^{1/2}/2m_{\rm C}. \tag{6.65}$$

Equation (6.64) is the result of an 'almost real life' calculation and a number of comments are in order. First, consider the question of dimensions. In our units $\hbar = c = 1$, Γ as an inverse time should have the dimensions of a mass (see appendix B), which can also be understood if we think of Γ as the width of an unstable resonance state. This requires 'g' to have the dimensions of a mass, i.e. $g \sim M$ in these units. Going back to our Hamiltonian (6.44) and (6.45), which must also have dimensions of a mass, we see from (6.44) that the scalar fields $\hat{\phi}_i \sim M$ (using $\mathrm{d}^3\boldsymbol{x} \sim M^{-3}$), and hence from (6.45) $g \sim M$ as required. It turns out that the dimensionality of the coupling constants (such as g) is of great significance in quantum field theory. In QED, the analogous quantity is the charge e, and this is dimensionless in our units ($\alpha = e^2/4\pi = 1/137$, see appendix C). However, we saw in (1.31) that Fermi's 'four-fermion' coupling constant G had dimensions $\sim M^{-2}$, while Yukawa's '$g_{\rm N}$' and 'g'' (see figure 1.4) were both dimensionless. In fact, as we shall explain in section 11.8, the dimensionality of a theory's coupling constant is an important guide as to whether the infinities generally present in the theory can be controlled by *renormalization* (see chapter 10) or not: in particular, theories in which the coupling constant has negative mass dimensions, such as the 'four-fermion' theory, are not renormalizable. Theories with dimensionless coupling constants, such as QED, are generally renormalizable, though not invariably so. Theories whose coupling constants have positive mass dimension, as in the ABC model, are 'super-renormalizable', meaning (roughly) that they have fewer basic divergences than ordinary renormalizable theories (see section 11.8).

In the present case, let us say that the mass of the decaying particle $m_{\rm C}$, 'sets the scale' for g, so that we write $g = \tilde{g}m_{\rm C}$ and then

$$\Gamma = \frac{\tilde{g}^2}{8\pi}|\boldsymbol{p}| \tag{6.66}$$

where $\tilde{g}$ is dimensionless. Equation (6.66) shows us nicely that Γ is simply proportional to the energy release in the decay, as determined by $|\boldsymbol{p}|$ (one often says that Γ is determined 'by the available phase space'). If $m_{\rm C}$ is exactly equal to $m_{\rm A} + m_{\rm B}$, then $|\boldsymbol{p}|$ vanishes and so does Γ. At the opposite extreme, if $m_{\rm A}$ and $m_{\rm B}$ are negligible compared to $m_{\rm C}$, we would have

$$\Gamma = \frac{\tilde{g}^2}{16\pi}m_{\rm C}. \tag{6.67}$$

Equation (6.67) shows that, even if $\tilde{g}^2/16\pi$ is small ($\sim$ 1/137 say) Γ can still be surprisingly large if $m_{\rm C}$ is, as in $\mathrm{W}^- \to \mathrm{e}^- + \bar{\nu}_{\rm e}$ for example.

6.3.2 A + B → A + B scattering: the amplitudes

We now consider the two-particle → two-particle process

$$\mathrm{A} + \mathrm{B} \to \mathrm{A} + \mathrm{B} \tag{6.68}$$

in which the initial 4-momenta are $p_{\rm A}$, $p_{\rm B}$ and the final 4-momenta are $p'_{\rm A}$, $p'_{\rm B}$ so that $p_{\rm A} + p_{\rm B} = p'_{\rm A} + p'_{\rm B}$. Our main task is to calculate the matrix element $\langle p'_{\rm A}, p'_{\rm B}|\hat{S}|p_{\rm A}, p_{\rm B}\rangle$ to lowest non-trivial order in g. The result will be the derivation of our first 'Feynman rules' for amplitudes in perturbative quantum field theory.

The first term in the $\hat{S}$-operator expansion (6.42) is '1', which does not involve g at all. Nevertheless, it is a useful exercise to evaluate and understand this contribution (which in the present case does not vanish), namely

$$\langle 0|\hat{a}_{\rm A}(p'_{\rm A})\hat{a}_{\rm B}(p'_{\rm B})\hat{a}^\dagger_{\rm A}(p_{\rm A})\hat{a}^\dagger_{\rm B}(p_{\rm B})|0\rangle(16E_{\rm A}E_{\rm B}E'_{\rm A}E'_{\rm B})^{1/2}. \tag{6.69}$$

We shall have to evaluate many such *vacuum expectation values* (vev) of products of $\hat{a}^\dagger$'s and $\hat{a}$'s. The general strategy is to commute the $\hat{a}^\dagger$'s to the left, and the $\hat{a}$'s to the right, and then make use of the facts

$$\langle 0|\hat{a}^\dagger_i = \hat{a}_i|0\rangle = 0 \tag{6.70}$$

for any $i = \mathrm{A}, \mathrm{B}, \mathrm{C}$. Thus, remembering that all 'A' operators commute with all 'B' ones, the vev in (6.69) is equal to

$$\begin{aligned}
&\langle 0|\hat{a}_{\rm A}(p'_{\rm A})\hat{a}^\dagger_{\rm A}(p_{\rm A})\{(2\pi)^3\delta^3(\boldsymbol{p}_{\rm B} - \boldsymbol{p}'_{\rm B}) + \hat{a}^\dagger_{\rm B}(p_{\rm B})\hat{a}_{\rm B}(p'_{\rm B})\}|0\rangle \\
&\quad = \langle 0|\{(2\pi)^3\delta^3(\boldsymbol{p}_{\rm A} - \boldsymbol{p}'_{\rm A}) + \hat{a}^\dagger_{\rm A}(p_{\rm A})\hat{a}_{\rm A}(p'_{\rm A})\}(2\pi)^3\delta^3(\boldsymbol{p}_{\rm B} - \boldsymbol{p}'_{\rm B})|0\rangle \\
&\quad = (2\pi)^3\delta^3(\boldsymbol{p}_{\rm A} - \boldsymbol{p}'_{\rm A})(2\pi)^3\delta^3(\boldsymbol{p}_{\rm B} - \boldsymbol{p}'_{\rm B}).
\end{aligned} \tag{6.71}$$

The δ-functions enforce $E_{\rm A} = E'_{\rm A}$ and $E_{\rm B} = E'_{\rm B}$ so that (6.69) becomes

$$2E_{\rm A}(2\pi)^3\delta^3(\boldsymbol{p}_{\rm A} - \boldsymbol{p}'_{\rm A})2E_{\rm B}(2\pi)^3\delta^3(\boldsymbol{p}_{\rm B} - \boldsymbol{p}'_{\rm B}), \tag{6.72}$$

a result which just expresses the normalization of the states, and the fact that, with no 'g' entering, the particles have not interacted at all, but have continued on their separate ways, quite unperturbed ($\boldsymbol{p}_{\rm A} = \boldsymbol{p}'_{\rm A}$, $\boldsymbol{p}_{\rm B} = \boldsymbol{p}'_{\rm B}$). This contribution can be represented diagrammatically as figure 6.1.

Next, consider the term of order g, which we used in $\mathrm{C} \to \mathrm{A} + \mathrm{B}$. This is

$$-\mathrm{i}g\int \mathrm{d}^4x\, \langle p'_{\rm A}, p'_{\rm B}|\hat{\phi}_{\rm A}(x)\hat{\phi}_{\rm B}(x)\hat{\phi}_{\rm C}(x)|p_{\rm A}, p_{\rm B}\rangle. \tag{6.73}$$

We have to remember, now, that all the $\hat{\phi}_i$ operators are in the interaction

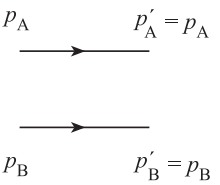


FIGURE 6.1
The order g^0 term in the perturbative expansion: the two particles do not interact.

picture and are therefore represented by standard mode expansions involving the *free* creation and annihilation operators $\hat{a}_i^\dagger$ and $\hat{a}_i$, i.e. the same ones used in defining the initial and final state vectors. It is then obvious that (6.73) must vanish, since no C-particle exists in either the initial or final state, and $\langle 0|\hat{\phi}_C|0\rangle = 0$.

So we move on to the term of order g^2, which will provide the real meat of this chapter. This term is

$$\begin{aligned}\frac{(-ig)^2}{2}\iint \mathrm{d}^4x_1\,\mathrm{d}^4x_2\,\langle 0|\hat{a}_A(p'_A)\hat{a}_B(p'_B)\\ \times T\{\hat{\phi}_A(x_1)\hat{\phi}_B(x_1)\hat{\phi}_C(x_1)\hat{\phi}_A(x_2)\hat{\phi}_B(x_2)\hat{\phi}_C(x_2)\}\\ \times \hat{a}_A^\dagger(p_A)\hat{a}_B^\dagger(p_B)|0\rangle(16E_AE_BE'_AE'_B)^{1/2}. \qquad (6.74)\end{aligned}$$

The vev here involves the product of *ten* operators, so it will pay us to pause and think how such things may be efficiently evaluated.

Consider the case of just four operators

$$\langle 0|\hat{A}\hat{B}\hat{C}\hat{D}|0\rangle \qquad (6.75)$$

where each of $\hat{A}$, $\hat{B}$, $\hat{C}$, $\hat{D}$ is an $\hat{a}_i$, an $\hat{a}_i^\dagger$ or a linear combination of these. Let $\hat{A}$ have the generic form $\hat{A} = \hat{a} + \hat{a}^\dagger$. Then (using $\langle 0|a^\dagger = a|0\rangle = 0$)

$$\begin{aligned}\langle 0|\hat{A}\hat{B}\hat{C}\hat{D}|0\rangle &= \langle 0|\hat{a}\hat{B}\hat{C}\hat{D}|0\rangle\\ &= \langle 0|[\hat{a},\hat{B}\hat{C}\hat{D}]|0\rangle. \qquad (6.76)\end{aligned}$$

Now it is an algebraic identity that

$$[\hat{a},\hat{B}\hat{C}\hat{D}] = [\hat{a},\hat{B}]\hat{C}\hat{D} + \hat{B}[\hat{a},\hat{C}]\hat{D} + \hat{B}\hat{C}[\hat{a},\hat{D}]. \qquad (6.77)$$

Hence

$$\langle 0|\hat{A}\hat{B}\hat{C}\hat{D}|0\rangle = [\hat{a},\hat{B}]\langle 0|\hat{C}\hat{D}|0\rangle + [\hat{a},\hat{C}]\langle 0|\hat{B}\hat{D}|0\rangle + [\hat{a},\hat{D}]\langle 0|\hat{B}\hat{C}|0\rangle, \qquad (6.78)$$

remembering that all the commutators – if non-vanishing – are just ordinary

numbers (see (6.46)). We can rewrite (6.78) in more suggestive form by noting that

$$[\hat{a}, \hat{B}] = \langle 0|[\hat{a}, \hat{B}]|0\rangle = \langle 0|\hat{a}\hat{B}|0\rangle = \langle 0|\hat{A}\hat{B}|0\rangle. \tag{6.79}$$

Thus the vev of a product of four operators is just the sum of the products of all the possible pairwise 'contractions' (the name given to the vev of the product of two fields):

$$\langle 0|\hat{A}\hat{B}\hat{C}\hat{D}|0\rangle = \langle 0|\hat{A}\hat{B}|0\rangle\langle 0|\hat{C}\hat{D}|0\rangle + \langle 0|\hat{A}\hat{C}|0\rangle\langle 0|\hat{B}\hat{D}|0\rangle + \langle 0|\hat{A}\hat{D}|0\rangle\langle 0|\hat{B}\hat{C}|0\rangle. \tag{6.80}$$

This result *generalizes* to the vev of the product of any number of operators; there is also a similar result for the vev of time-ordered products of operators, which is known as Wick's theorem (Wick 1950), and is indispensable for a general discussion of quantum field perturbation theory.

Consider then the application of (6.80), as generalized to ten operators, to the vev in (6.74). The only kind of non-vanishing contractions are of the form $\langle 0|\hat{a}_i\hat{a}_i^\dagger|0\rangle$. Thus the contractions of A-, B- and C-type operators can be considered separately. As far as the C-operators are concerned, then, we can immediately conclude that the only surviving contraction is

$$\langle 0|T(\hat{\phi}_{\rm C}(x_1)\hat{\phi}_{\rm C}(x_2))|0\rangle. \tag{6.81}$$

This quantity is, in fact, of fundamental importance: it is called the *Feynman propagator* (in coordinate space) for the spin-0 C-particle. We shall derive the mathematical formula for it in due course, but for the moment let us understand its physical significance. Each of the $\hat{\phi}_{\rm C}$'s in (6.81) can create or destroy C-quanta, but for the vev to be non-zero anything created in the 'initial' state must be destroyed in the 'final' one. Which of the times t_1 and t_2 is initial or final is determined by the T-ordering symbol: for $t_1 > t_2$, a C-quantum is created at x_2 and destroyed at x_1, while for $t_1 < t_2$ a C-quantum is created at x_1 and destroyed at x_2. Thus the amplitude (6.81) may be represented pictorially as in figure 6.2, where time increases to the right, and the vertical axis is a one-dimensional version of three-dimensional space. It seems reasonable, indeed, to call this object the 'propagator', since it clearly has to do with a quantum propagating between two space–time points.

We might now worry that this explicit time-ordering seems to introduce a Lorentz non-invariant element into the calculation, ultimately threatening the Lorentz invariance of the $\hat{S}$-operator (6.42). The reason that this is in fact not the case exposes an important property of quantum field theory. If the two points x_1 and x_2 are separated by a time-like interval (i.e. $(x_1 - x_2)^2 > 0$), then the time-ordering is Lorentz invariant; this is because no proper Lorentz transformation can alter the time-ordering of time-like separated events (here, the events are the creation/annihilation of particles/antiparticles at x_1 and x_2). By 'proper' is meant a transformation that does not reverse the sense of time; the behaviour of the theory under time-reversal is a different question altogether, discussed earlier in section 4.2.4. The fact that time-ordering is

FIGURE 6.2
C-quantum propagating (*a*) for $t_1 > t_2$ (from x_2 to x_1) and (*b*) $t_1 < t_2$ (from x_1 to x_2).

invariant for time-like separated events is what guarantees that we cannot influence our past, only our future. But what if the events are space-like separated, $(x_1 - x_2)^2 < 0$? We know that the scalar fields $\hat{\phi}_i(x_1)$ and $\hat{\phi}_i(x_2)$ commute for equal times: remarkably, one can show (problem 5.6(b)) that they *also* commute for $(x_1 - x_2)^2 < 0$; so in this sector of $x_1 - x_2$ space the time-ordering symbol is irrelevant. Thus, contrary to appearances, the T-product vev is Lorentz invariant. For the same reason, the $\hat{S}$ operator of (6.42) is also Lorentz invariant: see, for example, Weinberg (1995, section 3.5).

The property

$$[\hat{\phi}_i(x_1), \hat{\phi}_i(x_2)] = 0 \qquad \text{for } (x_1 - x_2)^2 < 0 \tag{6.82}$$

has an important physical interpretation. In quantum mechanics, if operators representing physical observables commute with each other, then measurements of either observable can be performed without interfering with each other; the observables are said to be 'compatible'. This is just what we would want for measurements done at two points which are space-like separated – no signal with speed less than or equal to light can connect them, and so we would expect them to be non-interfering. Condition (6.82) is often called a 'causality' condition.

More mathematically, the amplitude (6.81) is in fact a *Green function* for the KG operator $(\Box + m_C^2)$! (see appendix G, and problem 6.3). That is to say,

$$(\Box_{x_1} + m_C^2)\langle 0|T(\hat{\phi}_C(x_1)\hat{\phi}_C(x_2))|0\rangle = -\mathrm{i}\delta^4(x_1 - x_2). \tag{6.83}$$

Actually, problem 6.3 shows that (6.83) is true even when the $\langle 0|$ and $|0\rangle$ are removed, i.e. the operator quantity $T(\hat{\phi}_C(x_1)\hat{\phi}_C(x_2))$ is itself a KG Green function. The work of appendices G and H indicates the central importance of such Green functions in scattering theory, so we need not be surprised to find such a thing appearing here.

Now let us figure out what are all the surviving terms in the vev in (6.74). As far as contractions involving $\hat{a}_{\rm A}(p'_{\rm A})$ are concerned, we have only three non-zero possibilities:

$$\langle 0|\hat{a}_{\rm A}(p'_{\rm A})\hat{a}^\dagger_{\rm A}(p_{\rm A})|0\rangle \qquad \langle 0|\hat{a}_{\rm A}(p'_{\rm A})\hat{\phi}_{\rm A}(x_1)|0\rangle \qquad \langle 0|\hat{a}_{\rm A}(p'_{\rm A})\hat{\phi}_{\rm A}(x_2)|0\rangle. \quad (6.84)$$

There are similar possibilities for $\hat{a}^\dagger_{\rm A}(p_{\rm A})$, $\hat{a}_{\rm B}(p'_{\rm B})$ and $\hat{a}^\dagger_{\rm B}(p_{\rm B})$. The upshot is that we have only the following pairings to consider:

$$\begin{aligned}
&\langle 0|\hat{a}_{\rm A}(p'_{\rm A})\hat{a}^\dagger_{\rm A}(p_{\rm A})|0\rangle\langle 0|\hat{a}_{\rm B}(p'_{\rm B})\hat{a}^\dagger_{\rm B}(p_{\rm B})|0\rangle \\
&\qquad\times \langle 0|T(\hat{\phi}_{\rm A}(x_1)\hat{\phi}_{\rm A}(x_2))|0\rangle\langle 0|T(\hat{\phi}_{\rm B}(x_1)\hat{\phi}_{\rm B}(x_2))|0\rangle\langle 0|T(\hat{\phi}_{\rm C}(x_1)\hat{\phi}_{\rm C}(x_2))|0\rangle; \qquad (6.85)\\
&\langle 0|\hat{a}_{\rm A}(p'_{\rm A})\hat{a}^\dagger_{\rm A}(p_{\rm A})|0\rangle\langle 0|\hat{a}_{\rm B}(p'_{\rm B})\hat{\phi}_{\rm B}(x_1)|0\rangle \\
&\qquad\times \langle 0|\hat{\phi}_{\rm B}(x_2)\hat{a}^\dagger_{\rm B}(p_{\rm B})|0\rangle\langle 0|T(\hat{\phi}_{\rm C}(x_1)\hat{\phi}_{\rm C}(x_2))|0\rangle\langle 0|T(\hat{\phi}_{\rm A}(x_1)\hat{\phi}_{\rm A}(x_2))|0\rangle \\
&\qquad+ x_1 \leftrightarrow x_2; \qquad (6.86)\\
&\langle 0|\hat{a}_{\rm B}(p'_{\rm B})\hat{a}^\dagger_{\rm B}(p_{\rm B})|0\rangle\langle 0|\hat{a}_{\rm A}(p'_{\rm A})\hat{\phi}_{\rm A}(x_1)|0\rangle \\
&\qquad\times \langle 0|\hat{\phi}_{\rm A}(x_2)\hat{a}^\dagger_{\rm A}(p_{\rm A})|0\rangle\langle 0|T(\hat{\phi}_{\rm C}(x_1)\hat{\phi}_{\rm C}(x_2))|0\rangle\langle 0|T(\hat{\phi}_{\rm B}(x_1)\hat{\phi}_{\rm B}(x_2))|0\rangle \\
&\qquad+ x_1 \leftrightarrow x_2; \qquad (6.87)\\
&\langle 0|\hat{a}_{\rm A}(p'_{\rm A})\hat{\phi}_{\rm A}(x_1)|0\rangle\langle 0|\hat{\phi}_{\rm A}(x_2)\hat{a}^\dagger_{\rm A}(p_{\rm A})|0\rangle\langle 0|\hat{a}_{\rm B}(p'_{\rm B})\hat{\phi}_{\rm B}(x_1)|0\rangle \\
&\qquad\times \langle 0|\hat{\phi}_{\rm B}(x_2)\hat{a}^\dagger_{\rm B}(p_{\rm B})|0\rangle\langle 0|T(\hat{\phi}_{\rm C}(x_1)\hat{\phi}_{\rm C}(x_2))|0\rangle \\
&\qquad+ x_1 \leftrightarrow x_2; \qquad (6.88)\\
&\langle 0|\hat{a}_{\rm A}(p'_{\rm A})\hat{\phi}_{\rm A}(x_1)|0\rangle\langle 0|\hat{\phi}_{\rm A}(x_2)\hat{a}^\dagger_{\rm A}(p_{\rm A})|0\rangle\langle 0|\hat{a}_{\rm B}(p'_{\rm B})\hat{\phi}_{\rm B}(x_2)|0\rangle \\
&\qquad\times \langle 0|\hat{\phi}_{\rm B}(x_1)\hat{a}^\dagger_{\rm B}(p_{\rm B})|0\rangle\langle 0|T(\hat{\phi}_{\rm C}(x_1)\hat{\phi}_{\rm C}(x_2))|0\rangle \\
&\qquad+ x_1 \leftrightarrow x_2. \qquad (6.89)
\end{aligned}$$

We already know that quantities like $\langle 0|\hat{a}(p'_{\rm A})\hat{a}^\dagger_{\rm A}(p_{\rm A})|0\rangle$ yield something proportional to $\delta^3(\boldsymbol{p}_{\rm A} - \boldsymbol{p}'_{\rm A})$ and correspond to the initial A-particle going 'straight through'. The other factors in (6.85) which are new are quantities like $\langle 0|\hat{a}_{\rm A}(p'_{\rm A})\hat{\phi}_{\rm A}(x_1)|0\rangle$, which has the value (problem 6.4)

$$\langle 0|\hat{a}_{\rm A}(p'_{\rm A})\hat{\phi}_{\rm A}(x_1)|0\rangle = \frac{1}{\sqrt{2E'_{\rm A}}}\mathrm{e}^{\mathrm{i}p'_{\rm A}\cdot x_1} \qquad (6.90)$$

which is proportional (depending on the adopted normalization) to the wavefunction for an outgoing A-particle with 4-momentum $p'_{\rm A}$.

We are now in a position to give a diagrammatic interpretation of all of (6.85)–(6.89). In these diagrams, we shall *not* (as we did in figure 6.2) draw two separately time-ordered pieces for each propagator. We shall not indicate the time-ordering at all and we shall understand that *both* time-orderings are always included in each propagator line. Term (6.85) then has the structure shown in figure 6.3(a); term (6.86) that shown in figure 6.3(b); term (6.87) that in figure 6.3(c); term (6.88) that in figure 6.3(d); and term

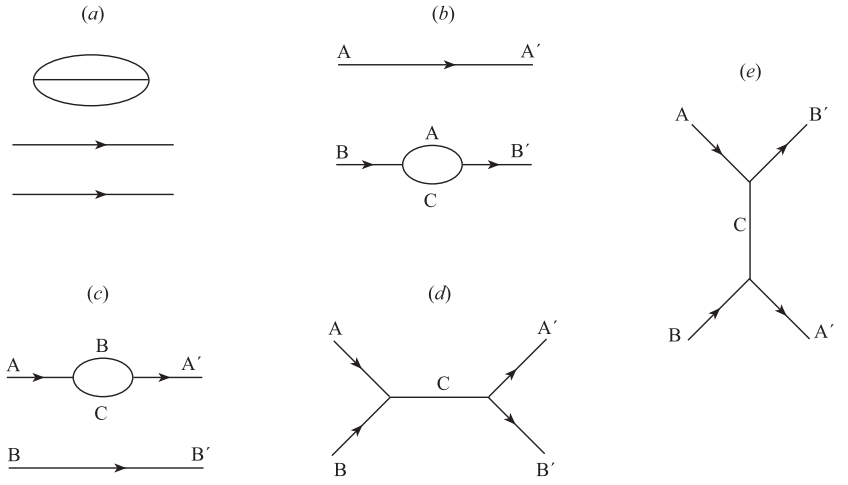


FIGURE 6.3
Graphical representation of (6.85)–(6.89): (*a*) (6.85); (*b*) (6.86); (*c*) (6.87); (*d*) (6.88); (*e*) (6.89).

(6.89) that in figure 6.3(*e*). We recognize in figure 6.3(*e*) the long-awaited Yukawa exchange process, which we shall shortly analyse in full – but the formalism has yielded much else besides! We shall come back to figures 6.3(*a*), (*b*) and (*c*) in section 6.3.5; for the moment we note that these processes do not represent true interactions between the particles, since at least one goes through unscattered in each case. So we shall concentrate on figures 6.3(*d*) and (*e*), and *derive the Feynman rules for them.*

First, consider figure 6.3(*e*), corresponding to the contraction (6.89). When this is inserted into (6.74), the two terms in which x_1 and x_2 are interchanged give identical results (interchanging x_1 and x_2 in the integral), so the contribution we are discussing is

$$(-\mathrm{i}g)^2 \iint \mathrm{d}^4x_1 \mathrm{d}^4x_2 \mathrm{e}^{\mathrm{i}(p'_\mathrm{A}-p_\mathrm{B})\cdot x_1} \mathrm{e}^{\mathrm{i}(p'_\mathrm{B}-p_\mathrm{A})\cdot x_2} \langle 0|T(\hat{\phi}_\mathrm{C}(x_1)\hat{\phi}_\mathrm{C}(x_2))|0\rangle. \quad (6.91)$$

We must now turn our attention, as promised, to the propagator of (6.81), $\langle 0|T(\hat{\phi}_\mathrm{C}(x_1)\hat{\phi}_\mathrm{C}(x_2))|0\rangle$. Inserting the mode expansion (6.52) for each of $\hat{\phi}_\mathrm{C}(x_1)$ and $\hat{\phi}_\mathrm{C}(x_2)$, and using the commutation relations (6.46) and the vacuum conditions (6.70) we find (problem 6.5)

$$\begin{aligned}\langle 0|T(\hat{\phi}_\mathrm{C}(x_1)\hat{\phi}_\mathrm{C}(x_2))|0\rangle \;&=\; \int \frac{\mathrm{d}^3\boldsymbol{k}}{(2\pi)^3 2\omega_k}[\theta(t_1-t_2)\mathrm{e}^{-\mathrm{i}\omega_k(t_1-t_2)+\mathrm{i}\boldsymbol{k}\cdot(\boldsymbol{x}_1-\boldsymbol{x}_2)} \\ &\quad + \theta(t_2-t_1)\mathrm{e}^{-\mathrm{i}\omega_k(t_2-t_1)+\mathrm{i}\boldsymbol{k}\cdot(\boldsymbol{x}_2-\boldsymbol{x}_1)}] \qquad (6.92)\end{aligned}$$

where $\omega_k = (\boldsymbol{k}^2 + m_\mathrm{C}^2)^{1/2}$. This expression is very 'uncovariant looking',

due to the presence of the θ-functions with time arguments. But the earlier discussion, after (6.81), has assured us that the left-hand side of (6.92) must be Lorentz invariant, and – by a clever trick – it is possible to recast the right-hand side in *manifestly* invariant form. We introduce an integral representation of the θ-function via

$$\theta(t) = \mathrm{i}\int_{-\infty}^{\infty} \frac{\mathrm{d}z}{2\pi}\frac{\mathrm{e}^{-\mathrm{i}zt}}{z+\mathrm{i}\epsilon} \tag{6.93}$$

where ϵ is an infinitesimally small positive quantity (see appendix F). Multiplying (6.93) by $\mathrm{e}^{-\mathrm{i}\omega_k t}$ and changing z to $z+\omega_k$ in the integral we have

$$\theta(t)\mathrm{e}^{-\mathrm{i}\omega_k t} = \mathrm{i}\int_{-\infty}^{\infty} \frac{\mathrm{d}z}{2\pi}\frac{\mathrm{e}^{-\mathrm{i}zt}}{z-(\omega_k-\mathrm{i}\epsilon)}. \tag{6.94}$$

Putting (6.94) into (6.92) then yields

$$\begin{aligned}\langle 0|T(\hat{\phi}_{\mathrm{C}}(x_1)\hat{\phi}_{\mathrm{C}}(x_2))|0\rangle \;&=\; \mathrm{i}\int \frac{\mathrm{d}^3\boldsymbol{k}\mathrm{d}z}{(2\pi)^4 2\omega_k}\left\{\frac{\mathrm{e}^{-\mathrm{i}z(t_1-t_2)+\mathrm{i}\boldsymbol{k}\cdot(\boldsymbol{x}_1-\boldsymbol{x}_2)}}{z-(\omega_k-\mathrm{i}\epsilon)}\right.\\ &\quad\left.+\;\frac{\mathrm{e}^{\mathrm{i}z(t_1-t_2)-\mathrm{i}\boldsymbol{k}\cdot(\boldsymbol{x}_1-\boldsymbol{x}_2)}}{z-(\omega_k-\mathrm{i}\epsilon)}\right\}.\end{aligned} \tag{6.95}$$

The exponentials and the volume element demand a more symmetrical notation: let us write $k_0 = z$ so that $(k_0 = z, \boldsymbol{k})$ form the components of a 4-vector k[1]. *Note very carefully, however, that* k_0 *is not* $(\boldsymbol{k}^2+m_{\mathrm{C}}^2)^{1/2}$! The variable k_0 is unrestricted, whereas it is ω_k that equals $(\boldsymbol{k}^2+m_{\mathrm{C}}^2)^{1/2}$. With this change of notation, (6.95) becomes

$$\langle 0|T(\hat{\phi}_{\mathrm{C}}(x_1)\hat{\phi}_{\mathrm{C}}(x_2))|0\rangle = \int \frac{\mathrm{d}^4k}{(2\pi)^4}\frac{\mathrm{i}}{2\omega_k}\left\{\frac{\mathrm{e}^{-\mathrm{i}k\cdot(x_1-x_2)}}{k_0-(\omega_k-\mathrm{i}\epsilon)}+\frac{\mathrm{e}^{\mathrm{i}k\cdot(x_1-x_2)}}{k_0-(\omega_k-\mathrm{i}\epsilon)}\right\}. \tag{6.96}$$

Changing $k\to -k$ ($k_0\to -k_0$, $\boldsymbol{k}\to -\boldsymbol{k}$) in the second term in (6.96), we finally have

$$\begin{aligned}&\langle 0|T(\hat{\phi}_{\mathrm{C}}(x_1)\hat{\phi}_{\mathrm{C}}(x_2))|0\rangle\\ &\quad=\int \frac{\mathrm{d}^4k}{(2\pi)^4}\mathrm{e}^{-\mathrm{i}k\cdot(x_1-x_2)}\frac{\mathrm{i}}{2\omega_k}\left\{\frac{1}{k_0-(\omega_k-\mathrm{i}\epsilon)}-\frac{1}{k_0+\omega_k-\mathrm{i}\epsilon}\right\}\\ &\quad=\int \frac{\mathrm{d}^4k}{(2\pi)^4}\mathrm{e}^{-\mathrm{i}k\cdot(x_1-x_2)}\frac{\mathrm{i}}{k_0^2-(\omega_k-\mathrm{i}\epsilon)^2},\end{aligned} \tag{6.97}$$

or

$$\boxed{\langle 0|T(\hat{\phi}_{\mathrm{C}}(x_1)\hat{\phi}_{\mathrm{C}}(x_2))|0\rangle = \int \frac{\mathrm{d}^4k}{(2\pi)^4}\mathrm{e}^{-\mathrm{i}k\cdot(x_1-x_2)}\frac{\mathrm{i}}{k_0^2-\boldsymbol{k}^2-m_{\mathrm{C}}^2+\mathrm{i}\epsilon}} \tag{6.98}$$

[1] We know that the left-hand side of (6.95) is Lorentz invariant, and that $(t_1-t_2, \boldsymbol{x}_1-\boldsymbol{x}_2)$ form the components of a 4-vector. The quantities $(k_0=z, \boldsymbol{k})$ must also form the components of a 4-vector, in order for the exponentials in (6.95) to be invariant.

where in the last step we have used $\omega_k^2 = \boldsymbol{k}^2 + m_{\rm C}^2$ and written 'iϵ' for '2i$\epsilon\omega_k$' since what matters is just the sign of the small imaginary part (note that ω_k is defined as the positive square root). In this final form, the Lorentz invariance of the scalar propagator is indeed manifest.

We shall have more to say about this propagator (Green function) in section 6.3.3. For the moment we simply note two points: first, it is the Fourier transform of $\mathrm{i}/k^2 - m_{\rm C}^2 + \mathrm{i}\epsilon$, as stated in appendix G, where $k^2 = k_0^2 - \boldsymbol{k}^2$; and second, it is a function of the coordinate difference $x_1 - x_2$, as it has to be since we do not expect physics to depend on the choice of origin. This second point gives us a clue as to how best to perform the $x_1 - x_2$ integral in (6.91). Let us introduce the new variables $x = x_1 - x_2$, $X = (x_1 + x_2)/2$. Then (problem 6.6) (6.91) reduces to

$$(-\mathrm{i}g)^2(2\pi)^4\delta^4(p_{\rm A} + p_{\rm B} - p'_{\rm A} - p'_{\rm B}) \int \mathrm{d}^4x\, \mathrm{e}^{\mathrm{i}q\cdot x} \int \frac{\mathrm{d}^4k}{(2\pi)^4} \mathrm{e}^{-\mathrm{i}k\cdot x} \frac{\mathrm{i}}{k^2 - m_{\rm C}^2 + \mathrm{i}\epsilon} \tag{6.99}$$

$$= (-\mathrm{i}g)^2(2\pi)^4\delta^4(p_{\rm A} + p_{\rm B} - p'_{\rm A} - p'_{\rm B}) \frac{\mathrm{i}}{q^2 - m_{\rm C}^2 + \mathrm{i}\epsilon} \tag{6.100}$$

where $q = p_{\rm A} - p'_{\rm B} = p'_{\rm A} - p_{\rm B}$ is the *4-momentum transfer* carried by the exchanged C-quantum in figure 6.4, and we have used the four-dimensional version of (E.26). We associate this single expression, which includes the two coordinate space processes of figure 6.2, with the *single momentum–space Feynman diagram* of figure 6.4. The arrows refer merely to the flow of 4-momentum, which is conserved at each 'vertex' (i.e. meeting of three lines). Thus although the arrow on the exchanged C-line is drawn as indicated, this has nothing to do with any presumed order of emission/absorption of the exchanged quantum. It cannot do so, after all, since in this diagram the states all have definite 4-momentum and hence are totally delocalized in space–time; equivalently, we recall from (6.91) that the amplitude in fact involves integrals over *all* space–time.

A similar analysis (problem 6.7) shows that the contribution of the contractions (6.88) to the S-matrix element (6.74) is

$$(-\mathrm{i}g)^2(2\pi)^4\delta^4(p_{\rm A} + p_{\rm B} - p'_{\rm A} - p'_{\rm B}) \frac{\mathrm{i}}{(p_{\rm A} + p_{\rm B})^2 - m_{\rm C}^2 + \mathrm{i}\epsilon} \tag{6.101}$$

which is represented by the momentum–space Feynman diagram of figure 6.5.

At this point we may start to write down the *Feynman rules for the* ABC *theory*, which enable us to associate a precise mathematical expression for an amplitude with a Feynman diagram such as figure 6.4 or figure 6.5. It is clear that we will always have a factor $(2\pi)^4\delta^4(p_{\rm A} + p_{\rm B} - p'_{\rm A} - p'_{\rm B})$ for all 'connected' diagrams, following from the flow of the conserved 4-momentum through the diagrams. It is conventional to extract this factor, and to define the invariant amplitude $\mathcal{M}_{\rm fi}$ via

$$S_{\rm fi} = \delta_{\rm fi} + \mathrm{i}(2\pi)^4\delta^4(p_{\rm f} - p_{\rm i})\mathcal{M}_{\rm fi} \tag{6.102}$$

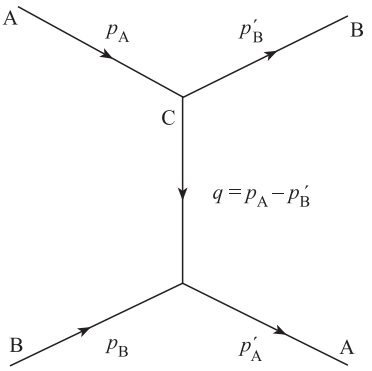


FIGURE 6.4
Momentum–space Feynman diagram corresponding to the $O(g^2)$ amplitude of (6.100).

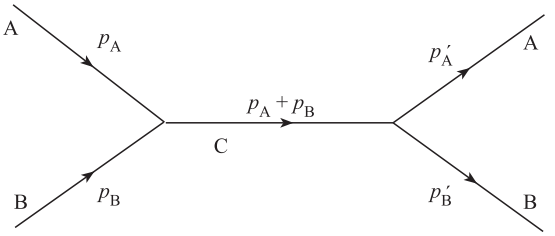


FIGURE 6.5
Momentum–space Feynman diagram corresponding to the $O(g^2)$ amplitude of (6.101).

in general (cf (6.57)). The rules reconstruct the invariant amplitude $\mathrm{i}\mathcal{M}_{\mathrm{fi}}$ corresponding to a given diagram, and for the present case they are:

(i) At each vertex, a factor $-\mathrm{i}g$.

(ii) For each internal line, a factor

$$\frac{\mathrm{i}}{q_i^2 - m_i^2 + \mathrm{i}\epsilon} \tag{6.103}$$

where $i = \mathrm{A}, \mathrm{B}$ or C and q_i is the 4-momentum carried by that line. The factor (6.103) is the *Feynman propagator in momentum space*, for the scalar particle 'i'.

Of course, it is no big deal to give a set of rules which will just reconstruct (6.100) and (6.101). The real power of the 'rules' is that they work for all diagrams we can draw by joining together vertices and propagators (except that we have not yet explained what to do if more than one particle appears 'internally' between two vertices, as in figures 6.3(a)–(c): see section 6.3.5).

6.3.3 A + B → A + B scattering: the Yukawa exchange mechanism, *s* and *u* channel processes

Referring back to section 1.3.3, equation (1.28), we see that the amplitude for the exchange process of figure 6.4 indeed has the form suggested there, namely $\sim g^2/(q^2 - m_{\mathrm{C}}^2)$ if C is exchanged. We have seen how, in the static limit, this may be interpreted as a Yukawa interaction of range $\hbar/m_{\mathrm{C}}c$ between the particles A and B, treated in the Born approximation. Expression (6.100), then, provides us with the correct relativistic formula for this Yukawa mechanism.

There is more to be said about this fundamental amplitude (6.100), which is essentially the C propagator in momentum space. While it is always true that $p_i^2 = m_i^2$ for a free particle of 4-momentum p_i and rest mass m_i, it is not the case that $q^2 = m_{\mathrm{C}}^2$ in (6.100). We emphasized after (6.95) that the variable k_0 introduced there was not equal to $(\boldsymbol{k}^2 + m_{\mathrm{C}}^2)^{1/2}$, and the result of the step (6.99) to (6.100) was to replace k_0 by q_0 and $\boldsymbol{k}$ by $\boldsymbol{q}$, so that $q_0 \neq (\boldsymbol{q}^2 + m_{\mathrm{C}}^2)^{1/2}$, i.e. $q^2 = q_0^2 - \boldsymbol{q}^2 \neq m_{\mathrm{C}}^2$. So the exchanged quantum in figure 6.4 does not satisfy the 'mass-shell condition' $p_i^2 = m_i^2$; it is said to be 'off-mass shell' or 'virtual' (see also problem 6.8). It is quite a different entity from a free quantum. Indeed, as we saw in more elementary physical terms in section 1.3.2, it has a fleeting existence, as sanctioned by the uncertainty relation.

It is convenient, at this point, to introduce some kinematic variables which will appear often in following chapters. These are the 'Mandelstam variables' (Mandelstam 1958, 1959)

$$s = (p_{\mathrm{A}} + p_{\mathrm{B}})^2 \qquad t = (p_{\mathrm{A}} - p'_{\mathrm{A}})^2 \qquad u = (p_{\mathrm{A}} - p'_{\mathrm{B}})^2. \tag{6.104}$$

They are clearly relativistically invariant. In terms of these variables the

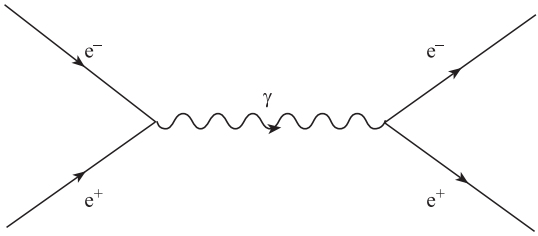


FIGURE 6.6
$O(e^2)$ contribution to $e^+e^- \to e^+e^-$ via annihilation to (and re-emission from) a virtual γ state.

amplitude (6.100) is essentially $\sim 1/(u - m_C^2 + i\epsilon)$, and the amplitude (6.101) is $\sim 1/(s - m_C^2 + i\epsilon)$. The first is said to be a 'u-channel process', the second an 's-channel process'. Amplitudes of the form $(t - m^2)^{-1}$ or $(u - m^2)^{-1}$ are basically one-quantum exchange (i.e. 'force') processes, while those of the form $(s - m_C^2)^{-1}$ have a rather different interpretation, as we now discuss.

Let us first ask: can $s = (p_A + p_B)^2$ ever equal m_C^2 in (6.101)? Since s is invariant, we can evaluate it in any frame we like, for example the centre-of-momentum (CM) frame in which

$$(p_A + p_B)^2 = (E_A + E_B)^2 \tag{6.105}$$

with $E_A = (m_A^2 + \boldsymbol{p}^2)^{1/2}$, $E_B = (m_B^2 + \boldsymbol{p}^2)^{1/2}$. It is then clear that if $m_C < m_A + m_B$ the condition $(p_A + p_B)^2 = m_C^2$ can never be satisfied, and the internal quantum in figure 6.5 is always virtual (note that $p_A + p_B$ is the 4-momentum of the C-quantum). Depending on the details of the theory with which we are dealing, such an s-channel process can have different interpretations. In QED, for example, in the process $e^+ + e^- \to e^+ + e^-$ we could have a virtual γ s-channel process as shown in figure 6.6. This would be called an 'annihilation process' for obvious reasons. In the process $\gamma + e^- \to \gamma + e^-$, however, we could have figure 6.7, which would be interpreted as an absorption and re-emission process (i.e. of a photon).

However, if $m_C > m_A + m_B$, then we can indeed satisfy $(p_A + p_B)^2 = m_C^2$, and so (remembering that ϵ is infinitesimal) we seem to have an infinite result when s (the square of the CM energy) hits the value m_C^2. In fact, this is not the case. If $m_C > m_A + m_B$, the C-particle is unstable against decay to A+B, as we saw in section 6.3.1. The s-channel process must then be interpreted as the formation of a resonance, i.e. of the transitory and decaying state consisting of the single C-particle. Such a process would be described non-relativistically by a Breit–Wigner amplitude of the form

$$\mathcal{M} \propto 1/(E - E_R + i\Gamma/2) \tag{6.106}$$

which produces a peak in $|\mathcal{M}|^2$ centred at $E = E_R$ and full width Γ at half-

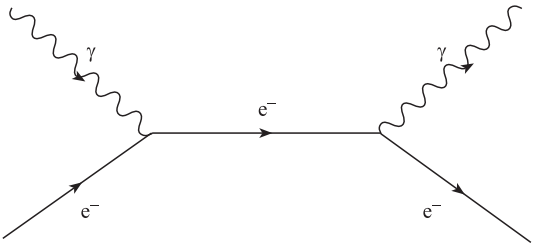


FIGURE 6.7
$O(e^2)$ contribution to $\gamma e^- \to \gamma e^-$ via absorption to (and re-emission from) a virtual e^- state.

height; Γ is, in fact, precisely the width calculated in section 6.3.1. The relativistic generalization of (6.106) is

$$\mathcal{M} \propto \frac{1}{s - M^2 + iM\Gamma} \tag{6.107}$$

where M is the mass of the unstable particle. Thus in the present case the prescription for avoiding the infinity in our amplitude is to replace the infinitesimal '$i\epsilon$' in (6.101) by the finite quantity $im_C\Gamma$, with Γ as calculated in section 6.3.1. We shall see examples of such s-channel resonances in section 9.5.

6.3.4 A + B → A + B scattering: the differential cross section

We complete this exercise in the 'ABC' theory by showing how to calculate the cross section for A+B→ A+B scattering in terms of the invariant amplitude $\mathcal{M}_{\rm fi}$ of (6.102). The discussion will closely parallel the calculation of the decay rate Γ in section 6.3.1.

As in (6.56), the transition rate per unit volume, in this case, is

$$\dot{P}_{\rm fi} = (2\pi)^4 \delta^4(p_{\rm A} + p_{\rm B} - p'_{\rm A} - p'_{\rm B}) |\mathcal{M}_{\rm fi}|^2. \tag{6.108}$$

In order to obtain a quantity which may be compared from experiment to experiment, we must remove the dependence of the transition rate on the incident flux of particles and on the number of target particles per unit volume. Now the flux of beam particles ('A' ones, let us say) incident on a stationary target is just the number of particles per unit area reaching the target in unit time which, with our normalization of '$2E$ particles per unit volume', is just

$$|\boldsymbol{v}| 2E_{\rm A} \tag{6.109}$$

where $\boldsymbol{v}$ is the velocity of the incident A in the rest frame of the target B.

The number of target particles per unit volume is $2E_{\rm B}$ ($= 2m_{\rm B}$ for B at rest, of course).

We must also include the 'density of final states' factors, as in (6.59). Putting all this together, the total cross section σ is given in terms of the differential cross section $\mathrm{d}\sigma$ by

$$\begin{aligned}\sigma &= \int \mathrm{d}\sigma = \frac{1}{2E_{\rm B}2E_{\rm A}|\boldsymbol{v}|}(2\pi)^4 \int \delta^4(p_{\rm A}+p_{\rm B}-p'_{\rm A}-p'_{\rm B}) \\ &\quad \times |\mathcal{M}_{\rm fi}|^2 \frac{\mathrm{d}^3\boldsymbol{p}'_{\rm A}}{(2\pi)^3 2E'_{\rm A}} \frac{\mathrm{d}^3\boldsymbol{p}'_{\rm B}}{(2\pi)^3 2E'_{\rm B}} \\ &\equiv \frac{1}{4E_{\rm A}E_{\rm B}|\boldsymbol{v}|} \int |\mathcal{M}_{\rm fi}|^2 \mathrm{dLips}(s;p'_{\rm A},p'_{\rm B}), \end{aligned} \tag{6.110}$$

where we have introduced the *Lorentz invariant phase space* $\mathrm{dLips}(s;p'_{\rm A},p'_{\rm B})$ defined by

$$\boxed{\mathrm{dLips}(s;p'_{\rm A},p'_{\rm B}) = \frac{1}{(4\pi)^2}\delta^4(p_{\rm A}+p_{\rm B}-p'_{\rm A}-p'_{\rm B})\frac{\mathrm{d}^3\boldsymbol{p}'_{\rm A}}{E'_{\rm A}}\frac{\mathrm{d}^3\boldsymbol{p}'_{\rm B}}{E'_{\rm B}}.} \tag{6.111}$$

We can write the flux factor for collinear collisions in invariant form using the relation (easily verified in a particular frame (problem 6.9))

$$E_{\rm A}E_{\rm B}|\boldsymbol{v}| = [(p_{\rm A}\cdot p_{\rm B})^2 - m_{\rm A}^2 m_{\rm B}^2]^{1/2}. \tag{6.112}$$

Everything in (6.110) is now written in invariant form.

It is a useful exercise to evaluate $\int \mathrm{d}\sigma$ in a given frame, and the simplest one is the centre-of-momentum (CM) frame defined by

$$\boldsymbol{p}_{\rm A}+\boldsymbol{p}_{\rm B} = \boldsymbol{p}'_{\rm A}+\boldsymbol{p}'_{\rm B} = \boldsymbol{0}. \tag{6.113}$$

However, before specializing to this frame, it is convenient to simplify our expression for dLips. Using the 3-momentum part of the δ-function in (6.110), we can eliminate the integral over $\mathrm{d}^3\boldsymbol{p}'_{\rm B}$:

$$\int \frac{\mathrm{d}^3\boldsymbol{p}'_{\rm B}}{E'_{\rm B}}\delta^4(p_{\rm A}+p_{\rm B}-p'_{\rm A}-p'_{\rm B}) = \frac{1}{E'_{\rm B}}\delta(E_{\rm A}+E_{\rm B}-E'_{\rm A}-E'_{\rm B}), \tag{6.114}$$

remembering also that now $\boldsymbol{p}'_{\rm B}$ has to be replaced by $\boldsymbol{p}_{\rm A}+\boldsymbol{p}_{\rm B}-\boldsymbol{p}'_{\rm A}$ in $\mathcal{M}_{\rm fi}$. On the right-hand side of (6.114), $\boldsymbol{p}'_{\rm B}$ and $E'_{\rm B}$ are no longer independent variables but are determined by the conditions

$$\boldsymbol{p}'_{\rm B} = \boldsymbol{p}_{\rm A}+\boldsymbol{p}_{\rm B}-\boldsymbol{p}'_{\rm A} \qquad E'_{\rm B} = (m_{\rm B}^2+\boldsymbol{p}'^2_{\rm B})^{1/2}. \tag{6.115}$$

Next, convert $\mathrm{d}^3\boldsymbol{p}'_{\rm A}$ to angular variables

$$\mathrm{d}^3\boldsymbol{p}'_{\rm A} = \boldsymbol{p}'^2_{\rm A}\,\mathrm{d}|\boldsymbol{p}'_{\rm A}|\,\mathrm{d}\Omega. \tag{6.116}$$

The energy $E'_{\rm A}$ is given by

$$E'_{\rm A} = (m_{\rm A}^2 + \boldsymbol{p}_{\rm A}'^2)^{1/2} \tag{6.117}$$

so that

$$E'_{\rm A}\,{\rm d}E'_{\rm A} = |\boldsymbol{p}'_{\rm A}|\,{\rm d}|\boldsymbol{p}'_{\rm A}|. \tag{6.118}$$

With all these changes we arrive at the result (valid in any frame)

$${\rm dLips}(s; p'_{\rm A}, p'_{\rm B}) = \frac{1}{(4\pi)^2}\frac{|\boldsymbol{p}'_{\rm A}|{\rm d}E'_{\rm A}}{E'_{\rm B}}\,{\rm d}\Omega\,\delta(E_{\rm A}+E_{\rm B}-E'_{\rm A}-E'_{\rm B}). \tag{6.119}$$

We now specialize to the CM frame for which $\boldsymbol{p}_{\rm A} = \boldsymbol{p} = -\boldsymbol{p}_{\rm B}$, $\boldsymbol{p}'_{\rm A} = \boldsymbol{p}' = -\boldsymbol{p}'_{\rm B}$, and

$$E'_{\rm A} = (m_{\rm A}^2 + \boldsymbol{p}'^2)^{1/2} \qquad E'_{\rm B} = (m_{\rm B}^2 + \boldsymbol{p}'^2)^{1/2} \tag{6.120}$$

so that

$$E'_{\rm A}\,{\rm d}E'_{\rm A} = |\boldsymbol{p}'|\,{\rm d}|\boldsymbol{p}'| = E'_{\rm B}\,{\rm d}E'_{\rm B}. \tag{6.121}$$

Introduce the variable $W' = E'_{\rm A} + E'_{\rm B}$ (note that W' is only constrained to equal the total energy $W = E_{\rm A} + E_{\rm B}$ after the integral over the energy-conserving δ-function has been performed). Then (as in (6.62))

$${\rm d}W' = {\rm d}E'_{\rm A} + {\rm d}E'_{\rm B} = \frac{W'|\boldsymbol{p}'|\,{\rm d}|\boldsymbol{p}'|}{E'_{\rm A}E'_{\rm B}} = \frac{W'}{E'_{\rm B}}\,{\rm d}E'_{\rm A} \tag{6.122}$$

where we have used (6.121) in each of the last two steps. Thus the factor

$$|\boldsymbol{p}'_{\rm A}|\frac{{\rm d}E'_{\rm A}}{E'_{\rm B}}\delta(E_{\rm A}+E_{\rm B}-E'_{\rm A}-E'_{\rm B}) \tag{6.123}$$

becomes

$$|\boldsymbol{p}'|\frac{{\rm d}W'}{W'}\delta(W-W') \tag{6.124}$$

which reduces to

$$|\boldsymbol{p}|/W$$

after integrating over W', since the energy-conservation relation forces $|\boldsymbol{p}'| = |\boldsymbol{p}|$. We arrive at the important result

$$\boxed{{\rm dLips}(s; p'_{\rm A}, p'_{\rm B}) = \frac{1}{(4\pi)^2}\frac{|\boldsymbol{p}|}{W}\,{\rm d}\Omega} \tag{6.125}$$

for the two-body phase space in the CM frame.

The last piece in the puzzle is the evaluation of the flux factor (6.112) in the CM frame. In the CM we have

$$p_{\rm A}\cdot p_{\rm B} = (E_{\rm A}, \boldsymbol{p})\cdot(E_{\rm B}, -\boldsymbol{p}) \tag{6.126}$$

$$= E_{\rm A}E_{\rm B} + \boldsymbol{p}^2 \tag{6.127}$$

and a straightforward calculation shows that

$$(p_{\rm A}\cdot p_{\rm B})^2 - m_{\rm A}^2 m_{\rm B}^2 = \boldsymbol{p}^2 W^2.$$

Hence we finally have

$$\sigma = \int {\rm d}\sigma = \frac{1}{4|\boldsymbol{p}|W}\frac{1}{(4\pi)^2}\frac{|\boldsymbol{p}|}{W}\int |\mathcal{M}_{\rm fi}|^2\,{\rm d}\Omega \tag{6.128}$$

and the *CM differential cross section* is

$$\boxed{\left.\frac{{\rm d}\sigma}{{\rm d}\Omega}\right|_{\rm CM} = \frac{1}{(8\pi W)^2}|\mathcal{M}_{\rm fi}|^2.} \tag{6.129}$$

6.3.5 A + B → A + B scattering: loose ends

We must now return to the amplitudes represented by figures 6.3(a)–(c), which we set aside earlier. Consider first figure 6.3(b). Here the A-particle has continued through without interacting, while the B-particle has made a virtual transition to the 'A + C' state, and then this state has reverted to the original B-state. So this is in the nature of a correction to the 'no-scattering' piece shown in figure 6.1, and does not contribute to $\mathcal{M}_{\rm fi}$. However, such a virtual transition B → A + C → B does represent a modification of the properties of the original single B state, due to its interactions with other fields as specified in $H_{\rm I}'$. We can easily imagine how, at order g^4, an amplitude will occur in which such a virtual process is inserted into the C propagator in figure 6.4 so as to arrive at figure 6.8, from which it is plausible that such emission and reabsorption processes by the same particle effectively modify the propagator for this particle. This, in turn, suggests that part, at least, of their effect will be to modify the mass of the affected particle, so as to change it from the original value specified in the Lagrangian. We may think of this physically as being associated, in some way, with a particle's carrying with it a 'cloud' of virtual particles, with which it is continually interacting; this will affect its mass, much as the mass of an electron in a solid becomes an 'effective' mass due to the various interactions experienced by the electron inside the solid.

We shall postpone the evaluation of amplitudes such as those represented by figures 6.3(b) and (c) to chapter 10. However, we note here just one feature: 4-momentum conservation applied at each vertex in figure 6.3(b) does not determine the individual 4-momenta of the intermediate A and C particles, only the sum of their 4-momenta, which is equal to $p_{\rm B}$ (and this is equal to $p_{\rm B}'$ also, so indeed no scattering has occurred). It is plausible that, if an internal 4-momentum in a diagram is undetermined in terms of the external (fixed) 4-momenta of the physical process, then that undetermined 4-momentum should be integrated over. This is the case, as can be verified straightforwardly by evaluating the amplitude (6.86), for example, as we evaluated (6.89); a similar calculation will be gone through in detail in chapter 10, section 10.1.1. The corresponding Feynman rule is

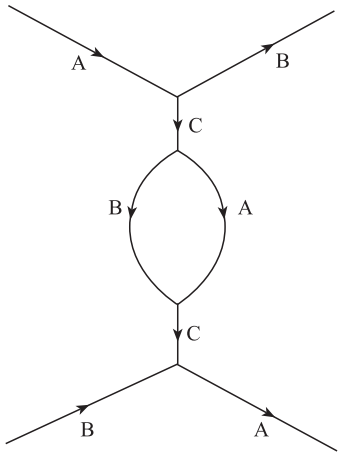


FIGURE 6.8
$O(g^4)$ contribution to the process $\mathrm{A}+\mathrm{B}\to\mathrm{A}+\mathrm{B}$, in which a virtual transition $\mathrm{C}\to\mathrm{A}+\mathrm{B}\to\mathrm{C}$ occurs in the C propagator.

(iii) For each internal 4-momentum k which is not fixed by 4-momentum conservation, carry out the integration $\int \mathrm{d}^4k/(2\pi)^4$. One such integration with respect to an internal 4-momentum occurs for each closed loop.

If we apply this new rule to figure 6.3(b), we find that we need to evaluate the integral

$$\int \frac{\mathrm{d}^4k}{(2\pi)^4}\,\frac{\mathrm{i}}{(k^2-m_{\mathrm{A}}^2)}\,\frac{\mathrm{i}}{((p_{\mathrm{B}}-k)^2-m_{\mathrm{C}}^2)} \tag{6.130}$$

which, by simple counting of powers of k in numerator and denominator, is logarithmically divergent. Thus we learn that, almost before we have started quantum field theory in earnest, we seem to have run into a serious problem, which is going to affect all higher-order processes containing loops. The procedure whereby these infinities are tamed is called renormalization, and we shall return to it in chapter 10.

Finally, what about figure 6.3(a)? In this case nothing at all has occurred to either of the scattering particles, and instead a virtual trio of $\mathrm{A}+\mathrm{B}+\mathrm{C}$ has appeared from the vacuum, and then disappeared back again. Such processes are called, obviously enough, vacuum diagrams. This particular one is in fact only (another) correction to figure 6.1, and it makes no contribution to $\mathcal{M}_{\mathrm{fi}}$. But as with figure 6.8, at $O(g^4)$ we can imagine such a vacuum process appearing 'alongside' figure 6.4 or figure 6.5, as in figures 6.9(a) and (b). These are called 'disconnected diagrams' and – since in them A and B have certainly interacted – they will contribute to $\mathcal{M}_{\mathrm{fi}}$ (note that they are in this respect quite different from the 'straight through' diagrams of figures 6.3(b)

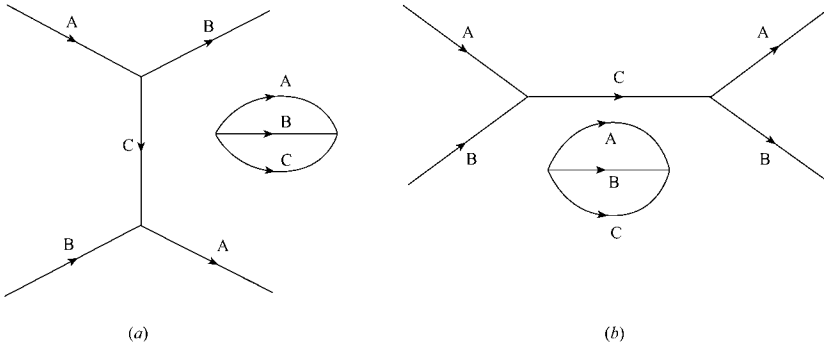


FIGURE 6.9
$O(g^4)$ disconnected diagrams in $\mathrm{A} + \mathrm{B} \to \mathrm{A} + \mathrm{B}$.

and (c)). However, it turns out, rather remarkably, that their effect is exactly compensated by another effect we have glossed over – namely the fact that the vacuum $|0\rangle$ we have used in our S-matrix elements is plainly the *unperturbed* vacuum (or ground state), whereas surely the introduction of interactions will perturb it. A careful analysis of this (Peskin and Schroeder 1995, section 7.2) shows that $\mathcal{M}_{\mathrm{fi}}$ is to be calculated from only the *connected* Feynman diagrams.

In this chapter we have seen how the Feynman rules for scattering and decay amplitudes in a simple scalar theory are derived, and also how cross sections and decay rates are calculated. A Yukawa (u-channel) exchange process has been found, in its covariant form, and the analogous s-channel process, together with a hint of the complications which arise when loops are considered, at higher order in g. Unfortunately, however, none of this applies directly to any real physical process, since we do not know of any physical 'scalar ABC' interaction. Rather, the interactions in the Standard Model are all gauge interactions similar to electrodynamics (with the exception of the Higgs sector, which has both cubic and quartic scalar interactions). The mediating quanta of these gauge interactions have spin-1, not zero; furthermore, the matter fields (again apart from the Higgs field) have spin-$\frac{1}{2}$. It is time to begin discussing the complications of spin and the particular form of dynamics associated with the 'gauge principle'.

Problems

6.1 Show that, for a quantum field $\hat{f}(t)$ (suppressing the space coordinates),

$$\int_{-\infty}^{\infty} \mathrm{d}t_1 \int_{-\infty}^{t_1} \mathrm{d}t_2\, \hat{f}(t_1)\hat{f}(t_2) = \tfrac{1}{2} \int_{-\infty}^{\infty} \mathrm{d}t_1 \int_{-\infty}^{\infty} \mathrm{d}t_2\, T(\hat{f}(t_1)\hat{f}(t_2))$$

where

$$\begin{aligned} T(\hat{f}(t_1)\hat{f}(t_2)) &= \hat{f}(t_1)\hat{f}(t_2) \qquad \text{for } t_1 > t_2 \\ &= \hat{f}(t_2)\hat{f}(t_1) \qquad \text{for } t_2 > t_1. \end{aligned}$$

6.2 Verify equation (6.65).

6.3 Let $\hat{\phi}(x,t)$ be a real scalar KG field in one space dimension, satisfying

$$(\Box_x + m^2)\hat{\phi}(x,t) \equiv \left(\frac{\partial^2}{\partial t^2} - \frac{\partial^2}{\partial x^2} + m^2\right)\hat{\phi}(x,t) = 0.$$

(a) Explain why

$$\begin{aligned} T(\hat{\phi}(x_1,t_1)\hat{\phi}(x_2,t_2)) &= \theta(t_1-t_2)\hat{\phi}(x_1,t_1)\hat{\phi}(x_2,t_2) \\ &\quad + \theta(t_2-t_1)\hat{\phi}(x_2,t_2)\hat{\phi}(x_1,t_1) \end{aligned}$$

(see equation (E.47) for a definition of the θ-function).

(b) Using equation (E.46), show that

$$\frac{\mathrm{d}}{\mathrm{d}x}\theta(x-a) = \delta(x-a).$$

(c) Using the result of (b) with appropriate changes of variable, and equation (5.118), show that

$$\begin{aligned} &\frac{\partial}{\partial t_1}\{T(\hat{\phi}(x_1,t_1)\hat{\phi}(x_2,t_2))\} \\ &\quad = \theta(t_1-t_2)\dot{\hat{\phi}}(x_1,t_1)\hat{\phi}(x_2,t_2) + \theta(t_2-t_1)\hat{\phi}(x_2,t_2)\dot{\hat{\phi}}(x_1,t_1). \end{aligned}$$

(d) Using (5.117) and (5.122) show that

$$\frac{\partial^2}{\partial {t_1}^2}\{T(\hat{\phi}(x_1,t_1)\hat{\phi}(x_2,t_2))\} = -\mathrm{i}\delta(x_1-x_2)\delta(t_1-t_2) + T(\ddot{\hat{\phi}}(x_1,t_1)\hat{\phi}(x_2,t_2))$$

and hence show that

$$\left(\frac{\partial^2}{\partial {t_1}^2} - \frac{\partial^2}{\partial {x_1}^2} + m^2\right) T(\hat{\phi}(x_1,t_1)\hat{\phi}(x_2,t_2)) = -\mathrm{i}\delta(x_1-x_2)\delta(t_1-t_2).$$

This shows that $T(\hat{\phi}(x_1,t_1)\hat{\phi}(x_2,t_2))$ is a Green function (see appendix G, equation (G.25) – the i is included here conventionally) for the KG operator

$$\frac{\partial^2}{\partial {t_1}^2} - \frac{\partial^2}{\partial {x_1}^2} + m^2.$$

The four-dimensional generalization is immediate.

6.4 Verify (6.90).

6.5 Verify (6.92).

6.6 Verify (6.99) and (6.100).

6.7 Show that the contribution of the contractions (6.88) to the S-matrix element (6.74) is given by (6.101).

6.8 Consider the case of equal masses $m_{\mathrm{A}} = m_{\mathrm{B}} = m_{\mathrm{C}}$. Evaluate u of (6.104) in the CM frame (compare section 1.3.6), and show that $u \leq 0$, so that u can never equal m_{C}^2 in (6.100). (This result is generally true for such single particle 'exchange' processes.)

6.9 Verify (6.112).


Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

7

Quantum Field Theory III: Complex Scalar Fields, Dirac and Maxwell Fields; Introduction of Electromagnetic Interactions

In the previous two chapters we have introduced the formalism of relativistic quantum field theory for the case of free real scalar fields obeying the Klein–Gordon (KG) equation of section 3.1, extended it to describe interactions between such quantum fields and shown how the Feynman rules for a simple Yukawa-like theory are derived. It is now time to return to the unfortunately rather more complicated real world of quarks and leptons interacting via gauge fields – in particular electromagnetism. For this, several generalizations of the formalism of chapter 5 are necessary.

First, a glance back at chapter 2 will remind the reader that the electromagnetic interaction has everything to do with the *phase* of wavefunctions, and hence presumably of their quantum field generalizations: fields which are real must be electromagnetically neutral. Indeed, as noted very briefly in section 5.3, the quanta of a real scalar field are their own antiparticles; for a given mass, there is only one type of particle being created or destroyed. However, physical particles and antiparticles have identical masses (e.g. e^- and e^+), and it is actually a deep result of quantum field theory that this is so (see section 4.2.5, and the end of section 7.1). In this case for a given mass m, there will have to be two distinct field degrees of freedom, one of which corresponds somehow to the 'particle', the other to the 'antiparticle'. This suggests that we will need a complex field if we want to distinguish particle from antiparticle, even in the absence of electromagnetism (for example, the $(K^0, \bar{K}^0)$ pair). Such a distinction will have to be made in terms of some conserved quantum number (or numbers), having opposite values for 'particle' and 'antiparticle'. This conserved quantum number must be associated with some symmetry. Now, referring again to chapter 2, we recall that electromagnetism is associated with invariance under *local* U(1) phase transformations. Even in the absence of electromagnetism, however, a theory with complex fields can exhibit a *global* U(1) phase invariance. As we shall show in section 7.1, such a symmetry indeed leads to the existence of a conserved quantum number, in terms of which we can distinguish the particle and antiparticle parts of a complex scalar field.

In section 7.2 we generalize the complex scalar field to the complex spinor (Dirac) field, suitable for charged spin-$\frac{1}{2}$ particles. Again we find an analogous

conserved quantum number, associated with a global U(1) phase invariance of the Lagrangian, which serves to distinguish particle from antiparticle. Central to the satisfactory physical interpretation of the Dirac field will be the requirement that it must be quantized with *anticommutation* relations – the famous 'spin-statistics' connection.

The electromagnetic field must then be quantized, and section 6.3 describes the considerable difficulties this poses. With all this in place, we can easily introduce (section 7.4) electromagnetic interactions via the 'gauge principle' of chapter 2. The resulting Lagrangians and Feynman rules will be applied to simple processes in the following chapter. In the final section of this chapter, we return to the discrete symmetries of chapter 4, and extend them from the single particle theory to quantum field theory.

7.1 The complex scalar field: global U(1) phase invariance, particles and antiparticles

Consider a Lagrangian for two free fields $\hat{\phi}_1$ and $\hat{\phi}_2$ having the same mass M:

$$\hat{\mathcal{L}} = \tfrac{1}{2}\partial_\mu\hat{\phi}_1\partial^\mu\hat{\phi}_1 - \tfrac{1}{2}M^2\hat{\phi}_1^2 + \tfrac{1}{2}\partial_\mu\hat{\phi}_2\partial^\mu\hat{\phi}_2 - \tfrac{1}{2}M^2\hat{\phi}_2^2. \tag{7.1}$$

We shall see how this is appropriate to a 'particle–antiparticle' situation.

In general 'particle' and 'antiparticle' are distinguished by having opposite values of one or more conserved additive quantum numbers. Since these quantum numbers are conserved, the operators corresponding to them commute with the Hamiltonian and are constant in time (in the Heisenberg formulation – see equation (5.59)); such operators are called *symmetry operators* and will be increasingly important in later chapters. For the present we consider the simplest case in which 'particle' and 'antiparticle' are distinguished by having opposite eigenvalues of just one symmetry operator. This situation is already realized in the simple Lagrangian of (7.1). The symmetry involved is just this: $\hat{\mathcal{L}}$ of (7.1) is left unchanged (is *invariant*) if $\hat{\phi}_1$ and $\hat{\phi}_2$ are replaced by $\hat{\phi}_1'$ and $\hat{\phi}_2'$, where (cf (2.64))

$$\begin{aligned}\hat{\phi}_1' &= (\cos\alpha)\hat{\phi}_1 - (\sin\alpha)\hat{\phi}_2\\ \hat{\phi}_2' &= (\sin\alpha)\hat{\phi}_1 + (\cos\alpha)\hat{\phi}_2\end{aligned} \tag{7.2}$$

where α is a real parameter. This is like a rotation of coordinates about the z-axis of ordinary space, but of course it mixes *field* degrees of freedom, not spatial coordinates. The symmetry transformation of (7.2) is sometimes called an 'O(2) transformation', referring to the two-dimensional rotation group O(2). We can easily check the invariance of $\hat{\mathcal{L}}$, i.e.

$$\hat{\mathcal{L}}(\hat{\phi}_1', \hat{\phi}_2') = \hat{\mathcal{L}}(\hat{\phi}_1, \hat{\phi}_2); \tag{7.3}$$

see problem 7.1.

Now let us see what is the conservation law associated with this symmetry. It is simpler (and sufficient) to consider an infinitesimal rotation characterized by the infinitesimal parameter ϵ, for which $\cos\epsilon \approx 1$ and $\sin\epsilon \approx \epsilon$ so that (7.2) becomes

$$\begin{aligned} \hat{\phi}_1' &= \hat{\phi}_1 - \epsilon\hat{\phi}_2 \\ \hat{\phi}_2' &= \hat{\phi}_2 + \epsilon\hat{\phi}_1 \end{aligned} \tag{7.4}$$

and we can define changes $\delta\hat{\phi}_i$ by

$$\begin{aligned} \delta\hat{\phi}_1 &\equiv \hat{\phi}_1' - \hat{\phi}_1 = -\epsilon\hat{\phi}_2 \\ \delta\hat{\phi}_2 &\equiv \hat{\phi}_2' - \hat{\phi}_2 = +\epsilon\hat{\phi}_1. \end{aligned} \tag{7.5}$$

Under this transformation $\hat{\mathcal{L}}$ is invariant, and so $\delta\hat{\mathcal{L}} = 0$. But $\hat{\mathcal{L}}$ is an explicit function of $\hat{\phi}_1$, $\hat{\phi}_2$, $\partial_\mu\hat{\phi}_1$ and $\partial_\mu\hat{\phi}_2$. Thus we can write

$$0 = \delta\hat{\mathcal{L}} = \frac{\partial\hat{\mathcal{L}}}{\partial(\partial_\mu\hat{\phi}_1)}\delta(\partial_\mu\hat{\phi}_1) + \frac{\partial\hat{\mathcal{L}}}{\partial(\partial_\mu\hat{\phi}_2)}\delta(\partial_\mu\hat{\phi}_2) + \frac{\partial\hat{\mathcal{L}}}{\partial\hat{\phi}_1}\delta\hat{\phi}_1 + \frac{\partial\hat{\mathcal{L}}}{\partial\hat{\phi}_2}\delta\hat{\phi}_2. \tag{7.6}$$

This is a bit like the manipulations leading up to the derivation of the Euler–Lagrange equations in section 5.2.4, but now the changes $\delta\hat{\phi}_i$ ($i \equiv 1, 2$) have nothing to do with space–time trajectories – they mix up the two fields. However, we can *use* the equations of motion for $\hat{\phi}_1$ and $\hat{\phi}_2$ to rewrite $\delta\hat{\mathcal{L}}$ as

$$\begin{aligned} 0 \;=\; & \frac{\partial\hat{\mathcal{L}}}{\partial(\partial_\mu\hat{\phi}_1)}\delta(\partial_\mu\hat{\phi}_1) + \frac{\partial\hat{\mathcal{L}}}{\partial(\partial_\mu\hat{\phi}_2)}\delta(\partial_\mu\hat{\phi}_2) \\ & + \left[\partial_\mu\left(\frac{\partial\hat{\mathcal{L}}}{\partial(\partial_\mu\hat{\phi}_1)}\right)\right]\delta\hat{\phi}_1 + \left[\partial_\mu\left(\frac{\partial\hat{\mathcal{L}}}{\partial(\partial_\mu\hat{\phi}_2)}\right)\right]\delta\hat{\phi}_2. \end{aligned} \tag{7.7}$$

Since $\delta(\partial_\mu\hat{\phi}_i) = \partial_\mu(\delta\hat{\phi}_i)$, the right-hand side of (7.7) is just a total divergence, and (7.7) becomes

$$0 = \partial_\mu\left[\frac{\partial\hat{\mathcal{L}}}{\partial(\partial_\mu\hat{\phi}_1)}\delta\hat{\phi}_1 + \frac{\partial\hat{\mathcal{L}}}{\partial(\partial_\mu\hat{\phi}_2)}\delta\hat{\phi}_2\right]. \tag{7.8}$$

These formal steps are actually perfectly general, and will apply whenever a certain Lagrangian depending on two fields $\hat{\phi}_1$ and $\hat{\phi}_2$ is invariant under $\hat{\phi}_i \to \hat{\phi}_i + \delta\hat{\phi}_i$. In the present case, with $\delta\hat{\phi}_i$ given by (7.5), we have

$$\begin{aligned} 0 \;&=\; \partial_\mu\left[-\frac{\partial\hat{\mathcal{L}}}{\partial(\partial_\mu\hat{\phi}_1)}\epsilon\hat{\phi}_2 + \frac{\partial\hat{\mathcal{L}}}{\partial(\partial_\mu\hat{\phi}_2)}\epsilon\hat{\phi}_1\right] \\ &=\; \epsilon\partial_\mu[(\partial^\mu\hat{\phi}_2)\hat{\phi}_1 - (\partial^\mu\hat{\phi}_1)\hat{\phi}_2] \end{aligned} \tag{7.9}$$

where the free-field Lagrangian (7.1) has been used in the second step. Since ϵ is arbitrary, we have proved that the 4-vector operator

$$\hat{N}^\mu_\phi = \hat{\phi}_1\partial^\mu\hat{\phi}_2 - \hat{\phi}_2\partial^\mu\hat{\phi}_1 \tag{7.10}$$

is conserved:

$$\partial_\mu \hat{N}^\mu_\phi = 0. \tag{7.11}$$

Such conserved 4-vector operators are called *symmetry currents*, often denoted generically by $\hat{J}^\mu$. There is a general theorem (due to Noether (1918) in the classical field case) to the effect that if a Lagrangian is invariant under a continuous transformation, then there will be an associated symmetry current. We shall consider Noether's theorem again in volume 2.

What does all this have to do with symmetry operators? Written out in full, (7.11) is

$$\partial \hat{N}^0_\phi / \partial t + \boldsymbol{\nabla} \cdot \hat{\boldsymbol{N}}_\phi = 0. \tag{7.12}$$

Integrating this equation over all space, we obtain

$$\frac{\mathrm{d}}{\mathrm{d}t} \int_{V\to\infty} \hat{N}^0_\phi \, \mathrm{d}^3\boldsymbol{x} + \int_{S\to\infty} \hat{\boldsymbol{N}}_\phi \cdot \mathrm{d}\boldsymbol{S} = 0 \tag{7.13}$$

where we have used the divergence theorem in the second term. Normally the fields may be assumed to die off sufficiently fast at infinity that the surface integral vanishes (by using wave packets, for example), and we can therefore deduce that the quantity $\hat{N}_\phi$ is constant in time, where

$$\hat{N}_\phi = \int \hat{N}^0_\phi \, \mathrm{d}^3\boldsymbol{x} \tag{7.14}$$

that is, *the volume integral of the $\mu = 0$ component of a symmetry current is a symmetry operator.*

In order to see how $\hat{N}_\phi$ serves to distinguish 'particle' from 'antiparticle' in the simple example we are considering, it turns out to be convenient to regard $\hat{\phi}_1$ and $\hat{\phi}_2$ as components of a single *complex* field

$$\begin{aligned} \hat{\phi} &= \tfrac{1}{\sqrt{2}}(\hat{\phi}_1 - \mathrm{i}\hat{\phi}_2) \\ \hat{\phi}^\dagger &= \tfrac{1}{\sqrt{2}}(\hat{\phi}_1 + \mathrm{i}\hat{\phi}_2). \end{aligned} \tag{7.15}$$

The plane-wave expansions of the form (5.155) for $\hat{\phi}_1$ and $\hat{\phi}_2$ imply that $\hat{\phi}$ has the expansion

$$\hat{\phi} = \int \frac{\mathrm{d}^3\boldsymbol{k}}{(2\pi)^3\sqrt{2\omega}} [\hat{a}(k)\mathrm{e}^{-\mathrm{i}k\cdot x} + \hat{b}^\dagger(k)\mathrm{e}^{\mathrm{i}k\cdot x}] \tag{7.16}$$

where

$$\begin{aligned} \hat{a}(k) &= \tfrac{1}{\sqrt{2}}(\hat{a}_1 - \mathrm{i}\hat{a}_2) \\ \hat{b}^\dagger(k) &= \tfrac{1}{\sqrt{2}}(\hat{a}_1^\dagger - \mathrm{i}\hat{a}_2^\dagger) \end{aligned} \tag{7.17}$$

and $\omega = (M^2 + \boldsymbol{k}^2)^{1/2}$. The operators $\hat{a}$, $\hat{a}^\dagger$, $\hat{b}$, $\hat{b}^\dagger$ obey the commutation relations

$$\begin{aligned} [\hat{a}(k), \hat{a}^\dagger(k')] &= (2\pi)^3\delta^3(\boldsymbol{k} - \boldsymbol{k}') \\ [\hat{b}(k), \hat{b}^\dagger(k')] &= (2\pi)^3\delta^3(\boldsymbol{k} - \boldsymbol{k}') \end{aligned} \tag{7.18}$$

with all others vanishing; this follows from the commutation relations

$$[\hat{a}_i(k), \hat{a}_j^\dagger(k')] = \delta_{ij}(2\pi)^3\delta(\boldsymbol{k}-\boldsymbol{k}') \qquad \text{etc} \tag{7.19}$$

for the $\hat{a}_i$ operators. Note that two distinct mode operators, $\hat{a}$ and $\hat{b}$, are appearing in the expansion (7.16) of the complex field.

In terms of this complex $\hat{\phi}$ the Lagrangian of (7.1) becomes

$$\hat{\mathcal{L}} = \partial_\mu\hat{\phi}^\dagger\partial^\mu\hat{\phi} - M^2\hat{\phi}^\dagger\hat{\phi} \tag{7.20}$$

and the Hamiltonian is (dropping the zero-point energy, i.e. normally ordering)

$$\hat{H} = \int \frac{\mathrm{d}^3\boldsymbol{k}}{(2\pi)^3}[\hat{a}^\dagger(k)\hat{a}(k) + \hat{b}^\dagger(k)\hat{b}(k)]\omega. \tag{7.21}$$

The O(2) transformation (7.2) becomes a simple phase change

$$\hat{\phi}' = \mathrm{e}^{-\mathrm{i}\alpha}\hat{\phi} \tag{7.22}$$

which (see comment (iii) of section 2.6) is called a global U(1) phase transformation; plainly the Lagrangian (7.20) is invariant under (7.22). The associated symmetry current $\hat{N}_\phi^\mu$ becomes

$$\hat{N}_\phi^\mu = \mathrm{i}(\hat{\phi}^\dagger\partial^\mu\hat{\phi} - \hat{\phi}\partial^\mu\hat{\phi}^\dagger) \tag{7.23}$$

and the symmetry operator $\hat{N}_\phi$ is (see problem 7.2)

$$\hat{N}_\phi = \int \frac{\mathrm{d}^3\boldsymbol{k}}{(2\pi)^3}[\hat{a}^\dagger(k)\hat{a}(k) - \hat{b}^\dagger(k)\hat{b}(k)]. \tag{7.24}$$

Note that $\hat{N}_\phi$ has been normally ordered in anticipation of our later vacuum definition (7.30), so that $\hat{N}_\phi|0\rangle = 0$.

We now observe that the Hamiltonian (7.21) involves the *sum* of the number operators for 'a' quanta and 'b' quanta, whereas $\hat{N}_\phi$ involves the *difference* of these number operators. Put differently, $\hat{N}_\phi$ counts $+1$ for each particle of type 'a' and -1 for each of type 'b'. This strongly suggests the interpretation that the b's are the antiparticles of the a's: $\hat{N}_\phi$ is the conserved symmetry operator whose eigenvalues serve to distinguish them. For a general state, the eigenvalue of $\hat{N}_\phi$ is the number of a's minus the number of anti-a's and it is a constant of the motion, as is the total energy, which is the sum of the a energies and anti-a energies.

We have here the simplest form of the particle–antiparticle distinction: only one additive conserved quantity is involved. A more complicated example would be the $(\mathrm{K}^+, \mathrm{K}^-)$ pair, which have opposite values of strangeness and of electric charge. Of course, in our simple Lagrangian (7.20) the electromagnetic interaction is absent, and so no electric charge can be defined (we shall remedy

this later); the complex field $\hat{\phi}$ would be suitable (in respect of strangeness) for describing the $(\mathrm{K}^0, \bar{\mathrm{K}}^0)$ pair.

The symmetry operator $\hat{N}_\phi$ has a number of further important properties. First of all, we have shown that $\mathrm{d}\hat{N}_\phi/\mathrm{d}t = 0$ from the general (Noether) argument, but we ought also to check that

$$[\hat{N}_\phi, \hat{H}] = 0 \tag{7.25}$$

as is required for consistency, and expected for a symmetry operator. This is indeed true (see problem 7.2(a)). We can also show

$$\begin{aligned} [\hat{N}_\phi, \hat{\phi}] &= -\hat{\phi} \\ [\hat{N}_\phi, \hat{\phi}^\dagger] &= \hat{\phi}^\dagger \end{aligned} \tag{7.26}$$

and, by expansion of the exponential (problem 7.2(b)), that

$$\hat{U}(\alpha)\hat{\phi}\hat{U}^{-1}(\alpha) = \mathrm{e}^{-\mathrm{i}\alpha}\hat{\phi} = \hat{\phi}' \tag{7.27}$$

with

$$\hat{U}(\alpha) = \mathrm{e}^{\mathrm{i}\alpha\hat{N}_\phi}. \tag{7.28}$$

This shows that the unitary operator $\hat{U}(\alpha)$ effects *finite* U(1) rotations.

Consider now a state $|N_\phi\rangle$ which is an eigenstate of $\hat{N}_\phi$ with eigenvalue N_ϕ. What is the eigenvalue of $\hat{N}_\phi$ for the state $\hat{\phi}|N_\phi\rangle$? It is easy to show, using (7.26), that

$$\hat{N}_\phi\hat{\phi}|N_\phi\rangle = (N_\phi - 1)\hat{\phi}|N_\phi\rangle \tag{7.29}$$

so the application of $\hat{\phi}$ to a state lowers its $\hat{N}_\phi$ eigenvalue by 1. This is consistent with our interpretation that the $\hat{\phi}$ field destroys particles 'a' via the $\hat{a}$ piece in (7.16). (This '$\hat{\phi}$ destroys particles' convention is the reason for choosing $\hat{\phi} = (\hat{\phi}_1 - \mathrm{i}\hat{\phi}_2)/\sqrt{2}$ in (7.15), which in turn led to the minus sign in the relation (7.26) and to the earlier eigenvalue $N_\phi - 1$.) That $\hat{\phi}$ lowers the $\hat{N}_\phi$ eigenvalue by 1 is also consistent with the interpretation that the same field $\hat{\phi}$ creates an antiparticle via the $\hat{b}^\dagger$ piece in (7.16). In the same way, by considering $\hat{\phi}^\dagger|N_\phi\rangle$, one easily verifies that $\hat{\phi}^\dagger$ increases N_ϕ by 1, by creating a particle via $\hat{a}^\dagger$ or destroying an antiparticle via $\hat{b}$. The vacuum state (no particles and no antiparticles present) is defined by

$$\hat{a}(k)|0\rangle = \hat{b}(k)|0\rangle = 0 \qquad \text{for all } k. \tag{7.30}$$

As anticipated, therefore, the complex field $\hat{\phi}$ contains two distinct kinds of mode operator, one having to do with particles (with positive N_ϕ), the other with antiparticles (negative N_ϕ). Which we choose to call 'particle' and which 'antiparticle' is of course purely a matter of convention: after all, the negatively charged electron is always regarded as the 'particle', while in the case of the pions we call the positively charged π^+ the particle.

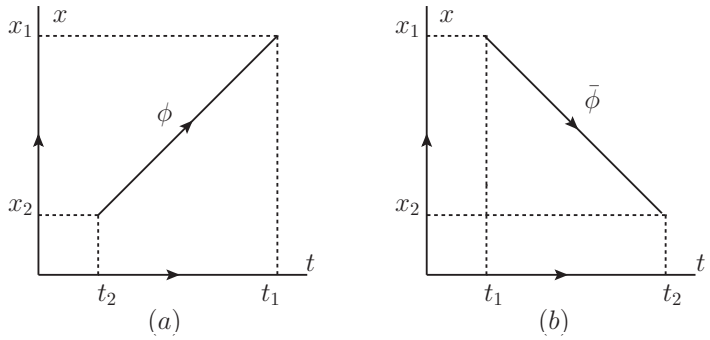


FIGURE 7.1
(*a*) For $t_1 > t_2$, a ϕ particle ($N_\phi = 1$) propagates from x_2 to x_1; (*b*) for $t_2 > t_1$ an anti-ϕ particle ($N_\phi = -1$) propagates from x_1 to x_2.

Feynman rules for theories involving complex scalar fields may be derived by a straightforward extension of the procedure explained in chapter 6. It is, however, worth pausing over the propagator . The only non-vanishing vev of the time-ordered product of two $\hat{\phi}$ fields is $\langle 0|T(\hat{\phi}(x_1)\hat{\phi}^\dagger(x_2))|0\rangle$ (the vev's of $T(\hat{\phi}\hat{\phi})$ and $T(\hat{\phi}^\dagger\hat{\phi}^\dagger)$ vanish with the vacuum defined as in (7.30)). In section 6.3.2 we gave a pictorial interpretation of the propagator for a real scalar field; let us now consider the analogous pictures for the complex field. For $t_1 > t_2$ the time-ordered product is $\hat{\phi}(x_1)\hat{\phi}^\dagger(x_2)$; using the expansion (7.16) and the vacuum conditions (7.30), the only surviving term in the vev is that in which an '$\hat{a}^\dagger$' creates a particle ($N_\phi = 1$) at $(\boldsymbol{x}_2, t_2)$ and an '$\hat{a}$' destroys it at $(\boldsymbol{x}_1, t_1)$; the '$\hat{b}$' operators in $\hat{\phi}(x_2)^\dagger$ give zero when acting on $|0\rangle$, as do the '$\hat{b}^\dagger$' operators in $\hat{\phi}^\dagger(x_1)$ when acting on $\langle 0|$. Thus for $t_1 > t_2$ we have the pictorial interpretation of figure 7.1(*a*). For $t_2 > t_1$, however, the time-ordered product is $\hat{\phi}^\dagger(x_2)\hat{\phi}(x_1)$. Here the surviving vev comes from the '$\hat{b}^\dagger$' in $\hat{\phi}(x_1)$ creating an antiparticle ($N_\phi = -1$) at x_1, which is then annihilated by the '$\hat{b}$' in $\hat{\phi}^\dagger(x_2)$. This $t_2 > t_1$ process is shown in figure 7.1(*b*). The inclusion of both processes shown in figure 7.1 makes sense physically, following considerations similar to those put forward 'intuitively' in section 3.5.4: the process of figure 7.1(*a*) creates (say) a positive unit of N_ϕ at x_2 and loses a positive unit at x_1, while another way of effecting the same 'N_ϕ transfer' is to create an antiparticle of unit negative N_ϕ at x_1, and propagate it to x_2 where it is destroyed, as in figure 7.1(*b*). It is important to be absolutely clear that the *Feynman propagator* $\langle 0|T(\hat{\phi}(x_1)\hat{\phi}^\dagger(x_2))|0\rangle$ includes *both* the processes in figures 7.1(*a*) and (*b*).

In practice, as we found in section 6.3.2, we want the momentum–space version of the propagator, i.e. its Fourier transform. As we also noted there

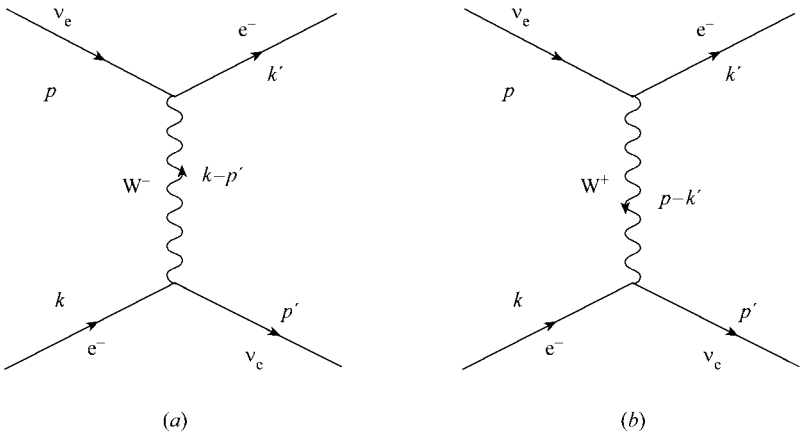


FIGURE 7.2
Equivalent Feynman graphs for single W-exchange in $\nu_e + e^- \to \nu_e + e^-$.

(cf also appendix G), the propagator is a Green function for the KG operator $(\Box + m^2)$ with mass parameter m ; in momentum–space this is just the inverse, $(-k^2 + m^2)^{-1}$. In the present case, since both $\hat{\phi}$ and $\hat{\phi}^\dagger$ obey the same KG equation, with mass parameter M, we expect that the momentum–space version of $\langle 0|T(\hat{\phi}(x_1)\hat{\phi}^\dagger(x_2))|0\rangle$ is also

$$\frac{\mathrm{i}}{k^2 - M^2 + \mathrm{i}\epsilon}. \tag{7.31}$$

This can be verified by inserting the expansion (7.16) into the vev of the T-product, and following the steps used in section 6.3.2 for the scalar case.

In this (momentum–space) version, it is the '$\mathrm{i}\epsilon$' which keeps track of the 'particles going from 2 to 1 if $t_1 > t_2$' and 'antiparticles going from 1 to 2 if $t_2 > t_1$' (recall its appearance in the representation (6.93) of the all-important θ-function). As in the scalar case, momentum–space propagators in Feynman diagrams carry no implied order of emission/absorption process; *both* the processes in figure 7.1 are always included in all propagators. Arrows showing 'momentum flow' now also show the flow of all conserved quantum numbers. Thus the process shown in figure 7.2(a) can equally well be represented as in figure 7.2(b).

There is one more bit of physics to be gleaned from $\langle 0|T(\hat{\phi}(x_1)\hat{\phi}^\dagger(x_2))|0\rangle$. As in the real scalar field case, the vanishing of the commutator at space-like separations

$$[\hat{\phi}(x_1), \hat{\phi}^\dagger(x_2)] = 0 \qquad \text{for } (x_1 - x_2)^2 < 0 \tag{7.32}$$

guarantees the Lorentz invariance of the propagator for the complex scalar field and of the S-matrix. But in this (complex) case there is a further twist

to the story. Evaluation of $[\hat{\phi}(x_1), \hat{\phi}^\dagger(x_2)]$ reveals (problem 7.3) that, in the region $(x_1 - x_2)^2 < 0$, the commutator is the difference of two functions (not field operators), one of which arises from the propagation of a particle from x_2 to x_1, the other of which comes from the propagation of an antiparticle from x_1 to x_2 (just as in figure 7.1). *Both* processes must exist for this difference to be zero, and furthermore for cancellations between them to occur in the space-like region the masses of the particle and antiparticle must be identical. In quantum field theory, therefore, 'causality' (in the sense of condition (7.32) – cf (6.82)) requires that every particle has to have a corresponding antiparticle, with the same mass and opposite quantum numbers. As we saw in chapter 4, these requirements are guaranteed by the **CPT** theorem, which is a consequence of very general principles of quantum field theory.

7.2 The Dirac field and the spin-statistics connection

> I remember that when someone had tried to teach me about creation and annihilation operators, that this operator creates an electron, I said 'how do you create an electron? It disagrees with the conservation of charge,' and in that way I blocked my mind from learning a very practical scheme of calculation.
>
> —From the lecture delivered by Richard Feynman in Stockholm, Sweden, on 11 December 1965, when he received the Nobel Prize in physics, which he shared with Sin-itiro Tomonaga and Julian Schwinger. (Feynman 1966).

We now turn to the problem of setting up a quantum field which, in its wave aspects, satisfies the Dirac equation (cf comment (5) in section 5.2.5), and in its 'particle' aspects creates or annihilates fermions and antifermions. Following the 'Heisenberg–Lagrange–Hamilton' approach of section 5.2.5, we begin by writing down the Lagrangian which, via the corresponding Euler–Lagrange equation, produces the Dirac equation as the 'field equation'. The answer (see problem 7.4) is

$$\mathcal{L}_{\mathrm{D}} = \mathrm{i}\psi^\dagger\dot{\psi} + \mathrm{i}\psi^\dagger\boldsymbol{\alpha}\cdot\boldsymbol{\nabla}\psi - m\psi^\dagger\beta\psi. \tag{7.33}$$

The relativistic invariance of this is more evident in γ-matrix notation (problem 4.3):

$$\mathcal{L}_{\mathrm{D}} = \bar{\psi}(\mathrm{i}\gamma^\mu\partial_\mu - m)\psi. \tag{7.34}$$

We can now attempt to 'quantize' the field ψ by making a mode expansion in terms of plane-wave solutions of the Dirac equation, in a fashion similar to that for the complex scalar field in (7.16). We obtain (see problem 3.8 for the definition of the spinors u and v, and the attendant normalization choice)

$$\hat{\psi} = \int \frac{\mathrm{d}^3\boldsymbol{k}}{(2\pi)^3\sqrt{2\omega}} \sum_{s=1,2} [\hat{c}_s(k)u(k,s)\mathrm{e}^{-\mathrm{i}k\cdot x} + \hat{d}_s^\dagger(k)v(k,s)\mathrm{e}^{\mathrm{i}k\cdot x}], \tag{7.35}$$

where $\omega = (m^2 + \boldsymbol{k}^2)^{1/2}$. We wish to interpret $\hat{c}_s^\dagger(k)$ as the creation operator for a Dirac particle of spin s and momentum k. By analogy with (7.16), we expect that $\hat{d}_s^\dagger(k)$ creates the corresponding antiparticle. Presumably we must define the vacuum by (cf (7.30))

$$\hat{c}_s(k)|0\rangle = \hat{d}_s(k)|0\rangle = 0 \qquad \text{for all } k \text{ and } s = 1, 2. \tag{7.36}$$

A two-fermion state is then

$$|k_1, s_1; k_2, s_2\rangle \propto \hat{c}_{s_1}^\dagger(k_1)\hat{c}_{s_2}^\dagger(k_2)|0\rangle. \tag{7.37}$$

But it is here that there must be a difference from the boson case. We require a state containing two identical fermions to be *antisymmetric* under the exchange of state labels $k_1 \leftrightarrow k_2$, $s_1 \leftrightarrow s_2$, and thus to be forbidden if the two sets of quantum numbers are the same, in accordance with the *Pauli exclusion principle*, responsible for so many well-established features of the structure of matter.

The solution to this dilemma is simple but radical: for fermions, commutation relations are replaced by *anticommutation* relations! The anticommutator of two operators $\hat{A}$ and $\hat{B}$ is written:

$$\{\hat{A}, \hat{B}\} \equiv \hat{A}\hat{B} + \hat{B}\hat{A}. \tag{7.38}$$

If two different $\hat{c}$'s anticommute, then

$$\hat{c}_{s_1}^\dagger(k_1)\hat{c}_{s_2}^\dagger(k_2) + \hat{c}_{s_2}^\dagger(k_2)\hat{c}_{s_1}^\dagger(k_1) = 0 \tag{7.39}$$

so that we have the desired antisymmetry

$$|k_1, s_1; k_2, s_2\rangle = -|k_2, s_2; k_1, s_1\rangle. \tag{7.40}$$

In general we postulate

$$\begin{aligned} \{\hat{c}_{s_1}(k_1), \hat{c}_{s_2}^\dagger(k_2)\} &= (2\pi)^3\delta^3(\boldsymbol{k}_1 - \boldsymbol{k}_2)\delta_{s_1 s_2} \\ \{\hat{c}_{s_1}(k_1), \hat{c}_{s_2}(k_2)\} &= \{\hat{c}_{s_1}^\dagger(k_1), \hat{c}_{s_2}^\dagger(k_2)\} = 0 \end{aligned} \tag{7.41}$$

and similarly for the $\hat{d}$'s and $\hat{d}^\dagger$'s. The factor in front of the δ-function depends on the convention for normalizing Dirac wavefunctions.

We must at once emphasize that in taking this 'replace commutators by anticommutators' step we now depart decisively from the intuitive, quasi-mechanical, picture of a quantum field given in chapter 5, namely as a system of quantized harmonic oscillators. Of course, the field expansion (7.35) is a linear superposition of 'modes' (plane-wave solutions), as for the complex scalar field in (7.16) for example; but the 'mode operators' $\hat{c}_s$ and $\hat{d}_s^\dagger$ are fermionic (obeying anticommutation relations) not bosonic (obeying commutation relations). As mentioned at the end of section 5.1, it does not seem possible to provide any mechanical model of a system (in three dimensions)

whose normal vibrations are fermionic. Correspondingly, there is no concept of a 'classical electron field', analogous to the classical electromagnetic field (which doubtless explains why we tend to think of fermions as basically 'more particle-like'). However, we can certainly recover a quantum mechanical wavefunction from (7.35) by considering, as in comment (5) of section 5.4, the vacuum-to-one-particle matrix element $\langle 0|\hat{\psi}(\boldsymbol{x},t)|k_1,s_1\rangle$.

In the bosonic case, we arrived at the commutation relations (5.130) for the mode operators by postulating the 'fundamental commutator of quantum field theory', equation (5.117), which was an extension to fields of the canonical commutation relations of quantum (particle) mechanics. For fermions, we have simply introduced the anticommutation relations (7.41) 'by hand', so as to satisfy the Pauli principle. We may ask: What then becomes of the analogous 'fundamental commutator' in the fermionic case? A plausible guess is that, as with the mode operators, the 'fundamental commutator' is to be replaced by a 'fundamental anticommutator', between the fermionic field $\hat{\psi}$ and its 'canonically conjugate momentum field' $\hat{\pi}_{\mathrm{D}}$, of the form:

$$\{\hat{\psi}(\boldsymbol{x},t),\hat{\pi}_{\mathrm{D}}(\boldsymbol{y},t)\}=\mathrm{i}\delta(\boldsymbol{x}-\boldsymbol{y}). \tag{7.42}$$

As far as $\hat{\pi}_{\mathrm{D}}$ is concerned, we may suppose that its definition is formally analogous to (5.122), which would yield

$$\hat{\pi}_{\mathrm{D}}=\frac{\partial\hat{\mathcal{L}}_{\mathrm{D}}}{\partial\dot{\hat{\psi}}}=\mathrm{i}\hat{\psi}^{\dagger}. \tag{7.43}$$

We must also not forget that both $\hat{\psi}$ and $\hat{\pi}_{\mathrm{D}}$ are four-component objects, carrying spinor indices. Thus we are led to expect the result

$$\boxed{\{\hat{\psi}_{\alpha}(\boldsymbol{x},t),\hat{\psi}^{\dagger}_{\beta}(\boldsymbol{y},t)\}=\delta(\boldsymbol{x}-\boldsymbol{y})\delta_{\alpha\beta},} \tag{7.44}$$

where α and β are spinor indices. It is a good exercise to check, using (7.41), that this is indeed the case (problem 7.5). We also find

$$\{\hat{\psi}(\boldsymbol{x},t),\hat{\psi}(\boldsymbol{y},t)\}=\{\hat{\psi}^{\dagger}(\boldsymbol{x},t),\hat{\psi}^{\dagger}(\boldsymbol{y},t)\}=0. \tag{7.45}$$

In this (anticommutator) sense, then, we have a 'canonical' formalism for fermions.

The Dirac Hamiltonian density is then (cf (5.123))

$$\hat{\mathcal{H}}_{\mathrm{D}}=\hat{\pi}_{\mathrm{D}}\dot{\hat{\psi}}-\hat{\mathcal{L}}_{\mathrm{D}}=\hat{\psi}^{\dagger}\boldsymbol{\alpha}\cdot-\mathrm{i}\boldsymbol{\nabla}\hat{\psi}+m\hat{\psi}^{\dagger}\beta\hat{\psi} \tag{7.46}$$

using (7.43) and (7.33), and the Hamiltonian is

$$\hat{H}_{\mathrm{D}}=\int[\hat{\psi}^{\dagger}\boldsymbol{\alpha}\cdot-\mathrm{i}\boldsymbol{\nabla}\hat{\psi}+m\hat{\psi}^{\dagger}\beta\hat{\psi}]\;\mathrm{d}^3\boldsymbol{x}. \tag{7.47}$$

One may well wonder *why* things have to be this way – 'bosons commute, fermions anticommute'. To gain further insight, we turn again to a consideration of symmetries and the question of particle and antiparticle – this time for the Dirac *field*, rather than the Dirac wavefunction discussed in chapter 4.

The Dirac field $\hat{\psi}$ is a complex field, as is reflected in the two distinct mode operators in the expansion (7.35); as in the complex scalar field case, there is only one mass parameter and we expect the quanta to be interpretable as particle and antiparticle. The symmetry operator which distinguishes them is found by analogy with the complex scalar field case. We note that $\hat{\mathcal{L}}_{\mathrm{D}}$ (the quantized version of (7.34)) is invariant under the global U(1) transformation

$$\hat{\psi} \to \hat{\psi}' = \mathrm{e}^{-\mathrm{i}\alpha}\hat{\psi} \tag{7.48}$$

which is

$$\hat{\psi} \to \hat{\psi}' = \hat{\psi} - \mathrm{i}\epsilon\hat{\psi} \tag{7.49}$$

in infinitesimal form. The corresponding (Noether) symmetry current can be calculated as

$$\hat{N}^{\mu}_{\psi} = \bar{\hat{\psi}}\gamma^{\mu}\hat{\psi} \tag{7.50}$$

and the associated symmetry operator is

$$\hat{N}_{\psi} = \int \hat{\psi}^{\dagger}\hat{\psi}\,\mathrm{d}^3\boldsymbol{x}. \tag{7.51}$$

$\hat{N}_{\psi}$ is clearly a number operator for the fermion case. As for the complex scalar field, invariance under a global U(1) phase transformation is associated with a number conservation law.

Inserting the plane-wave expansion (7.35), we obtain, after some effort (problem 7.6),

$$\hat{N}_{\psi} = \int \frac{\mathrm{d}^3\boldsymbol{k}}{(2\pi)^3} \sum_{s=1,2} [\hat{c}^{\dagger}_s(k)\hat{c}_s(k) + \hat{d}_s(k)\hat{d}^{\dagger}_s(k)]. \tag{7.52}$$

Similarly the Dirac Hamiltonian may be shown to have the form (problem 7.6)

$$\hat{H}_{\mathrm{D}} = \int \frac{\mathrm{d}^3\boldsymbol{k}}{(2\pi)^3} \sum_{s=1,2} [\hat{c}^{\dagger}_s(k)\hat{c}_s(k) - \hat{d}_s(k)\hat{d}^{\dagger}_s(k)]\omega. \tag{7.53}$$

It is important to state that in obtaining (7.52) and (7.53), we have *not* assumed either commutation or anticommutation relations for the mode operators $\hat{c}$, $\hat{c}^{\dagger}$, $\hat{d}$ and $\hat{d}^{\dagger}$, only properties of the Dirac spinors; in particular, neither (7.52) nor (7.53) has been normally ordered. Suppose now that we assume commutation relations, so as to rewrite the last terms in (7.52) and (7.53) in normally ordered form as $\hat{d}^{\dagger}_s(k)\hat{d}_s(k)$. We see that $\hat{H}_{\mathrm{D}}$ will then contain the *difference* of two number operators for 'c' and 'd' particles, and is therefore not positive-definite as we require for a sensible theory. Moreover, we suspect

that, as in the $\hat{\phi}$ case, the 'd's' ought to be the antiparticles of the 'c's', carrying opposite $\hat{N}_\psi$ value: but $\hat{N}_\psi$ is then (with the previous assumption about commutation relations) just proportional to the *sum* of 'c' and 'd' number operators, counting $+1$ for each type, which does not fit this interpretation. However, if *anticommutation* relations are assumed, both these problems disappear: dropping the usual infinite terms, we obtain the normally ordered forms

$$\hat{N}_\psi = \int \frac{\mathrm{d}^3\boldsymbol{k}}{(2\pi)^3} \sum_{s=1,2} [\hat{c}_s^\dagger(k)\hat{c}_s(k) - \hat{d}_s^\dagger(k)\hat{d}_s(k)] \tag{7.54}$$

$$\hat{H}_\mathrm{D} = \int \frac{\mathrm{d}^3\boldsymbol{k}}{(2\pi)^3} \sum_{s=1,2} [\hat{c}_s^\dagger(k)\hat{c}_s(k) + \hat{d}_s^\dagger(k)\hat{d}_s(k)]\omega \tag{7.55}$$

which are satisfactory, and allow us to interpret the 'd' quanta as the antiparticles of the 'c' quanta. Similar difficulties would have occurred in the complex scalar field case if we had assumed anticommutation relations for the boson operators, and the 'causality' discussion at the end of the preceding section would not have worked either (instead of a difference of terms we would have had a sum). It is in this way that quantum field theory enforces the connection between spin and statistics.

Our discussion here is only a part of a more general approach leading to the same conclusion, first given by Pauli (1940); see also Streater *et al.* (1964).

As in the complex scalar case, the other crucial ingredient we need is the Dirac propagator $\langle 0|T(\hat{\psi}(x_1)\bar{\hat{\psi}}(x_2))|0\rangle$. We shall see in section 7.4 why it is $\bar{\hat{\psi}}$ here rather than $\hat{\psi}^\dagger$ – the reason is essentially to do with Lorentz covariance (see section 4.1.2). Because the $\hat{\psi}$ fields are anticommuting, the T-symbol now has to be understood as

$$T(\hat{\psi}(x_1)\bar{\hat{\psi}}(x_2)) = \hat{\psi}(x_1)\bar{\hat{\psi}}(x_2) \qquad \text{for } t_1 > t_2 \tag{7.56}$$

$$= -\bar{\hat{\psi}}(x_2)\hat{\psi}(x_1) \qquad \text{for } t_2 > t_1. \tag{7.57}$$

Once again, this propagator is proportional to a Green function, this time for the Dirac equation, of course. Using γ-matrix notation (problem 4.3) the Dirac equation is (cf (7.34))

$$(\mathrm{i}\gamma^\mu\partial_\mu - m)\hat{\psi} = 0. \tag{7.58}$$

The momentum–space version of the propagator is proportional to the inverse of the operator in (7.58), when written in k-space, namely to $(\not{k} - m)^{-1}$ where

$$\boxed{\not{k} = \gamma^\mu k_\mu} \tag{7.59}$$

is an important shorthand notation (pronounced 'k-slash'). In fact, the Feynman propagator for Dirac fields is

$$\frac{\mathrm{i}}{\not{k} - m + \mathrm{i}\epsilon}. \tag{7.60}$$

As in (7.31), the $\mathrm{i}\epsilon$ takes care of the particle/antiparticle, emission/absorption business. Formula (7.60) is the fermion analogue of 'rule (ii)' in (6.103).

The reader should note carefully one very important difference between (7.60) and (7.31), which is that (7.60) is a 4×4 *matrix*. What we are really saying (cf (6.98)) is that the Fourier transform of $\langle 0|T(\hat{\psi}_\alpha(x_1)\bar{\hat{\psi}}_\beta(x_2))|0\rangle$, where α and β run over the four components of the Dirac field, is equal to the (α,β) matrix element of the matrix $\mathrm{i}(\not{k}-m+\mathrm{i}\epsilon)^{-1}$:

$$\boxed{\int \mathrm{d}^4(x_1-x_2)\,\mathrm{e}^{\mathrm{i}k\cdot(x_1-x_2)}\langle 0|T(\hat{\psi}_\alpha(x_1)\bar{\hat{\psi}}_\beta(x_2))|0\rangle = \mathrm{i}(\not{k}-m+\mathrm{i}\epsilon)^{-1}_{\alpha\beta}.} \quad (7.61)$$

The form (7.61) can be made to look more like (7.31) by making use of the result (problem 7.7)

$$(\not{k}-m)(\not{k}+m) = (k^2-m^2) \quad (7.62)$$

(where the 4×4 unit matrix is understood on the right-hand side) so as to write (7.61) as

$$\frac{\mathrm{i}(\not{k}+m)}{k^2-m^2+\mathrm{i}\epsilon}. \quad (7.63)$$

As in the scalar case, (7.61) can be directly verified by inserting the field expansion (7.35) into the left-hand side, and following steps analogous to those in equations (6.92)–(6.98). In following this through one will meet the expressions $\sum_s u(k,s)\bar{u}(k,s)$ and $\sum_s v(k,s)\bar{v}(k,s)$, which are also 4×4 matrices. Problem 7.8 shows that these quantities are given by

$$\sum_s u_\alpha(k,s)\bar{u}_\beta(k,s) = (\not{k}+m)_{\alpha\beta} \qquad \sum_s v_\alpha(k,s)\bar{v}_\beta(k,s) = (\not{k}-m)_{\alpha\beta}. \quad (7.64)$$

With these results, and remembering the minus sign in (7.57), one can check (7.63) (problem 7.9).

One might now worry that the adoption of anticommutation relations for Dirac fields might spoil 'causality', in the sense of the discussion after (7.32). One finds, indeed, that the fields $\hat{\psi}$ and $\bar{\hat{\psi}}$ anticommute at space-like separation, but this is enough to preserve causality for physical observables, which will involve an even number of fermionic fields.

We now turn to the problem of quantizing the Maxwell (electromagnetic) field.

7.3 The Maxwell field $A^\mu(x)$

7.3.1 The classical field case

Following the now familiar procedure, our first task is to find the classical field Lagrangian which, via the corresponding Euler–Lagrangian equations, yields

the Maxwell equation for the electromagnetic potential A^ν, namely (cf (2.22))

$$\Box A^\nu - \partial^\nu(\partial_\mu A^\mu) = j^\nu_{\rm em}. \tag{7.65}$$

The answer is (see problem 7.10)

$$\mathcal{L}_{\rm em} = -\frac{1}{4}F_{\mu\nu}\ F^{\mu\nu} - j^\nu_{\rm em}A_\nu \tag{7.66}$$

where $F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu$. So the pure A-field part is the *Maxwell Lagrangian*

$$\mathcal{L}_A = -\frac{1}{4}F_{\mu\nu}\ F^{\mu\nu}. \tag{7.67}$$

Before proceeding to try to quantize (7.67), we need to understand some important aspects of the free *classical* field $A^\nu(x)$.

When $j_{\rm em}$ is set equal to zero, A^ν satisfies the equation

$$\partial_\mu F^{\mu\nu} = \Box A^\nu - \partial^\nu(\partial^\mu A_\mu) = 0. \tag{7.68}$$

As we have seen in section 2.3, these equations are left unchanged if we perform the gauge transformation

$$A^\mu \to A'^\mu = A^\mu - \partial^\mu\chi. \tag{7.69}$$

We can use this freedom to *choose* the A^μ with which we work to satisfy the condition

$$\boxed{\partial_\mu A^\mu = 0.} \tag{7.70}$$

This is called the *Lorentz condition*. The process of choosing a particular condition on A^μ so as to define it (ultimately) uniquely is called 'choosing a gauge'; actually the condition (7.70) does not yet define A^μ uniquely, as we shall see shortly. The Lorentz condition is a very convenient one, since it decouples the different components of A^μ in Maxwell's equations (7.68) – in a covariant way, moreover, leaving the very simple equation

$$\Box A^\mu = 0. \tag{7.71}$$

This has plane-wave solutions of the form

$$A^\mu = N\epsilon^\mu \mathrm{e}^{-\mathrm{i}k\cdot x} \tag{7.72}$$

with $k^2 = 0$ (i.e. $k_0^2 = \boldsymbol{k}^2$), where N is a normalization factor and ϵ^μ is a *polarization vector* for the wave. The gauge condition (7.70) now reduces to a condition on ϵ^μ:

$$k\cdot\epsilon = 0. \tag{7.73}$$

However, we have not yet exhausted all the gauge freedom. We are still free to make another shift in the potential

$$A^\mu \to A^\mu - \partial^\mu\tilde{\chi} \tag{7.74}$$

provided $\tilde{\chi}$ satisfies the massless KG equation

$$\Box\tilde{\chi} = 0. \tag{7.75}$$

This condition on $\tilde{\chi}$ ensures that, even after the further shift, the resulting potential still satisfies $\partial_\mu A^\mu = 0$. For our plane-wave solutions, this residual gauge freedom corresponds to changing ϵ^μ by a multiple of k^μ:

$$\epsilon^\mu \to \epsilon^\mu + \beta k^\mu \equiv \epsilon'^\mu \tag{7.76}$$

which still satisfies $\epsilon'^\mu \cdot k = 0$ *since* $k^2 = 0$ *for these free-field solutions.* The condition $k^2 = 0$ is, of course, the statement that a free photon is massless.

This freedom has important consequences. Consider a solution with

$$k^\mu = (k^0, \boldsymbol{k}) \qquad (k^0)^2 = \boldsymbol{k}^2 \tag{7.77}$$

and polarization vector

$$\epsilon^\mu = (\epsilon^0, \boldsymbol{\epsilon}) \tag{7.78}$$

satisfying the Lorentz condition

$$k \cdot \epsilon = 0. \tag{7.79}$$

Gauge invariance now implies that we can add multiples of k^μ to ϵ^μ and still have a satisfactory polarization vector.

It is therefore clear that we can arrange for the time component of ϵ^μ to vanish so that the Lorentz condition reduces to the 3-vector condition

$$\boldsymbol{k} \cdot \boldsymbol{\epsilon} = 0. \tag{7.80}$$

This means that there are only two independent polarization vectors, both transverse to $\boldsymbol{k}$, i.e. to the propagation direction. For a wave travelling in the z-direction ($k^\mu = (k^0, 0, 0, k^0)$) these may be chosen to be

$$\boldsymbol{\epsilon}_{(1)} = (1, 0, 0) \tag{7.81}$$

$$\boldsymbol{\epsilon}_{(2)} = (0, 1, 0). \tag{7.82}$$

Such a choice corresponds to *linear polarization* of the associated $\boldsymbol{E}$ and $\boldsymbol{B}$ fields – which can be easily calculated from (2.10) and (2.11), given

$$A^\mu_{(i)} = N(0, \boldsymbol{\epsilon}_{(i)})\mathrm{e}^{-\mathrm{i}k\cdot x} \qquad i = 1, 2. \tag{7.83}$$

A commonly used alternative choice is

$$\boldsymbol{\epsilon}(\lambda = +1) = -\frac{1}{\sqrt{2}}(1, \mathrm{i}, 0) \tag{7.84}$$

$$\boldsymbol{\epsilon}(\lambda = -1) = \frac{1}{\sqrt{2}}(1, -\mathrm{i}, 0) \tag{7.85}$$

(linear combinations of (7.81) and (7.82)), which correspond to circularly polarized radiation. The phase convention in (7.84) and (7.85) is the standard one in quantum mechanics for states of definite spin projection ('helicity') $\lambda = \pm 1$ along the direction of motion (the z-axis here). We may easily check that

$$\boldsymbol{\epsilon}^*(\lambda) \cdot \boldsymbol{\epsilon}(\lambda') = \delta_{\lambda\lambda'} \tag{7.86}$$

or, in terms of the corresponding 4-vectors $\epsilon^\mu = (0, \boldsymbol{\epsilon})$,

$$\epsilon^*(\lambda) \cdot \epsilon(\lambda') = -\delta_{\lambda\lambda'}. \tag{7.87}$$

We have therefore arrived at the result, familiar in classical electromagnetic theory, that the free electromagnetic fields are purely *transverse*. Though they are described in this formalism by a vector potential with apparently four independent components $(V, \boldsymbol{A})$, the condition (7.70) reduces this number by one, and the further gauge freedom exploited in (7.74)–(7.76) reduces it by one more.

A crucial point to note is that the reduction to only *two* independent field components (polarization states) can be traced back to the fact that the free photon is massless: see the remark after (7.76). By contrast, for massive spin-1 bosons, such as the $\mathrm{W}^\pm$ and Z^0, all *three* expected polarization states are indeed present. However, weak interactions are described by a gauge theory, and the $\mathrm{W}^\pm$ and Z^0 particles are gauge-field quanta, analogous to the photon. How gauge invariance can be reconciled with the existence of massive gauge quanta with three polarization states will be explained in volume 2.

We may therefore write the plane-wave mode expansion for the classical $A^\mu(x)$ field in the form

$$A^\mu(x) = \int \frac{\mathrm{d}^3\boldsymbol{k}}{(2\pi)^3\sqrt{2\omega}} \sum_\lambda [\epsilon^\mu(k,\lambda)\alpha(k,\lambda)\mathrm{e}^{-\mathrm{i}k\cdot x} + \epsilon^{\mu *}(k,\lambda)\alpha^*(k,\lambda)\mathrm{e}^{\mathrm{i}k\cdot x}] \tag{7.88}$$

where the sum is over the two possible polarization states λ, for given k, as described by the suitable polarization vector $\epsilon^\mu(k,\lambda)$ and $\omega = |\boldsymbol{k}|$.

It would seem that all we have to do now, in order to 'quantize' (7.88), is to promote α and α^* to operators $\hat{\alpha}$ and $\hat{\alpha}^\dagger$, as usual. However, things are actually not nearly so simple.

7.3.2 Quantizing $A^\mu(x)$

Readers familiar with Lagrangian mechanics may already suspect that quantizing A^ν is not going to be straightforward. The problem is that, clearly, $A^\nu(x)$ has four (Lorentz) components – but, equally clearly in view of the previous section, they are not all *independent* field components or field degrees of freedom. In fact, there are only two independent degrees of freedom, both transverse. Thus there are *constraints* on the four fields, for instance the gauge condition (7.70). Constrained systems are often awkward to handle in

classical mechanics (see for example Goldstein 1980) or classical field theory; and they present major problems when it comes to canonical quantization. It is actually at just this point that the 'path-integral' approach to quantization, alluded to briefly at the end of section 5.2.2, comes into its own. This is basically because it does not involve non-commuting (or anticommuting) operators and it is therefore to that extent closer to the classical case. This means that the relatively straightforward procedures available for constrained classical mechanics systems can – when suitably generalized! – be efficiently brought to bear on the quantum problem. For an introduction to these ideas, we refer to Swanson (1992).

However, we do not wish at this stage to take what would be a very long detour, in setting up the path-integral quantization of QED. We shall continue along the 'canonical' route. To see the kind of problems we encounter, let us try and repeat for the A^ν field the 'canonical' procedure we introduced in section 5.2.5. This was based, crucially, on obtaining from the Lagrangian the momentum π conjugate to ϕ, and then imposing the commutation relation (5.117) on the corresponding operators $\hat{\pi}$ and $\hat{\phi}$. But inspection of our Maxwell Lagrangian (7.67) quickly reveals that

$$\frac{\partial \mathcal{L}_A}{\partial \dot{A}^0} = 0 \tag{7.89}$$

and hence there is *no* canonical momentum π^0 conjugate to A^0. We appear to be stymied before we can even start.

There is another problem as well. Following the procedure explained in chapter 6, we expect that the Feynman propagator for the $\hat{A}^\mu$ field, namely $\langle 0|T(\hat{A}^\mu(x_1)\hat{A}^\nu(x_2))|0\rangle$, will surely appear, describing the propagation of a photon between x_1 and x_2. In the case of real scalar fields, problem 6.3 showed that the analogous quantity was actually a Green function for the KG differential operator, $(\Box + m^2)$. It turned out, in that case, that what we really wanted was the Fourier transform of the Green function, which was essentially (apart from the tricky 'iϵ prescription' and a trivial $-$i factor) the inverse of the momentum–space operator corresponding to $(\Box + m^2)$, namely $(-k^2 + m^2)^{-1}$ (see equation (6.98) and appendix G, and also (7.58)–(7.60) for the Dirac case). Suppose, then, that we try to follow this route to obtaining the propagator for the $\hat{A}^\nu$ field. For this it is sufficient to consider the classical equations (7.68) with $j_{\rm em} = 0$, written in k space (problem 7.11(a)):

$$(-k^2 g^{\nu\mu} + k^\nu k^\mu)\tilde{A}_\mu(k) \equiv M^{\nu\mu}\tilde{A}_\mu(k) = 0 \tag{7.90}$$

where $\tilde{A}_\mu(k)$ is the Fourier transform of $A_\mu(x)$. We therefore require the inverse

$$(-k^2 g^{\nu\mu} + k^\nu k^\mu)^{-1} \equiv (M^{-1})^{\nu\mu}. \tag{7.91}$$

Unfortunately it is easy to show that this inverse does not exist. From Lorentz covariance, it has to transform as a second-rank tensor, and the only

ones available are $g^{\mu\nu}$ and $k^\mu k^\nu$. So the general form of $(M^{-1})^{\nu\mu}$ must be

$$(M^{-1})^{\nu\mu} = A(k^2)g^{\nu\mu} + B(k^2)k^\nu k^\mu. \tag{7.92}$$

Now the inverse is defined by

$$(M^{-1})^{\nu\mu}M_{\mu\sigma} = g^\nu_\sigma. \tag{7.93}$$

Putting (7.92) and (7.90) into (7.93) yields (problem 7.11(b))

$$-k^2A(k^2)g^\nu_\sigma + A(k^2)k^\nu k_\sigma = g^\nu_\sigma \tag{7.94}$$

which cannot be satisfied. So we are thwarted again.

Nothing daunted, the attentive reader may have an answer ready for the propagator problem. Suppose that, instead of (7.68), we start from the much simpler equation

$$\Box A^\nu = 0 \tag{7.95}$$

which results from *imposing* the Lorentz condition (7.70). Then, in momentum–space, (7.95) becomes

$$-k^2\tilde{A}^\nu = 0. \tag{7.96}$$

The '$-k^2$' on the left-hand side certainly has an inverse, implying that the Feynman propagator for the photon is (proportional to) $g_{\mu\nu}/k^2$. This form is indeed plausible, as it is very much what we would expect by taking the massless limit of the spin-0 propagator and tacking on $g_{\mu\nu}$ to account for the Lorentz indices in $\langle 0|T(\hat{A}_\mu(x_1)\hat{A}_\nu(x_2))|0\rangle$ (but then why no term in $k_\mu k_\nu$? – see the final two paragraphs of this section!).

Perhaps this approach helps with the 'no canonical momentum π^0' problem too. Let us ask: What Lagrangian leads to the field equation (7.95)? The answer is (problem 7.12)

$$\mathcal{L}_L = -\frac{1}{4}F_{\mu\nu}F^{\mu\nu} - \tfrac{1}{2}(\partial_\mu A^\mu)^2. \tag{7.97}$$

This form does seem to offer better prospects for quantization, since at least all our π^μ's are non-zero; in particular

$$\pi^0 = \frac{\partial\mathcal{L}}{\partial\dot{A}^0} = -\partial_\mu A^\mu. \tag{7.98}$$

The other π's are unchanged by the addition of the extra term in (7.97) and are given by

$$\pi^i = -\dot{A}^i + \partial^i A^0. \tag{7.99}$$

Interestingly, these are precisely the electric fields E^i (see (2.10)). Let us see, then, if all our problems are solved with $\mathcal{L}_L$.

Now that we have at least got four non-zero π^μ's, we can write down a plausible set of commutation relations between the corresponding operator quantities $\hat{\pi}^\mu$ and $\hat{A}^\nu$:

$$[\hat{A}_\mu(\boldsymbol{x},t),\hat{\pi}_\nu(\boldsymbol{y},t)] = \mathrm{i}g_{\mu\nu}\delta^3(\boldsymbol{x}-\boldsymbol{y}). \tag{7.100}$$

Again, the $g_{\mu\nu}$ is there to give the same Lorentz transformation character on both sides of the equation. But we must now remember that, in the classical case, our development rested on imposing the condition $\partial_\mu A^\mu = 0$ (7.70). Can we, in the quantum version we are trying to construct, simply impose $\partial_\mu \hat{A}^\mu = 0$? We certainly cannot do so in $\hat{\mathcal{L}}_L$, or we are back to $\hat{\mathcal{L}}_A$ again (besides, constraints cannot be 'substituted back' into Lagrangians, in general). Furthermore, if we set $\mu = \nu = 0$ in (7.100), then the right-hand side is non-zero while the left-hand side is zero if $\partial_\mu \hat{A}^\mu = 0 = \hat{\pi}^0$. So it is inconsistent simply to set $\partial_\mu \hat{A}^\mu = 0$.

We will return to the treatment of '$\partial_\mu \hat{A}^\mu = 0$' eventually. First, let us press on with (7.97) and see if we can get as far as a (quantized) mode expansion, of the form (7.88), for $\hat{A}^\mu(x)$.

To set this up, we need to massage the commutator (7.100) into a form as close as possible to the canonical '$[\phi,\dot{\phi}] = \mathrm{i}\delta$' form. Assuming the other commutation relations (cf (5.118))

$$[\hat{A}_\mu(\boldsymbol{x},t),\hat{A}_\nu(\boldsymbol{y},t)] = [\hat{\pi}_\mu(\boldsymbol{x},t),\hat{\pi}_\nu(\boldsymbol{y},t)] = 0 \tag{7.101}$$

we see that the spatial derivatives of the $\hat{A}$'s commute with the $\hat{A}$'s, and with each other, at equal times. This implies that we can rewrite the (quantum) $\hat{\pi}$'s as

$$\hat{\pi}_\mu = -\dot{\hat{A}}_\mu + \text{pieces that commute}. \tag{7.102}$$

Hence (7.100) can be rewritten as

$$[\hat{A}_\mu(\boldsymbol{x},t),\dot{\hat{A}}_\nu(\boldsymbol{y},t)] = -\mathrm{i}g_{\mu\nu}\delta^3(\boldsymbol{x}-\boldsymbol{y}) \tag{7.103}$$

and (7.101) remains the same. Now (7.103) is indeed very much the same as '$[\phi,\dot{\phi}] = \mathrm{i}\delta$' for the *spatial* component $\hat{A}^i$ – but the sign is wrong in the $\mu = \nu = 0$ case. We are not out of the maze yet.

Nevertheless, proceeding onwards on the basis of (7.103), we write the quantum mode expansion as (cf (7.88))

$$\hat{A}^\mu(x) = \sum_{\lambda=0}^{3}\int \frac{\mathrm{d}^3\boldsymbol{k}}{(2\pi)^3\sqrt{2\omega}}[\epsilon^\mu(k,\lambda)\hat{\alpha}_\lambda(k)\mathrm{e}^{-\mathrm{i}k\cdot x} + \epsilon^{*\mu}(k,\lambda)\hat{\alpha}^\dagger_\lambda(k)\mathrm{e}^{\mathrm{i}k\cdot x}] \tag{7.104}$$

where the sum is over *four* independent polarization states $\lambda = 0,1,2,3$, since all four fields are still in play. Before continuing, we need to say more about these ϵ's (previously, we only had two of them, now we have four and they are 4-vectors). We take $\boldsymbol{k}$ to be along the z-direction, as in our discussion of

the ϵ's in section 7.3.1, and choose two transverse polarization vectors as (cf (7.81), (7.82))

$$\begin{aligned}\epsilon^\mu(k,\lambda=1)&=(0,1,0,0)\\ \epsilon^\mu(k,\lambda=2)&=(0,0,1,0)\end{aligned}\qquad\text{'transverse polarizations'.}\tag{7.105}$$

The other two ϵ's are

$$\epsilon^\mu(k,\lambda=0)=(1,0,0,0)\qquad\text{'time-like polarization'}\tag{7.106}$$

and

$$\epsilon^\mu(k,\lambda=3)=(0,0,0,1)\qquad\text{'longitudinal polarization'.}\tag{7.107}$$

Making (7.104) consistent with (7.103) then requires

$$[\hat\alpha_\lambda(k),\hat\alpha^\dagger_{\lambda'}(k')]=-g_{\lambda\lambda'}(2\pi)^3\delta^3(\boldsymbol{k}-\boldsymbol{k}').\tag{7.108}$$

This is where the wrong sign in (7.103) has come back to haunt us: we have the wrong sign in (7.108) for the case $\lambda=\lambda'=0$ (time-like modes).

What is the consequence of this? It seems natural to assume that the vacuum is defined by

$$\hat\alpha_\lambda(k)|0\rangle=0\qquad\text{for all }\lambda=0,1,2,3.\tag{7.109}$$

But suppose we use (7.108) and (7.109) to calculate the normalization overlap of a 'one time-like photon' state; this is

$$\begin{aligned}\langle\boldsymbol{k}',\lambda=0|\boldsymbol{k},\lambda=0\rangle&=\langle0|\hat\alpha_0(\boldsymbol{k})\hat\alpha_0^\dagger(\boldsymbol{k}')|0\rangle\\&=-(2\pi)^3\delta^3(\boldsymbol{k}-\boldsymbol{k}')\end{aligned}\tag{7.110}$$

and the state effectively has a negative norm (the $\boldsymbol{k}=\boldsymbol{k}'$ infinity is the standard plane-wave artefact). Such states would threaten fundamental properties such as the conservation of total probability if they contributed, uncancelled, in physical processes.

At this point we would do well to recall the condition '$\partial_\mu\hat A^\mu=0$', which still needs to be taken into account, somehow, and it does indeed save us. Gupta (1950) and Bleuler (1950) proposed that, rather than trying (unsuccessfully) to impose it as an operator condition, one should replace it by the weaker condition

$$\partial_\mu\hat A^{\mu(+)}(x)|\Psi\rangle=0\tag{7.111}$$

where the $(+)$ signifies the positive frequency part of $\hat A$, i.e. the part involving annihilation operators, and $|\Psi\rangle$ is any physical state (including $|0\rangle$). From (7.111) and its Hermitian conjugate

$$\langle\Psi|\partial_\mu\hat A^{\mu(-)}(x)=0\tag{7.112}$$

we can deduce that the Lorentz condition (7.70) does hold for all expectation values:

$$\langle\Psi|\partial_\mu\hat{A}^\mu|\Psi\rangle = \langle\Psi|\partial_\mu\hat{A}^{\mu(+)} + \partial_\mu\hat{A}^{\mu(-)}|\Psi\rangle = 0, \tag{7.113}$$

and so the classical limit of this quantization procedure will recover the classical Maxwell theory in Lorentz gauge.

Using (7.104), (7.106) and (7.107) with $k^\mu = (|\boldsymbol{k}|, 0, 0, |\boldsymbol{k}|)$, condition (7.111) becomes

$$[\hat{\alpha}_0(k) - \hat{\alpha}_3(k)]|\Psi\rangle = 0. \tag{7.114}$$

To see the effect of this condition, consider the expression for the Hamiltonian of this theory. In normally ordered form, it turns out to be

$$\hat{H} = \int \frac{\mathrm{d}^3\boldsymbol{k}}{(2\pi)^3}(\hat{\alpha}_1^\dagger\hat{\alpha}_1 + \hat{\alpha}_2^\dagger\hat{\alpha}_2 + \hat{\alpha}_3^\dagger\hat{\alpha}_3 - \hat{\alpha}_0^\dagger\hat{\alpha}_0)\omega \tag{7.115}$$

so the contribution from the time-like modes looks dangerously negative. However, for any physical state $|\Psi\rangle$, we have

$$\begin{aligned}\langle\Psi|(\hat{\alpha}_3^\dagger\hat{\alpha}_3 - \hat{\alpha}_0^\dagger\hat{\alpha}_0)|\Psi\rangle &= \langle\Psi|(\hat{\alpha}_3^\dagger\hat{\alpha}_3 - \hat{\alpha}_3^\dagger\hat{\alpha}_0)|\Psi\rangle \\ &= \langle\Psi|\hat{\alpha}_3^\dagger(\hat{\alpha}_3 - \hat{\alpha}_0)|\Psi\rangle \\ &= 0,\end{aligned} \tag{7.116}$$

so that only the transverse modes survive.

We hope that by now the reader will have at least begun to develop a healthy respect for quantum gauge fields – and the non-Abelian versions in volume 2 are even worse! The fact is that the canonical approach has a difficult time coping with these constrained systems. Indeed, the complete Feynman rules in the non-Abelian case were found by an alternative quantization procedure ('path integral' quantization). This, however, is outside the scope of the present volume. The important points for our purposes are as follows. It is possible to carry out a consistent quantization in the Gupta–Bleuler formalism, which is the quantum version of the Maxwell theory constrained by the Lorentz condition. The propagator for the photon in this theory is

$$-\mathrm{i}g^{\mu\nu}/k^2 + \mathrm{i}\epsilon \tag{7.117}$$

which is the expected massless limit of the KG propagator as far as the spatial components are concerned (the time-like component has that negative sign).

As in all the other cases we have dealt with so far, the Feynman propagator $\langle 0|T(\hat{A}^\mu(x_1)\hat{A}^\nu(x_2))|0\rangle$ can be evaluated using the expansion (7.104) and the commutation relations (7.108). One finds that it is indeed equal to the Fourier transform of $-\mathrm{i}g^{\mu\nu}/k^2 + \mathrm{i}\epsilon$ just as asserted in (7.117). For this result, we need the 'pseudo completeness relation' (problem 7.13)

$$\begin{aligned}&-\epsilon^\mu(k,\lambda=0)\epsilon^\nu(k,\lambda=0) + \epsilon^\mu(k,\lambda=1)\epsilon^\nu(k,\lambda=1) \\ &\quad + \epsilon^\mu(k,\lambda=2)\epsilon^\nu(k,\lambda=2) + \epsilon^\mu(k,\lambda=3)\epsilon^\nu(k,\lambda=3) = -g^{\mu\nu}.\end{aligned} \tag{7.118}$$

We call this a pseudo completeness relation because of the minus sign appearing in the first term: its origin in the evaluation of this vev is precisely the 'wrong sign commutator' for the $\hat{\alpha}_0$ mode, (7.108).

Thus the gauge choice (7.70) can be made to work in quantum field theory via the condition (7.111). *But other choices are possible too.* In particular, a useful generalization of the Lagrangian (7.97) is

$$\mathcal{L}_\xi = -\frac{1}{4}F_{\mu\nu}F^{\mu\nu} - \frac{1}{2\xi}(\partial_\mu A^\mu)^2 \tag{7.119}$$

where ξ is a constant, the 'gauge parameter'. $\mathcal{L}_\xi$ leads to the equation of motion (problem 7.14)

$$\left(\Box g_{\mu\nu} - \partial_\mu\partial_\nu + \frac{1}{\xi}\partial_\mu\partial_\nu\right) A^\nu = 0. \tag{7.120}$$

In momentum–space this becomes (problem 7.14)

$$\left(-k^2 g_{\mu\nu} + k_\mu k_\nu - \frac{1}{\xi}k_\mu k_\nu\right) \tilde{A}^\nu = 0. \tag{7.121}$$

The inverse of the matrix acting on $\tilde{A}^\nu$ exists, and gives us the more general photon propagator (or Green function)

$$\boxed{\frac{\mathrm{i}[-g^{\mu\nu} + (1-\xi)k^\mu k^\nu/k^2]}{k^2 + \mathrm{i}\epsilon}} \tag{7.122}$$

as shown in problem 7.14. The previous case is recovered as $\xi \to 1$. Confusingly, the choice $\xi = 1$ is often called the 'Feynman gauge', though in classical terms it corresponds to the Lorentz gauge choice. For some purposes the 'Landau gauge' $\xi = 0$ (which is well defined in (7.122)) is convenient. In any event, it is important to be clear that *the photon propagator depends on the choice of gauge.* Formula (7.122) is the photon analogue of 'rule (ii)' in (6.103).

This may seem to imply that when we use the photon propagator (7.122) in Feynman amplitudes we will not get a definite answer, but rather one that depends on the arbitrary parameter ξ. This is a serious worry. But the propagator is not by itself a physical quantity – it is only one part of a physical amplitude. In the following chapter we shall derive the amplitudes for some simple processes in scalar and spinor electrodynamics, and one can verify that they are gauge invariant – either in the sense (for external photons) of being invariant under the replacement (7.76), or (in the case of internal photons) of being independent of ξ. It can be shown (Weinberg 1995, section 10.5) that at a given order in perturbation theory the sum of all diagrams contributing to the S-matrix is gauge invariant.

7.4 Introduction of electromagnetic interactions

After all these preliminaries, the job of introducing the first of our gauge field interactions, namely electromagnetism, into our non-interacting theory of complex scalar fields, and of Dirac fields, is very easy. From our discussion in chapter 2, we have a strong indication of how to introduce electromagnetic interactions into our theories. The 'gauge principle' in quantum mechanics consisted in elevating a global (space–time-independent) U(1) phase invariance into a local (space–time-dependent) U(1) invariance – the compensating fields being then identified with the electromagnetic ones. In quantum field theory, exactly the same principle exists and leads to the form of the electromagnetic interactions. Indeed, in the field theory formalism we have a true local U(1) phase (gauge) *in*variance of the Lagrangian (rather than a gauge *co*variance of a wave equation) and we shall be able to exhibit explicitly the symmetry current, and symmetry operator, associated with the U(1) invariance – and identify them precisely with the electromagnetic current and charge.

We have seen that for both the complex scalar and the Dirac fields the free Lagrangian is invariant under U(1) transformations (see (7.22) and (7.48)) which, we once again emphasize, are *global*. Let us therefore promote these global invariances into local ones in the way learned in chapter 2 – namely by invoking the 'gauge principle' replacement

$$\partial^\mu \to \hat{D}^\mu = \partial^\mu + \mathrm{i}q\hat{A}^\mu \tag{7.123}$$

for a particle of charge q, this time written in terms of the quantum field $\hat{A}^\mu$. In the case of the Dirac Lagrangian

$$\hat{\mathcal{L}}_{\mathrm{D}} = \bar{\hat{\psi}}(\mathrm{i}\gamma^\mu\partial_\mu - m)\hat{\psi} \tag{7.124}$$

we expect to be able to 'promote' it to one which is invariant under the *local* U(1) phase transformation[1]

$$\hat{\psi}(\boldsymbol{x},t) \to \hat{\psi}'(\boldsymbol{x},t) = \mathrm{e}^{-\mathrm{i}q\hat{\chi}(\boldsymbol{x},t)}\hat{\psi}(\boldsymbol{x},t) \tag{7.125}$$

provided we make the replacement (7.123) and demand that the (quantized) 4-vector potential transforms as (cf (2.15) with the sign change for $\hat{\chi}$)

$$\hat{A}^\mu \to \hat{A}'^\mu = \hat{A}^\mu + \partial^\mu\hat{\chi}. \tag{7.126}$$

Thus the locally U(1)-invariant Dirac Lagrangian is expected to be

$$\hat{\mathcal{L}}_{\mathrm{D\ local}} = \bar{\hat{\psi}}(\mathrm{i}\gamma^\mu\hat{D}_\mu - m)\hat{\psi}. \tag{7.127}$$

[1]Note that the classical field $\chi(\boldsymbol{x},t)$ of (2.34) has become a quantum field $\hat{\chi}(\boldsymbol{x},t)$ in (7.125); the sign change of $\hat{\chi}$ compared with χ is conventional in qft.

The invariance of (7.127) under (7.125) is easy to check, using the crucial property (2.43), which clearly carries over to the quantum field case:

$$\hat{D}'_\mu \hat{\psi}' = \mathrm{e}^{-\mathrm{i}q\hat{\chi}}(\hat{D}_\mu \hat{\psi}). \tag{7.128}$$

Equation (7.128) implies at once that

$$(\mathrm{i}\gamma^\mu \hat{D}'_\mu - m)\hat{\psi}' = \mathrm{e}^{-\mathrm{i}q\hat{\chi}}(\mathrm{i}\gamma^\mu \hat{D}_\mu - m)\hat{\psi}, \tag{7.129}$$

while taking the conjugate of (7.125) yields

$$\bar{\hat{\psi}}' = \bar{\hat{\psi}}\mathrm{e}^{\mathrm{i}q\hat{\chi}}. \tag{7.130}$$

Thus we have

$$\bar{\hat{\psi}}'(\mathrm{i}\gamma^\mu \hat{D}'_\mu - m)\hat{\psi}' = \bar{\hat{\psi}}\mathrm{e}^{\mathrm{i}q\hat{\chi}}\mathrm{e}^{-\mathrm{i}q\hat{\chi}}(\mathrm{i}\gamma^\mu \hat{D}_\mu - m)\hat{\psi} \tag{7.131}$$

$$= \bar{\hat{\psi}}(\mathrm{i}\gamma^\mu \hat{D}_\mu - m)\hat{\psi} \tag{7.132}$$

and the invariance is proved.

The Lagrangian has therefore gained an interaction term

$$\hat{\mathcal{L}}_{\mathrm{D}} \to \hat{\mathcal{L}}_{\mathrm{D\ local}} = \hat{\mathcal{L}}_{\mathrm{D}} + \hat{\mathcal{L}}_{\mathrm{int}} \tag{7.133}$$

where

$$\boxed{\hat{\mathcal{L}}_{\mathrm{int}} = -q\bar{\hat{\psi}}\gamma^\mu \hat{\psi}\hat{A}_\mu.} \tag{7.134}$$

Since the addition of $\mathcal{L}_{\mathrm{int}}$ has not changed the canonical momenta, the Hamiltonian then becomes $\hat{\mathcal{H}} = \hat{\mathcal{H}}_{\mathrm{D}} + \hat{\mathcal{H}}'_{\mathrm{D}}$, where

$$\hat{\mathcal{H}}'_{\mathrm{D}} = -\hat{\mathcal{L}}_{\mathrm{int}} = q\bar{\hat{\psi}}\gamma^\mu \hat{\psi}\hat{A}_\mu = q\hat{\psi}^\dagger \hat{\psi}\hat{A}_0 - q\hat{\psi}^\dagger \boldsymbol{\alpha}\hat{\psi} \cdot \hat{\boldsymbol{A}} \tag{7.135}$$

which is the field theory analogue of the potential in (3.102). It has the expected form '$\rho A_0 - \boldsymbol{j}\cdot\boldsymbol{A}$' if we identify the electromagnetic charge density operator with $q\hat{\psi}^\dagger\hat{\psi}$ (the charge times the number density operator) and the electromagnetic current density operator with $q\hat{\psi}^\dagger\boldsymbol{\alpha}\hat{\psi}$. The electromagnetic 4-vector current operator $\hat{j}^\mu_{\mathrm{em}}$ is thus identified as

$$\hat{j}^\mu_{\mathrm{em}} = q\bar{\hat{\psi}}\gamma^\mu \hat{\psi}, \tag{7.136}$$

which is gauge invariant and a Lorentz 4-vector. The Lagrangian (7.134) is manifestly Lorentz invariant.

We now note that $\hat{j}^\mu_{\mathrm{em}}$ is just q times the symmetry current $\hat{N}^\mu_\psi$ of section 7.2 (see equation (7.50)). Conservation of $\hat{j}^\mu_{\mathrm{em}}$ would follow from *global* U(1) invariance alone (i.e. $\hat{\chi}$ a constant in equation (7.125)); but many Lagrangians, including interactions, could be constructed obeying this global U(1) invariance. The force of the *local* U(1) invariance requirement is that it

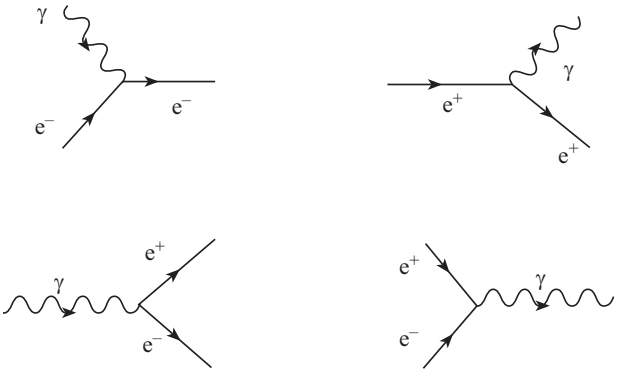


FIGURE 7.3
Possible basic '*vertices*' associated with the interaction density $e\bar{\hat{\psi}}\gamma^\mu\hat{\psi}\hat{A}_\mu$; these cannot occur as physical processes due to energy–momentum constraints.

has specified a unique form of the interaction (i.e. $\hat{\mathcal{L}}_{\rm int}$ of equation (7.134)). Indeed, this is just $-\hat{j}^\mu_{\rm em}\hat{A}_\mu$, so that in this type of theory the current $\hat{j}^\mu_{\rm em}$ is not only a symmetry current, but also determines the precise way in which the vector potential $\hat{A}^\mu$ couples to the matter field $\hat{\psi}$. Adding the Lagrangian for the $\hat{A}^\mu$ field then completes the theory of a charged fermion field interacting with the Maxwell field. In a general gauge, the $\hat{A}^\mu$ field Lagrangian is the operator form of (7.119), $\hat{\mathcal{L}}_\xi$.

The interaction term $\hat{H}'_{\rm D} = q\bar{\hat{\psi}}\gamma^\mu\hat{\psi}\hat{A}_\mu$ is a 'three-fields-at-a-point' kind of interaction just like our 3-scalar interaction $g\hat{\phi}_{\rm A}\hat{\phi}_{\rm B}\hat{\phi}_{\rm C}$ in chapter 6. We know, by now, exactly what all the operators in $\hat{H}'_{\rm D}$ are capable of: some of the possible emission and absorption processes are shown in figure 7.3. Unlike the 'ABC' model with $m_{\rm C} > m_{\rm A} + m_{\rm B}$ however, none of these elementary 'vertex' processes can occur as a real physical process, because all are forbidden by the requirement of overall 4-momentum conservation. However, they will of course contribute as virtual transitions when 'paired up' to form Feynman diagrams, such as those in figure 7.4 (compare figures 6.4 and 6.5).

It is worth remarking on the fact that the 'coupling constant' q is dimensionless, in our units. Of course, we know this from its identification with the electromagnetic charge in this case (see appendix C). But it is instructive to check it as follows. A Lagrangian density has mass dimension M^4, since the action is dimensionless (with $\hbar = 1$). Referring then to (7.33) we see that the (mass) dimension of the $\hat{\psi}$ field is $M^{3/2}$, while (7.67) shows that that of $\hat{A}^\mu$ is M. It follows that $\bar{\hat{\psi}}\gamma^\mu\hat{\psi}\hat{A}_\mu$ has mass dimension M^4, and hence q must be dimensionless.

The application of the Dyson formalism of chapter 6 to fermions interacting via $\hat{H}'_{\rm D}$ leads directly to the Feynman rules for associating precise mathemat-

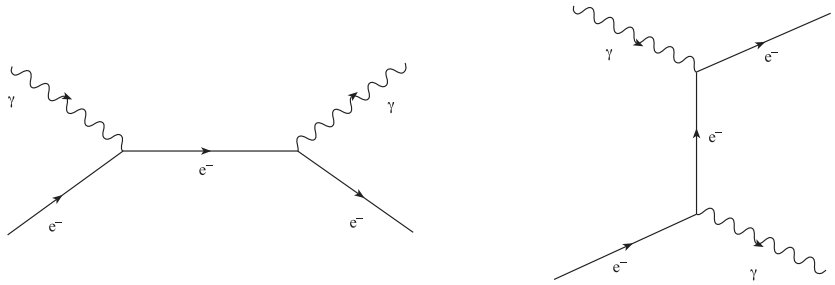


FIGURE 7.4
Lowest-order contributions to $\gamma e^- \to \gamma e^-$.

ical formulae with diagrams such as those in figure 7.4, as usual. This will be presented in the following chapter: see comment (3) in section 8.3.1 and appendix L. We may simply note here that a '$\hat{\psi}$' appears along with a '$\bar{\hat{\psi}}$' in $\hat{H}'_{\rm D}$, so that the process of 'contraction' (cf chapter 6) will lead to the form $\langle 0|T(\hat{\psi}(x_1)\bar{\hat{\psi}}(x_2))|0\rangle$ of the Dirac propagator, as stated in section 7.2.

In the same way, the global U(1) invariance (7.22) of the complex scalar field may be generalized to a local U(1) invariance incorporating electromagnetism. We have

$$\hat{\mathcal{L}}_{\rm KG} \to \hat{\mathcal{L}}_{\rm KG} + \hat{\mathcal{L}}_{\rm int} \tag{7.137}$$

where

$$\hat{\mathcal{L}}_{\rm KG} = \partial_\mu \hat{\phi}^\dagger \partial^\mu \hat{\phi} - m^2 \hat{\phi}^\dagger \hat{\phi} \tag{7.138}$$

and (under $\partial_\mu \to \hat{D}_\mu$)

$$\hat{\mathcal{L}}_{\rm int} = -{\rm i}q(\hat{\phi}^\dagger \partial^\mu \hat{\phi} - (\partial^\mu \hat{\phi}^\dagger)\hat{\phi})\hat{A}_\mu + q^2 \hat{A}^\mu \hat{A}_\mu \hat{\phi}^\dagger \hat{\phi} \tag{7.139}$$

which is the field theory analogue of the interaction in (3.100). The electromagnetic current is

$$\hat{j}^\mu_{\rm em} = -\partial \hat{\mathcal{L}}_{\rm int} / \partial \hat{A}_\mu \tag{7.140}$$

as before, which from (7.139) is

$$\hat{j}^\mu_{\rm em} = {\rm i}q(\hat{\phi}^\dagger \partial^\mu \hat{\phi} - (\partial^\mu \hat{\phi}^\dagger)\hat{\phi}) - 2q^2 \hat{A}^\mu \hat{\phi}^\dagger \hat{\phi}. \tag{7.141}$$

We note that for the boson case the electromagnetic current is *not* just q times the (number) current $\hat{N}_\phi$ appropriate to the global phase invariance. This has its origin in the fact that the boson current involves a derivative, and so the gauge invariant boson current must develop a term involving $\hat{A}^\mu$ itself, as is evident in (7.141), and as we also saw in the wavefunction case (cf equation (2.40)). The full scalar QED Lagrangian is completed by the inclusion of $\hat{\mathcal{L}}_\xi$ as before.

The application of the formalism of chapter 6 is not completely straightforward in this scalar case. The problem is that $\hat{\mathcal{L}}_{\rm int}$ of (7.139) involves derivatives of the fields and, in particular, their time derivatives. Hence the canonical momenta will be changed from their non-interacting forms. This, in turn, implies that the additional (interaction) term in the Hamiltonian is not just $-\hat{\mathcal{L}}_{\rm int}$, as in the Dirac case, but is given by (problem 7.15)

$$\hat{\mathcal{H}}'_S = -\hat{\mathcal{L}}_{\rm int} - q^2(\hat{A}^0)^2\hat{\phi}^\dagger\hat{\phi}. \tag{7.142}$$

The problem here is that the Hamiltonian and $-\hat{\mathcal{L}}_{\rm int}$ differ by a term which is non-covariant (only $\hat{A}^0$ appears).This seems to threaten the whole approach of chapter 6. Fortunately, another subtlety rescues the situation. There is a second source of non-covariance arising from the time-ordering of terms involving time derivatives, which will occur when (7.142) is used in the Dyson series (6.42). In particular, one can show (problem 7.16) that

$$\begin{aligned}&\langle 0|T(\partial_{1\mu}\hat{\phi}(x_1)\partial_{2\nu}\hat{\phi}^\dagger(x_2))|0\rangle \\ &\quad = \partial_{1\mu}\partial_{2\nu}\langle 0|T(\hat{\phi}(x_1)\hat{\phi}^\dagger(x_2))|0\rangle - {\rm i}g_{\mu 0}g_{\nu 0}\delta^4(x_1-x_2)\end{aligned} \tag{7.143}$$

which also exhibits a non-covariant piece. A careful analysis (Itzykson and Zuber 1980, section 6.1.4) shows that the two covariant effects exactly compensate, so that in the Dyson series we may use $\hat{\mathcal{H}}'_S = -\hat{\mathcal{L}}_{\rm int}$ after all. The Feynman rules for charged scalar electrodynamics are given in appendix L.

7.5 P, C and T in quantum field theory

We end this chapter by completing the discussion of the discrete symmetries which we began in section 4.2, extending it from the single particle (wavefunction) theory to quantum fields. We begin with the parity transformation.

7.5.1 Parity

The algebraic manipulations of section 4.2.1 apply equally well to the equations of motion for the quantum field, and we can take over the results by replacing a transformed wavefunction such as $\psi_{\mathbf{P}}(\boldsymbol{x},t)$ by the corresponding transformed field $\hat{\psi}_{\mathbf{P}}(\boldsymbol{x},t) = \hat{\mathbf{P}}\hat{\psi}(\boldsymbol{x},t)\hat{\mathbf{P}}^{-1}$ where $\hat{\mathbf{P}}$ is a unitary quantum field operator (which we shall not need to calculate explicitly). Thus we have

$$\hat{\phi}_{\mathbf{P}}(\boldsymbol{x},t) = \hat{\phi}(-\boldsymbol{x},t) \tag{7.144}$$

$$\hat{\psi}_{\mathbf{P}}(\boldsymbol{x},t) = \beta\hat{\psi}(-\boldsymbol{x},t), \tag{7.145}$$

for the KG and Dirac fields, and

$$\hat{\boldsymbol{A}}_{\mathbf{P}}(\boldsymbol{x},t) = -\hat{\boldsymbol{A}}(-\boldsymbol{x},t), \qquad \hat{A}^0_{\mathbf{P}}(\boldsymbol{x},t) = \hat{A}^0(-\boldsymbol{x},t) \tag{7.146}$$

for the electromagnetic fields. In (7.144) - (7.146) a simple choice of phase factor has been made.

There is however one new feature in the quantum field case, which is that the commutation or anticommutation relations must be left unchanged by the transformation, if it is to be an invariance of the theory. Evidently for **P** the only non-trivial case is the Dirac field, and it is easy to check that the anticommutation relations (7.44) and (7.45) are invariant under (7.145).

Let us see the effect of **P** on the free particle expansion (7.35). Equation (7.145) becomes

$$\begin{aligned}\hat{\psi}_{\mathbf{P}}(\boldsymbol{x},t) &= \int \frac{\mathrm{d}^3\boldsymbol{k}}{(2\pi)^3\sqrt{2\omega}} \sum_{s=1,2} [\hat{\mathbf{P}}\hat{c}_s(k)\hat{\mathbf{P}}^{-1}u(k,s)\mathrm{e}^{-\mathrm{i}\omega t+\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{x}} \\ &\qquad + \hat{\mathbf{P}}\hat{d}_s^\dagger(k)\hat{\mathbf{P}}^{-1}v(k,s)\mathrm{e}^{\mathrm{i}\omega t-\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{x}}] \\ &= \int \frac{\mathrm{d}^3\boldsymbol{k}}{(2\pi)^3\sqrt{2\omega}} \sum_{s=1,2} [\hat{c}_s(k)\beta u(k,s)\mathrm{e}^{-\mathrm{i}\omega t-\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{x}} \\ &\qquad + \hat{d}_s^\dagger(k)\beta v(k,s)\mathrm{e}^{\mathrm{i}\omega t+\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{x}}]. \qquad (7.147)\end{aligned}$$

Changing $\boldsymbol{k}$ to $-\boldsymbol{k}$ in the second integral and using the spinor properties

$$\beta u((\omega,-\boldsymbol{k}),s) = u(k,s), \qquad \beta v((\omega,-\boldsymbol{k}),s) = -v(k,s) \qquad (7.148)$$

in the right hand side of (7.147), we obtain the conditions

$$\hat{\mathbf{P}}\hat{c}_s(k)\hat{\mathbf{P}}^{-1} = \hat{c}(\omega,-\boldsymbol{k}), \qquad \hat{\mathbf{P}}\hat{d}_s^\dagger(k)\hat{\mathbf{P}}^{-1} = -\hat{d}_s^\dagger(\omega,-\boldsymbol{k}) \qquad (7.149)$$

with similar ones for $\hat{c}_s^\dagger$ and $\hat{d}_s$. Since $\hat{c}_s^\dagger$ creates a fermion from the vacuum and $\hat{d}_s^\dagger$ creates its antiparticle, it follows that a fermion and its antiparticle have opposite intrinsic parities. Similarly, equation (7.146) shows, when applied to the expansion (7.104), that a physical (transverse) photon has negative intrinsic parity.

Turning now to the electromagnetic interaction, it is clear that $\hat{j}^\mu_{\mathrm{em}}(x) = q\bar{\hat{\psi}}(x)\gamma^\mu\hat{\psi}(x)$ has exactly the same transformation properties under **P** as $\bar{\psi}\gamma^\mu\psi(x)$ had – namely $\hat{j}^0_{\mathrm{em}}(x)$ is a scalar and $\hat{\boldsymbol{j}}_{\mathrm{em}}(x)$ is a polar vector. Since this is also the way $\hat{A}^\mu$ transforms, according to (7.146), it follows that the interaction $-\hat{j}^\mu_{\mathrm{em}}\hat{A}_\mu$ is parity invariant, as we expect for QED. The scalar interaction (7.139) is also parity invariant.

7.5.2 Charge conjugation

The discussion of **C** proceeds similarly, the transformation being represented by a unitary quantum field operator $\hat{\mathbf{C}}$ such that

$$\hat{\mathbf{C}}\ \hat{\phi}\ \hat{\mathbf{C}}^{-1} = \hat{\phi}^\dagger \qquad (7.150)$$

$$\hat{\mathbf{C}}\ \hat{\psi}\ \hat{\mathbf{C}}^{-1} = \mathrm{i}\gamma^2\hat{\psi}^{\dagger\mathrm{T}} \qquad (7.151)$$

$$\hat{\mathbf{C}}\ \hat{A}^\mu\ \hat{C}^{-1} = -\hat{A}^\mu \qquad (7.152)$$

in the three cases of interest. Note that in terms of the decomposition (7.15) of the complex field $\hat{\phi}$ into the two real fields $\hat{\phi}_1$ and $\hat{\phi}_2$, (7.150) reads

$$\hat{\mathbf{C}}(\hat{\phi}_1 - \mathrm{i}\hat{\phi}_2)\hat{\mathbf{C}}^{-1} = \hat{\phi}_1 + \mathrm{i}\hat{\phi}_2. \tag{7.153}$$

The reader may check (problem 7.17(a)) that the Dirac field anticommutation relations are invariant under (7.151).

Applying (7.150) to the free field expansion (7.16), we easily find

$$\hat{\mathbf{C}}\hat{a}(k)\hat{\mathbf{C}}^{-1} = \hat{b}(k), \quad \hat{\mathbf{C}}\hat{b}^\dagger(k)\hat{\mathbf{C}}^{-1} = \hat{a}^\dagger(k), \tag{7.154}$$

so that particle and antiparticle operators are interchanged. The conditions (7.154) are of course consistent with (7.153). It follows that the normally ordered $\hat{H}$ of (7.21) is even under $\mathbf{C}$, while the normally ordered number density (7.24) is odd – the ordering being with Bose commutation relations. Carrying out the same steps for the Dirac field, and using the spinor relations (4.95) and (4.96), we obtain

$$\hat{\mathbf{C}}\hat{c}_s(k)\hat{\mathbf{C}}^{-1} = \hat{d}_s(k), \quad \hat{\mathbf{C}}\hat{d}_s^\dagger(k)\hat{\mathbf{C}}^{-1} = \hat{c}_s^\dagger(k); \tag{7.155}$$

particle and antiparticle operators are again interchanged. We particularly note that the Dirac Hamiltonian (7.55) is even under $\mathbf{C}$, while the Dirac number operator (7.54) is odd, in both cases after normal ordering with anticommutation relations (Fermi statistics). The reader may check (problem 7.17(b)) that the electromagnetic current density $q\bar{\hat{\psi}}(x)\gamma^\mu\hat{\psi}(x)$ is odd under $\mathbf{C}$, when normally ordered, and so the interaction $-\hat{j}^\mu_{\mathrm{em}}\hat{A}_\mu$ is $\mathbf{C}$-invariant. The same is true for the KG case, after normal ordering using Bose statistics.

In section 4.2.2 we introduced self-conjugate (Majorana) spinors. In extending that discussion to quantum field theory, it is again convenient to use the alternative representation (3.40) for the Dirac matrices, since we can then read off the Lorentz transformation properties from the results of section 4.1.2. Consider the 4-component Majorana field

$$\hat{\psi}_{\mathrm{M}}(x) = \begin{pmatrix} -\mathrm{i}\sigma_2\hat{\chi}^{\dagger\mathrm{T}}(x) \\ \hat{\chi}(x) \end{pmatrix}. \tag{7.156}$$

It is easy to check from (4.19) and (4.42) that the quantity $\sigma_2\chi^*(x)$ transforms like a ϕ-type spinor, and so the construction (7.156) is consistent with Lorentz covariance. The $\mathbf{C}$-conjugate field is

$$\hat{\psi}_{\mathrm{M}\mathbf{C}}(x) = \mathrm{i}\gamma^2\hat{\psi}^{\dagger\mathrm{T}}_{\mathrm{M}}(x) = \begin{pmatrix} 0 & -\mathrm{i}\sigma_2 \\ \mathrm{i}\sigma_2 & 0 \end{pmatrix}\begin{pmatrix} -\mathrm{i}\sigma_2\hat{\chi}(x) \\ \hat{\chi}^{\dagger\mathrm{T}}(x) \end{pmatrix} = \hat{\psi}_{\mathrm{M}}(x), \tag{7.157}$$

showing that it is self-conjugate. It is clear that the Majorana field has only two independent degrees of freedom – those in $\hat{\chi}(x)$ – in contrast to the Dirac field which has four (we could of course have equally well constructed a Majorana field using a ϕ-type spinor field instead of a χ-type one). The latter

corresponds physically to fermion and antifermion, spin up and down, but the Majorana fermion is the same as its antiparticle. The free field expansion corresponding to (7.35) for a Majorana field is

$$\hat{\psi}_{\rm M}(x) = \int \frac{{\rm d}^3\boldsymbol{k}}{(2\pi)^3\sqrt{2\omega}} \sum_{\lambda=1,2} [\hat{c}_\lambda(k)u(k,\lambda){\rm e}^{-{\rm i}k\cdot x} + \hat{c}^\dagger_\lambda(k)v(k,\lambda){\rm e}^{{\rm i}k\cdot x}]. \quad (7.158)$$

The Lagrangian for a free Majorana field may be taken to be $\bar{\hat{\psi}}_{\rm M}({\rm i}\partial\!\!\!/ - m)\hat{\psi}_{\rm M}$, which the reader can rewrite in terms of $\hat{\chi}$. For example, the mass term is

$$-m\bar{\hat{\psi}}_{\rm M}\hat{\psi}_{\rm M} = -m\hat{\chi}^{\rm T}{\rm i}\sigma_2\hat{\chi} + \text{Hermitian conjugate.} \quad (7.159)$$

We note that this expression will vanish unless the components $\hat{\chi}_1$ and$\hat{\chi}_2$ anticommute with each other.

7.5.3 Time reversal

In section 4.2.4 we found that the time reversal transformation for the single particle theories was not represented by a unitary operator, but rather by the product of a unitary operator and the complex conjugation operator. We can see that the same must be true in quantum field theory by considering the equation of motion (6.18) for a scalar field (for simplicity), in the interaction picture:

$$\frac{\partial\hat{\phi}(\boldsymbol{x},t)}{\partial t} = {\rm i}[\hat{H}_0, \hat{\phi}(\boldsymbol{x},t)]. \quad (7.160)$$

Suppose the field $\hat{\phi}_{\mathbf{T}}$ in the time reversed frame were related to $\hat{\phi}$ by a unitary quantum field operator $\hat{\mathbf{U}}_{\mathbf{T}}$ so that (suppressing the spatial argument) $\hat{\mathbf{U}}_{\mathbf{T}}\hat{\phi}(t)\hat{\mathbf{U}}^\dagger_{\mathbf{T}} = \hat{\phi}_{\mathbf{T}}(t')$. Then applying $\hat{\mathbf{U}}_{\mathbf{T}}\ldots\hat{\mathbf{U}}^\dagger_{\mathbf{T}}$ to equation (7.160) we would obtain

$$\frac{\partial\hat{\phi}_{\mathbf{T}}(t')}{\partial t} = {\rm i}[\hat{\mathbf{U}}_{\mathbf{T}}\hat{H}_0\hat{\mathbf{U}}^\dagger_{\mathbf{T}}, \hat{\phi}_{\mathbf{T}}(t')] \quad (7.161)$$

or equivalently

$$\frac{\partial\hat{\phi}_{\mathbf{T}}(t')}{\partial t'} = -{\rm i}[\hat{\mathbf{U}}_{\mathbf{T}}\hat{H}_0\hat{\mathbf{U}}^\dagger_{\mathbf{T}}, \hat{\phi}_{\mathbf{T}}(t')]. \quad (7.162)$$

To restore (7.162) to the form (7.160) – i.e. for covariance to hold – would require that $\hat{\mathbf{U}}_{\mathbf{T}}$ transforms $\hat{H}_0$ to $-\hat{H}_0$. But this is unacceptable on physical grounds, because the eigenvalues of $\hat{H}_0$ must be positive relative to the vacuum, both before and after the transformation. We must therefore write the transformation as

$$\hat{\mathbf{T}} = \hat{\mathbf{U}}_{\mathbf{T}}\mathbf{K} \quad (7.163)$$

where, as in section 4.2.4, $\mathbf{K}$ takes the complex conjugate of ordinary numbers and functions (i.e. it replaces i by -i). The operator $\hat{\mathbf{U}}_{\rm T}$ depends on the field involved, but we shall not need to exhibit it explicitly.

We must now decide how the fields transform under $\hat{\mathbf{T}}$. We can be guided by our work in section 4.2.4 in the single particle theory, remembering that a wavefunction is the vacuum to one particle matrix element of the corresponding quantum field operator (see Comment (5) in section 5.2.5), and also that matrix elements of operators and their time-reversed transforms are related by (4.126). In the case of the KG field, for example, let us take in (4.126) $< \psi_2 \mid =< 0|$, $\hat{O} = \hat{\phi}(x)$, and $|\psi_1 >= |\mathrm{a}; p >$ for the state of one 'a' particle with 4-momentum p. Then (4.126) gives

$$\phi(x) =< 0|\hat{\phi}(x)|\mathrm{a}; E, \boldsymbol{p} >=< 0_{\mathrm{T}}|\hat{\mathbf{T}}\hat{\phi}(x)\hat{\mathbf{T}}^{-1}|\mathrm{a}; E, -\boldsymbol{p} >^{*}, \tag{7.164}$$

where $\phi(x)$ is the free particle solution $\exp(-\mathrm{i}Et + \mathrm{i}\boldsymbol{p}\cdot\boldsymbol{x})/(2E)^{1/2}$. Now in section 4.2.4 we found the result $\phi_{\mathbf{T}}(\boldsymbol{x}, t) = \phi^{*}(\boldsymbol{x}, -t)$, for the time-reversed solution. This will be consistent with (7.164) if we take, in the quantum field case,

$$\hat{\mathbf{T}}\hat{\phi}(\boldsymbol{x}, t)\hat{\mathbf{T}}^{-1} = \hat{\phi}(\boldsymbol{x}, -t), \tag{7.165}$$

assuming that the vacuum is invariant. Applying (7.165) to the free field expansion (4.5) gives

$$\hat{\mathbf{T}}\hat{\phi}(\boldsymbol{x}, t)\hat{\mathbf{T}}^{-1} = \int \frac{\mathrm{d}^3\boldsymbol{k}}{(2\pi)^3\sqrt{2\omega}}[\hat{\mathbf{U}}_{\mathbf{T}}\hat{a}(k)\hat{\mathbf{U}}_{\mathbf{T}}^{\dagger}\mathrm{e}^{\mathrm{i}\omega t - \mathrm{i}\boldsymbol{k}\cdot\boldsymbol{x}} + \hat{\mathbf{U}}_{\mathbf{T}}\hat{b}^{\dagger}(k)\hat{\mathbf{U}}_{\mathbf{T}}^{\dagger}\mathrm{e}^{-\mathrm{i}\omega t + \mathrm{i}\boldsymbol{k}\cdot\boldsymbol{x}}] \tag{7.166}$$

$$= \hat{\phi}(\boldsymbol{x}, -t) = \int \frac{\mathrm{d}^3\boldsymbol{k}}{(2\pi)^3\sqrt{2\omega}}[\hat{a}(k)\mathrm{e}^{\mathrm{i}\omega t + \mathrm{i}\boldsymbol{k}\cdot\boldsymbol{x}} + \hat{b}^{\dagger}(k)\mathrm{e}^{-\mathrm{i}\omega t - \mathrm{i}\boldsymbol{k}\cdot\boldsymbol{x}}]. \tag{7.167}$$

Note that the plane wave functions have been complex conjugated in (7.166), because $\hat{\mathbf{T}}$ contains $\mathbf{K}$. Changing $\boldsymbol{k}$ to $-\boldsymbol{k}$ in the integral in (7.167), we obtain the conditions

$$\hat{\mathbf{U}}_{\mathbf{T}}\hat{a}(\omega, \boldsymbol{k})\hat{\mathbf{U}}_{\mathbf{T}}^{\dagger} = \hat{a}(\omega, -\boldsymbol{k}), \qquad \hat{\mathbf{U}}_{\mathbf{T}}\hat{b}^{\dagger}(\omega, \boldsymbol{k})\hat{\mathbf{U}}_{\mathbf{T}}^{\dagger} = \hat{b}^{\dagger}(\omega, -\boldsymbol{k}). \tag{7.168}$$

The transformation preserves particle and antiparticle, and reverses the 3-momentum in the creation and annihilation operators.

For the Dirac theory, we take, similarly,

$$\hat{\mathbf{T}}\hat{\psi}(\boldsymbol{x}, t)\hat{\mathbf{T}}^{-1} = \mathrm{i}\alpha_1\alpha_3\hat{\psi}(\boldsymbol{x}, -t) \tag{7.169}$$

as suggested by (4.118). The reader may check that the anticommutation relations are left invariant by (7.169). Applying (7.169) to the free field expansion (7.35), and taking the spinors to be helicity eigenstates as in section 4.2.5, we obtain the conditions

$$\hat{\mathbf{U}}_{\mathbf{T}}\hat{c}_{\lambda}(\omega, \boldsymbol{k})\hat{\mathbf{U}}_{\mathbf{T}}^{\dagger} = \hat{c}_{\lambda}(\omega, -\boldsymbol{k}), \qquad \hat{\mathbf{U}}_{\mathbf{T}}\hat{d}_{\lambda}^{\dagger}(\omega, \boldsymbol{k})\hat{\mathbf{U}}_{\mathbf{T}}^{\dagger} = \hat{d}_{\lambda}^{\dagger}(\omega, -\boldsymbol{k}). \tag{7.170}$$

Once again, the 3-momentum has been reversed in the creation and annihilation operators.

Let us check the behaviour of the current density $\hat{j}^\mu_{\rm em}(x) = q\bar{\hat{\psi}}(x)\gamma^\mu\hat{\psi}(x)$ under the transformation (7.169). Recalling that in the standard representation $i\alpha_1\alpha_3 = \Sigma_2$, we find

$$\begin{aligned}\hat{\mathbf{T}}\hat{j}^0_{\rm em}(\boldsymbol{x},t)\hat{\mathbf{T}}^{-1} &= \hat{j}^0_{\rm em}(\boldsymbol{x},-t)\\ \hat{\mathbf{T}}\hat{\boldsymbol{j}}_{\rm em}(\boldsymbol{x},t)\hat{\mathbf{T}}^{-1} &= q\hat{\psi}^\dagger(\boldsymbol{x},-t)\Sigma_2\boldsymbol{\alpha}^*\Sigma_2\hat{\psi}(\boldsymbol{x},-t) = -\hat{\boldsymbol{j}}_{\rm em}(\boldsymbol{x},-t). \quad (7.171)\end{aligned}$$

This is exactly how $A^\mu(x)$, and hence $\hat{A}^\mu(x)$, transforms, and hence the electromagnetic interaction $-\hat{j}^\mu_{\rm em}\hat{A}_\mu$ is **T**-invariant. The same is true in the KG case.

We may now proceed to look at some simple processes in scalar and spinor electrodynamics, in the following two chapters.

Problems

7.1 Verify that the Lagrangian $\hat{\mathcal{L}}$ of (7.1) is invariant (i.e. $\hat{\mathcal{L}}(\hat{\phi}_1,\hat{\phi}_2) = \hat{\mathcal{L}}(\hat{\phi}'_1,\hat{\phi}'_2)$) under the transformation (7.2) of the fields $(\hat{\phi}_1,\hat{\phi}_2) \to (\hat{\phi}'_1,\hat{\phi}'_2)$.

7.2

(a) Verify that, for $\hat{N}^\mu_\phi$ given by (7.23), the corresponding $\hat{N}_\phi$ of (7.14) reduces to the form (7.24); and that, with $\hat{H}$ given by (7.21),

$$[\hat{N}_\phi, \hat{H}] = 0.$$

(b) Verify equation (7.27).

7.3 Show that

$$[\hat{\phi}(x_1), \hat{\phi}^\dagger(x_2)] = 0 \qquad \text{for } (x_1 - x_2)^2 < 0$$

[*Hint*: insert expression (7.16) for the $\hat{\phi}$'s and use the commutation relations (7.18) to express the commutator as the difference of two integrals; in the second integral, $x_1 - x_2$ can be transformed to $-(x_1 - x_2)$ by a Lorentz transformation – the time-ordering of space-like separated events is frame-dependent!].

7.4 Verify that varying $\psi^\dagger$ in the action principle with Lagrangian (7.34) gives the Dirac equation.

7.5 Verify (7.44).

7.6 Verify equations (7.52) and (7.53).

7.7 Verify (7.62).

7.8 Verify the expression given in (7.64) for $\sum_s u(k,s)\bar{u}(k,s)$. [*Hint*: first, note that u is a four-component Dirac spinor arranged as a column, while $\bar{u}$ is another four-component spinor but this time arranged as a row because of the transpose in the † symbol. So '$u\bar{u}$' has the form

$$\begin{pmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \end{pmatrix} \begin{pmatrix} \bar{u}_1 & \bar{u}_2 & \bar{u}_3 & \bar{u}_4 \end{pmatrix} = \begin{pmatrix} u_1\bar{u}_1 & u_1\bar{u}_2 & \cdots \\ u_2\bar{u}_1 & u_2\bar{u}_2 & \cdots \\ \vdots & \vdots & \end{pmatrix}$$

and is therefore a 4×4 matrix. Use the expression (3.73) for the u's, and take

$$\phi^1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \qquad \phi^2 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}.$$

Verify that

$$\phi^1\phi^{1\dagger} + \phi^2\phi^{2\dagger} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}. \quad]$$

Similarly, verify the expression for $\sum_s v(k,s)\bar{v}(k,s)$.

7.9 Verify the result quoted in (7.63) for the Feynman propagator for the Dirac field.

7.10 Verify that if $\mathcal{L} = -\frac{1}{4}F_{\mu\nu}F^{\mu\nu} - j^\mu_{\rm em}A_\mu$, where $F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu$, the Euler–Lagrange equations for A_μ yield the Maxwell form

$$\Box A^\mu - \partial^\mu(\partial_\nu A^\nu) = j^\mu_{\rm em}.$$

[*Hint*: it is helpful to use antisymmetry of $F_{\mu\nu}$ to rewrite the '$F \cdot F$' term as $-\frac{1}{2}F_{\mu\nu}\partial^\mu A^\nu$.]

7.11

(a) Show that the Fourier transform of the free-field equation for A_μ (i.e. the one in the previous question with $j^\mu_{\rm em}$ set to zero) is given by (7.90).

(b) Verify (7.94).

7.12 Show that the equation of motion for A_μ, following from the Lagrangian $\mathcal{L}_L$ of (7.97) is

$$\Box A^\mu = 0.$$

7.13 Verify equation (7.118).

7.14 Verify equations (7.120), (7.121) and (7.122).

7.15 Verify the form (7.142) of the interaction Hamiltonian, $\mathcal{H}'_S$, in charged spin-0 electrodynamics.

7.16 Verify equation (7.143).

7.17

(a) Check that the anticommutation relations (7.44) and (7.45) are left invariant under (7.151).

(b) Check that the Dirac electromagnetic current density $\bar{\hat{\psi}}(x)\gamma^\mu\hat{\psi}(x)$ is odd under **C** when normally ordered. [*Hint*: the normally ordered current can be written as $\frac{1}{2}[\bar{\hat{\psi}}(x), \gamma^\mu\hat{\psi}(x)]$.]


Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

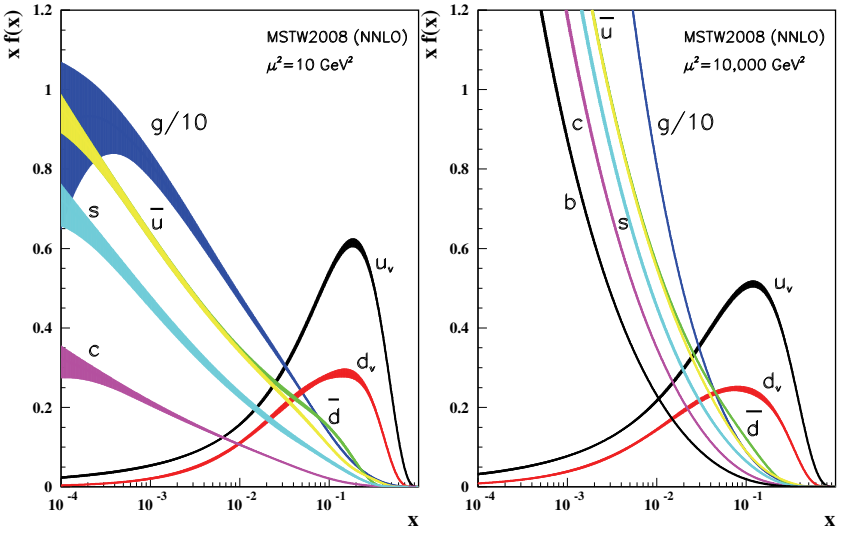


Plate I
Distributions of x times the unpolarized parton distribution functions $f(x)$ (where $f = u_V, d_V, \bar{u}, \bar{d}, s, c, b, g$) and their associated uncertainties using the MSTW2008 parametrization (Martin *et al.* 2009) at a scale $\mu^2 = 10$ GeV^2 and $\mu^2 = 10,000$ GeV^2. [Figure reproduced courtesy Michael Barnett, for the Particle Data Group, from the review of Structure Functions by B F Foster, A D Martin and M G Vincter, section 16 in the *Review of Particle Physics*, K Nakamura *et al.*(Particle Data Group) *Journal of Physics* G **37** (2010) 075021, IOP Publishing Limited.] (See figure 9.9 on page 283.)

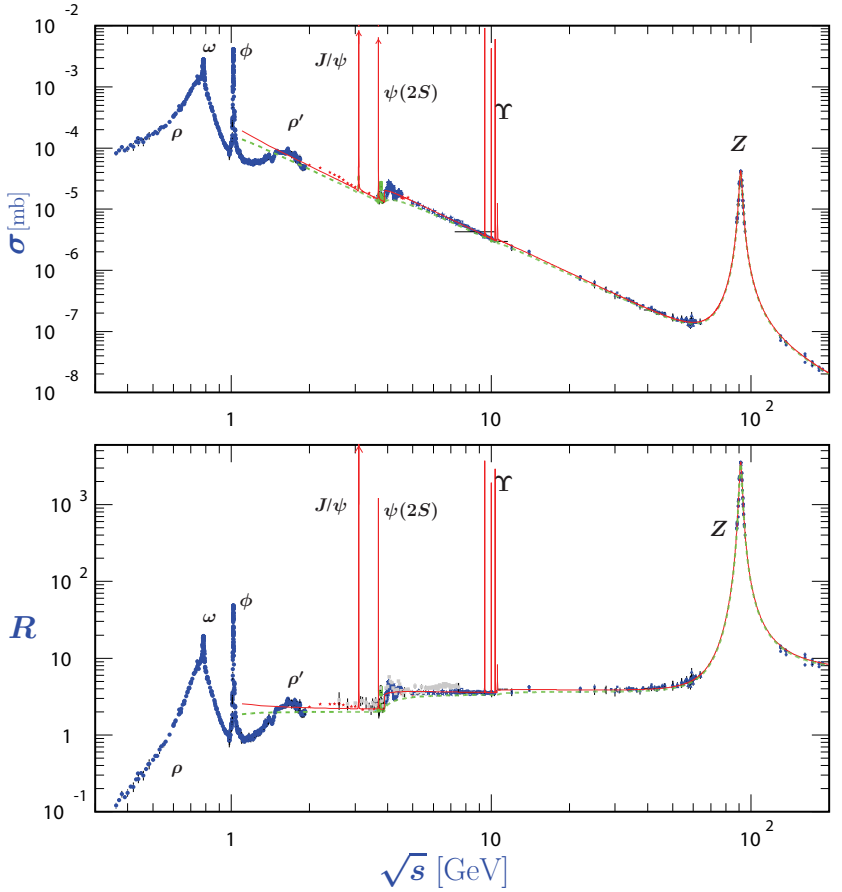


Plate II

The cross section σ for the annihilation process $e^+e^- \to$ hadrons, and the ratio R (see equation (9.100)), as a function of cm energy. [Figure reproduced courtesy Michael Barnett, for the Particle Data Group, from the *Review of Particle Physics*, K Nakamura *et al.* (Particle Data Group) *Journal of Physics* G **37** (2010) 075021 IOP Publishing Limited.] (See figure 9.16 on page 290.)

Part III

Tree-Level Applications in QED


Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

8

Elementary Processes in Scalar and Spinor Electrodynamics

8.1 Coulomb scattering of charged spin-0 particles

We begin our study of electromagnetic interactions by considering the simplest case, that of the scattering of a (hypothetical) positively charged spin-0 particle 's^+' by a fixed Coulomb potential, treated as a classical field. This will lead us to the relativistic generalization of the Rutherford formula for the cross section. We shall use this example as an exercise to gain familiarity with the quantum field-theoretic approach of chapter 6, since it can also be done straightforwardly using the 'wavefunction' approach familiar from non-relativistic quantum mechanics, when supplemented by the work of chapter 3. We shall also look at 's^-' Coulomb scattering, to test the antiparticle prescriptions of chapter 3. Incidentally, we call these scalar particles $s^\pm$ to emphasize that they are not to be identified with, for instance, the physical pions $\pi^\pm$, since the latter are composite ($q\bar{q}$) systems, and hence their interactions are more complicated than those of our hypothetical 'point-like' $s^\pm$ (as we shall see in section 8.4). No point-like charged scalar particles have been discovered, as yet.

8.1.1 Coulomb scattering of s^+ (wavefunction approach)

Consider the scattering of a spin-0 particle of charge e and mass M, the 's^+', in an electromagnetic field described by the classical potential A^μ. The process we are considering is

$$s^+(p) \to s^+(p') \tag{8.1}$$

as shown in figure 8.1, where p and p' are the initial and final 4-momenta respectively. The appropriate potential for use in the KG equation has been given in section 3.5:

$$\hat{V}_{\mathrm{KG}} = \mathrm{i}e(\partial_\mu A^\mu + A^\mu \partial_\mu) - e^2 A^2. \tag{8.2}$$

As we shall see in more detail as we go along, the parameter characterizing each order of perturbation theory based on this potential is found to be $e^2/4\pi$.

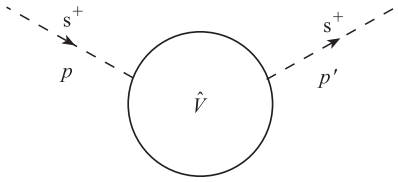


FIGURE 8.1
Coulomb scattering of s^+.

In natural units (see appendices B and C) this has the value

$$\alpha = e^2/4\pi \approx \frac{1}{137} \tag{8.3}$$

for the elementary charge e. α is called the fine structure constant. The smallness of α is the reason why a perturbation approach has been very successful for QED.

To lowest order in α we can neglect the e^2A^2 term and the perturbing potential is then

$$\hat{V} = \mathrm{i}e(\partial_\mu A^\mu + A^\mu \partial_\mu). \tag{8.4}$$

For a scattering process we shall assume[1] the same formula for the transition amplitude as in non-relativistic quantum mechanics (NRQM) time-dependent perturbation theory (see appendix A, equations (A.23) and (A.24)):

$$\mathcal{A}_{\mathrm{s}^+} = -\mathrm{i}\int \mathrm{d}^4x\, \phi'^*\hat{V}\phi \tag{8.5}$$

where ϕ and ϕ' are the initial and final state free-particle solutions. The latter are (recall equation (3.11))

$$\phi = N\mathrm{e}^{-\mathrm{i}p\cdot x} \tag{8.6}$$

$$\phi' = N'\mathrm{e}^{-\mathrm{i}p'\cdot x} \tag{8.7}$$

and we shall fix the normalization factors later. Inserting the expression for $\hat{V}$ into (8.5), and doing some integration by parts (problem 8.1), we obtain

$$\mathcal{A}_{\mathrm{s}^+} = -\mathrm{i}\int \mathrm{d}^4x\, \{\mathrm{i}e[\phi'^*(\partial_\mu\phi) - (\partial_\mu\phi'^*)\phi]\}A^\mu. \tag{8.8}$$

The expression inside the braces is very reminiscent of the probability current expression (3.20). Indeed we can write (8.8) as

$$\mathcal{A}_{\mathrm{s}^+} = -\mathrm{i}\int \mathrm{d}^4x\, j^\mu_{\mathrm{em},\mathrm{s}^+}(x)A_\mu(x) \tag{8.9}$$

[1] Justification may be found in chapter 9 of Bjorken and Drell (1964).

where

$$j^{\mu}_{\mathrm{em,s^+}}(x) = \mathrm{i}e(\phi'^{*}\partial^{\mu}\phi - (\partial^{\mu}\phi'^{*})\phi) \tag{8.10}$$

can be regarded as an electromagnetic 'transition current', analogous to the simple probability current for a single state. In the following section we shall see the exact meaning of this idea, using quantum field theory. Meanwhile, we insert the plane-wave free-particle solutions (8.6) and (8.7) for ϕ and ϕ' into (8.10) to obtain

$$j^{\mu}_{\mathrm{em,s^+}}(x) = NN'e(p+p')^{\mu}\mathrm{e}^{-\mathrm{i}(p-p')\cdot x} \tag{8.11}$$

so that (8.9) becomes

$$\mathcal{A}_{\mathrm{s^+}} = -\mathrm{i}NN'\int \mathrm{d}^4x\, e(p+p')_{\mu}\mathrm{e}^{-\mathrm{i}(p-p')\cdot x}A^{\mu}(x). \tag{8.12}$$

In the case of Coulomb scattering from a static point charge Ze ($e>0$), the vector potential A^{μ} is given by

$$A^0 = \frac{Ze}{4\pi|\boldsymbol{x}|} \qquad \boldsymbol{A} = 0. \tag{8.13}$$

Inserting (8.13) into (8.12) we obtain

$$\mathcal{A}_{\mathrm{s^+}} = -\mathrm{i}NN'Ze^2(E+E')\int \mathrm{e}^{-\mathrm{i}(E-E')t}\,\mathrm{d}t\int \frac{\mathrm{e}^{\mathrm{i}(\boldsymbol{p}-\boldsymbol{p}')\cdot\boldsymbol{x}}}{4\pi|\boldsymbol{x}|}\,\mathrm{d}^3\boldsymbol{x}. \tag{8.14}$$

The initial and final 4-momenta are

$$p = (E,\boldsymbol{p}) \qquad p' = (E',\boldsymbol{p}')$$

with $E = \sqrt{M^2+\boldsymbol{p}^2}, E' = \sqrt{M^2+\boldsymbol{p}'^2}$. The first (time) integral in (8.14) gives an energy-conserving δ-function $2\pi\delta(E-E')$ (see appendix E), as is expected for a static (non-recoiling) scattering centre. The second (spatial) integral is the Fourier transform of $1/4\pi|\boldsymbol{x}|$, which can be obtained from (1.13), (1.26) and (1.27) by setting $m_{\mathrm{U}}=0$; the result is $1/\boldsymbol{q}^2$ where $\boldsymbol{q}=\boldsymbol{p}-\boldsymbol{p}'$. Hence

$$\mathcal{A}_{\mathrm{s^+}} = -\mathrm{i}NN'2\pi\delta(E-E')\frac{Ze^2}{\boldsymbol{q}^2}2E \tag{8.15}$$

$$\equiv -\mathrm{i}(2\pi)\delta(E-E')V_{\mathrm{s^+}} \qquad \text{(cf equation (A.25))} \tag{8.16}$$

where in (8.15) we have used $E=E'$ in the matrix element. This is in the standard form met in time-dependent perturbation theory (cf equations (A.25) and (A.26)).

The transition probability per unit time is then (appendix H, equation (H.18))

$$\dot{P}_{\mathrm{s^+}} = 2\pi|V_{\mathrm{s^+}}|^2\rho(E') \tag{8.17}$$

where $\rho(E')$ is the density of final states per energy interval $\mathrm{d}E'$. This will depend on the normalization adopted for ϕ, ϕ' via the factors N, N'. We choose these to be unity, which means that we are adopting the 'covariant' normalization of $2E$ particles per unit volume. Then (cf equation (H.22))

$$\rho(E')\,\mathrm{d}E' = \frac{|\boldsymbol{p}'|^2}{(2\pi)^3}\frac{\mathrm{d}|\boldsymbol{p}'|}{2E'}\,\mathrm{d}\Omega. \tag{8.18}$$

Using $E' = (M^2 + \boldsymbol{p}'^2)^{1/2}$ one easily finds

$$\rho(E') = \frac{|\boldsymbol{p}'|\,\mathrm{d}\Omega}{16\pi^3}. \tag{8.19}$$

Note that this differs from equation (H.22) since here we are using relativistic kinematics.

To obtain the cross section, we need to divide $\dot{P}_{\mathrm{s}^+}$ by the incident flux, which is $2|\boldsymbol{p}|$ in our normalization. Hence

$$\mathrm{d}\sigma = (4Z^2e^4E^2/16\pi^2\boldsymbol{q}^4)\,\mathrm{d}\Omega. \tag{8.20}$$

Finally, since $\boldsymbol{q}^2 = (\boldsymbol{p} - \boldsymbol{p}')^2 = 4|\boldsymbol{p}|^2\sin^2\theta/2$ (cf section 1.3.4) where θ is the angle between $\boldsymbol{p}$ and $\boldsymbol{p}'$, we obtain

$$\boxed{\frac{\mathrm{d}\sigma}{\mathrm{d}\Omega} = (Z\alpha)^2\frac{E^2}{4|\boldsymbol{p}|^4}\frac{1}{\sin^4\theta/2}.} \tag{8.21}$$

This is the Rutherford formula with relativistic kinematics, showing the characteristic $\sin^{-4}\theta/2$ angular dependence (cf figure 1.8). This deservedly famous formula will serve as a 'reference point' for all the subsequent calculations in this chapter, as we proceed to add in various complications, such as spin, recoil and structure. The non-relativistic form may be retrieved by replacing E by M.

8.1.2 Coulomb scattering of s^+ (field-theoretic approach)

We follow steps closely similar to those in section 6.3.1, making use of the result quoted in section 7.4, that the appropriate interaction Hamiltonian for use in the Dyson series (6.42) is $\hat{\mathcal{H}}'_{\mathrm{s}} = -\hat{\mathcal{L}}_{\mathrm{int}}$ where $\hat{\mathcal{L}}_{\mathrm{int}}$ is given by (7.139), with $q = e$. As in the step from (8.2) to (8.4) we discard the e^2 term to first order and use

$$\hat{\mathcal{H}}'_{\mathrm{s}}(x) = \mathrm{i}e(\hat{\phi}^\dagger(x)\partial^\mu\hat{\phi}(x) - (\partial^\mu\hat{\phi}^\dagger(x))\hat{\phi}(x))A_\mu(x). \tag{8.22}$$

Equation (8.22) can be written as $\hat{j}^\mu_{\mathrm{em,s}}A_\mu$ where

$$\hat{j}^\mu_{\mathrm{em,s}} = \mathrm{i}e(\hat{\phi}^\dagger\partial^\mu\hat{\phi} - (\partial^\mu\hat{\phi}^\dagger)\hat{\phi}). \tag{8.23}$$

Note that the field A_μ is *not* quantized: it is being treated as an 'external' classical potential. The expansion for the field $\hat{\phi}$ is given in (7.16). As in (6.48), the lowest-order amplitude is

$$\mathcal{A}_{\mathrm{s}^+} = -\mathrm{i}\langle \mathrm{s}^+, p'| \int \mathrm{d}^4x\, \hat{H}'_{\mathrm{s}}(x)|\mathrm{s}^+, p\rangle \tag{8.24}$$

where (cf (6.49))

$$|\mathrm{s}^+, p\rangle = \sqrt{2E}\hat{a}^\dagger(p)|0\rangle. \tag{8.25}$$

We are, of course, anticipating in our notation that (8.24) will indeed be the same as (8.12). The required amplitude is then

$$\mathcal{A}_{\mathrm{s}^+} = -\mathrm{i}\int \mathrm{d}^4x\, \langle \mathrm{s}^+, p'|\hat{j}^\mu_{\mathrm{em,s}}(x)|\mathrm{s}^+, p\rangle A_\mu(x). \tag{8.26}$$

Using the expansion (7.16), the definition (8.25) and the vacuum conditions (7.30), and following the method of section 6.3.1, it is a good exercise to check that the value of the matrix element in (8.26) is (problem 8.2)

$$\langle \mathrm{s}^+, p'|\hat{j}^\mu_{\mathrm{em,s}}(x)|\mathrm{s}^+, p\rangle = e(p+p')^\mu \mathrm{e}^{-\mathrm{i}(p-p')\cdot x}. \tag{8.27}$$

This is exactly the same as the expression we obtained in (8.11) for the wave mechanical transition current in this case, using the normalization $N = N' = 1$, which is consistent with the field-theoretic normalization in (8.25). Thus our *wave mechanical transition current is indeed the matrix element of the field-theoretical electromagnetic current operator*:

$$j^\mu_{\mathrm{em,s}^+}(x) = \langle \mathrm{s}^+, p'|\hat{j}^\mu_{\mathrm{em,s}}(x)|\mathrm{s}^+, p\rangle. \tag{8.28}$$

Combining all these results, we have therefore connected the 'wavefunction' amplitude and the 'field-theory' amplitude via

$$\begin{aligned} \mathcal{A}_{\mathrm{s}^+} &= -\mathrm{i}\int \mathrm{d}^4x\, j^\mu_{\mathrm{em,s}^+}(x)A_\mu(x) \\ &= -\mathrm{i}\int \mathrm{d}^4x\, \langle \mathrm{s}^+, p'|\hat{j}^\mu_{\mathrm{em,s}}(x)|\mathrm{s}^+, p\rangle A_\mu(x). \end{aligned} \tag{8.29}$$

We note that because of the static nature of the potential, and the non-covariant choice of A^μ (only $A^0 \neq 0$), our answer in either case cannot be expected to yield a Lorentz invariant amplitude.

8.1.3 Coulomb scattering of s^-

The physical process is (figure 8.2(a))

$$\mathrm{s}^-(p) \to \mathrm{s}^-(p') \tag{8.30}$$

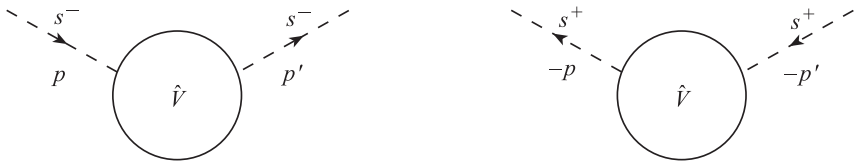


FIGURE 8.2
Coulomb scattering of s^-: (a) the physical process with antiparticles of positive 4-momentum, and (b) the related unphysical process with particles of negative 4-momentum, using the Feynman prescription.

where, of course, E and E' are both positive ($E = (M^2 + \boldsymbol{p}^2)^{1/2}$ and similarly for E'). Since the charge on the antiparticle s^- is $-e$, the amplitude for this process can, in fact, be immediately obtained from (8.12) by merely changing the sign of e. Because of the way e and the 4-momenta p and p' enter (8.12), however, this in turn is the same as letting $p \to -p'$ and $p' \to -p$: this changes the sign of the '$e(p+p')_\mu$' part as required, and leaves the exponential unchanged. Hence we see in action here (admittedly in a very simple example) the Feynman interpretation of the negative 4-momentum solutions, described in section 3.4.4: the amplitude for $s^-(p) \to s^-(p')$ is the same as the amplitude for $s^+(-p') \to s^+(-p)$. The latter process is shown in figure 8.2(b).

The same conclusion can be *derived* from the field-theory formalism. In this case we need to evaluate the matrix element

$$\langle s^-, p' | \hat{j}^\mu_{\mathrm{em,s}}(x) | s^-, p \rangle, \tag{8.31}$$

where the *same* $\hat{j}_{\mathrm{em,s}}$ of equation (8.23) enters: $\hat{\phi}$ of (7.16) contains the antiparticle operator too! It is again a good exercise to check, using

$$|s^-, p\rangle = \sqrt{2E}\, \hat{b}^\dagger(p)|0\rangle \tag{8.32}$$

and remembering to *normally order* the operators in $\hat{j}^\mu_{\mathrm{em,s}}$, that (8.31) is given by the expected result, namely, (8.27) with $e \to -e$ (problem 8.3).

Since the matrix elements only differ by a sign, the cross sections for s^+ and s^- Coulomb scattering will be the same to this (lowest) order in α.

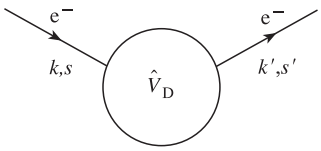


FIGURE 8.3
Coulomb scattering of e^-.

8.2 Coulomb scattering of charged spin-$\frac{1}{2}$ particles

8.2.1 Coulomb scattering of e^- (wavefunction approach)

We shall call the particle an electron, of charge $-e(e>0)$ and mass m; note that by convention it is the negatively charged fermion that is the 'particle', but the positively charged boson. The process we are considering is (figure 8.3)

$$e^-(k,s) \to e^-(k',s') \tag{8.33}$$

where k, s are the 4-momentum and spin of the incident e^-, and similarly for k', s', with $k=(E,\boldsymbol{k})$ and $E=(m^2+\boldsymbol{k}^2)^{1/2}$ and similarly for k'.

The appropriate potential to use in the Dirac equation has been given in section 3.5:

$$\hat{V}_{\rm D} = -eA^0\mathbf{1} + e\boldsymbol{\alpha}\cdot\boldsymbol{A} = -e\begin{pmatrix} A^0 & \boldsymbol{\sigma}\cdot\boldsymbol{A} \\ \boldsymbol{\sigma}\cdot\boldsymbol{A} & A^0 \end{pmatrix} \tag{8.34}$$

for a particle of charge $-e$. This potential is a 4×4 matrix and to obtain an amplitude in the form of a single complex number, we must use $\psi^\dagger$ instead of ψ^* in the matrix element. The first-order amplitude (figure 8.3) is therefore

$$\mathcal{A}_{e^-} = -\mathrm{i}\int \mathrm{d}^4x\, \psi^\dagger(k',s')\hat{V}_{\rm D}\psi(k,s) \tag{8.35}$$

where s and s' label the spin components. The spin labels are necessary since the spin configuration may be changed by the interaction. In (8.35), ψ and ψ' are free-particle positive-energy solutions of the Dirac equation, as in (3.74), with u given by equation (3.73) and normalized to $u^\dagger u = 2E$, $E=(m^2+\boldsymbol{k}^2)^{1/2}$.

The Lorentz properties of (8.35) become much clearer if we use the γ-matrix notation of problem 4.3. For convenience we re-state the definitions here:

$$\gamma^0 = \beta \qquad (\gamma^0)^2 = \mathbf{1} \tag{8.36}$$

$$\gamma^i = \beta\alpha_i \qquad (\gamma^i)^2 = -\mathbf{1} \qquad i=1,2,3. \tag{8.37}$$

The Dirac equation may then be written (problem 4.3) as

$$(\mathrm{i}\partial\!\!\!/ - m)\psi = 0 \tag{8.38}$$

where the 'slash' notation introduced in (7.59) has been used ($\mathrm{i}\partial\!\!\!/ = \mathrm{i}\gamma^\mu\partial_\mu$). Defining $\bar{\psi} = \psi^\dagger\gamma^0$, (8.35) becomes

$$\mathcal{A}_{\mathrm{e}^-} = -\mathrm{i}\int \mathrm{d}^4x\,(-e\bar{\psi}'(x)\gamma^\mu\psi(x))A_\mu(x) \tag{8.39}$$

$$\equiv -\mathrm{i}\int \mathrm{d}^4x\, j^\mu_{\mathrm{em},\mathrm{e}^-}(x)A_\mu(x) \tag{8.40}$$

where we have defined an electromagnetic transition current for a negatively charged fermion:

$$j^\mu_{\mathrm{em},\mathrm{e}^-}(x) = -e\bar{\psi}'(x)\gamma^\mu\psi(x), \tag{8.41}$$

exactly analogous to the one for a positively charged boson introduced in section 8.1.1. We know from section 4.1.2 that $\bar{\psi}'\gamma^\mu\psi$ is a 4-vector, showing that $\mathcal{A}_{\mathrm{e}^-}$ of (8.40) is Lorentz invariant.

Inserting free-particle solutions for ψ and $\psi'^\dagger$ in (8.41), we obtain

$$j^\mu_{\mathrm{em},\mathrm{e}^-}(x) = -e\bar{u}(k',s')\gamma^\mu u(k,s)\mathrm{e}^{-\mathrm{i}(k-k')\cdot x} \tag{8.42}$$

so that (8.39) becomes

$$\mathcal{A}_{\mathrm{e}^-} = -\mathrm{i}\int \mathrm{d}^4x\,(-e\bar{u}'\gamma^\mu u\mathrm{e}^{-\mathrm{i}(k-k')\cdot x})A_\mu(x) \tag{8.43}$$

where $u = u(k,s)$ and similarly for u'. Note that the u's do not depend on x. For the case of the Coulomb potential in equation (8.13), $\mathcal{A}_{\mathrm{e}^-}$ becomes

$$\mathcal{A}_{\mathrm{e}^-} = \mathrm{i}2\pi\delta(E-E')\frac{Ze^2}{\boldsymbol{q}^2}u'^\dagger u \tag{8.44}$$

just as in (8.15), where $\boldsymbol{q} = \boldsymbol{k} - \boldsymbol{k}'$ and we have used $\bar{u}'\gamma^0 = u'^\dagger$. Comparing (8.44) with (8.15), we see that (using the covariant normalization $N = N' = 1$) the amplitude in the spinor case is obtained from that for the scalar case by the replacement '$2E \to u'^\dagger u$' and the sign of the amplitude is reversed as expected for e^- rather than s^+ scattering.

We now have to understand how to define the cross section for particles with spin and then how to calculate it. Clearly the cross section is proportional to $|\mathcal{A}_{\mathrm{e}^-}|^2$, which involves $|u^\dagger(k',s')u(k,s)|^2$ here. Usually the incident beam is *unpolarized*, which means that it is a random mixture of both spin states s ('up' or 'down'). It is important to note that this is an *incoherent* average, in the sense that we average the *cross section* rather than the amplitude. Furthermore, most experiments usually measure only the direction and energy

of the scattered electron and are not sensitive to the spin state s'. Thus what we wish to calculate, in this case, is the unpolarized cross section defined by

$$\begin{aligned} \mathrm{d}\bar{\sigma} &\equiv \tfrac{1}{2}(\mathrm{d}\sigma_{\uparrow\uparrow} + \mathrm{d}\sigma_{\uparrow\downarrow} + \mathrm{d}\sigma_{\downarrow\uparrow} + \mathrm{d}\sigma_{\downarrow\downarrow}) \\ &= \tfrac{1}{2}\sum_{s',s} \mathrm{d}\sigma_{s's} \end{aligned} \tag{8.45}$$

where $\mathrm{d}\sigma_{s',s} \propto |u^\dagger(k',s')u(k,s)|^2$. In (8.45), we are averaging over the two possible initial spin polarizations and summing over the final spin states arising from each initial spin state.

It is possible to calculate the quantity

$$S = \tfrac{1}{2}\sum_{s',s} |u'^\dagger u|^2 \tag{8.46}$$

by brute force, using (3.73) and taking the two-component spinors to be, say,

$$\phi^1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \qquad \phi^2 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}. \tag{8.47}$$

One finds (problem 8.4)

$$S = (2E)^2(1 - v^2 \sin^2\theta/2) \tag{8.48}$$

where $v = |\boldsymbol{k}|/E$ is the particle's speed and θ is the scattering angle. If we now recall that (i) the matrix element (8.44) can be obtained from (8.15) by the replacement '$2E \to u'^\dagger u$' and (ii) the normalization of our spinor states is the same ('$\rho = 2E$') as in the scalar case, so that the flux and density of states factors are unchanged, we may infer from (8.21) that

$$\boxed{\frac{\mathrm{d}\bar{\sigma}}{\mathrm{d}\Omega} = (Z\alpha)^2 \frac{E^2}{4|\boldsymbol{k}|^4} \frac{(1 - v^2\sin^2\theta/2)}{\sin^4\theta/2}.} \tag{8.49}$$

This is the Mott cross section (Mott 1929). Comparing this with the basic Rutherford formula (8.21), we see that the factor $(1 - v^2\sin^2\theta/2)$ (which comes from the spin summation) represents the effect of replacing spin-0 scattering particles by spin-$\frac{1}{2}$ ones.

Indeed, this factor has an important physical interpretation. Consider the extreme relativistic limit ($v \to 1, m \to 0$), when the factor becomes $\cos^2\theta/2$, which vanishes in the backward direction $\theta = \pi$. This may be understood as follows. In the $m \to 0$ limit, it is appropriate to use the representation (3.40) of the Dirac matrices and, in this case equations (4.14) and (4.15) show that the Dirac spinor takes the form

$$u = \begin{pmatrix} u_{\mathrm{R}} \\ u_{\mathrm{L}} \end{pmatrix} \tag{8.50}$$

where $u_{\rm R}$ and $u_{\rm L}$ have positive and negative helicity respectively. The spinor part of the matrix element (8.44) then becomes ${u'_{\rm R}}^\dagger u_{\rm R} + {u'_{\rm L}}^\dagger u_{\rm L}$, from which it is clear that *helicity is conserved*: the helicity of the u' spinors equals that of the u spinors; in particular there are no helicity mixing terms of the form ${u'_{\rm R}}^\dagger u_{\rm L}$ or ${u'_{\rm L}}^\dagger u_{\rm R}$. Consider then an initial state electron with positive helicity, and take the z-axis to be along the incident momentum. The z-component of angular momentum is then $+\frac{1}{2}$. Suppose the electron is scattered through an angle of π. Since helicity is conserved, the scattered electron's helicity will still be positive, but since the direction of its momentum has been reversed, its angular momentum along the original axis will be $-\frac{1}{2}$. Hence this configuration is forbidden by angular momentum conservation – and similarly for an incoming negative helicity state. The spin labels s', s in (8.46) can be taken to be helicity labels and so it follows that the quantity S must vanish for $\theta = \pi$ in the $m \to 0$ limit. The 'R' and 'L' states are mixed by a mass term in the Dirac equation (see (4.14) and (4.15)) and hence we expect backward scattering to be increasingly allowed as m/E increases (recall that $v = (1 - m^2/E^2)^{1/2}$ so that $1 - v^2 \sin^2\theta/2 = \cos^2\theta/2 + (m^2/E^2)\sin^2\theta/2$).

8.2.2 Coulomb scattering of e^- (field-theoretic approach)

Once again, the interaction Hamiltonian has been given in section 7.4, namely

$$\hat{H}'_{\rm D} = -e\bar{\hat{\psi}}\gamma^\mu\hat{\psi}A_\mu \equiv \hat{j}^\mu_{\rm em,e}A_\mu \tag{8.51}$$

where the current operator $\hat{j}^\mu_{\rm em,e}$ is just $-e\bar{\hat{\psi}}\gamma^\mu\hat{\psi}$ in this case. The lowest-order amplitude is then

$$\mathcal{A}_{\rm e^-} = -{\rm i}\langle {\rm e}^-, k', s'| \int {\rm d}^4x\, \hat{H}'_{\rm D}(x)|{\rm e}^-, k, s\rangle \tag{8.52}$$

$$= -{\rm i}\int {\rm d}^4x\, \langle {\rm e}^-, k', s'|\hat{j}^\mu_{\rm em,e}(x)|{\rm e}^-, k, s\rangle A_\mu(x). \tag{8.53}$$

With our normalization, and referring to the fermionic expansion (7.35), the states are defined by

$$|{\rm e}^-, k, s\rangle = \sqrt{2E}\hat{c}^\dagger_s(k)|0\rangle \tag{8.54}$$

and similarly for the final state. We then find (problem 8.5) that the current matrix element in (8.53) takes the form

$$\langle {\rm e}^-, k', s'|\hat{j}^\mu_{\rm em,e}(x)|{\rm e}^-, k, s\rangle = -e\bar{u}'\gamma^\mu u{\rm e}^{-{\rm i}(k-k')\cdot x} = j^\mu_{\rm em,e^-}(x) \tag{8.55}$$

exactly as in (8.42). Thus once again, the 'wavefunction' and 'field-theoretic' approaches have been shown to be equivalent, in a simple case.

8.2.3 Trace techniques for spin summations

The calculation of cross sections involving fermions rapidly becomes laborious following the 'brute force' method of section 8.2.1, in which the explicit forms

for u and $u'^{\dagger}$ were used. Fortunately we can avoid this by using a powerful labour-saving device due to Feynman, in which the γ's come into their own.

We need to calculate the quantity S given in (8.46). This will turn out to be just the first in a series of such objects. With later needs in mind, we shall here calculate a more general quantity than (8.46), namely the *lepton tensor*

$$L^{\mu\nu}(k',k) = \tfrac{1}{2}\sum_{s',s} \bar{u}(k',s')\gamma^{\mu}u(k,s)[\bar{u}(k',s')\gamma^{\nu}u(k,s)]^{*} \tag{8.56}$$

$$= \frac{1}{2e^2}\sum_{s',s}\langle \mathrm{e}^-,k',s'|\hat{j}^{\mu}_{\mathrm{em,e}}(0)|\mathrm{e}^-,k,s\rangle\langle \mathrm{e}^-,k',s'|\hat{j}^{\nu}_{\mathrm{em,e}}(0)|\mathrm{e}^-,k,s\rangle^{*}. \tag{8.57}$$

Clearly this will be relevant to the more general case in which A^{μ} contains non-zero spatial components, for example. For our present application, we shall need only L^{00}.

We first note that $L^{\mu\nu}$ is correctly called a *tensor* (a contravariant second-rank one, in fact – see appendix D), because the two '$\bar{u}\gamma^{\mu}u$, $\bar{u}\gamma^{\nu}u$' factors are each 4-vectors, as we have seen. (We might worry a little over the complex conjugation of the second factor, but this will disappear after the next step.) Consider therefore the factor $[\bar{u}(k',s')\gamma^{\nu}u(k,s)]^{*}$. For each value of the index ν, this is just a number (the corresponding component of the 4-vector), and so it can make no difference if we take its transpose, in a matrix sense (the transpose of a 1×1 matrix is certainly equal to itself!). In that case the complex conjugate becomes the Hermitian conjugate, which is:

$$[\bar{u}(k',s')\gamma^{\nu}u(k,s)]^{\dagger} = u^{\dagger}(k,s)\gamma^{\nu\dagger}\gamma^{0\dagger}u(k',s') \tag{8.58}$$

$$= \bar{u}(k,s)\gamma^{\nu}u(k',s') \tag{8.59}$$

since (problem 8.6)

$$\gamma^0\gamma^{\nu\dagger}\gamma^0 = \gamma^{\nu} \tag{8.60}$$

and $\gamma^0 = \gamma^{0\dagger}$. Thus $L^{\mu\nu}$ may be written in the more streamlined form

$$L^{\mu\nu} = \tfrac{1}{2}\sum_{s',s}\bar{u}(k',s')\gamma^{\mu}u(k,s)\bar{u}(k,s)\gamma^{\nu}u(k',s') \tag{8.61}$$

which is, moreover, evidently the (tensor) product of two 4-vectors. However, there is more to this than saving a few symbols. We have seen the expression

$$\sum_{s} u(k,s)\bar{u}(k,s) \tag{8.62}$$

before! (See (7.64) and problem 7.8.) Thus we can replace the sum (8.62) over spin states 's' by the corresponding matrix $(\not{k}+m)$:

$$L^{\mu\nu} = \tfrac{1}{2}\sum_{s'}\bar{u}_{\alpha}(k',s')(\gamma^{\mu})_{\alpha\beta}(\not{k}+m)_{\beta\gamma}(\gamma^{\nu})_{\gamma\delta}u_{\delta}(k',s') \tag{8.63}$$

where we have made the matrix indices explicit, and *summation on all repeated matrix indices is understood*. In particular, note that every matrix index is repeated, so that each one is in fact summed over: there are no 'spare' indices. Now, since we can reorder matrix elements as we wish, we can bring the u_δ to the front of the expression, and use the same trick to perform the second spin sum:

$$\sum_{s'} u_\delta(k',s')\bar{u}_\alpha(k',s') = (\not{k}' + m)_{\delta\alpha}. \tag{8.64}$$

Thus $L^{\mu\nu}$ takes the form of a matrix product, summed over the diagonal elements:

$$\begin{aligned} L^{\mu\nu} &= \tfrac{1}{2}(\not{k}' + m)_{\delta\alpha}(\gamma^\mu)_{\alpha\beta}(\not{k} + m)_{\beta\gamma}(\gamma^\nu)_{\gamma\delta} && (8.65)\\ &= \tfrac{1}{2}\sum_\delta [(\not{k}' + m)\gamma^\mu(\not{k} + m)\gamma^\nu]_{\delta\delta} && (8.66) \end{aligned}$$

where we have explicitly reinstated the sum over δ. The right-hand side of (8.66) is the *trace* (i.e. the sum of the diagonal elements) of the matrix formed by the product of the four indicated matrices:

$$L^{\mu\nu} = \tfrac{1}{2}\mathrm{Tr}[(\not{k}' + m)\gamma^\mu(\not{k} + m)\gamma^\nu]. \tag{8.67}$$

Such matrix traces have some useful properties which we now list. Denote the trace of a matrix $\mathbf{A}$ by

$$\mathrm{Tr}\mathbf{A} = \sum_i A_{ii}. \tag{8.68}$$

Consider now the trace of a matrix product,

$$\mathrm{Tr}(\mathbf{AB}) = \sum_{i,j} A_{ij}B_{ji} \tag{8.69}$$

where we have written the summations in explicitly. We can (as before) freely exchange the order of the matrix elements A_{ij} and B_{ji}, to rewrite (8.69) as

$$\mathrm{Tr}(\mathbf{AB}) = \sum_{i,j} B_{ji}A_{ij}. \tag{8.70}$$

But the right-hand side is precisely $\mathrm{Tr}(\mathbf{BA})$; hence we have shown that

$$\mathrm{Tr}(\mathbf{AB}) = \mathrm{Tr}(\mathbf{BA}). \tag{8.71}$$

Similarly it is easy to show that

$$\mathrm{Tr}(\mathbf{ABC}) = \mathrm{Tr}(\mathbf{CAB}). \tag{8.72}$$

We may now return to (8.67). The advantage of the trace form is that we can invoke some powerful results about *traces of products of γ-matrices*. Here

we shall just list the trace 'theorems' that we shall use to evaluate $L^{\mu\nu}$: more complete statements of trace theorems and γ-matrix algebra, together with proofs of these theorems, are given in appendix J .

We need the following results:

$$\text{(i)} \qquad \mathrm{Tr}\mathbf{1} = 4 \tag{8.73}$$

$$\text{(ii)} \qquad \mathrm{Tr}\ (\text{odd number of } \gamma\text{'s}) = 0 \tag{8.74}$$

$$\text{(iii)} \qquad \mathrm{Tr}(\not{a}\not{b}) = 4(a\cdot b) \tag{8.75}$$

$$\text{(iv)} \qquad \mathrm{Tr}(\not{a}\not{b}\not{c}\not{d}) = 4[(a\cdot b)(c\cdot d) + (a\cdot d)(b\cdot c) - (a\cdot c)(b\cdot d)]. \tag{8.76}$$

Then

$$\begin{aligned}\mathrm{Tr}[(\not{k}' + m)\gamma^\mu(\not{k} + m)\gamma^\nu] \quad &= \quad \mathrm{Tr}(\not{k}'\gamma^\mu\not{k}\gamma^\nu) + m\mathrm{Tr}(\gamma^\mu\not{k}\gamma^\nu) \\ &\quad + m\mathrm{Tr}(\not{k}'\gamma^\mu\gamma^\nu) + m^2\mathrm{Tr}(\gamma^\mu\gamma^\nu)\end{aligned} \tag{8.77}$$

The terms linear in m are zero by theorem (ii), and using (iii) in the form

$$\mathrm{Tr}(\gamma_\mu\gamma_\nu)a^\mu b^\nu = 4g_{\mu\nu}a^\mu b^\nu = 4a\cdot b \tag{8.78}$$

and (iv) in a similar form, we obtain (problem 8.7)

$$L^{\mu\nu} = \tfrac{1}{2}\mathrm{Tr}[(\not{k}' + m)\gamma^\mu(\not{k} + m)\gamma^\nu] = 2[k'^\mu k^\nu + k'^\nu k^\mu - (k'\cdot k)g^{\mu\nu}] + 2m^2 g^{\mu\nu}. \tag{8.79}$$

In the present case we simply want L^{00}, which is found to be (problem 7.9)

$$L^{00} = 4E^2(1 - v^2\sin^2\theta/2) \tag{8.80}$$

where $v = |\boldsymbol{k}|/E$, just as in (8.48).

8.2.4 Coulomb scattering of e^+

The physical process is

$$e^+(k,s) \to e^+(k',s') \tag{8.81}$$

where, as usual, we emphasize that E and E' are both positive. In the wave-function approach, we saw in section 3.4.4. that, because $\rho \geq 0$ always for a Dirac particle, we had to introduce a minus sign 'by hand', according to the rule stated at the end of section 3.4.4. This rule gives us, in the present case,

$$\begin{aligned}&\text{amplitude } (e^+(k,s) \to e^+(k',s')) \\ &\quad = -\text{amplitude } (e^-(-k',-s') \to e^-(-k,-s)).\end{aligned} \tag{8.82}$$

Referring to (8.43), therefore, the required amplitude for the process (8.81) is

$$\mathcal{A}_{e^+} = -\mathrm{i}\int \mathrm{d}^4x\,(e\bar{v}(k,s)\gamma^\mu v(k',s')\mathrm{e}^{-\mathrm{i}(k-k')\cdot x})A_\mu(x) \tag{8.83}$$

since the 'v' solutions have been set up precisely to correspond to the '$-k, -s$' situation. In evaluating the cross section from (8.83), the only difference from the e^- case is the appearance of the spinors 'v' rather than 'u'; the lepton tensor in this case is

$$L^{\mu\nu} = \tfrac{1}{2}\mathrm{Tr}[(\not{k} - m)\gamma^\mu(\not{k}' - m)\gamma^\nu] \tag{8.84}$$

using the result (7.64) for $\sum_s v(k,s)\bar{v}(k,s)$. Expression (8.84) differs from (8.67) by the sign of m and by $k \leftrightarrow k'$, but the result (8.79) for the trace is insensitive to these changes. Thus the positron Coulomb scattering cross section is equal to the electron one to lowest order in α.

In the field-theoretic approach, the *same* interaction Hamiltonian $\hat{H}'_{\mathrm{D}}$ which we used for e^- scattering will again automatically yield the e^+ matrix element (recall the discussion at the end of section 8.1.3). In place of (8.53), the amplitude we wish to calculate is

$$\begin{aligned} \mathcal{A}_{\mathrm{e}^+} &= -\mathrm{i}\int \mathrm{d}^4x\, \langle \mathrm{e}^+, k', s'|\hat{j}^\mu_{\mathrm{em,e}}(x)|\mathrm{e}^+, k, s\rangle A_\mu(x) \\ &= -\mathrm{i}\int \mathrm{d}^4x\, \langle \mathrm{e}^+, k', s'| - e\bar{\hat{\psi}}(x)\gamma^\mu\hat{\psi}(x)|\mathrm{e}^+, k, s\rangle A_\mu(x) \end{aligned} \tag{8.85}$$

where, referring to the fermionic expansion (7.35),

$$|\mathrm{e}^+, k, s\rangle = \sqrt{2E}d^\dagger_s(k)|0\rangle, \tag{8.86}$$

and similarly for the final state. In evaluating the matrix element in (8.85) we must again remember to *normally order* the fields, according to the discussion in section 7.2. Bearing this in mind, and inserting the expansion (7.35), one finds (problem 8.9)

$$\langle \mathrm{e}^+, k', s'|\hat{j}^\mu_{\mathrm{em,e}}(x)|\mathrm{e}^+, k, s\rangle = +e\bar{v}(k,s)\gamma^\mu v(k',s')\mathrm{e}^{-\mathrm{i}(k-k')\cdot x} \tag{8.87}$$

$$\equiv j^\mu_{\mathrm{em,e}^+}(x) \tag{8.88}$$

just as required in (8.83). Note especially that the correct sign has emerged naturally without having to be put in 'by hand', as was necessary in the wavefunction approach when applied to an antifermion.

We are now ready to look at some more realistic (and covariant) processes.

8.3 $\mathrm{e}^-\mathrm{s}^+$ scattering

8.3.1 The amplitude for $\mathrm{e}^-\mathrm{s}^+ \to \mathrm{e}^-\mathrm{s}^+$

We consider the two-body scattering process

$$\mathrm{e}^-(k,s) + \mathrm{s}^+(p) \to \mathrm{e}^-(k',s') + \mathrm{s}^+(p') \tag{8.89}$$

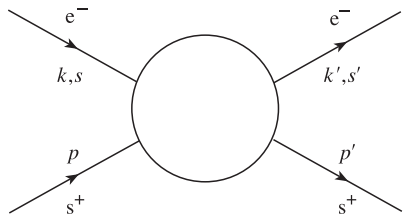


FIGURE 8.4
e^-s^+ scattering amplitude.

where the 4-momenta and spins are as indicated in figure 8.4. How will the e^- and s^+ interact? In this case, there is no 'external' classical electromagnetic potential in the problem. Instead, each of e^- and s^+, as charged particles, act as sources for the electromagnetic field, with which they in turn interact. We can picture the process as one in which each particle scatters off the 'virtual' field produced by the other (we shall make this more precise in comment (2) after equation (8.102)). The formalism of quantum field theory is perfectly adapted to account for such effects, as we shall see. It is very significant that no *new* interaction is needed to describe the process (8.89) beyond what we already have: the complete Lagrangian is now simply the free-field Lagrangians for the spin-$\frac{1}{2}$ e^-, the spin-0 s^+ and the Maxwell field, together with the sum of the lowest order scalar electromagnetic interaction Hamiltonian of (8.22), and the Dirac interaction Hamiltonian of (7.135) with $q = -e$. The full interaction Hamiltonian is then

$$\hat{H}'(x) = [\mathrm{i}e(\hat{\phi}^\dagger(x)\partial^\mu\hat{\phi}(x) - \partial^\mu\hat{\phi}^\dagger(x)\hat{\phi}(x)) - e\bar{\hat{\psi}}(x)\gamma^\mu\hat{\psi}(x)]\hat{A}_\mu(x) \quad (8.90)$$

$$\equiv (\hat{j}^\mu_{\mathrm{em,s}}(x) + \hat{j}^\mu_{\mathrm{em,e}}(x))\hat{A}_\mu(x) \quad (8.91)$$

where the 'total current' in (8.91) is just the indicated sum of the $\hat{\phi}$ (scalar) and $\hat{\psi}$ (spinor) currents. This $\hat{H}'$ must now be used in the Dyson expansion (6.42), in a perturbative calculation of the $e^-s^+ \to e^-s^+$ amplitude.

Note now that, in contrast to our Coulomb scattering 'warm-ups', the electromagnetic field *is* quantized in (8.90). We first observe that, since there are no free photons in either the initial or final states in our process $e^-s^+ \to e^-s^+$, the first-order matrix element of $\hat{H}'$ must vanish (as did the corresponding first-order amplitude in AB $\to$ AB scattering, in section 6.3.2). The first non-vanishing scattering processes arise at second order (cf (6.74)):

$$\mathcal{A}_{e^-s^+} = \frac{(-\mathrm{i})^2}{2}\int\!\!\int \mathrm{d}^4x_1\,\mathrm{d}^4x_2\,\langle 0|\hat{c}_{s'}(k')\hat{a}(p')T\{\hat{H}'(x_1)\hat{H}'(x_2)\}\hat{a}^\dagger(p)\hat{c}^\dagger_s(k)|0\rangle \times (16E_kE_{k'}E_pE_{p'})^{1/2}. \quad (8.92)$$

Just as for AB $\to$ AB and the $\hat{C}$ field in the 'ABC' model (cf (6.81)), as far

as the $\hat{A}_\mu$ operators in (8.92) are concerned the only surviving contraction is

$$\langle 0|T(\hat{A}_\mu(x_1)\hat{A}_\nu(x_2))|0\rangle \tag{8.93}$$

which is the Feynman propagator for the photon, in coordinate space. As regards the rest of the matrix element (8.92), since the $\hat{a}$'s and $\hat{c}$'s commute the 's^+' and 'e^-' parts are quite independent, and (8.92) reduces to

$$\frac{(-\mathrm{i})^2}{2}\iint \mathrm{d}^4x_1\,\mathrm{d}^4x_2\,\{\langle \mathrm{s}^+,p'|\hat{j}^\mu_{\mathrm{em,s}}(x_1)|\mathrm{s}^+,p\rangle\langle 0|T(\hat{A}_\mu(x_1)\hat{A}_\nu(x_2)|0\rangle \\ \times\langle \mathrm{e}^-,k',s'|\hat{j}^\nu_{\mathrm{em,e}}(x_2)|\mathrm{e}^-,k,s\rangle + (x_1\leftrightarrow x_2)\}. \tag{8.94}$$

But we know the explicit form of the current matrix elements in (8.94), from (8.27) and (8.55). Inserting these expressions into (8.94), and noting that the term with $x_1 \leftrightarrow x_2$ is identical to the first term, one finds (cf (6.102) and problem 8.10)

$$\mathcal{A}_{\mathrm{e}^-\mathrm{s}^+} = \mathrm{i}(2\pi)^4\delta^4(p+k-p'-k')\mathcal{M}_{\mathrm{e}^-\mathrm{s}^+} \tag{8.95}$$

where (using the general form (7.122) of the photon propagator)

$$\begin{aligned}\mathrm{i}\mathcal{M}_{\mathrm{e}^-\mathrm{s}^+} &= (-\mathrm{i})^2(e(p+p')^\mu)\left(\frac{\mathrm{i}[-g_{\mu\nu}+(1-\xi)q_\mu q_\nu/q^2]}{q^2}\right) \\ &\quad\times(-e\bar{u}(k',s')\gamma^\nu u(k,s)) \qquad (8.96)\\ &\equiv (-\mathrm{i})^2 j^\mu_{\mathrm{s}^+}(p,p')\left(\frac{\mathrm{i}[-g_{\mu\nu}+(1-\xi)q_\mu q_\nu/q^2]}{q^2}\right) j^\nu_{\mathrm{e}^-}(k,k') \qquad (8.97)\end{aligned}$$

and $q=(k-k')=(p'-p)$. We have introduced here the 'momentum–space' currents

$$j^\mu_{\mathrm{s}^+}(p,p') = e(p+p')^\mu \tag{8.98}$$

and

$$j^\mu_{\mathrm{e}^-}(k,k') = -e\bar{u}(k',s')\gamma^\mu u(k,s) \tag{8.99}$$

shortening the notation by dropping the 'em' suffix, which is understood.

Before proceeding to calculate the cross section, some comments on (8.97) are in order:

Comment (1)

The $j^\mu_{\mathrm{s}^+}(p,p')$ and $j^\nu_{\mathrm{e}^-}(k,k')$ in (8.98) and (8.99) are the momentum–space versions of the x-dependent current matrix elements in (8.27) and (8.55); they are, in fact, simply those matrix elements evaluated at $x=0$. The x-dependent matrix elements (8.27) and (8.55) both satisfy the current conservation equations $\partial_\mu j^\mu(x)=0$ as is easy to check (problem 8.11). Correspondingly, it follows from (8.98) and (8.99) that we have

$$q_\mu j^\mu_{\mathrm{s}^+}(p,p') = q_\mu j^\mu_{\mathrm{e}^-}(k,k') = 0 \tag{8.100}$$

where $q = p' - p = k - k'$, and we have used the mass-shell conditions $p^2 = p'^2 = M^2$, $\not{k}u = mu$, $\not{k}'u' = mu'$; the relations (8.100) are the momentum–space versions of current conservation. The ξ-dependent part of the photon propagator, which is proportional to $q^\mu q^\nu$, therefore vanishes in the matrix element (8.97). This shows that the amplitude is independent of the gauge parameter ξ – in other words, it is *gauge invariant* and proportional simply to

$$j^\mu_{s^+} \frac{g_{\mu\nu}}{q^2} j^\nu_{e^-}. \tag{8.101}$$

Comment (2)

The amplitude (8.97) has the appealing form of two currents 'hooked together' by the photon propagator. In the form (8.101), it has a simple 'semi-classical' interpretation. Suppose we regard the process $e^-s^+ \to e^-s^+$ as the scattering of the e^-, say, in the field produced by the s^+ (we can see from (8.101) that the answer is going to be symmetrical with respect to whichever of e^- and s^+ is singled out in this way). Then the amplitude will be, as in (8.43),

$$\mathcal{A}_{e^-s^+} = -i \int d^4x \, j^\nu_{e^-}(k, k') e^{-i(k-k')\cdot x} A_\nu(x) \tag{8.102}$$

where now the classical field $A_\nu(x)$ is not an 'external' Coulomb field but the field caused by the motion of the s^+. It seems very plausible that this $A_\nu(x)$ should be given by the solution of the Maxwell equations (2.22), with the $j_{\nu\mathrm{em}}(x)$ on the right-hand side given by the transition current (8.11) (with $N = N' = 1$) appropriate to the motion $s^+(p) \to s^+(p')$:

$$\Box A^\nu - \partial^\nu(\partial^\mu A_\mu) = j^\nu_{s^+}(x) \tag{8.103}$$

where

$$j^\nu_{s^+}(x) = e(p + p')^\nu e^{-i(p-p')\cdot x}. \tag{8.104}$$

Equation (8.103) will be much easier to solve if we can decouple the components of A^ν by using the Lorentz condition $\partial^\mu A_\mu = 0$. We are aware of the problems with this condition in the field-theory case (cf section 7.3.2) but we are here treating A^ν classically. Although A^ν is not a free field in (8.103), it is easy to see that we may consistently take $\partial^\mu A_\mu = 0$ provided that the current is conserved, $\partial_\nu j^\nu_{s^+}(x) = 0$, which we know to be the case. Thus we have to solve

$$\Box A^\nu(x) = e(p + p')^\nu e^{-i(p-p')\cdot x}. \tag{8.105}$$

Noting that

$$\Box e^{-i(p-p')\cdot x} = -(p - p')^2 e^{-i(p-p')\cdot x} \tag{8.106}$$

we obtain, by inspection,

$$A^\nu(x) = -\frac{1}{q^2} e(p + p')^\nu e^{-i(p-p')\cdot x} \tag{8.107}$$

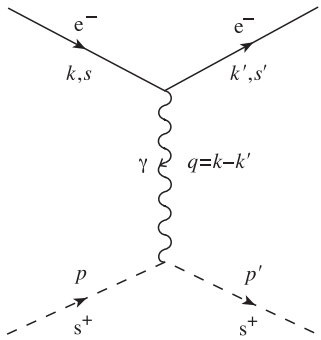


FIGURE 8.5
Feynman diagram for e^-s^+ scattering in the one-photon exchange approximation.

where $q = p' - p$. Inserting this expression into the amplitude (8.102) we find

$$\mathcal{A}_{e^-s^+} = i(2\pi)^4\delta^4(p + k - p' - k')\mathcal{M}_{e^-s^+} \tag{8.108}$$

where

$$i\mathcal{M}_{e^-s^+} = j^\mu_{s^+}(p, p')\frac{ig_{\mu\nu}}{q^2}j^\nu_{e^-}(k, k') \tag{8.109}$$

exactly as in (8.97) for $\xi = 1$ (the gauge appropriate to '$\partial_\mu A^\mu = 0$').

Comment (3)

From the work of chapter 6, it is clear that we can give a *Feynman graph interpretation* of the amplitude (8.109), as shown in figure 8.5, and set out the corresponding *Feynman rules*:

(i) At a vertex where a photon is emitted or absorbed by an s^+ particle, the factor is $-ie(p + p')^\mu$ where p, p' are the incident and outgoing 4-momenta of the s^+; the vertex for s^- has the opposite sign.

(ii) At a vertex where a photon is emitted or absorbed by an e^-, the factor is $ie\gamma^\mu (e > 0)$; for an e^+ it is $-ie\gamma^\mu$. (This and the previous rule arise from associating one '$(-i)$' factor in (8.94) or (8.97) with each current.)

(iii) For each initial state fermion line a factor $u(k, s)$ and for each final state fermion line a factor $\bar{u}(k', s')$; for each initial state antifermion a factor $\bar{v}(k, s)$ and for each final state antifermion line a factor $v(k', s')$ (these rules reconstruct the e^+ Coulomb amplitudes of section 8.2.4).

(iv) For an internal photon of 4-momentum q, there is a factor $-ig_{\mu\nu}/q^2$ in the gauge $\xi = 1$.

(v) Multiplying these factors together gives the quantity $\mathrm{i}\mathcal{M}$; multiplying the result by an overall 4-momentum-conserving δ-function factor $(2\pi)^4\delta(p' + k' + \cdots - p - k - \cdots)$ gives the quantity $\mathcal{A}$.

Comment (4)

We know that our amplitude is proportional to

$$j^\mu_{\mathrm{s}^+}\frac{g_{\mu\nu}}{q^2}j^\nu_{\mathrm{e}^-}. \tag{8.110}$$

Choosing the coordinate system such that $q = (q^0, 0, 0, |\boldsymbol{q}|)$, the current conservation equations $q \cdot j_{\mathrm{s}^+} = q \cdot j_{\mathrm{e}^-} = 0$ read:

$$j^3 = q^0 j^0/|\boldsymbol{q}| \tag{8.111}$$

for both currents. Expression (8.101) can then be written as

$$\begin{aligned}&(j^1_{\mathrm{s}^+}j^1_{\mathrm{e}^-} + j^2_{\mathrm{s}^+}j^2_{\mathrm{e}^-})/q^2 + (j^3_{\mathrm{s}^+}j^3_{\mathrm{e}^-} - j^0_{\mathrm{s}^+}j^0_{\mathrm{e}^-})/q^2\\ &\quad = (j^1_{\mathrm{s}^+}j^1_{\mathrm{e}^-} + j^2_{\mathrm{s}^+}j^2_{\mathrm{e}^-})/q^2 + j^0_{\mathrm{s}^+}j^0_{\mathrm{e}^-}/\boldsymbol{q}^2\end{aligned} \tag{8.112}$$

using (8.111). The first term may be interpreted as being due to the exchange of a transversely polarized photon (only the $1, 2$ components enter, perpendicular to $\boldsymbol{q}$). For *real* photons $q^2 \to 0$, so that this term will completely dominate the second. The latter, however, must obviously be included when $q^2 \neq 0$, as of course is the case for this *virtual* γ (cf section 6.3.3). We note that the second term depends on the 3-momentum squared, $\boldsymbol{q}^2$, rather than the 4-momentum squared q^2, and that it involves the charge densities $j^0_{\mathrm{s}^+}$ and $j^0_{\mathrm{e}^-}$. Referring back to section 7.1, we can interpret it as the *instantaneous Coulomb interaction* between these charge densities, since

$$\int \mathrm{d}^4x\, \mathrm{e}^{\mathrm{i}q\cdot x}\delta(t)/r = \int \mathrm{d}^3\boldsymbol{x}\, \mathrm{e}^{\mathrm{i}\boldsymbol{q}\cdot\boldsymbol{x}}/r = 4\pi/\boldsymbol{q}^2. \tag{8.113}$$

Thus, in summary, the single covariant amplitude (8.109) includes contributions from the exchange of transversely polarized photons *and* from the familiar Coulomb potential. This is the true relativistic extension of the static Coulomb results of (8.15) and (8.44).

8.3.2 The cross section for $\mathrm{e}^-\mathrm{s}^+ \to \mathrm{e}^-\mathrm{s}^+$

The invariant amplitude $\mathcal{M}_{\mathrm{e}^-\mathrm{s}^+}(s, s')$ for our process is given by (8.109) as

$$\mathcal{M}_{\mathrm{e}^-\mathrm{s}^+}(s, s') = e\bar{u}(k', s')\gamma^\mu u(k, s)(-g_{\mu\nu}/q^2)e(p + p')^\nu \tag{8.114}$$

where we have now included the spin dependence of the amplitude $\mathcal{M}_{\mathrm{e}^-\mathrm{s}^+}$ in the notation. The steps to the cross sections are now exactly as for the spin-0 case (section 6.3.4), as modified by the spin summing and averaging already

met in sections 8.2.1 and 8.2.3, particularly the latter. The cross section for the scattering of an electron in spin state s to one in spin state s' is (cf (6.110))

$$\begin{aligned} \mathrm{d}\sigma_{ss'} &= \frac{1}{4E\omega|\boldsymbol{v}|}|\mathcal{M}_{\mathrm{e}^-\mathrm{s}^+}(s,s')|^2(2\pi)^4\delta^4(k'+p'-k-p) \\ &\quad\times\frac{1}{(2\pi)^6}\frac{\mathrm{d}^3k'}{2\omega'}\frac{\mathrm{d}^3p'}{2E'} \end{aligned} \tag{8.115}$$

where we have defined

$$\begin{aligned} k^\mu &= (\omega,\boldsymbol{k}) \qquad k'^\mu = (\omega',\boldsymbol{k}') \\ p^\mu &= (E,\boldsymbol{p}) \qquad p'^\mu = (E',\boldsymbol{p}'). \end{aligned} \tag{8.116}$$

For the unpolarized cross section we are required, as in (8.46), to evaluate the quantity

$$\begin{aligned} \frac{1}{2}\sum_{s,s'}|\mathcal{M}_{\mathrm{e}^-\mathrm{s}^+}(s,s')|^2 &= \left(\frac{e^2}{q^2}\right)^2\frac{1}{2}\sum_{s,s'}\bar{u}(k',s')\gamma^\mu u(k,s)\bar{u}(k,s)\gamma^\nu u(k',s') \\ &\quad\times(p+p')_\mu(p+p')_\nu \qquad (8.117) \\ &\equiv \left(\frac{e^2}{q^2}\right)^2 L^{\mu\nu}(k,k')T_{\mu\nu}(p,p') \qquad (8.118) \end{aligned}$$

where the boson tensor $T_{\mu\nu}$ is just $(p+p')_\mu(p+p')_\nu$ and the lepton tensor $L^{\mu\nu}$ has been evaluated in (8.79). Using $q^2=(k-k')^2=2m^2-2k\cdot k'$, the expression (8.79) can be rewritten as

$$\boxed{L^{\mu\nu}(k,k')=2[k'^\mu k^\nu+k'^\nu k^\mu+(q^2/2)g^{\mu\nu}].} \tag{8.119}$$

We then find (problem 8.12)

$$L^{\mu\nu}T_{\mu\nu}=8[2(p\cdot k)(p\cdot k')+(q^2/2)M^2] \tag{8.120}$$

since $k'\cdot p'=k\cdot p$ and $k\cdot p'=k'\cdot p$ from 4-momentum conservation, and $p^2=p'^2=M^2$ (we are using m for the e^- mass and M for the s^+ mass).

We can now give the differential cross section in the CM frame by taking over the formula (6.129) with

$$|\mathcal{M}|^2\to\tfrac{1}{2}\sum_{s,s'}|\mathcal{M}_{\mathrm{e}^-\mathrm{s}^+}(s,s')|^2$$

so as to obtain

$$\left(\frac{\mathrm{d}\bar{\sigma}}{\mathrm{d}\Omega}\right)_{\mathrm{CM}}=\frac{2\alpha^2}{W^2(q^2)^2}[2(p\cdot k)(p\cdot k')+(q^2/2)M^2] \tag{8.121}$$

where $\alpha=e^2/4\pi$ and $W^2=(k+p)^2$.

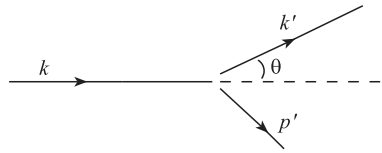


FIGURE 8.6
Two-body scattering in the 'laboratory' frame.

A somewhat more physically meaningful formula is found if we ask for the cross section in the 'laboratory' frame which we define by the condition $p^\mu = (M, \mathbf{0})$. The evaluation of the phase space integral requires some care and this is detailed in appendix K. The result is

$$\frac{\mathrm{d}\bar{\sigma}}{\mathrm{d}\Omega} = \frac{\alpha^2}{4k^2 \sin^4(\theta/2)} \cos^2(\theta/2) \frac{k'}{k}. \tag{8.122}$$

In this formula we have neglected the electron mass in the kinematics so that

$$k \equiv |\boldsymbol{k}| = \omega \tag{8.123}$$
$$k' \equiv |\boldsymbol{k}'| = \omega' \tag{8.124}$$

and

$$q^2 = -4kk' \sin^2(\theta/2) \tag{8.125}$$

where θ is the electron scattering angle in this frame, as shown in figure 8.6, and

$$(k/k') = 1 + (2k/M) \sin^2(\theta/2) \tag{8.126}$$

from equation (K.20). Note that there is a slight abuse of notation here: in the context of results for such laboratory frame calculations, 'k' and 'k'' are not 4-vectors, but rather the moduli of 3-vectors, as defined in equations (8.123) and (8.124).

We shall denote the cross section (8.122) by

$$\left(\frac{\mathrm{d}\sigma}{\mathrm{d}\Omega}\right)_{\mathrm{ns}} \qquad \text{'no-structure' cross section.} \tag{8.127}$$

It describes essentially the 'kinematics' of a relativistic electron scattering from a pointlike spin-0 target which recoils. Comparing the result (8.122) with equation (8.49), and remembering that here $Z = 1$ and we are taking $v \to 1$ for the electron, we see that the effect of recoil is contained in the factor (k'/k), in this limit. We recover the 'no-recoil' result (8.49) in the limit $M \to \infty$, as expected. In particular, referring to (8.125), we understand Rutherford's '$\sin^{-4}\theta/2$' factor in terms of the exchange of a massless quantum, via the propagator factor $(1/q^2)^2$.

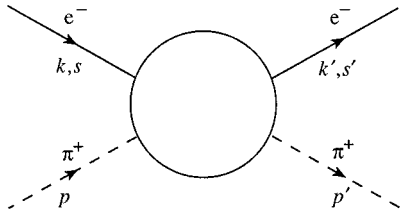


FIGURE 8.7
$e^-\pi^+$ scattering amplitude.

This 'no-structure' cross section also occurs in the cross section for the scattering of electrons by protons or muons: the appellation 'no-structure' will be made clearer in the discussion of form factors which follows. As in the case of e^+ Coulomb scattering, the cross sections for e^-s^+ and for e^+s^+ scattering are identical at this (lowest) order of perturbation theory.

8.4 Scattering from a non-point-like object: the pion form factor in $e^-\pi^+ \to e^-\pi^+$

As remarked earlier, we have been careful not to call the 's^+' particle a π^+, because the latter is a composite system which cannot be expected to have point-like interactions with the electromagnetic field, as has been assumed for the s^+; rather, in the case of the π^+ it is the quark constituents which interact locally with the electromagnetic field. The quarks also, of course, interact *strongly* with each other via the interactions of QCD, and since these are strong they cannot (in this case) be treated perturbatively. Indeed, a full understanding of the electromagnetically probed 'structure' of hadrons has not yet been achieved. Instead, we must describe the e^- scattering from physical π^+'s in terms of a phenomenological quantity – the *pion form-factor* – which encapsulates in a relativistically invariant manner the 'non-point-like' aspect of the hadronic state π^+.

The physical process is

$$e^-(k,s) + \pi^+(p) \to e^-(k',s') + \pi^+(p') \tag{8.128}$$

which we represent, in general, by figure 8.7. To lowest order in α, the amplitude is represented diagrammatically by a generalization of figure 8.5, shown in figure 8.8, in which the point-like $ss\gamma$ vertex is replaced by the $\pi\pi\gamma$ 'blob', which signifies all the unknown strong interaction corrections.

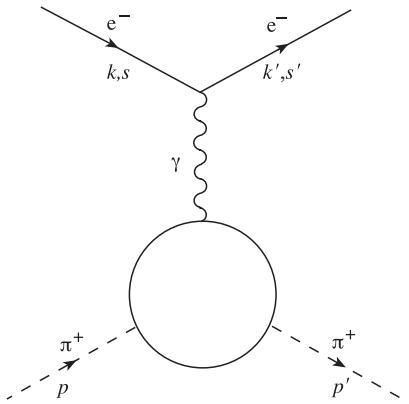


FIGURE 8.8
One-photon exchange amplitude in $e^-\pi^+$ scattering, including hadronic corrections at the $\pi\pi\gamma$ vertex.

8.4.1 e^- scattering from a charge distribution

It is helpful to begin the discussion by returning to e^- Coulomb scattering again, but this time let us consider the case in which the potential $A^0(\boldsymbol{x})$ corresponds, not to a point charge, but to a spread-out charge density $\rho(\boldsymbol{x})$. Then $A^0(\boldsymbol{x})$ satisfies Poisson's equation

$$\nabla^2 A^0(\boldsymbol{x}) = -Ze\rho(\boldsymbol{x}). \tag{8.129}$$

Note that if $A^0(\boldsymbol{x}) = Ze/4\pi|\boldsymbol{x}|$ as in (8.13) then $\rho(\boldsymbol{x}) = \delta(\boldsymbol{x})$ (see appendix G) and we recover the point-like source. The calculation of the Coulomb matrix element will proceed as before, except that now we require, at equation (8.43), the Fourier transform

$$\tilde{A^0}(\boldsymbol{q}) = \int \mathrm{e}^{\mathrm{i}\boldsymbol{q}\cdot\boldsymbol{x}} A^0(\boldsymbol{x}) \mathrm{d}^3x \tag{8.130}$$

where $q = \boldsymbol{k} - \boldsymbol{k}'$. To evaluate (8.130), note first that from the definition of $A^0(\boldsymbol{x})$, we can write

$$\begin{aligned} \int \mathrm{e}^{-\mathrm{i}\boldsymbol{q}\cdot\boldsymbol{x}} \boldsymbol{\nabla}^2 A^0(\boldsymbol{x})\, \mathrm{d}^3\boldsymbol{x} &= -Ze \int \mathrm{e}^{-\mathrm{i}\boldsymbol{q}\cdot\boldsymbol{x}} \rho(\boldsymbol{x})\, \mathrm{d}^3\boldsymbol{x} \\ &\equiv -ZeF(\boldsymbol{q}) \end{aligned} \tag{8.131}$$

where the (static) form factor $F(\boldsymbol{q})$ has been introduced, the Fourier transform of $\rho(\boldsymbol{x})$, satisfying

$$F(0) = \int \rho(\boldsymbol{x})\, \mathrm{d}^3\boldsymbol{x} = 1. \tag{8.132}$$

Condition (8.132) simply means that the total charge is Ze. The left-hand side of (8.131) can be transformed by two (three-dimensional) partial integrations to give

$$\int (\boldsymbol{\nabla}^2 \mathrm{e}^{-\mathrm{i}\boldsymbol{q}\cdot\boldsymbol{x}}) A^0(\boldsymbol{x})\,\mathrm{d}^3\boldsymbol{x} = -\boldsymbol{q}^2 \int \mathrm{e}^{-\mathrm{i}\boldsymbol{q}\cdot\boldsymbol{x}} A^0(\boldsymbol{x})\,\mathrm{d}^3\boldsymbol{x}. \tag{8.133}$$

Using this result in (8.131), we find

$$\tilde{A}^0(\boldsymbol{q}) = \frac{F(\boldsymbol{q})}{\boldsymbol{q}^2} Ze. \tag{8.134}$$

Thus referring to equation (8.44) for example, the net result of the non-point-like charge distribution is to multiply the 'point-like' amplitude $Ze^2/\boldsymbol{q}^2$ by the form factor $F(\boldsymbol{q})$ which in this simple static case has the interpretation of the Fourier transform of the charge distribution. So, for this (infinitely heavy π^+ case), the 'blob' in figure 8.8 would be represented by $F(\boldsymbol{q})$.

To gain some idea of what $F(\boldsymbol{q}^2)$ might look like, consider a simple exponential shape for $\rho(\boldsymbol{x})$:

$$\rho(\boldsymbol{x}) = \frac{1}{(8\pi a^3)} \mathrm{e}^{-|\boldsymbol{x}|/a} \tag{8.135}$$

which has been normalized according to (8.132). Then $F(\boldsymbol{q}^2)$ is (problem 8.13)

$$F(\boldsymbol{q}^2) = \frac{1}{(\boldsymbol{q}^2 a^2 + 1)^2}. \tag{8.136}$$

We see that $F(\boldsymbol{q}^2)$ decreases smoothly away from unity at $\boldsymbol{q}^2 = 0$. The characteristic scale of the fall-off in $|\boldsymbol{q}|$ is $\sim a^{-1}$ from (8.136), which, as expected from Fourier transform theory, is the reciprocal of the spatial fall-off, which is approximately a from (8.135); the root mean square radius of the distribution (8.135) is actually $\sqrt{12}a$ (problem 8.13). Since $\boldsymbol{q}^2 = 4\boldsymbol{k}^2 \sin^2\theta/2$, a larger $\boldsymbol{q}^2$ means a larger θ: hence, in scattering from an extended charge distribution, the cross section at larger angles will drop below the point-like value. This is, of course, how Rutherford deduced that the nucleus had a spatial extension.

We now seek a Lorentz-invariant generalization of this static form factor. In the absence of a fundamental understanding of the π^+ structure coming from QCD, we shall rely on Lorentz invariance and electromagnetic current conservation (one aspect of gauge invariance) to restrict the general form of the $\pi\pi\gamma$ vertex shown in figure 8.8. The use of invariance arguments to place restrictions on the form of amplitudes is an extremely general and important tool, in the absence of a complete theory.

8.4.2 Lorentz invariance

First, consider Lorentz invariance. We seek to generalize the point-like ssγ vertex (cf (8.98) and comment (1) after (8.99))

$$j^\mu_{\mathrm{s}^+}(p,p') = \langle \mathrm{s}^+, p' | \hat{j}^\mu_{\mathrm{em,s}}(0) | \mathrm{s}^+, p\rangle = e(p+p')^\mu \tag{8.137}$$

to $j^{\mu}_{\pi^+}(p,p')$, which will include strong interaction effects. Whatever these effects are, they cannot destroy the 4-vector character of the current. To construct the general form of $j^{\mu}_{\pi^+}(p,p')$ therefore, we must first enumerate the independent momentum 4-vectors we have at our disposal to parametrize the 4-vector nature of the current. These are just

$$p \qquad p' \qquad \text{and} \qquad q \tag{8.138}$$

subject to the condition

$$p' = p + q. \tag{8.139}$$

There are two independent combinations; these we can choose to be the linear combinations

$$(p' + p)_{\mu} \tag{8.140}$$

and

$$(p' - p)_{\mu} = q_{\mu}. \tag{8.141}$$

Both of these 4-vectors can, in general, parametrize the 4-vector nature of the electromagnetic current of a real pion. Moreover, they can be multiplied by an unknown scalar function of the available Lorentz scalar products for this process. Since

$$p^2 = p'^2 = M^2 \tag{8.142}$$

and

$$q^2 = 2M^2 - 2p \cdot p' \tag{8.143}$$

there is only one independent scalar in the problem, which we may take to be q^2, the 4-momentum transfer to the vertex. Thus, from Lorentz invariance, we are led to write the electromagnetic vertex of a pion in the form

$$j^{\mu}_{\pi^+}(p,p') = \langle \pi^+, p' | \hat{j}^{\mu}_{\mathrm{em},\pi}(0) | \pi^+, p \rangle = e[F(q^2)(p' + p)^{\mu} + G(q^2)q^{\mu}]. \tag{8.144}$$

The functions F and G are called 'form factors'.

This is as far as Lorentz invariance can take us. To identify the pion form factor, we must consider our second symmetry principle, gauge invariance – in the form of current conservation.

8.4.3 Current conservation

The Maxwell equations (7.65) reduce, in the Lorentz gauge

$$\partial_{\mu} A^{\mu} = 0 \tag{8.145}$$

to the simple form

$$\Box A^{\mu} = j^{\mu} \tag{8.146}$$

and the gauge condition is consistent with the familiar current conservation condition

$$\partial_{\mu} j^{\mu} = 0. \tag{8.147}$$

As we have seen in (8.100), the current conservation condition is equivalent to the condition

$$q_\mu \langle \pi^+(p')|\hat{j}^\mu_{\mathrm{em},\pi}(0)|\pi^+(p)\rangle = 0 \tag{8.148}$$

on the pion electromagnetic vertex.

In the case of the point-like s^+ this is clearly satisfied since

$$q \cdot (p' + p) = 0 \tag{8.149}$$

with the aid of (8.142). In the general case we obtain the condition

$$q_\mu [F(q^2)(p' + p)^\mu + G(q^2)q^\mu] = 0. \tag{8.150}$$

The first term vanishes as before, but $q^2 \neq 0$ in general, and we therefore conclude that current conservation implies that

$$G(q^2) = 0. \tag{8.151}$$

In other words, all the virtual strong interaction effects at the $\pi^+\pi^+\gamma$ vertex are described by one scalar function of the virtual photon's squared 4-momentum:

$$\boxed{\begin{array}{ccc} e(p'+p)^\mu & \rightarrow & eF(q^2)(p'+p)^\mu. \\ \text{'point pion'} & & \text{'real pion'} \end{array}} \tag{8.152}$$

$F(q^2)$ is the *electromagnetic form factor of the pion*, which generalizes the static form factor $F(\boldsymbol{q}^2)$ of section 8.4.1. The pion electromagnetic vertex is then

$$j^\mu_{\pi+}(p, p') = eF(q^2)(p + p')^\mu. \tag{8.153}$$

The electric charge is defined to be the coupling at zero momentum transfer, so the form factor is normalized by the condition (cf (8.132))

$$F(0) = 1. \tag{8.154}$$

To lowest order in α, the invariant amplitude for $\mathrm{e}^-\pi^+ \rightarrow \mathrm{e}^-\pi^+$ is therefore given by replacing $j^\mu_{\mathrm{s}^+}(p, p')$ in (8.97) or (8.109) by $j^\mu_{\pi^+}(p, p')$:

$$\mathrm{i}\mathcal{M}_{\mathrm{e}^-\pi^+} = -\mathrm{i}e(p + p')^\mu F((p' - p)^2) \left(\frac{-\mathrm{i}g_{\mu\nu}}{(p' - p)^2}\right) [+\mathrm{i}e\bar{u}(k', s')\gamma_\nu u(k, s)]. \tag{8.155}$$

It is clear that the effect of the pion structure is simply to multiply the 'no-structure' cross section (8.122) by the square of the form factor, $F(q^2 = (p' - p)^2)$.

For $\mathrm{e}^-\pi^+ \rightarrow \mathrm{e}^-\pi^+$ in the CM frame we may take $p = (E, \boldsymbol{p})$ and $p' = (E, \boldsymbol{p}')$ with $|\boldsymbol{p}| = |\boldsymbol{p}'|$ and $E = (m_\pi^2 + \boldsymbol{p}^2)^{1/2}$. Then

$$q^2 = (p' - p)^2 = -4\boldsymbol{p}^2 \sin^2 \theta/2 \tag{8.156}$$

as in section 8.1, where θ is now the CM scattering angle between $\boldsymbol{p}$ and $\boldsymbol{p}'$.

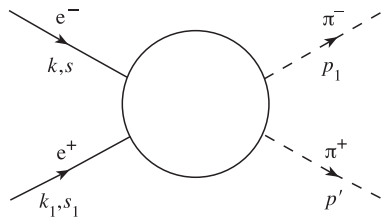


FIGURE 8.9
$e^+e^- \to \pi^+\pi^-$ scattering amplitude.

Hence $F(q^2)$ can be probed for negative (space-like) values of q^2, in the process $e^-\pi^+ \to e^-\pi^+$. As in the static case, we expect the form factor to fall off as $-q^2$ increases since, roughly speaking, it represents the amplitude for the target to remain intact when probed by the electromagnetic current. As $-q^2$ increases, the amplitudes of inelastic processes which involve the creation of extra particles become greater, and the elastic amplitude is correspondingly reduced. We shall consider inelastic scattering in the following chapter.

Interestingly, $F(q^2)$ may also be measured at positive (time-like) q^2, in the related reaction $e^+e^- \to \pi^+\pi^-$ as we now discuss.

8.5 The form factor in the time-like region: $e^+e^- \to \pi^+\pi^-$ and crossing symmetry

The physical process is

$$e^+(k_1, s_1) + e^-(k, s) \to \pi^+(p') + \pi^-(p_1) \tag{8.157}$$

as shown in figure 8.9. We can use this as an instructive exercise in the Feynman interpretation of section 3.4.4. From that section, we know that the invariant amplitude for (8.157) is equal to minus the amplitude for a process in which the ingoing antiparticle e^+ with (k_1, s_1) becomes an outgoing particle e^- with $(-k_1, -s_1)$, and the outgoing antiparticle π^- with p_1 becomes an ingoing particle π^+ with $-p_1$. In this way the 'physical' (positive 4-momentum) antiparticle states (e^+ and π^-) are replaced by appropriate 'unphysical' (negative 4-momentum) particle states (e^- and π^+). These changes transform figure 8.9 to figure 8.10.

If we now look at figure 8.10 'from the top downwards' (instead of from left to right – remember that Feynman diagrams are *not* in coordinate space!), we see a process of $e^-\pi^+$ scattering, namely

$$e^-(k, s) + \pi^+(-p_1) \to e^-(-k_1, -s_1) + \pi^+(p'). \tag{8.158}$$

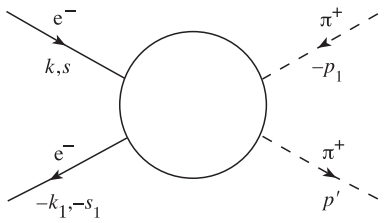


FIGURE 8.10
The amplitude of figure 8.9, with positive 4-momentum antiparticles replaced by negative 4-momentum particles.

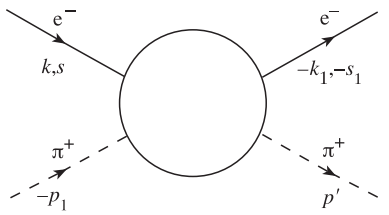


FIGURE 8.11
The amplitude of figure 8.10 redrawn so as to obtain a reaction in which the initial state has only 'ingoing' lines and the final state has only 'outgoing' lines.

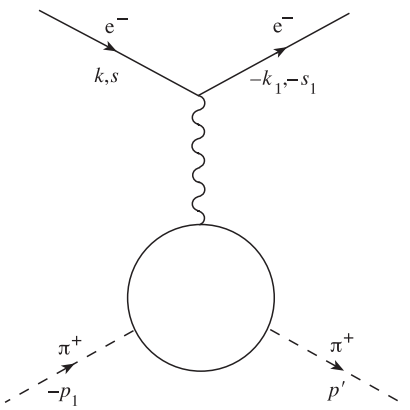


FIGURE 8.12
One-photon exchange amplitude for the process of figure 8.11.

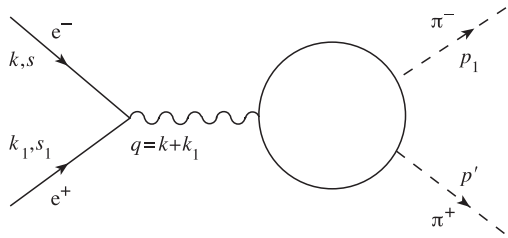


FIGURE 8.13
One-photon exchange amplitude for the process of figure 8.9.

But (8.158) is something we have already calculated! (Though we shall have to substitute a negative-energy spinor v for a positive energy one u.) In fact, let us redraw figure 8.10 as figure 8.11 to make it look more like figure 8.7. Then, to lowest order in α, the amplitude for figure 8.11 is shown in figure 8.12 (compare figure 8.8). To obtain the corresponding mathematical expression for the amplitude $\mathrm{i}\mathcal{M}_{e^+e^-\to\pi^+\pi^-}$, we simply need to modify (8.155): (i) by inserting a minus sign; (ii) by replacing p by $-p_1$ and k' by $-k_1$ as in figure 8.12; and (iii) by replacing $\bar{u}(k',s')$ by $\bar{v}(k_1,s_1)$. This yields the invariant amplitude for figure 8.12 as

$$\begin{aligned}\mathrm{i}\mathcal{M}_{e^+e^-\to\pi^+\pi^-} &= -\mathrm{i}e(-p_1+p')^\mu F((p_1+p')^2)\left(\frac{-\mathrm{i}g_{\mu\nu}}{(p_1+p')^2}\right)\\ &\quad\times[-\mathrm{i}e\bar{v}(k_1,s_1)\gamma^\nu u(k,s)] \qquad (8.159)\end{aligned}$$

which is represented by the Feynman diagram of figure 8.13 for the *original* process of (8.157) and figure 8.9.

In the language introduced in section 6.3.3, figure 8.13 is an 's-channel process' ($s=(k+k_1)^2=(p_1+p')^2$) for $e^+e^-\to\pi^+\pi^-$, whereas figure 8.8 is a 't-channel process' ($t=(k-k')^2=(p'-p)^2$) for $e^-\pi^+\to e^-\pi^+$. However, we have seen that the amplitude for the $e^+e^-\to\pi^+\pi^-$ process can be obtained from the $e^-\pi^+\to e^-\pi^+$ amplitude by making the replacement $k'\to -k_1, p\to -p_1$ (together with the sign, and $\bar{u}\to\bar{v}$). Under these replacements of the 4-momenta, the variable $t=(k-k')^2=(p-p')^2$ of figure 8.8 becomes the variable $s=(k+k_1)^2=(p_1+p')^2$ of figure 8.13. In particular, as is evident in the formula (8.159), the *same* form factor F is a function of the invariant $s=(p_1+p')^2$ in process (8.157), and of $t=(p-p')^2$ in process (8.128). The interesting thing is that whereas (as we have seen) 't' is negative in process (8.128), 's' for process (8.157) is the square of the total CM energy, which is $\geq 4M^2$ where M is the pion mass ($2M$ is the threshold energy for the reaction to proceed in the CM system). Thus the form factor can be probed at negative values of its argument in the process $e^-\pi^+\to e^-\pi^+$, and at positive values $\geq 4M^2$ in the process $e^+e^-\to\pi^+\pi^-$.

In the next chapter (section 9.5) we shall see how, in the latter process, meson resonances dominate $F(s)$.

The procedure whereby an ingoing/outgoing antiparticle is switched to an outgoing/ingoing particle is called 'crossing' (the state is being 'crossed' from one side of the reaction to the other). By an extension of this language, $e^+e^- \to \pi^+\pi^-$ is called the crossed process relative to $e^-\pi^+ \to e^-\pi^+$ (or vice versa). The fact that the amplitude for a given process and its 'crossed' analogue are directly related via the Feynman interpretation (or by quantum field theory!) is called 'crossing symmetry'. In the example studied here, what is an s-channel process for one reaction becomes a t-channel process for the crossed reaction. Essentially, little more is involved than looking in the one case from left to right and, in the other, from top to bottom!

8.6 Electron Compton scattering

8.6.1 The lowest-order amplitudes

We proceed to explore some other elementary electromagnetic processes. So far we have not considered a reaction with external photons, so let us now discuss electron Compton scattering

$$\gamma(k,\lambda) + e^-(p,s) \to \gamma(k',\lambda') + e^-(p',s') \tag{8.160}$$

where the λ's stand for the polarizations of the photons. Since only the γ's and e^-'s are involved, the interaction Hamiltonian is simply $\hat{H}'_D$, and it is clear that this must act at least twice in the reaction (8.160). By following the method of section 6.3.2 one can formally derive what we are here going to assume is by now obvious, which is that to order e^2 (i.e. α in the amplitude) there are two contributing Feynman graphs, as shown in figures 8.14(a) and (b). The first is an s-channel process, the second a u-channel process. We already know the factors for the vertices and for the external electron lines; we need to know the factors for the internal electron lines (propagators) and the external photon lines. The fermion propagator was given in section 7.2 and is $i/(\not{q} - m + i\epsilon)$ for a line carrying 4-momentum q. As regards the 'external-γ' factor, this will arise from contractions of the form (cf (6.90))

$$\sqrt{2E_{k'}}\langle 0|\alpha(k',\lambda')\hat{A}^\mu(x_1)|0\rangle = \epsilon^{\mu *}(k',\lambda')e^{ik'\cdot x_1} \tag{8.161}$$

where the evaluation of the vev has used the mode expansion (7.104) and the commutation relations (7.108), as usual; note, however, that only *transverse* polarization states ($\lambda, \lambda' = 1$ and 2) enter in the external (physical) photon lines in figures 8.14(a) and (b).

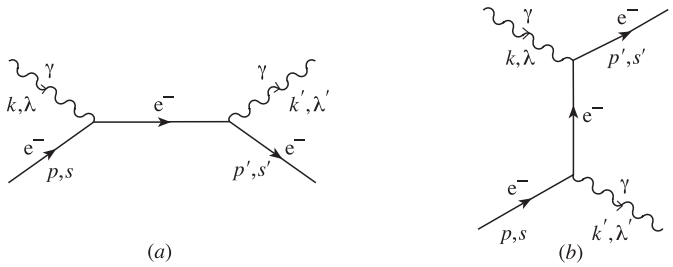


FIGURE 8.14
$O(e^2)$ contributions to electron Compton scattering.

Thus we add two more rules to the (i)–(v) of section 8.3.1:

(vi) For an incoming photon of 4-momentum k and polarization λ, there is a factor $\epsilon^\mu(k,\lambda)$; for an outgoing one, $\epsilon^{\mu *}(k',\lambda')$.

(vii) For an internal spin-$\frac{1}{2}$ particle carrying 4-momentum q, there is a factor $\mathrm{i}/(\not{q} - m + \mathrm{i}\epsilon) = \mathrm{i}(\not{q} + m)/(q^2 - m^2 + \mathrm{i}\epsilon)$.

The invariant amplitude $\mathcal{M}_{\gamma \mathrm{e}^-}$ corresponding to figures 8.14(a) and (b) is therefore

$$\begin{aligned}\mathcal{M}_{\gamma \mathrm{e}^-} &= -e^2 \epsilon^*_\nu(k',\lambda')\epsilon_\mu(k,\lambda)\bar{u}(p',s')\gamma^\nu \frac{(\not{p} + \not{k} + m)}{(p+k)^2 - m^2}\gamma^\mu u(p,s) \\ &\quad - e^2 \epsilon^*_\nu(k',\lambda')\epsilon_\mu(k,\lambda)\bar{u}(p',s')\gamma^\mu \frac{(\not{p} - \not{k}' + m)}{(p-k')^2 - m^2}\gamma^\nu u(p,s). \quad (8.162)\end{aligned}$$

To get the spinor factors in expressions such as these, the rule is to start at the ingoing fermion line ('$u(p,s)$') and follow the line through until the end, inserting vertices and propagators in the right order, until you reach the outgoing state ('$\bar{u}$'). Note that here $s = (p+k)^2$ and $u = (p-k')^2$.

8.6.2 Gauge invariance

We learned in section 7.3.1 that the gauge symmetry ($A^\mu \to A^\mu - \partial^\mu \chi$) of electromagnetism, as applied to real free photons, implied that any photon polarization vector $\epsilon^\mu(k,\lambda)$ could be replaced by

$$\epsilon'^\mu(k\lambda) = \epsilon^\mu(k,\lambda) + \beta k^\mu \quad (8.163)$$

where β is an arbitrary constant. Such a transformation amounted to a change of gauge, always remaining within the Lorentz gauge for which $\epsilon \cdot k = \epsilon' \cdot k = 0$. Thus our amplitude (8.162) must be unchanged if we make either or both the replacements $\epsilon \to \epsilon + \beta k$ and $\epsilon^* \to \epsilon^* + \beta k'$ indicated in (8.163). This means that if in (8.162) we replace either or both of $\epsilon_\mu(k,\lambda)$ and $\epsilon^*_\nu(k',\lambda')$ by k_μ

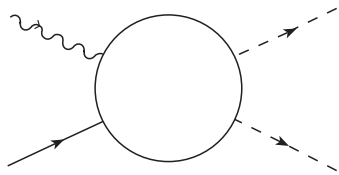


FIGURE 8.15
General one-photon process.

and k'_ν, respectively, the result has to be zero. This can indeed be verified (problem 8.14).

A similar result is generally true and very important. Consider a process, shown in figure 8.15, involving a photon of momentum k^μ, whose polarization state is described by the vector ϵ^μ. The amplitude $\mathcal{A}_\gamma$ for this process must be linear in the photon polarization vector and thus we may write

$$\mathcal{A}_\gamma = \epsilon^\mu T_\mu \tag{8.164}$$

where T_μ depends on the particular process under consideration. With the Lorentz choice for ϵ^μ we have

$$k \cdot \epsilon = 0. \tag{8.165}$$

But gauge invariance implies that if we replace ϵ^μ in (8.164) by k^μ we must get zero:

$$\boxed{k^\mu T_\mu = 0.} \tag{8.166}$$

This important condition on T_μ is known as a *Ward identity* (Ward 1950).

8.6.3 The Compton cross section

The calculation of the cross section is of considerable interest, since it is required when considering lowest-order QCD corrections to the parton model for deep inelastic scattering of leptons from nucleons (see the following chapter and volume 2). We must average $|\mathcal{M}_{\gamma e^-}|^2$ over initial electron spins and photon polarizations and sum over final ones. Consider first the s-channel process of figure 8.14(a), with amplitude $\mathcal{M}^{(s)}_{\gamma e^-}$. For this contribution we must evaluate

$$\frac{e^4}{4(s-m^2)^2} \cdot \sum_{\lambda,\lambda',s,s'} \epsilon'^*_\nu \epsilon_\mu \epsilon^*_\rho \epsilon'_\sigma \bar{u}' \gamma^\nu (\not{p} + \not{k} + m) \gamma^\mu u \bar{u} \gamma^\rho (\not{p} + \not{k} + m) \gamma^\sigma u' \tag{8.167}$$

where we have shortened the notation in an obvious way and introduced the invariant Mandelstam variable (section 6.3.3) $s = (p+k)^2$. We know how to write the spin sums in a convenient form, as a trace. We need to find a similar trick for the polarization sum.

Consider the general 'one-photon' process shown in figure 8.15, with amplitude $\mathcal{A}_\gamma = \epsilon^\mu(k,\lambda)T_\mu$, where $\epsilon^\mu(k,1) = (0,1,0,0)$ and $\epsilon^\mu(k,2) = (0,0,1,0)$, and $k^\mu = (k,0,0,k)$. Then the required polarization sum would be

$$\sum_{\lambda=1,2} \epsilon^\mu(k,\lambda)T_\mu \epsilon^{\nu *}(k,\lambda)T^*_\nu = |T_1|^2 + |T_2|^2. \tag{8.168}$$

However, we also know that $k^\mu T_\mu = 0$ from the Ward identity (8.166). This tells us that

$$kT_0 - kT_3 = 0 \tag{8.169}$$

and hence $T_0 = T_3$. It follows that we may write (8.168) as

$$\sum_{\lambda=1,2} \epsilon^\mu(k,\lambda)\epsilon^{\nu *}(k,\lambda)T_\mu T^*_\nu = |T_1|^2 + |T_2|^2 + |T_3|^2 - |T_0|^2 \tag{8.170}$$

$$= -g^{\mu\nu}T_\mu T^*_\nu. \tag{8.171}$$

Thus we may replace the non-covariant expression '$\sum_{\lambda=1,2}\epsilon^\mu(k,\lambda)\epsilon^{\nu *}(k,\lambda)$' by the covariant one '$-g^{\mu\nu}$'. The reader may here recall equation (7.118), where the 'pseudo-completeness' relation involving all *four* ϵ's was given, a similarly covariant expression. This relation corresponds exactly to the right-hand side of (8.170), which (in these terms) shows that the $\lambda = 0$ state enters with negative norm.

Using this result, the term (8.167) becomes

$$\frac{e^4}{4(s-m^2)^2}\sum_{s,s'} \bar{u}'\gamma^\nu(\not{p}+\not{k}+m)\gamma^\mu u \bar{u}\gamma_\mu(\not{p}+\not{k}+m)\gamma_\nu u'$$

$$= \frac{e^4}{4(s-m^2)^2}\mathrm{Tr}[\gamma_\nu(\not{p}'+m)\gamma^\nu(\not{p}+\not{k}+m)\gamma^\mu(\not{p}+m)\gamma_\mu(\not{p}+\not{k}+m)] \tag{8.172}$$

where, in the second step, we have moved the γ_ν to the front of the trace, using (8.71). Expression (8.172) involves the trace of eight γ matrices, which is beyond the power of the machinery given so far. However, it simplifies greatly if we neglect the electron mass – that is, if we are interested in the high-energy limit, as we shall be in parton model applications. In that case, (8.172) becomes

$$\frac{e^4}{4s^2}\mathrm{Tr}[\gamma_\nu \not{p}'\gamma^\nu(\not{p}+\not{k})\gamma^\mu \not{p}\gamma_\mu(\not{p}+\not{k})] \tag{8.173}$$

which we can simplify using the result (J.3) to

$$\frac{e^4}{s^2}\mathrm{Tr}[\not{p}'(\not{p}+\not{k})\not{p}(\not{p}+\not{k})] \tag{8.174}$$

$$= \frac{e^4}{s^2}\mathrm{Tr}[\not{p}'\not{k}\not{p}\not{k}] \qquad \text{using } \not{p}^2 = p^2 = 0 \tag{8.175}$$

$$= \frac{4e^4}{s^2}\cdot 2(p'\cdot k)(p\cdot k) \qquad \text{using (8.76) and } k^2 = 0 \tag{8.176}$$

$$= -2e^4 u/s \tag{8.177}$$

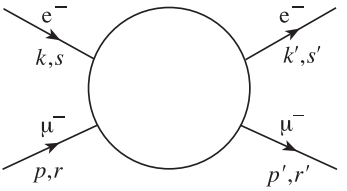


FIGURE 8.16
$e^-\mu^-$ scattering amplitude.

where $u = (p - k')^2$. Problem 8.15 finishes the calculation, with the result that the spin-averaged squared amplitude is

$$\frac{1}{4}\sum_{s,s',\lambda,\lambda'} |\mathcal{M}_{\gamma e^-}|^2 = -2e^4\left(\frac{u}{s} + \frac{s}{u}\right). \tag{8.178}$$

The cross section in the CMS is then (cf (6.129))

$$\frac{\mathrm{d}\sigma}{\mathrm{d}(\cos\theta)} = \frac{2\pi 2e^4}{64\pi^2 s}\left(\frac{-u}{s} - \frac{s}{u}\right) = \frac{\pi\alpha^2}{s}\left(\frac{-u}{s} - \frac{s}{u}\right). \tag{8.179}$$

For parton model calculations, what is actually required is the analogous quantity calculated for the case in which the initial photon is virtual (see section 9.2). However, the discussion of section 7.3.2 shows that we may still use the polarization sum (8.170). A difference will arise in passing from (8.175) to (8.176) where we must remember that $k^2 \neq 0$. Since k^2 will be space-like, we put $k^2 = -Q^2$ and find (problem 8.16) that the spin-averaged squared amplitude for the virtual Compton process

$$\gamma^*(k^2 = -Q^2) + \mathrm{e}^- \to \gamma + \mathrm{e}^- \tag{8.180}$$

is given by

$$-2e^4\left(\frac{u}{s} + \frac{s}{u} - \frac{2Q^2 t}{su}\right). \tag{8.181}$$

8.7 Electron muon elastic scattering

Our final examples of electrodynamic processes are ones in which two fermions interact electromagnetically. In this section we discuss the scattering of two point-like fermions (i.e. leptons); in the following one we look at the change (analogous to those for the π^+ as compared to the s^+) necessitated when one fermion is a hadron, for example the proton.

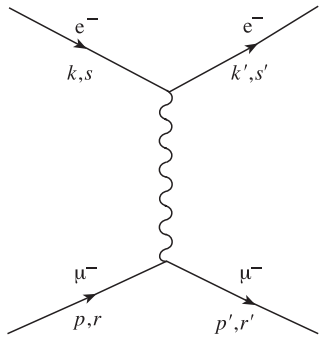


FIGURE 8.17
One-photon exchange amplitude in $e^-\mu^-$ scattering.

We shall consider $e^-\mu^-$ elastic scattering: our notation is indicated in figure 8.16. In the lowest order of perturbation theory – the one-photon exchange approximation – we can draw the relevant Feynman graph for this process. This is shown in figure 8.17. All the elements for the graph have been met before and so we can immediately write down the invariant amplitude which now depends on four spin labels:

$$\mathcal{M}_{e^-\mu^-}(r,s;r',s') = e\bar{u}(k',s')\gamma_\mu u(k,s)(g^{\mu\nu}/q^2)e\bar{u}(p',r')\gamma_\nu u(p,r). \quad (8.182)$$

Although experiments with polarized leptons are not uncommon, we shall only be concerned with the unpolarized cross section

$$d\bar{\sigma} \sim \tfrac{1}{4}\sum_{r,r';s,s'} |\mathcal{M}_{e^-\mu^-}(r,s;r',s')|^2. \quad (8.183)$$

We perform the same manipulations as in our e^-s^+ example and the cross section reduces to a factorized form involving two traces:

$$\begin{aligned}\frac{1}{4}\sum_{r,r';s,s'} |\mathcal{M}_{e^-\mu^-}(r,s;r',s')|^2 &= \left(\frac{e^2}{q^2}\right)^2 \left\{\frac{1}{2}\mathrm{Tr}[(\not{k}'+m)\gamma_\mu(\not{k}+m)\gamma_\nu]\right\} \\ &\quad \times \{\tfrac{1}{2}\mathrm{Tr}[(\not{p}'+M)\gamma^\mu(\not{p}+M)\gamma^\nu]\} \quad (8.184) \\ &= (e^2/q^2)^2 L_{\mu\nu}M^{\mu\nu} \quad (8.185)\end{aligned}$$

where $L_{\mu\nu}$ is the 'electron tensor' calculated before (see (8.119)):

$$L_{\mu\nu} = 2[k'_\mu k_\nu + k'_\nu k_\mu + (q^2/2)g_{\mu\nu}] \quad (8.186)$$

but now $M^{\mu\nu}$ is the appropriate tensor for the muon coupling, with the same structure as $L_{\mu\nu}$:

$$M^{\mu\nu} = 2[p'^\mu p^\nu + p'^\nu p^\mu + (q^2/2)g^{\mu\nu}]. \quad (8.187)$$

To evaluate the cross section we must perform the 'contraction' $L_{\mu\nu}M^{\mu\nu}$. A useful trick to simplify this calculation is to use current conservation for the electron tensor $L_{\mu\nu}$. For the electron transition current, the electromagnetic current conservation condition is (cf equation (8.100))

$$q^\mu[\bar{u}(k',s')\gamma_\mu u(k,s)] = 0 \tag{8.188}$$

i.e. independent of the particular spin projections s and s'. Since $L_{\mu\nu}$ is the product of two such currents, summed and averaged over polarizations, current conservation implies the conditions

$$q^\mu L_{\mu\nu} = q^\nu L_{\mu\nu} = 0 \tag{8.189}$$

which can be explicitly checked using our result for $L_{\mu\nu}$. The usefulness of this result is that in the contraction $L_{\mu\nu}M^{\mu\nu}$ we can replace p' in $M^{\mu\nu}$ by $(p+q)$ and then drop all the terms involving q's, i.e.

$$L_{\mu\nu}M^{\mu\nu} = L_{\mu\nu}M^{\mu\nu}_{\text{eff}} \tag{8.190}$$

where

$$M^{\mu\nu}_{\text{eff}} = 2[2p^\mu p^\nu + (q^2/2)g^{\mu\nu}]. \tag{8.191}$$

The calculation of the cross section is now straightforward. In the 'laboratory' system, defined (unrealistically) by the target muon at rest

$$p^\mu = (M,0,0,0) \tag{8.192}$$

with M now the muon mass, the result is (problem 8.17(a))

$$\frac{\mathrm{d}\sigma}{\mathrm{d}\Omega} = \left(\frac{\mathrm{d}\sigma}{\mathrm{d}\Omega}\right)_{\text{ns}} \left(1 - \frac{q^2\tan^2(\theta/2)}{2M^2}\right). \tag{8.193}$$

Note the following points:

Comment (a)

The 'no-structure' cross section (8.122) for e^-s^+ scattering now appears modified by an additional term proportional to $\tan^2(\theta/2)$. This is due to the spin-$\frac{1}{2}$ nature of the muon which gives rise to scattering from both the charge *and* the magnetic moment of the muon.

Comment (b)

In the kinematics the electron mass has been neglected, which is usually a good approximation at high energies. We should add a word of explanation for the 'laboratory' cross sections we have calculated, with the target muon unrealistically at rest. The form of the cross section, $(\mathrm{d}\sigma/\mathrm{d}\Omega)_{\text{ns}}$, and of the cross section for the scattering of two Dirac point particles, will be of great value in our discussion of the quark parton model in the next chapter.

Comment (c)

The crossed version of this process, namely $e^+e^- \to \mu^+\mu^-$, is a very important monitoring reaction for electron–positron colliding beam machines. It is also basic to a discussion of the predictions of the quark parton model for $e^+e^- \to$ hadrons, which will be discussed in section 9.5. An instructive calculation similar to this one leads to the result (see problem 8.18)

$$\frac{d\sigma}{d\Omega} = \frac{\alpha^2}{4q^2}(1 + \cos^2\theta) \tag{8.194}$$

where all variables are defined in the e^+e^- CM frame, q^2 is now the square of the CM energy, and the electron and muon masses have been neglected. The total cross section, in the one-photon exchange approximation, is then

$$\sigma = 4\pi\alpha^2/3q^2 = 86.8 \text{ nb}/q^2(\text{GeV}^2), \tag{8.195}$$

where we have made use of equation (B.18) of appendix B.

The energy dependence of this cross section ($\propto 1/q^2$) is important, and can be understood by a simple dimensional argument. A cross section has dimensions of a squared length, or in natural units (appendix B) inverse squared mass or energy. Here both colliding particles are taken to be pointlike, with no form factors involving a length parameter, and the mediating quantum is massless. At energies much larger than the lepton masses, the only available dimensional quantity is the CM energy. It follows that the cross section must be inversely proportional to the square of the CM energy, in this 'pointlike, high energy' limit. By the same token, deviations from this behaviour would be evidence for non-pointlike leptonic structure.

8.8 Electron–proton elastic scattering and nucleon form factors

In the one-photon exchange approximation, the Feynman diagram for elastic electron–proton scattering may be drawn as in figure 8.18, where the 'blob' at the ppγ vertex signifies the expected modification of the point coupling due to strong interactions. The structure of the proton vertex can be analysed using symmetry principles in the same way as for the pion vertex. The presence of Dirac spinors and γ-matrices makes this a somewhat involved procedure: problem 8.20 is an example of the type of complication that arises. Full details of such an analysis can be found in Bernstein (1968), for example. Here, however, we shall proceed in a different way, in order to generalize more easily to inelastic scattering in the following chapter. We focus directly on the 'proton tensor' $B^{\mu\nu}$, which is the product of two proton current matrix elements,

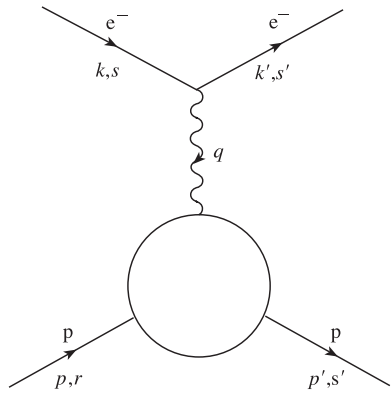


FIGURE 8.18
One-photon exchange amplitude in e^-p scattering, including hadronic corrections at the ppγ vertex.

summed and averaged over polarizations, as is required in the calculation of the unpolarized cross section (cf (8.57)):

$$B^{\mu\nu} = \frac{1}{2e^2}\sum_{s,s'}\langle \mathrm{p};p',s'|\hat{j}^{\mu}_{\mathrm{em,p}}(0)|\mathrm{p};p,s\rangle(\langle \mathrm{p};p',s'|\hat{j}^{\nu}_{\mathrm{em,p}}(0)|\mathrm{p};p,s\rangle)^*. \quad (8.196)$$

We remarked in comment (a) after equation (8.193) that for e^- scattering from a point-like charged fermion an additional term in the cross section was present, corresponding to scattering from the target's magnetic moment. Since a real proton is not a point particle, the virtual strong interaction effects will modify both the charge and the magnetic moment distribution. Hence we may expect that *two* form factors will be needed to describe the deviation from point-like behaviour. This is in fact the case, as we now show using symmetry arguments similar to those of section 8.4.

8.8.1 Lorentz invariance

$B^{\mu\nu}$ must retain its tensor character: this must be made up using the available 4-vectors and tensors at our disposal. For the spin-averaged case we have only

$$p,\ q \text{ and } g_{\mu\nu} \quad (8.197)$$

since $p' = p + q$. The antisymmetric tensor $\epsilon_{\mu\nu\alpha\beta}$ (see appendix J) must actually be ruled out using parity invariance: the tensor $B^{\mu\nu}$ is not a pseudo tensor since $\hat{j}^{\mu}_{\mathrm{em,p}}$ is a vector. It is helpful to remember that $\epsilon_{\mu\nu\alpha\beta}$ is the generalization of ϵ_{ijk} in three dimensions, and that the vector product of two 3-vectors – a pseudo vector – may be written

$$(\boldsymbol{a}\times\boldsymbol{b})_i = \epsilon_{ijk}a_jb_k. \quad (8.198)$$

8.8.2 Current conservation

For a real proton, current conservation gives the condition (cf (8.148))

$$q_\mu \langle \mathrm{p}; p', s' | \hat{j}^\mu_{\mathrm{em,p}}(0) | \mathrm{p}; p, s \rangle = 0 \tag{8.199}$$

which translates to the conditions (cf (8.189))

$$q_\mu B^{\mu\nu} = q_\nu B^{\mu\nu} = 0 \tag{8.200}$$

on the tensor $B^{\mu\nu}$.

There are only two possible tensors we can make that satisfy both these requirements. One involves p and is constructed to be orthogonal to q. We introduce a vector

$$\tilde{p}_\mu = p_\mu + \alpha q_\mu \tag{8.201}$$

and require

$$q \cdot \tilde{p} = 0. \tag{8.202}$$

Hence we find

$$\tilde{p}_\mu = p_\mu - (p \cdot q/q^2) q_\mu \tag{8.203}$$

and thus the tensor

$$\tilde{p}^\mu \tilde{p}^\nu = [p^\mu - (p \cdot q/q^2) q^\mu][p^\nu - (p \cdot q/q^2) q^\nu] \tag{8.204}$$

satisfies all our requirements. The second tensor must involve $g^{\mu\nu}$ and may be chosen to be

$$-g^{\mu\nu} + q^\mu q^\nu / q^2 \tag{8.205}$$

which again satisfies our conditions. Thus from invariance arguments alone, the tensor $B^{\mu\nu}$ for the proton vertex may be parametrized by these two tensors, each multiplied by an unknown function of q^2. If we define

$$\begin{aligned} B^{\mu\nu} \quad = \quad & 4A(q^2)[p^\mu - (p \cdot q/q^2) q^\mu][p^\nu - (p \cdot q/q^2) q^\nu] \\ & + 2M^2 B(q^2)(-g^{\mu\nu} + q^\mu q^\nu / q^2) \end{aligned} \tag{8.206}$$

the cross section in the laboratory frame is (problem 8.19)

$$\frac{\mathrm{d}\sigma}{\mathrm{d}\Omega} = \left(\frac{\mathrm{d}\sigma}{\mathrm{d}\Omega} \right)_{\mathrm{ns}} [A + B \tan^2(\theta/2)]. \tag{8.207}$$

Formula (8.207) implies that a plot of $(\mathrm{d}\sigma/\mathrm{d}\Omega)/(\mathrm{d}\sigma/\mathrm{d}\Omega)_{\mathrm{ns}}$ versus $\tan^2 \theta/2$, at fixed q^2, will be a straight line with slope B and intercept A.

The functions A and B may be related to the 'charge' and 'magnetic' form factors of the proton. The Dirac 'charge' and Pauli 'anomalous magnetic moment' form factors, $\mathcal{F}_1$ and $\mathcal{F}_2$ respectively, are defined by

$$\begin{aligned} & \langle \mathrm{p}; p', s' | \hat{j}^\mu_{\mathrm{em,p}}(0) | \mathrm{p}; p, s \rangle \\ & \quad = (+e) \bar{u}(p', s') \left[\gamma^\mu \mathcal{F}_1(q^2) + \frac{\mathrm{i}\kappa \mathcal{F}_2(q^2)}{2M} \sigma^{\mu\nu} q_\nu \right] u(p, s) \end{aligned} \tag{8.208}$$

with the normalization

$$\mathcal{F}_1(0) = 1 \tag{8.209}$$
$$\mathcal{F}_2(0) = 1 \tag{8.210}$$

and the magnetic moment of the proton is not one (nuclear) magneton, as for an electron or muon (neglecting higher-order corrections), but rather $\mu_{\rm p} = 1 + \kappa$ with $\kappa = 1.79$. Problem 8.20 shows that the $\bar{u}\gamma^\mu u$ piece in (8.208) can be rewritten in terms of $\bar{u}(p+p')^\mu u/2M$ and $\bar{u}\mathrm{i}\sigma^{\mu\nu}q_\nu u/2M$. The first of these is analogous to the interaction of a charged spin-0 particle. As regards the second, we note that $\sigma^{\mu\nu}$ is just

$$\sigma^{\mu\nu} = \tfrac{1}{2}\mathrm{i}[\gamma^\mu, \gamma^\nu] \tag{8.211}$$

which reduces to the Pauli spin matrices for the space-like components

$$\sigma^{ij} = \begin{pmatrix} \sigma^k & 0 \\ 0 & \sigma^k \end{pmatrix} \tag{8.212}$$

with our representation of γ-matrices (σ^{ij} is a 4×4 matrix, σ^k is 2×2, and i, j and k are in cyclic order). The second term in this 'Gordon decomposition' of $\bar{u}\gamma^\mu u$ thus corresponds to an interaction via the spin magnetic moment – with, in fact, $g = 2$. Thus the addition of the κ term in (8.208) corresponds to an 'anomalous' magnetic moment piece. In terms of $\mathcal{F}_1$ and $\mathcal{F}_2$ one can show that

$$A = \mathcal{F}_1^2 + \tau\kappa^2\mathcal{F}_2^2 \tag{8.213}$$
$$B = 2\tau(\mathcal{F}_1 + \kappa\mathcal{F}_2)^2 \tag{8.214}$$

where

$$\tau = -q^2/4M^2. \tag{8.215}$$

The point-like cross section (8.193) is recovered from (8.207) by setting $\mathcal{F}_1 = 1$ and $\kappa = 0$ in (8.213) and (8.214).

The functions $\mathcal{F}_1$ and $\mathcal{F}_2$ are, in turn, usually expressed in terms of the electric and magnetic form factors $G_{\rm E}$ and $G_{\rm M}$, defined by $G_{\rm E} = \mathcal{F}_1 - \tau\kappa\mathcal{F}_2$, $G_{\rm M} = \mathcal{F}_1 + \kappa\mathcal{F}_2$. We then find $A = (G_{\rm E}^2 + \tau G_{\rm M}^2)/(1+\tau)$ and $B = 2\tau G_{\rm M}^2$. The cross section formula (8.207), written in terms of $G_{\rm E}$ and $G_{\rm M}$, is known as the 'Rosenbluth' cross section.

Experimental data indicate that the q^2-dependences of $G_{\rm E}$ and $G_{\rm M}$ for the proton, and of $G_{\rm M}$ for the neutron, are all quite well represented by the function $F(\boldsymbol{q}^2)$ of (8.136) with $\boldsymbol{q}^2$ replaced by $-q^2$ and with $a \sim 0.84$ GeV^{-1}, at least for values of $-q^2$ up to a few GeV^2 (see, for example, Perkins 1987, section 6.5).

Before we leave elastic scattering it is helpful to look in some more detail at the kinematics. It will be sufficient to consider the 'point-like' case, which

we shall call $e^-\mu^+$, for definiteness. Energy and momentum conservation at the μ^+ vertex gives the condition

$$p + q = p' \tag{8.216}$$

with the mass-shell conditions (M is the μ^+ mass)

$$p^2 = p'^2 = M^2. \tag{8.217}$$

Hence for elastic scattering we have the relation

$$2p \cdot q = -q^2. \tag{8.218}$$

It is conventional to relate these invariants to the corresponding laboratory frame ($p^\mu = (M, \mathbf{0})$) expressions. Neglecting the electron mass so that[2]

$$k \equiv |\boldsymbol{k}| = \omega \tag{8.219}$$

$$k' \equiv |\boldsymbol{k}'| = \omega' \tag{8.220}$$

we have

$$q^2 = -2kk'(1 - \cos\theta) = -4kk' \sin^2(\theta/2) \tag{8.221}$$

and

$$p \cdot q = M(k - k') = M\nu \tag{8.222}$$

where ν is the energy transfer q^0 in this frame. To avoid unnecessary minus signs, it is convenient to define

$$Q^2 = -q^2 = 4kk' \sin^2(\theta/2) \tag{8.223}$$

and the elastic scattering relation between $p \cdot q$ and q^2 reads

$$\nu = Q^2/2M \tag{8.224}$$

or

$$\frac{k'}{k} = \frac{1}{1 + (2k/M)\sin^2(\theta/2)}. \tag{8.225}$$

Remembering, therefore, that for elastic scattering k' and θ are not independent variables, we can perform a change of variables (see appendix K) in the laboratory frame

$$\mathrm{d}\Omega = 2\pi\, \mathrm{d}(\cos\theta) = (\pi/k'^2)\, \mathrm{d}Q^2 \tag{8.226}$$

and write the differential cross section for $e^-\mu^+$ scattering as

$$\frac{\mathrm{d}\sigma}{\mathrm{d}Q^2} = \frac{\pi\alpha^2}{4k^2 \sin^4(\theta/2)} \frac{1}{kk'} [\cos^2(\theta/2) + 2\tau \sin^2(\theta/2)]. \tag{8.227}$$

[2] As after equation (8.126), note again that in the present context 'k' and 'k'' are not 4-vectors but the moduli of 3-vectors.

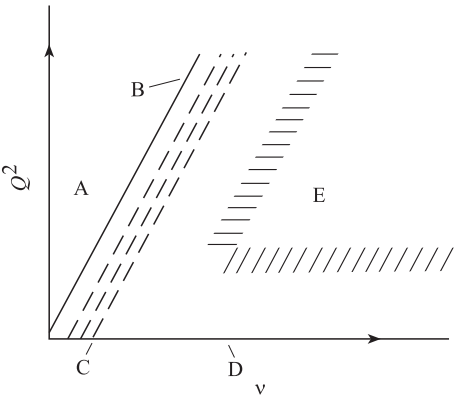


FIGURE 8.19
Physical regions for e^- p scattering in the Q^2, ν variables: A, kinematically forbidden region; B, line of elastic scattering ($Q^2 = 2M\nu$); C, lines of resonance electroproduction; D, photoproduction; E, deep inelastic region (Q^2 and ν large).

For elastic scattering ν is not independent of Q^2 but we may formally write this as a double-differential cross section by inserting the δ-function to ensure this condition is satisfied:

$$\boxed{\frac{\mathrm{d}^2\sigma}{\mathrm{d}Q^2\mathrm{d}\nu} = \frac{\pi\alpha^2}{4k^2\sin^4(\theta/2)}\frac{1}{kk'}\left[\cos^2(\theta/2) + \left(\frac{Q^2}{2M^2}\right)\sin^2(\theta/2)\right]\delta\left(\nu - \frac{Q^2}{2M}\right).} \tag{8.228}$$

This is the cross section for the scattering of an electron from a point-like fermion target of charge e and mass M.

It is illuminating to plot out the physically allowed regions of Q^2 and ν (figure 8.19). Elastic e^-p scattering corresponds to the line $Q^2 = 2M\nu$. Resonance production $e^-p \to e^-N^*$ with $p'^2 = M'^2$ corresponds to lines parallel to the elastic line, shifted to the right by $M'^2 - M^2$ since

$$2M\nu = Q^2 + M'^2 - M^2. \tag{8.229}$$

Experiments with real photons, $Q^2 = 0$, correspond to exploring along the ν-axis. In the next chapter we switch our attention to so-called deep inelastic electron scattering – the region of large Q^2 and large ν.

Problems

8.1 Consider a matrix element of the form

$$M = \int \mathrm{d}^3\boldsymbol{x} \int \mathrm{d}t\, \mathrm{e}^{+\mathrm{i}p_\mathrm{f}\cdot x}\partial_\mu A^\mu \mathrm{e}^{-\mathrm{i}p_\mathrm{i}\cdot x}.$$

Assuming the integration is over all space–time and that

$$A^0 \to 0 \qquad \text{as } t \to \pm\infty$$

and

$$|\boldsymbol{A}| \to 0 \qquad \text{as } |\boldsymbol{x}| \to \infty$$

use integration by parts to show

(a) $$\int \mathrm{d}t\, \mathrm{e}^{+\mathrm{i}p_\mathrm{f}\cdot x}\partial_0 A^0 \mathrm{e}^{-\mathrm{i}p_\mathrm{i}\cdot x} = (-\mathrm{i}p_{\mathrm{f}0}) \int \mathrm{d}t\, \mathrm{e}^{+\mathrm{i}p_\mathrm{f}\cdot x} A^0 \mathrm{e}^{-\mathrm{i}p_\mathrm{i}\cdot x}$$

(b) $$\int \mathrm{d}^3\boldsymbol{x}\, \mathrm{e}^{+\mathrm{i}p_\mathrm{f}\cdot x}\boldsymbol{\nabla}\cdot\boldsymbol{A}\mathrm{e}^{-\mathrm{i}p_\mathrm{i}\cdot x} = +\mathrm{i}\boldsymbol{p}_\mathrm{f}\cdot\left(\int \mathrm{d}^3\boldsymbol{x}\, \mathrm{e}^{+\mathrm{i}p_\mathrm{f}\cdot x}\boldsymbol{A}\mathrm{e}^{-\mathrm{i}p_\mathrm{i}\cdot x}\right).$$

Hence show that

$$\int \mathrm{d}^3\boldsymbol{x} \int \mathrm{d}t\, \mathrm{e}^{+\mathrm{i}p_\mathrm{f}\cdot x}(\partial_\mu A^\mu + A^\mu\partial_\mu)\mathrm{e}^{-\mathrm{i}p_\mathrm{i}\cdot x} = -\mathrm{i}(p_\mathrm{f}+p_\mathrm{i})_\mu \int \mathrm{d}^3\boldsymbol{x} \int \mathrm{d}t\, \mathrm{e}^{+\mathrm{i}p_\mathrm{f}\cdot x} A^\mu \mathrm{e}^{-\mathrm{i}p_\mathrm{i}\cdot x}.$$

8.2 Verify equation (8.27).

8.3 Evaluate (8.31) and interpret the result physically (i.e. compare it with (8.27)).

8.4

(a) Using the u-spinors normalized as in (3.73), the $\phi^{1,2}$ of (8.47), and the result for $\boldsymbol{\sigma}\cdot\boldsymbol{A}\boldsymbol{\sigma}\cdot\boldsymbol{B}$ from problem 3.4(b), show that

$$u^\dagger(k', s'=1)u(k, s=1) = (E+m)\left\{1 + \frac{\boldsymbol{k}'\cdot\boldsymbol{k}}{(E+m)^2} + \frac{\mathrm{i}\phi^{1\dagger}\boldsymbol{\sigma}\cdot\boldsymbol{k}'\times\boldsymbol{k}\phi^1}{(E+m)^2}\right\}.$$

(b) For any vector $\boldsymbol{A} = (A^1, A^2, A^3)$, show that $\phi^{1\dagger}\boldsymbol{\sigma}\cdot\boldsymbol{A}\phi^1 = A^3$. Find similar expressions for $\phi^{1\dagger}\boldsymbol{\sigma}\cdot\boldsymbol{A}\phi^2, \phi^{2\dagger}\boldsymbol{\sigma}\cdot\boldsymbol{A}\phi^1, \phi^{2\dagger}\boldsymbol{\sigma}\cdot\boldsymbol{A}\phi^2$.

(c) Show that the S of (8.46) is equal to

$$S = (E+m)^2\left\{\left[1 + \frac{\boldsymbol{k}'\cdot\boldsymbol{k}}{(E+m)^2}\right]^2 + \frac{(\boldsymbol{k}'\times\boldsymbol{k})^2}{(E+m)^4}\right\}.$$

(d) Using $\cos\theta = \boldsymbol{k}\cdot\boldsymbol{k}'/(|\boldsymbol{k}||\boldsymbol{k}'|)$, $|\boldsymbol{k}| = |\boldsymbol{k}'|$ and $v = |\boldsymbol{k}|/E$, show that

$$S = (2E)^2(1 - v^2\sin^2\theta/2).$$

8.5 Verify equation (8.55).

8.6 Check that $\gamma^0\gamma^{\mu\dagger}\gamma^0 = \gamma^\mu$.

8.7 Verify equation (8.79) for the lepton tensor $L^{\mu\nu}$.

8.8 Evaluate L^{00} as in equation (8.80).

8.9 Verify equation (8.87).

8.10 Verify equation (8.96) for the $e^-s^+ \to e^-s^+$ amplitude to $O(e^2)$.

8.11 Check that both the scalar and the spinor current matrix elements (8.27) and (8.55), satisfy $\partial_\mu j^\mu(x) = 0$.

8.12 Verify equation (8.120).

8.13 Verify equation (8.136) for the Fourier transform of $\rho(\boldsymbol{x})$ given by (8.135). Show that the mean square radius of the distribution (8.135) is $12a^2$.

8.14 Check the gauge invariance of $\mathcal{M}_{\gamma e^-}$ given by (8.162), by showing that if ϵ_μ is replaced by k_μ, or ϵ^*_ν by k'_ν, the result is zero.

8.15

(a) The spin-averaged squared amplitude for lowest-order electron Compton scattering contains the interference term

$$\sum_{\lambda,\lambda',s,s'} \mathcal{M}^{(s)}_{\gamma e^-}\mathcal{M}^{(u)*}_{\gamma e^-}$$

where (s) and (u) refer to the s- and u-channel processes of figure 8.14(a) and (b) respectively. Obtain an expression analogous to (8.172) for this term, and prove that it is, in fact, zero. [*Hint*: work in the massless limit, and use relations (J.4) and (J.5).]

(b) Explain why the term

$$\sum_{\lambda,\lambda',s,s'} \mathcal{M}^{(u)}_{\gamma e^-}\mathcal{M}^{(u)*}_{\gamma e^-}$$

is given by (8.177) with s and u interchanged.

8.16 Recalculate the interference term of problem 8.16(a) for the case $k^2 = -Q^2$ (but with $k'^2 = p^2 = p'^2 = 0$), and hence verify (8.181).

8.17

(a) Derive an expression for the spin-averaged differential cross section for lowest-order $e^-\mu^-$ scattering in the laboratory frame, defined by $p^\mu = (M, \mathbf{0})$ where M is now the muon mass, and show that it may be written in the form

$$\frac{d\sigma}{d\Omega} = \left(\frac{d\sigma}{d\Omega}\right)_{ns} [1 - (q^2/2M^2)\tan^2(\theta/2)]$$

where the 'no-structure' cross section is that of e^-s^+ scattering (appendix K) and the electron mass has been neglected.

(b) Neglecting *all* masses, evaluate the spin-averaged expression (8.184) in terms of s, t and u and use the result

$$\frac{d\sigma}{dt} = \frac{1}{16\pi s^2}\frac{1}{4}\sum_{r,r';s,s'} |\mathcal{M}_{e^-\mu^-}(r,s;r',s')|^2$$

to show that the $e^-\mu^-$ cross section may be written in the form

$$\frac{d\sigma}{dt} = \frac{4\pi\alpha^2}{t^2}\frac{1}{2}\left(1 + \frac{u^2}{s^2}\right).$$

Show also that by introducing the variable y, defined in terms of laboratory variables by $y = (k - k')/k$, this reduces to the result

$$\frac{d\sigma}{dy} = \frac{4\pi\alpha^2}{t^2} s \frac{1}{2}[1 + (1-y)^2].$$

8.18 Consider the process $e^+e^- \to \mu^+\mu^-$ in the CM frame.

(a) Draw the lowest-order Feynman diagram and write down the corresponding amplitude.

(b) Show that the spin-averaged squared matrix element has the form

$$\overline{|\mathcal{M}|^2} = \frac{(4\pi\alpha)^2}{q^4} L(e)_{\mu\nu} L(\mu)^{\mu\nu}$$

where q^2 is the square of the total CM energy, and $L(e)$ depends on the e^- and e^+ momenta and $L(\mu)$ on those of the μ^+, μ^-.

(c) Evaluate the traces and the tensor contraction (neglecting lepton masses): (i) directly, using the trace theorems; and (ii) by using crossing symmetry and the results of section 8.7 for $e^-\mu^-$ scattering. Hence show that

$$\overline{|\mathcal{M}|^2} = (4\pi\alpha)^2(1 + \cos^2\theta)$$

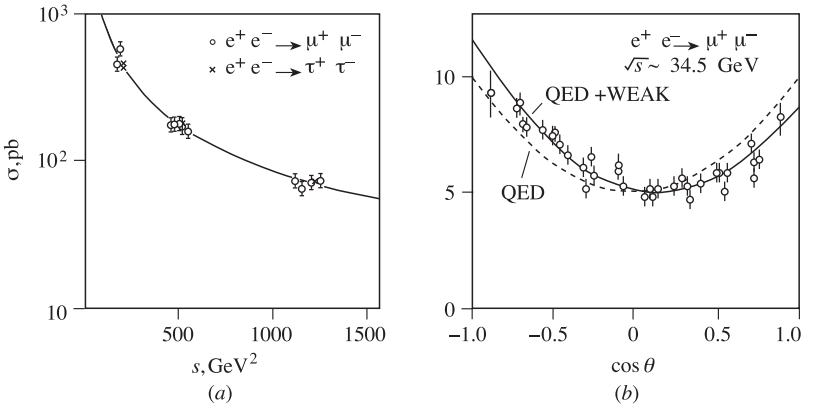


FIGURE 8.20
(a) Total cross sections for $e^+e^- \to \mu^+\mu^-$ and $e^+e^- \to \tau^+\tau^-$; (b) differential cross section for $e^+e^- \to \mu^+\mu^-$. (From D H Perkins 2000 *Introduction to High Energy Physics* 4th edn, courtesy Cambridge University Press.)

where θ is the CM scattering angle, and that the CM differential cross section is

$$\frac{\mathrm{d}\sigma}{\mathrm{d}\Omega} = \frac{\alpha^2}{4q^2}(1 + \cos^2\theta).$$

(d) Hence show that the total cross section is (see equation (B.18) of appendix B)

$$\sigma = 4\pi\alpha^2/3q^2 = 86.8 \text{ nb}/q^2(\mathrm{GeV}^2).$$

Figure 8.20 shows data (a) for σ in $e^+e^- \to \mu^+\mu^-$ and $e^+e^- \to \tau^+\tau^-$ and (b) for the angular distribution in $e^+e^- \to \mu^+\mu^-$. Note that $s = q^2$. The data in figure 8.20(a) agree well with the prediction above for σ. The broken curve in figure 8.20(b) shows the pure QED prediction of part (c) for $\frac{\mathrm{d}\sigma}{\mathrm{d}\Omega}$.

It is clear that, while the distribution has the general $1+\cos^2\theta$ form as predicted, there is a small but definite forward–backward asymmetry. This arises because, in addition to the γ-exchange amplitude there is also a Z^0-exchange amplitude (see section 22.3 of volume 2) which we have neglected. Such asymmetries are an important test of the electroweak theory. They are too small to be visible in the total cross sections in figure 8.20(a).

8.19 Verify equation (8.207). [*Hint*: as in equation (8.191) the terms in q^μ and q^ν in $B^{\mu\nu}$ may be neglected because of the conditions (8.189).]

8.20 Starting from the expression

$$\bar{u}(p')\mathrm{i}\frac{\sigma^{\mu\nu}}{2M}q_\nu u(p)$$

where $q = p' - p$ and $\sigma^{\mu\nu} = \frac{1}{2}\mathrm{i}[\gamma^\mu, \gamma^\nu]$, use the Dirac equation and properties of γ-matrices to prove the 'Gordon decomposition' of the current

$$\bar{u}(p')\gamma^\mu u(p) = \bar{u}(p')\left(\frac{(p+p')^\mu}{2M} + \mathrm{i}\frac{\sigma^{\mu\nu}q_\nu}{2M}\right)u(p).$$


Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

9

Deep Inelastic Electron–Nucleon Scattering and the Parton Model

We have obtained the rules for doing calculations of simple processes in quantum electrodynamics for particles of spin-0 and spin-$\frac{1}{2}$, and many explicit examples have been considered. In this chapter we build on these results to give an (admittedly brief) introduction to a topic of central importance in particle physics, the structure of hadrons as revealed by deep inelastic scattering experiments (the equally important neutrino scattering experiments will be discussed in volume 2). We do this partly because the necessary calculations involve straightforward, illustrative and eminently practical applications of the rules already obtained, but, more particularly, because it is from a comparison of these calculations with experiment that compelling evidence was obtained for the existence of the point-like constituents of hadrons – quarks and gluons – the interactions of which are described by QCD.

9.1 Inelastic electron–proton scattering: kinematics and structure functions

At large momentum transfers there is very little elastic scattering: inelastic scattering, in which there is more than just the electron and proton in the final state, is much more probable. The simplest inelastic cross section to measure is the so-called 'inclusive' cross section, for which only the final electron is observed. This is therefore a sum over the cross sections for all the possible hadronic final states: no attempt is made to select any particular state from the hadronic debris created at the proton vertex. This process may be represented by the diagram of figure 9.1, assuming that the one-photon exchange amplitude dominates. The 'blob' at the proton vertex indicates our ignorance of the detailed structure: X indicates a sum over all possible hadronic final states. However, the assumption of one-photon exchange, which is known experimentally to be a very good approximation, means that, as in our previous examples (cf (8.118) and (8.185)), the cross section must factorize into a leptonic tensor contracted with a tensor describing the hadron vertex:

$$\mathrm{d}\sigma \sim L_{\mu\nu} W^{\mu\nu}(q, p). \tag{9.1}$$

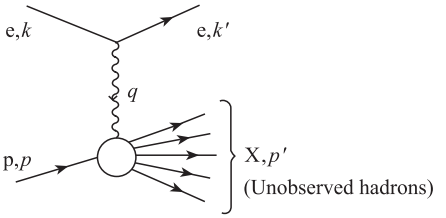


FIGURE 9.1
Inelastic electron–proton scattering, in one-photon exchange approximation.

The lepton vertex is well described by QED and takes the same form as before:

$$L_{\mu\nu} = 2[k'_\mu k_\nu + k'_\nu k_\mu + (q^2/2)g_{\mu\nu}]. \tag{9.2}$$

For the hadron tensor, however, we expect strong interactions to play an important role and we must deduce its general structure by our powerful invariance arguments. We will only consider unpolarized scattering and therefore perform an average over the initial proton spins. The sum over final states, X, includes all possible quantum numbers for each hadronic state with total momentum p'. For an inclusive cross section, the final phase space involves only the scattered electron. Moreover, since we are not restricting the scattering process by picking out any specific state of X, the energy k' and the scattering angle θ of the final electron are now independent variables. In $W^{\mu\nu}(q,p)$ the sum over X includes the phase space for each hadronic state restricted by the usual 4-momentum-conserving δ-function to ensure that each state in X has momentum p'. Including some conventional factors, we define $W^{\mu\nu}(q,p)$ by (see problem 9.1)

$$\begin{aligned} e^2 W^{\mu\nu}(q,p) &= \frac{1}{4\pi M}\frac{1}{2}\sum_s \sum_{\mathrm{X}} \langle \mathrm{p};p,s|\hat{j}^\mu_{\mathrm{em,p}}(0)|\mathrm{X};p'\rangle\langle \mathrm{X};p'|\hat{j}^\nu_{\mathrm{em,p}}(0)|\mathrm{p};p,s\rangle \\ &\quad \times (2\pi)^4\delta^4(p+q-p'). \end{aligned} \tag{9.3}$$

How do we parametrize the tensor structure of $W^{\mu\nu}$? As usual, Lorentz invariance and current conservation come to our aid. There is one important difference compared with the elastic form factor case of section 8.8. For inclusive inelastic scattering there are now two independent scalar variables. The relation

$$p' = p + q \tag{9.4}$$

leads to

$$p'^2 = M^2 + 2p\cdot q + q^2 \tag{9.5}$$

where M is the proton mass. In this case, the invariant mass of the hadronic final state is a *variable*

$$p'^2 \equiv W^2 \tag{9.6}$$

and is related to the other two scalar variables

$$p \cdot q = M\nu \tag{9.7}$$

and (cf (8.223))

$$q^2 = -Q^2 \tag{9.8}$$

by the condition (cf (8.229))

$$2M\nu = Q^2 + W^2 - M^2. \tag{9.9}$$

Our invariance arguments lead us to the same *tensor* structure as for *elastic* electron–proton scattering, but now the functions $A(q^2)$, $B(q^2)$ are replaced by 'structure functions' which are functions of two variables, usually taken to be ν and Q^2. The conventional definition of the proton structure functions W_1 and W_2 is

$$\boxed{\begin{aligned} W^{\mu\nu}(q,p) = {} & (-g^{\mu\nu} + q^\mu q^\nu/q^2)W_1(Q^2,\nu) \\ & + [p^\mu - (p\cdot q/q^2)q^\mu][p^\nu - (p\cdot q/q^2)q^\nu]M^{-2}W_2(Q^2,\nu). \end{aligned}} \tag{9.10}$$

Inserting the usual flux factor together with the final electron phase space leads to the following expression for the inclusive differential cross section for inelastic electron–proton scattering (see problem 9.1):

$$\mathrm{d}\sigma = \left(\frac{4\pi\alpha}{q^2}\right)^2 \frac{1}{4[(k\cdot p)^2 - m^2M^2]^{1/2}} 4\pi M L_{\mu\nu} W^{\mu\nu} \frac{\mathrm{d}^3k'}{2\omega'(2\pi)^3}. \tag{9.11}$$

In terms of 'laboratory' variables, neglecting electron mass effects, this yields (problem 9.2(a))

$$\boxed{\frac{\mathrm{d}^2\sigma}{\mathrm{d}\Omega \mathrm{d}k'} = \frac{\alpha^2}{4k^2\sin^4(\theta/2)}[W_2\cos^2(\theta/2) + 2W_1\sin^2(\theta/2)].} \tag{9.12}$$

Remembering now that $\cos\theta$ and k' are independent variables for inelastic scattering, we can change variables from $\cos\theta$ and k' to Q^2 and ν, assuming azimuthal symmetry for the unpolarized cross section. We have

$$Q^2 = 2kk'(1-\cos\theta) \tag{9.13}$$

$$\nu = k - k' \tag{9.14}$$

so that (problem 9.2(b))

$$\mathrm{d}(\cos\theta)\,\mathrm{d}k' = \frac{1}{2kk'}\mathrm{d}Q^2\,\mathrm{d}\nu \tag{9.15}$$

and

$$\boxed{\frac{\mathrm{d}^2\sigma}{\mathrm{d}Q^2\mathrm{d}\nu} = \frac{\pi\alpha^2}{4k^2\sin^4(\theta/2)}\frac{1}{kk'}[W_2\cos^2(\theta/2) + 2W_1\sin^2(\theta/2)].} \tag{9.16}$$

Yet another choice of variables is sometimes used instead of these, namely the dimensionless variables

$$x = Q^2/2M\nu \tag{9.17}$$

whose significance we shall see in the next section, and

$$y = \nu/k \tag{9.18}$$

which is the fractional energy transfer in the 'laboratory' frame. Note that relation (8.224) shows that $x = 1$ for elastic scattering. The Jacobian for the transformation from Q^2 and ν to x and y is (see problem 9.2(b))

$$\mathrm{d}Q^2\,\mathrm{d}\nu = 2Mk^2 y\,\mathrm{d}x\,\mathrm{d}y. \tag{9.19}$$

We emphasize that the foregoing – in particular (9.3), (9.12) and (9.16) – is all completely general, given the initial one-photon approximation. The physics is all contained in the ν and Q^2 dependence of the two structure functions W_1 and W_2.

A priori, one might expect W_1 and W_2 to be complicated functions of ν and Q^2, reflecting the complexity of the inelastic scattering process. However, in 1969 Bjorken predicted that in the 'deep inelastic region' – large ν and Q^2, but Q^2/ν finite – there should be a very simple behaviour. He predicted that the structure functions should *scale*, i.e. become functions not of Q^2 and ν independently but only of their ratio Q^2/ν. It was the verification of approximate 'Bjorken scaling' that led to the development of the modern parton model. We therefore specialize our discussion of inelastic scattering to the deep inelastic region.

9.2 Bjorken scaling and the parton model

From considerations based on the quark model current algebra of Gell-Mann (1962), Bjorken (1969) was led to propose the following 'scaling hypothesis': in the limit

$$\left.\begin{aligned} Q^2 &\to \infty \\ \nu &\to \infty \end{aligned}\right\} \qquad \text{with } x = Q^2/2M\nu \text{ fixed} \tag{9.20}$$

the structure functions scale as

$$MW_1(Q^2,\nu) \;\to\; F_1(x) \tag{9.21}$$

$$\nu W_2(Q^2,\nu) \;\to\; F_2(x). \tag{9.22}$$

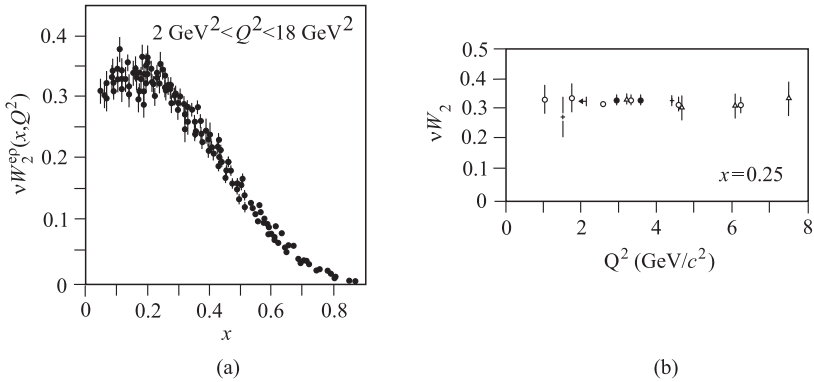


FIGURE 9.2
Bjorken scaling: the structure function νW_2 (a) plotted against x for different Q^2 values (Attwood 1980, courtesy SLAC) and (b) plotted against Q^2 for the single x value, $x = 0.25$ (Friedman and Kendall 1972).

We must emphasize that the *physical* content of Bjorken's hypothesis is that the functions $F_1(x)$ and $F_2(x)$ are *finite*[1].

Early experimental support for these predictions (figure 9.2) led initially to an examination of the theoretical basis of Bjorken's arguments and to the formulation of the simple intuitive picture provided by the parton model. Closer scrutiny of figure 9.2(a) will encourage the (correct) suspicion that, in fact, there is a small but significant spread in the data for any given x value. In volume 2 we shall give an introduction to the way in which QCD corrections to the parton model lead to predictions for logarithmic (in Q^2) violations of simple scaling behaviour, which are in excellent agreement with experiment. These violations are particularly large at small values of x; for x greater than about 0.1, the structure functions are substantially independent of Q^2, for a given x. The scaling predicted by Bjorken is certainly the most immediate gross feature of the data, and an understanding of it is of fundamental importance.

How can the scaling be understood? Feynman, when asked to explain Bjorken's arguments, gave an intuitive explanation in terms of elastic scattering from free point-like constituents of the nucleon, which he dubbed 'partons' (Feynman 1969). The essence of the argument lies in the *kinematics of elastic scattering of electrons by free point-like charged partons*: we will therefore be able to use the results of the previous chapters to derive the parton model results. At high Q^2 and ν it is intuitively reasonable (and in fact the basis for

[1]It is always possible to write $W(Q^2, \nu) = f(x, Q^2)$, say, where $f(x, Q^2)$ will tend to some function $F(x)$ as $Q^2 \to \infty$ with x fixed. $F(x)$ may, however, be zero, finite or infinite. The physics lies in the hypothesis that, in this limit, a finite part remains.

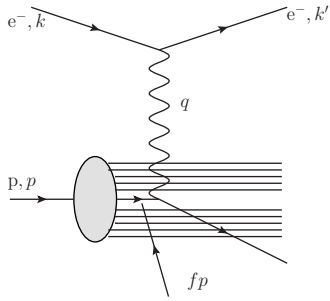


FIGURE 9.3
Photon–parton interaction.

the light-cone and short-distance operator approach (Wilson 1969) to scaling) that the virtual photon is probing very short distances and time scales within the proton. In this situation, Feynman supposed that the photon interacts with small (point-like) constituents within the proton, which carry only a certain fraction f of the proton's energy and momentum (figure 9.3). Over the short time scales involved in the transfer of a large amount of energy ν, and at the short distances probed at large Q^2, the struck constituents can perhaps be treated as effectively free and independent. (This is in sharp contrast to the case of elastic scattering, where the constituents are acting coherently.) We then have the idealized elastic scattering process shown in figure 9.4. It is the kinematics of the elastic scattering condition for the partons that leads directly to a relation between Q^2 and ν and hence to the observed scaling behaviour. The original discussion of the parton model took place in the infinite-momentum frame of the proton. While this has the merit that it eliminates the need for explicit statements about parton masses and so on, it also obscures the simple kinematic origin of the scaling. For this reason, at the expense of some theoretical niceties, we prefer to perform a direct calculation of electron–parton scattering in close analogy with our previous examples.

We first show that the fraction f is none other than Bjorken's variable x. For a parton of type i we write

$$p_i^\mu \approx f p^\mu \tag{9.23}$$

and, roughly speaking[2], we can imagine that the partons have mass

$$m_i \approx f M. \tag{9.24}$$

Then, exactly as in (8.216) and (8.217), energy and momentum conservation

[2]Explicit statements about parton transverse momenta and masses, such as those made in equations (9.23) and (9.24), are unnecessary in a rigorous treatment, where such quantities can be shown to give rise to non-leading scaling behaviour (Sachrajda 1983).

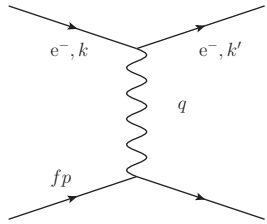


FIGURE 9.4
Elastic electron–parton scattering.

at the parton vertex, together with the assumption that the struck parton remains on-shell (as indicated by the fact that in figure 9.4 the partons are free), imply that

$$(q + fp)^2 = m_i^2 \tag{9.25}$$

which, using (9.8), (8.222) and (9.24), gives

$$f = Q^2/2M\nu \equiv x. \tag{9.26}$$

Thus the fact that the nucleon structure functions do seem to depend (to a good approximation) only on the variable x is interpreted physically as showing that the scattering is dominated by the 'quasi-free' electron–parton process shown in figure 9.4. In section 11.5.3 we shall see how the 'asymptotic freedom' property of QCD suggests a dynamical understanding of this picture, as will be discussed further in chapter 15 of volume 2.

What sort of values for x do we expect? Consider an analogous situation – electron scattering from deuterium. Here the target (the deuteron) is undoubtedly composite, and its 'partons' are, to a first approximation, just the two nucleons. Since $m_{\rm N} \simeq \frac{1}{2}m_{\rm D}$, we expect to see the value $x \simeq \frac{1}{2}$ (cf (9.24)) favoured; $x = 1$ here would correspond to elastic scattering from the deuteron. A peak at $x \approx \frac{1}{2}$ is indeed observed (figure 9.5) in quasi-elastic $\mathrm{e}^-\mathrm{d}$ scattering (the broadening of the peak is due to the fact that the constituent nucleons have some motion within the deuteron). By 'quasi-elastic' here we mean that the incident electron scatters off 'quasi-free' nucleons, an approximation we expect to be good for incident energies significantly greater than the binding energy of the n and p in the deuteron (∼2 MeV). What about the nucleon itself, then? A simple three-quark model would, on this analogy, lead us to expect a peak at $x \simeq \frac{1}{3}$, but the data already shown (figure 9.2(a)) do not look much like that. Perhaps there is something else present too – which we shall uncover as our story proceeds.

Certainly it seems sensible to suppose that a nucleon contains *at least* some quarks (and also antiquarks) of the type introduced in the simple composite models of the nucleon (section 1.2.2). If quarks are supposed to have spin-$\frac{1}{2}$, then the scattering of an electron from a quark or antiquark – generically a

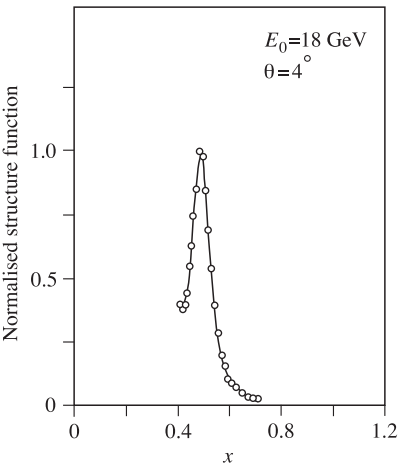


FIGURE 9.5
Structure function for quasi-elastic ed scattering, plotted against x (Attwood 1980, courtesy SLAC).

charged parton – of type i, charge e_i (in units of e) is just given by the eμ scattering cross section (8.228), with obvious modifications:

$$\frac{\mathrm{d}^2\sigma^i}{\mathrm{d}Q^2\mathrm{d}\nu} = \frac{\pi\alpha^2}{4k^2\sin^4(\theta/2)}\frac{1}{kk'}\left(e_i^2\cos^2(\theta/2) + e_i^2\frac{Q^2}{4m_i^2}2\sin^2(\theta/2)\right) \times \delta(\nu - Q^2/2m_i). \tag{9.27}$$

This is to be compared with the general inclusive inelastic cross section formula written in terms of W_1 and W_2:

$$\frac{\mathrm{d}^2\sigma}{\mathrm{d}Q^2\mathrm{d}\nu} = \frac{\pi\alpha^2}{4k^2\sin^4(\theta/2)}\frac{1}{kk'}[W_2\cos^2(\theta/2) + W_1 2\sin^2(\theta/2)]. \tag{9.28}$$

Thus the contribution to W_1 and W_2 from one parton of type i is immediately seen to be

$$W_1^i = e_i^2\frac{Q^2}{4M^2x^2}\delta(\nu - Q^2/2Mx) \tag{9.29}$$

$$W_2^i = e_i^2\delta(\nu - Q^2/2Mx) \tag{9.30}$$

where we have set $m_i = xM$. At large ν and Q^2 it is assumed that the contributions from different partons add incoherently in cross section. Thus, to obtain the total contribution from all quark partons, we must sum over the contributions from all types of partons, i, and integrate over all values of x, the momentum fraction carried by the parton. The integral over x must be

weighted by the probability $f_i(x)$ for the parton of type i to have a fraction x of momentum. These probability distributions – or *parton distribution functions* (PDFs) – are not predicted by the model and are, in this parton picture, fundamental parameters of the proton. The structure function W_2 becomes

$$W_2(\nu, Q^2) = \sum_i \int_0^1 \mathrm{d}x\, f_i(x) e_i^2 \delta(\nu - Q^2/2Mx). \tag{9.31}$$

Using the result for the Dirac δ-function (see appendix E, equation (E.34))

$$\delta(g(x)) = \frac{\delta(x - x_0)}{|\mathrm{d}g/\mathrm{d}x|_{x=x_0}} \tag{9.32}$$

where x_0 is defined by $g(x_0) = 0$, we can rewrite

$$\delta(\nu - Q^2/2Mx) = (x/\nu)\delta(x - Q^2/2M\nu) \tag{9.33}$$

under the x integral. Hence we obtain

$$\nu W_2(\nu, Q^2) = \sum_i e_i^2 x f_i(x) \equiv F_2(x) \tag{9.34}$$

which is the desired scaling behaviour. Similar manipulations lead to

$$MW_1(\nu, Q^2) = F_1(x) \tag{9.35}$$

where

$$2xF_1(x) = F_2(x). \tag{9.36}$$

This relation between F_1 and F_2 is called the Callan–Gross relation (see Callan and Gross 1969): it is a direct consequence of our assumption of spin-$\frac{1}{2}$ partons. The physical origin of this relation is best discussed in terms of virtual photon total cross sections for transverse ($\lambda = \pm 1$) virtual photons and for a longitudinal/scalar ($\lambda = 0$) virtual photon contribution. The longitudinal/scalar photon is present because $q^2 \neq 0$ for a virtual photon (see comment (4) in section 8.3.1). However, in the discussion of polarization vectors a slight difference occurs for space-like q^2. In a frame in which

$$q^\mu = (q^0, 0, 0, q^3) \tag{9.37}$$

the transverse polarization vectors are as before

$$\epsilon^\mu(\lambda = \pm 1) = \mp 2^{-1/2}(0, 1, \pm \mathrm{i}, 0) \tag{9.38}$$

with normalization (see equation (7.87))

$$\epsilon^* \cdot \epsilon = -1. \tag{9.39}$$

To construct the longitudinal/scalar polarization vector, we must satisfy

$$q \cdot \epsilon = 0 \tag{9.40}$$

and so are led to the result

$$\epsilon^\mu(\lambda = 0) = (1/\sqrt{Q^2})(q^3, 0, 0, q^0) \tag{9.41}$$

with

$$\epsilon^2(\lambda = 0) = +1. \tag{9.42}$$

The precise definition of a virtual photon cross section is obviously just a convention. It is usually taken to be

$$\sigma_\lambda(\gamma \mathrm{p} \to \mathrm{X}) = (4\pi^2\alpha/K)\epsilon^*_\mu(\lambda)\epsilon_\nu(\lambda)W^{\mu\nu} \tag{9.43}$$

by analogy with the total cross section for real photons of polarization λ incident on an unpolarized proton target. Note the presence of the factor $W^{\mu\nu}$ defined in (9.3). The factor K is the flux factor; for real photons, producing a final state of mass W, this is just the photon energy in the rest frame of the target nucleon:

$$K = (W^2 - M^2)/2M. \tag{9.44}$$

In the so-called 'Hand convention', this same factor is used for virtual photons which produce a final state of mass W. With these definitions we find (see problem 9.3) that the transverse ($\lambda = \pm 1$) photon cross section

$$\sigma_\mathrm{T} = \left(\frac{4\pi^2\alpha}{K}\right)\frac{1}{2}\sum_{\lambda=\pm1}\epsilon^*_\mu(\lambda)\epsilon_\nu(\lambda)W^{\mu\nu} \tag{9.45}$$

is given by

$$\sigma_\mathrm{T} = (4\pi^2\alpha/K)W_1 \tag{9.46}$$

and the longitudinal/scalar cross section

$$\sigma_\mathrm{S} = (4\pi^2\alpha/K)\epsilon^*_\mu(\lambda = 0)\epsilon_\nu(\lambda = 0)W^{\mu\nu} \tag{9.47}$$

by

$$\sigma_\mathrm{S} = (4\pi^2\alpha/K)[(1 + \nu^2/Q^2)W_2 - W_1]. \tag{9.48}$$

In fact these expressions give an intuitive explanation of the positivity properties of W_1 and W_2, namely

$$W_1 \geq 0 \tag{9.49}$$

$$(1 + \nu^2/Q^2)W_2 - W_1 \geq 0. \tag{9.50}$$

The combination in the $\lambda = 0$ cross section is sometimes denoted by W_L:

$$W_\mathrm{L} = (1 + \nu^2/Q^2)W_2 - W_1. \tag{9.51}$$

The scaling limit of these expressions can be taken using

$$\nu W_2 \quad \to \quad F_2 \tag{9.52}$$

$$MW_1 \quad \to \quad F_1 \tag{9.53}$$

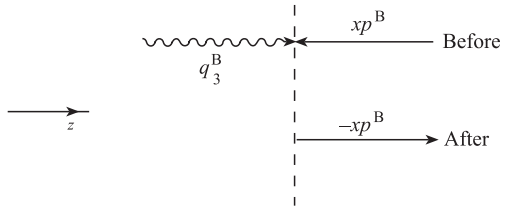


FIGURE 9.6
Photon–parton interaction in the Breit frame.

and $x = Q^2/2M\nu$ finite, as Q^2 and ν grow large. We find

$$\sigma_{\mathrm{T}} \to \frac{4\pi^2\alpha}{MK} F_1(x) \tag{9.54}$$

and

$$\sigma_{\mathrm{S}} \to (4\pi^2\alpha/MK)(1/2x)(F_2 - 2xF_1) \tag{9.55}$$

where we have neglected a term of order MF_2/ν in the last expression. Thus the Callan–Gross relation corresponds to the result

$$\sigma_{\mathrm{S}}/\sigma_{\mathrm{T}} \to 0 \tag{9.56}$$

in terms of photon cross sections.

A parton calculation using point-like spin-0 partons shows the opposite result, namely

$$\sigma_{\mathrm{T}}/\sigma_{\mathrm{S}} \to 0. \tag{9.57}$$

Both these results may be understood by considering the helicities of partons and photons in the so-called parton Breit or 'brick-wall' frame. The particular frame is the one in which the photon and parton are collinear and the 3-momentum of the parton is exactly reversed by the collision (see figure 9.6). In this frame, the photon transfers no energy, only 3-momentum. The vanishing of transverse photon cross sections for scalar partons is now obvious. The transverse photons bring in ± 1 units of the z-component of angular momentum: spin-0 partons cannot absorb this. Thus only the scalar $\lambda = 0$ cross section is non-zero. For spin-$\frac{1}{2}$ partons the argument is slightly more complicated in that it depends on the helicity properties of the γ_μ coupling of the parton to the photon. As is shown in problem 9.4, for massless spin-$\frac{1}{2}$ particles the γ_μ coupling conserves helicity – i.e. the projection of spin along the direction of motion of the particle. Thus in the Breit frame, and neglecting parton masses, conservation of helicity necessitates a change in the z-component of the parton's angular momentum by ± 1 unit, thereby requiring the absorption of a transverse photon (figure 9.7). The Lorentz transformation from the parton Breit frame to the 'laboratory' frame does not affect the ratio of transverse to longitudinal photons, if we neglect the parton transverse momenta.

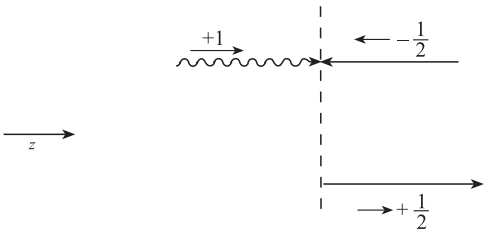


FIGURE 9.7
Angular momentum balance for absorption of photon by helicity-conserving spin-$\frac{1}{2}$ parton.

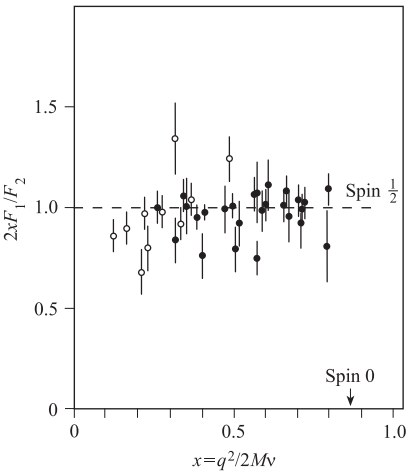


FIGURE 9.8
The ratio $2xF_1/F_2$: ○, $1.5 < Q^2 < 4$ GeV^2; •, $0.5 < Q^2 < 11$ GeV^2; ×, $12 < Q^2 < 16$ GeV^2. (Figure from D H Perkins *Introduction to High Energy Physics* 3rd edn, copyright 1987; reprinted by permission of Pearson Education, Inc., Upper Saddle River, NJ.)

These arguments therefore make clear the origin of the Callan–Gross relation. Experimentally, the Callan–Gross relation is reasonably well satisfied in that $R = \sigma_S/\sigma_T$ is small for most, if not all, of the deep inelastic regime (figure 9.8). This leads us to suppose that the *electrically charged partons coupling to photons have spin-*$\frac{1}{2}$.

9.3 Partons as quarks and gluons

We now proceed a stage further, with the idea that the charged partons *are* quarks (and antiquarks). If we assume that the photon only couples to these objects, we can make more specific scaling predictions. The quantum numbers of the quarks have been given in Table 1.2. For a proton we have the result (cf (9.34))

$$F_2^{\mathrm{ep}}(x) = x\{\tfrac{4}{9}[u(x) + \bar{u}(x)] + \tfrac{1}{9}[d(x) + \bar{d}(x) + s(x) + \bar{s}(x)] + \cdots\} \qquad (9.58)$$

where $u(x)$ is the probability distribution for u quarks in the proton, $\bar{u}(x)$ for u antiquarks and so on in an obvious notation, and the dots indicate further possible flavours. So far we do not seem to have gained much, replacing one unknown function by six or more unknown functions. The full power of the quark parton model lies in the fact that the same distribution functions appear, in different combinations, for neutron targets, and in the analogous scaling functions for deep inelastic scattering with neutrino and antineutrino beams (see volume 2). For electron scattering from neutron targets we can use I-spin invariance (see for example Close 1979, or Leader and Predazzi 1996) to relate the distribution of u and d quarks in a neutron to the distributions in a proton, and similarly for the antiquarks. The results are

$$u^{\mathrm{p}}(x) = d^{\mathrm{n}}(x) \equiv u(x) \quad d^{\mathrm{p}}(x) = u^{\mathrm{n}}(x) \equiv d(x) \qquad (9.59)$$

$$\bar{d}^{\mathrm{p}}(x) = \bar{u}^{\mathrm{n}}(x) \equiv \bar{d}(x) \quad \bar{u}^{\mathrm{p}}(x) = \bar{d}^{\mathrm{n}}(x) \equiv \bar{u}(x) \qquad (9.60)$$

$$s^{\mathrm{p}}(x) = s^{\mathrm{n}}(x) \equiv s(x) \quad \bar{s}^{\mathrm{p}}(x) = \bar{s}^{\mathrm{n}}(x) \equiv \bar{s}(x). \qquad (9.61)$$

Hence the scaling function for en scattering may be written

$$F_2^{\mathrm{en}}(x) = x\{\tfrac{4}{9}[d(x) + \bar{d}(x)] + \tfrac{1}{9}[u(x) + \bar{u}(x) + s(x) + \bar{s}(x)] + \cdots\}. \qquad (9.62)$$

The quark distributions inside the proton and neutron must satisfy some constraints. Since both proton and neutron have strangeness zero, we have a *sum rule* (treating only u, d and s flavours from now on)

$$\int_0^1 \mathrm{d}x\,[s(x) - \bar{s}(x)] = 0. \qquad (9.63)$$

Similarly, from the proton and neutron charges we obtain two other sum rules:

$$\int_0^1 \mathrm{d}x\,\{\tfrac{2}{3}[u(x) - \bar{u}(x)] - \tfrac{1}{3}[d(x) - \bar{d}(x)]\} = 1 \qquad (9.64)$$

$$\int_0^1 \mathrm{d}x\,\{\tfrac{2}{3}[d(x) - \bar{d}(x)] - \tfrac{1}{3}[u(x) - \bar{u}(x)]\} = 0. \qquad (9.65)$$

These are equivalent to the sum rules

$$2 = \int_0^1 \mathrm{d}x\,[u(x) - \bar{u}(x)] \tag{9.66}$$

$$1 = \int_0^1 \mathrm{d}x\,[d(x) - \bar{d}(x)] \tag{9.67}$$

which are, of course, just the excess of u and d quarks over antiquarks inside the proton. Testing these sum rules requires neutrino data to separate the various structure functions, as we shall explain in volume 2, chapter 20.

One can gain some further insight if one is prepared to make a model. For example, one can introduce the idea of 'valence' quarks (those of the elementary constituent quark model) and 'sea' quarks ($\mathrm{q\bar{q}}$ pairs created virtually). Then, in a proton, the u and d quark distributions would be parametrized by the sum of valence and sea contributions

$$u = u_{\mathrm{V}} + q_{\mathrm{S}} \tag{9.68}$$

$$d = d_{\mathrm{V}} + q_{\mathrm{S}} \tag{9.69}$$

while the antiquark and strange quark distributions are taken to be pure sea

$$\bar{u} = \bar{d} = s = \bar{s} = q_{\mathrm{S}} \tag{9.70}$$

where we have assumed that the 'sea' is flavour-independent. Such a model replaces the six unknown functions now in play by three, and is consequently more predictive. The strangeness sum rule (9.63) is now satisfied automatically, while (9.66) and (9.67) are satisfied by the valence distributions alone:

$$\int_0^1 \mathrm{d}x\, u_{\mathrm{V}}(x) = 2 \tag{9.71}$$

$$\int_0^1 \mathrm{d}x\, d_{\mathrm{V}}(x) = 1. \tag{9.72}$$

One more important sum rule emerges from the picture of $xf_i(x)$ as the fractional momentum carried by quark i. This is the *momentum sum rule*

$$\int_0^1 \mathrm{d}x\, x[u(x) + \bar{u}(x) + d(x) + \bar{d}(x) + s(x) + \bar{s}(x)] = 1 - \epsilon \tag{9.73}$$

where ϵ is interpreted as the fraction of the proton momentum that is not carried by quarks and antiquarks. The integral in (9.73) is directly related to ν and $\bar{\nu}$ cross sections, and its evaluation implies $\epsilon \simeq \frac{1}{2}$ (the CHARM (1981) result was $1 - \epsilon = 0.44 \pm 0.02$). This suggests that about half the total momentum is carried by *uncharged* objects. These remaining partons are identified with the gluons of QCD. They have their own PDF, $g(x)$.

An enormous effort, both experimental and theoretical, has gone into determining the parton distribution functions. The subject is regularly reviewed

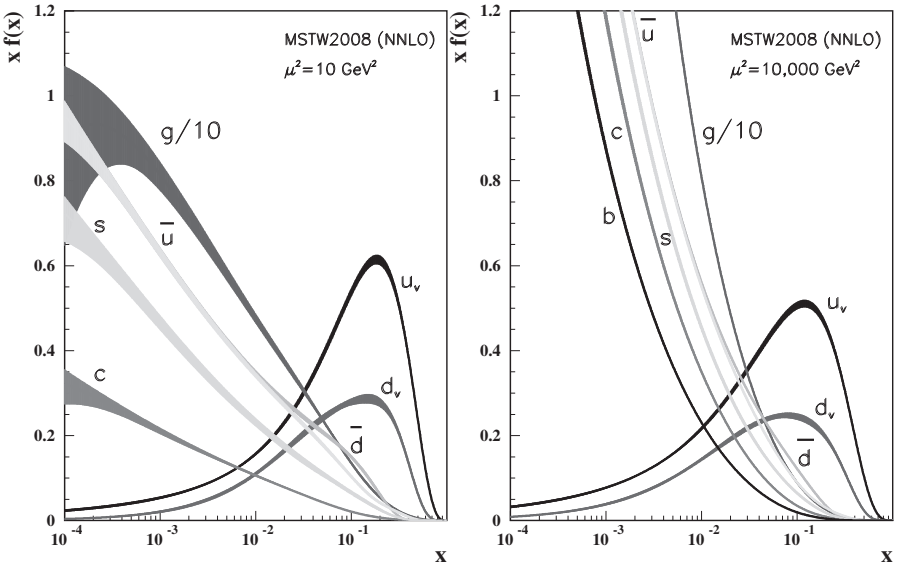


FIGURE 9.9
Distributions of x times the unpolarized parton distribution functions $f(x)$ (where $f = u_V, d_V, \bar{u}, \bar{d}, s, c, b, g$) and their associated uncertainties using the MSTW2008 parametrization (Martin *et al.* 2009) at a scale $\mu^2 = 10$ GeV^2 and $\mu^2 = 10,000$ GeV^2. [Figure reproduced courtesy Michael Barnett, for the Particle Data Group, from the review of Structure Functions by B F Foster, A D Martin and M G Vincter, section 16 in the *Review of Particle Physics*, K Nakamura *et al.* (Particle Data Group) *Journal of Physics* G **37** (2010) 075021, IOP Publishing Limited.] (See color plate I.)

by the Particle Data Group (currently Nakamura *et al.* 2010). Figure 9.9 shows the result of one analysis. In this much more sophisticated approach, which includes higher order QCD corrections, it is necessary to specify a particular value of Q^2 (here denoted by $Q^2 = \mu^2$) at which the distributions are defined, as explained in chapter 15 of volume 2. The distributions at this value are quantities to be determined from experiment. The distributions at other values of Q^2 are then predicted by perturbative QCD.

The main features of the PDFs shown in figure 9.9 are: the valence quark distributions are peaked at around $x = 0.2$, and go to zero for $x \to 0$ and $x \to 1$; the sea quarks, on the other hand, have a high probability of carrying very low momentum fractions, as do the gluons – in fact, the gluons dominate for x below about 0.1. This is then the picture of 'what nucleons are made of', as revealed by some 40 years of research.

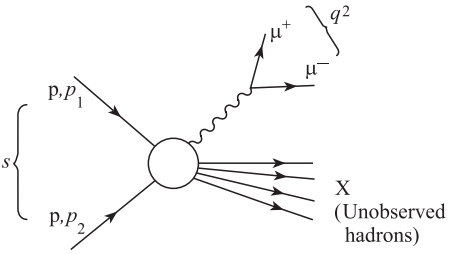


FIGURE 9.10
Drell–Yan process.

9.4 The Drell–Yan process

Much of the importance of the parton model lies outside its original domain of deep inelastic scattering. In deep inelastic scattering it is possible to provide a more formal basis for the parton model in terms of light-cone and short-distance operator expansions (see chapter 18 of Peskin and Schroeder 1995). The advantage of the parton formulation lies in the fact that it suggests other processes for which a parton description may be relevant but for which formal operator arguments are not possible. One such example is the Drell–Yan process (Drell and Yan 1970)

$$\mathrm{p} + \mathrm{p} \to \mu^+\mu^- + \mathrm{X} \tag{9.74}$$

in which a $\mu^+\mu^-$ pair is produced in proton–proton collisions along with unobserved hadrons X, as shown in figure 9.10. The assumption of the parton model is that in the limit

$$s \to \infty \qquad \text{with } \tau = q^2/s \text{ finite} \tag{9.75}$$

the dominant process is that shown in figure 9.11: a quark and antiquark from different hadrons are assumed to annihilate to a virtual photon which then decays to a $\mu^+\mu^-$ pair (compare figures 9.3 and 9.4), the remaining quarks and antiquarks subsequently emerging as hadrons.

Let us work in the CM system and neglect all masses. In this case we have

$$p_1^\mu = (P, 0, 0, P) \qquad p_2^\mu = (P, 0, 0, -P) \tag{9.76}$$

and

$$s = 4P^2. \tag{9.77}$$

Neglecting quark masses and transverse momenta, we have quark momenta

$$p_{q_1}^\mu = x_1(P, 0, 0, P) \tag{9.78}$$
$$p_{q_2}^\mu = x_2(P, 0, 0, -P) \tag{9.79}$$

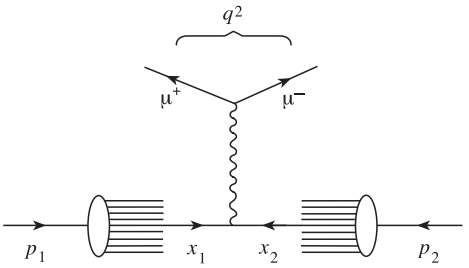


FIGURE 9.11
Parton model amplitude for the Drell–Yan process.

and the photon momentum

$$q = p_{q_1} + p_{q_2} \tag{9.80}$$

has non-zero components

$$q^0 = (x_1 + x_2)P \tag{9.81}$$
$$q^3 = (x_1 - x_2)P. \tag{9.82}$$

Thus we find

$$q^2 = 4x_1x_2P^2 \tag{9.83}$$

and hence

$$\boxed{\tau = q^2/s = x_1x_2.} \tag{9.84}$$

The cross section for the basic process

$$\mathrm{q\bar{q}} \to \mu^+\mu^- \tag{9.85}$$

is calculated using the result of problem 8.18. Since the QED process

$$\mathrm{e^+e^-} \to \mu^+\mu^- \tag{9.86}$$

has the cross section (neglecting all masses)

$$\sigma(\mathrm{e^+e^-} \to \mu^+\mu^-) = 4\pi\alpha^2/3q^2 \tag{9.87}$$

we expect the result for a quark of type a with charge e_a (in units of e) to be

$$\sigma(q_a\bar{q}_a \to \mu^+\mu^-) = (4\pi\alpha^2/3q^2)e_a^2. \tag{9.88}$$

To obtain the parton model prediction for proton–proton collisions, one merely multiplies this cross section by the probabilities for finding a quark of type a with momentum fraction x_1, and an antiquark of the same type with fraction x_2, namely

$$q_a(x_1)\,\mathrm{d}x_1\,\bar{q}_a(x_2)\,\mathrm{d}x_2. \tag{9.89}$$

There is, of course, another contribution for which the antiquark has fraction x_1 and the quark x_2:

$$\bar{q}_a(x_1)\,\mathrm{d}x_1\,q_a(x_2)\,\mathrm{d}x_2. \tag{9.90}$$

Thus the Drell–Yan prediction is

$$\boxed{\begin{aligned}&\mathrm{d}^2\sigma(\mathrm{pp}\to\mu^+\mu^-+\mathrm{X})\\&\qquad=\frac{4\pi\alpha^2}{9q^2}\sum_a e_a^2[q_a(x_1)\bar{q}_a(x_2)+\bar{q}_a(x_1)q_a(x_2)]\,\mathrm{d}x_1\,\mathrm{d}x_2\end{aligned}} \tag{9.91}$$

where we have included a factor $\frac{1}{3}$ to account for the *colour* of the quarks: in order to make a colour singlet photon, one needs to match the colours of quark and antiquark. Equation (9.91) is the master formula. Its importance lies in the fact that the *same* quark distribution functions are measured in deep inelastic lepton scattering so one can make absolute predictions.[3] For example, if the photon in figure 9.11 is replaced by a W(Z), one can predict W(Z) production cross sections, as we shall see in volume 2.

We would expect some 'scaling' property to hold for this cross section, following from the point-like constituent cross section (9.88). One way to exhibit this is to use the variables q^2 and $x_\mathrm{F}=x_1-x_2$ as discussed in problem 9.6. There it is shown that the dimensionless quantity

$$q^4\frac{\mathrm{d}^2\sigma}{\mathrm{d}q^2\,\mathrm{d}x_\mathrm{F}} \tag{9.92}$$

should be a function of x_F and the *ratio* $\tau=q^2/s$. The data bear out this prediction well – see figure 9.12.

Furthermore, the assumption that the lepton pair is produced via quark–antiquark annihilation to a virtual photon can be checked by observing the angular distribution of either lepton in the dilepton rest frame, relative to the incident proton beam direction. This distribution is expected to be the same as in $\mathrm{e}^+\mathrm{e}^-\to\mu^+\mu^-$, namely (cf (8.194))

$$\mathrm{d}\sigma/\mathrm{d}\Omega\propto(1+\cos^2\theta) \tag{9.93}$$

as is indeed observed (figure 9.13). Note that figure 9.13 provides evidence that the quarks have spin-$\frac{1}{2}$: if they are assumed to have spin-0, the angular distribution would be (see problem 9.7) proportional to $(1-\cos^2\theta)$, and this is clearly ruled out.

[3] QCD corrections make the connection more complicated, but still perturbatively computable.

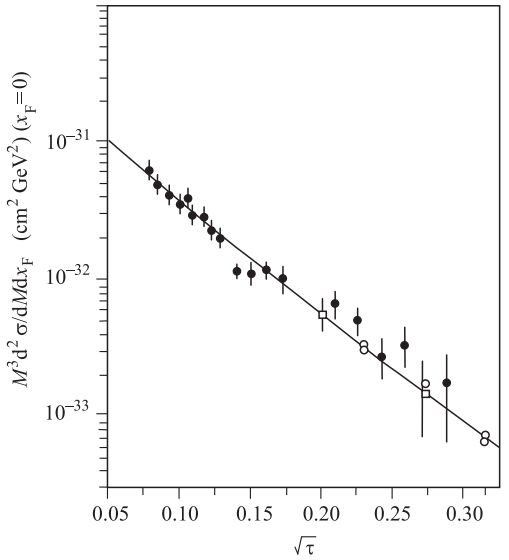


FIGURE 9.12
The dimensionless cross section $M^3 d^2\sigma/dM dx_F$ ($M = \sqrt{q^2}$) at $x_F = 0$ for pN scattering, plotted against $\sqrt{\tau} = M/\sqrt{s}$ (Scott 1985): •, $\sqrt{s} = 62$ GeV; □, 44; □, 27.4; ◯, 23.8.

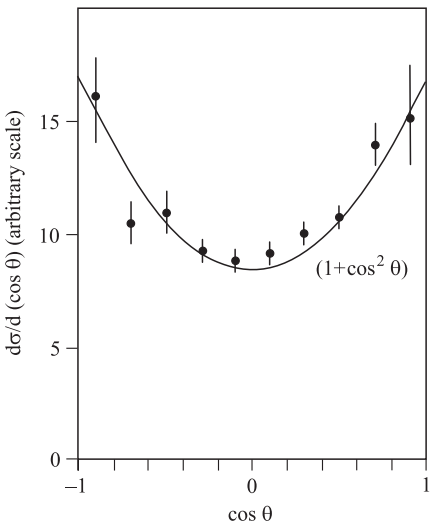


FIGURE 9.13
Angular distribution of muons, measured in the $\mu^+\mu^-$ rest frame, relative to the incident beam direction, in the Drell–Yan process. (Figure from D H Perkins *Introduction to High Energy Physics* 3rd edn, copyright 1987; reprinted by permission of Pearson Education, Inc., Upper Saddle River, NJ.)

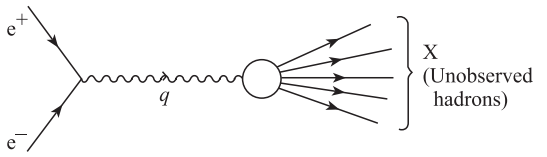


FIGURE 9.14
e^+e^- annihilation to hadrons in one-photon approximation.

9.5 e^+e^- annihilation into hadrons

The last electromagnetic process we wish to consider is electron–positron annihilation into hadrons (figure 9.14):

$$e^+e^- \to X. \tag{9.94}$$

As usual, the dominance of the one-photon intermediate state is assumed. Figure 9.14 is clearly a generalization of figure 8.9, the latter describing the particular case in which the final hadronic state is $\pi^+\pi^-$. As a preliminary to discussing (9.94), let us therefore revisit $e^+e^- \to \pi^+\pi^-$ first.

The $O(e^2)$ amplitude is given in equation (8.159). We shall simplify the calculation by neglecting both the electron and the pion masses. The spinor part of the amplitude is then $-2\bar{v}(k_1)\not{p}_1 u(k)$, and the '$L \cdot T$' product is $16(k \cdot p_1)(k_1 \cdot p_1)$. Borrowing the general CM cross section formula (6.129) from chapter 6 as in (8.121), and including the pion form factor, we obtain for the unpolarized CM differential cross section

$$\left(\frac{d\bar{\sigma}}{d\Omega}\right)_{CM} = \frac{F^2(q^2)\alpha^2}{4q^2}\,(1 - \cos^2\theta) \tag{9.95}$$

and the total unpolarized cross section is

$$\bar{\sigma} = F^2(q^2)\frac{2\pi\alpha^2}{3q^2}. \tag{9.96}$$

The cross section $\bar{\sigma}$ contains a $1/q^2$ factor, just like that for $e^+e^- \to \mu^+\mu^-$ as in (9.87), but this 'pointlike' behaviour is modified by the square of the form-factor, evaluated at time-like q^2. When the measured $\bar{\sigma}$ is plotted against q^2 for $q^2 \leq 1\ (\text{GeV})^2$, a pronounced resonance is seen at $q^2 \approx m_\rho^2$, superimposed on the smooth $1/q^2$ background, where m_ρ is the mass of the rho resonance ($J^P = 1^-$ $q\bar{q}$ state). The interpretation of this is shown in figure 9.15. $F(q^2)$ should therefore be parametrized as a resonance, as in (6.107) – or a more sophisticated version to take account of the fact that the π's are emitted in an

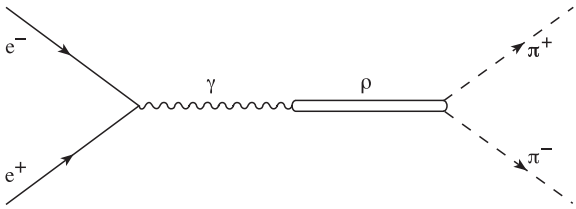


FIGURE 9.15
ρ-dominance of the pion electromagnetic form factor in the time-like ($q^2 > 0$) region.

$\ell = 1$ state. Just as $F^2(q^2)$ modified the point-like cross section in the space-like region for $e^-\pi^+ \to e^-\pi^+$, so here it modifies the point-like ($\sim 1/q^2$) behaviour in the time-like region.

Returning now to the process (9.94), the cross section for it is shown as a function of CM energy $(q^2)^{1/2}$ in figure 9.16. The general point-like fall-off as $1/q^2$ is seen, with peaks due to a succession of boson resonances superimposed ($\rho, J/\psi, \Upsilon, Z^0, \ldots$). The $1/q^2$ fall-off is suggestive of a (point-like) parton picture and indeed the process (9.94) is similar to the Drell–Yan one:

$$\mathrm{pp} \to \mu^+\mu^- + \mathrm{X}. \tag{9.97}$$

It is natural to imagine that at large q^2 the basic subprocess is quark–antiquark pair creation (figure 9.17). The total cross section for $\mathrm{q\bar{q}}$ pair production is then (cf (9.88))

$$\sigma(e^+e^- \to q_a\bar{q}_a) = (4\pi\alpha^2/3q^2)e_a^2. \tag{9.98}$$

In the vicinity of mesonic resonances such as the ρ, we can infer that the dominant component in the final state is that in which the $\mathrm{q\bar{q}}$ pair is strongly bound into a mesonic state, which then decays into hadrons. Away from resonances, and increasingly at larger values of q^2, the produced q and $\bar{\mathrm{q}}$ seek to separate from the interaction region. As they draw apart, however, the interaction between them increases (recall section 1.3.6), producing more $\mathrm{q\bar{q}}$ pairs, together with radiated gluons. In this process, the coloured quarks and gluons eventually must form colourless hadrons, since we know that no coloured particles have been observed ('confinement of colour'). If one assumes that the presumed colour confinement mechanism does not affect the prediction (9.98), then we arrive at the result

$$\sigma(e^+e^- \to \mathrm{hadrons}) = (4\pi\alpha^2/3q^2)\sum_a e_a^2 \tag{9.99}$$

at large q^2, where 'a' includes all flavours produced at that energy.

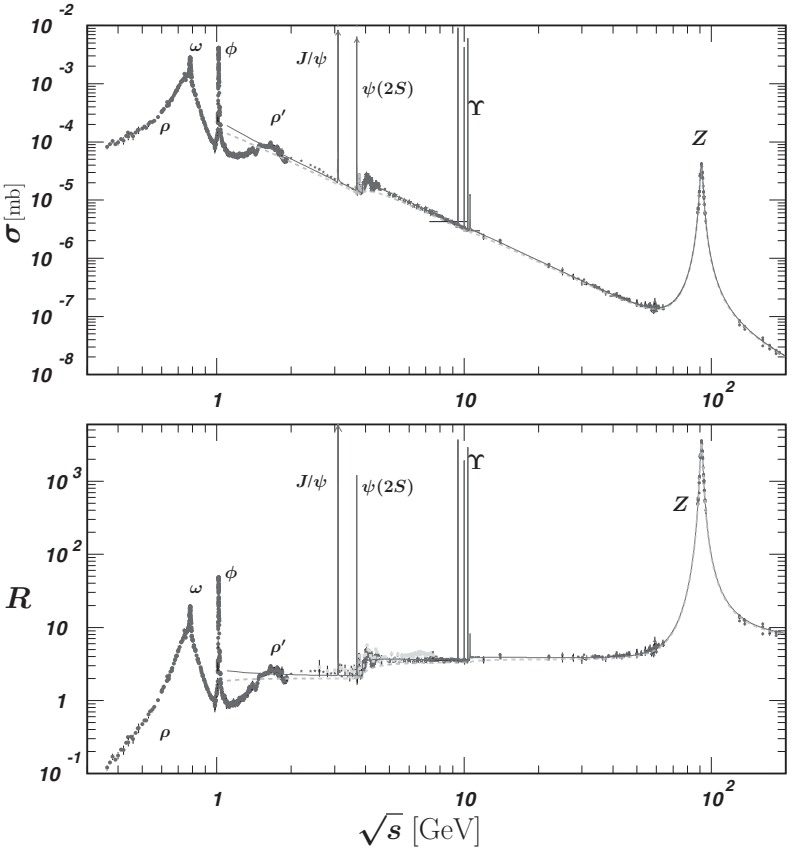


FIGURE 9.16
The cross section σ for the annihilation process $e^+e^- \to$ hadrons, and the ratio R (see equation (9.100)), as a function of cm energy. [Figure reproduced courtesy Michael Barnett, for the Particle Data Group, from the *Review of Particle Physics*, K Nakamura *et al.* (Particle Data Group) *Journal of Physics* G **37** (2010) 075021 IOP Publishing Limited.] (See color plate II.)

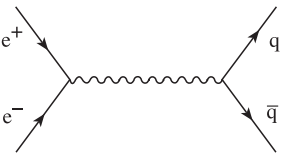


FIGURE 9.17
Parton model subprocess in $e^+e^- \to$ hadrons.

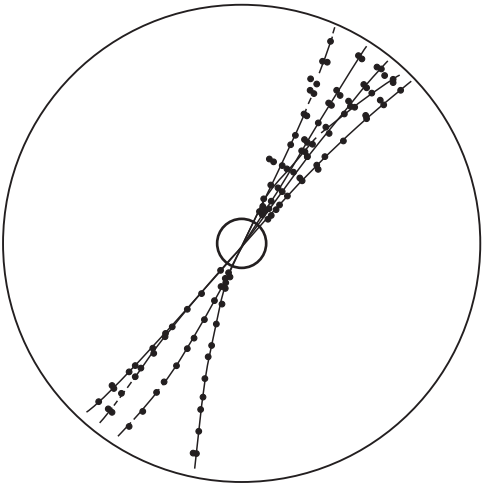


FIGURE 9.18
Two-jet event in e^+e^- annihilation from the TASSO detector at the e^+e^- storage ring PETRA.

This model is best tested by taking out the dominant $1/q^2$ behaviour and plotting the ratio

$$R = \frac{\sigma(e^+e^- \to \text{hadrons})}{\sigma(e^+e^- \to \mu^+\mu^-)} = \sum_a e_a^2. \tag{9.100}$$

For the light quarks u, d and s occurring in three colours, we therefore predict

$$R = 3[(\tfrac{2}{3})^2 + (-\tfrac{1}{3})^2 + (-\tfrac{1}{3})^2] = 2. \tag{9.101}$$

Above the c threshold but below the b threshold we expect $R = \frac{10}{3}$, and above the b threshold $R = \frac{11}{3}$. These expectations are in reasonable accord with experiment, especially at energies well beyond the resonance region and the b threshold, as figure 9.16 shows. In this figure the dotted curve is the prediction of the quark-parton model, equation (9.99). The solid curve includes perturbative QCD corrections, which we will return to in chapter 15 of volume 2.

The success of this prediction leads one to consider more detailed consequences of the picture. For example, the angular distribution of massless spin-$\frac{1}{2}$ quarks is expected to be (cf (8.194) again)

$$\mathrm{d}\sigma/\mathrm{d}\Omega = (\alpha^2/4q^2)e_a^2(1 + \cos^2\theta) \tag{9.102}$$

just as for the $\mu^+\mu^-$ process. However, in this case there is an important difference: the quarks are not observed! Nevertheless a remarkable 'memory' of (9.102) is retained by the observed final-state hadrons. Experimentally one

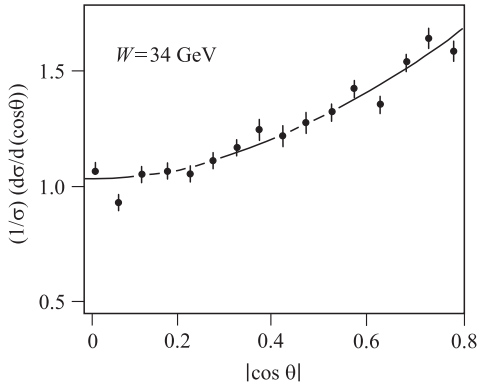


FIGURE 9.19
Angular distribution of jets in two-jet events, measured in the two-jet rest frame, relative to the incident beam direction, in the process $e^+e^- \to$ two jets (Althoff *et al.* 1984). The full curve is the $(1 + \cos^2\theta)$ distribution. Since it is not possible to say which jet corresponded to the quark and which to the antiquark, only half the angular distribution can be plotted. The asymmetry visible in figure 8.20(b) is therefore not apparent.

observes events in which hadrons emerge from the interaction region in two relatively well-collimated cones or 'jets' – see figure 9.18. The distribution of events as a function of the (inferred) angle of the jet axis is shown in figure 9.19 and is in good agreement with (9.102). The interpretation is that the primary process is $e^+e^- \to q\bar{q}$, the quark and the antiquark then turning into hadrons as they separate and experience the very strong colour forces, but without losing the memory of the original quark angular distribution. We shall discuss jets more fully in chapter 14 of volume 2, in the context of QCD.

Problems

9.1 The various normalization factors in equations (9.3) and (9.11) may be checked in the following way. The cross section for inclusive electron–proton scattering may be written (equation (9.11)):

$$\mathrm{d}\sigma = \left(\frac{4\pi\alpha}{q^2}\right)^2 \frac{1}{4[(k\cdot p)^2 - m^2M^2]^{1/2}} 4\pi M L_{\mu\nu} W^{\mu\nu} \frac{\mathrm{d}^3\boldsymbol{k}'}{2\omega'(2\pi)^3} \qquad (9.103)$$

in the usual one-photon exchange approximation, and the tensor $W^{\mu\nu}$ is related to hadronic matrix elements of the electromagnetic current operator by

equation (9.3):

$$e^2 W^{\mu\nu}(q,p) = \frac{1}{4\pi M}\frac{1}{2}\sum_s \sum_{\mathrm{X}} \langle \mathrm{p};p,s|\hat{j}^{\mu}_{\mathrm{em}}(0)|\mathrm{X};p'\rangle \times \langle \mathrm{X};p'|\hat{j}^{\nu}_{\mathrm{em}}(0)|\mathrm{p};p,s\rangle (2\pi)^4\delta^4(p+q-p')$$

where the sum X is over all possible hadronic final states. If we consider the special case of elastic scattering, the sum over X is only over the final proton's degrees of freedom:

$$e^2 W^{\mu\nu}_{\mathrm{el}} = \frac{1}{4\pi M}\frac{1}{2}\sum_s \sum_{s'} \langle \mathrm{p};p,s|\hat{j}^{\mu}_{\mathrm{em}}(0)|\mathrm{p};p',s'\rangle\langle \mathrm{p};p',s'|\hat{j}^{\nu}_{\mathrm{em}}(0)|\mathrm{p};p,s\rangle \times (2\pi)^4\delta^4(p+q-p')\frac{1}{(2\pi)^3}\frac{\mathrm{d}^3p'}{2E'}.$$

Now use equation (8.208) with $\mathcal{F}_1 = 1$ and $\kappa = 0$ (i.e. the electromagnetic current matrix element for a 'point' proton) to show that the resulting cross section is identical to that for elastic eμ scattering.

9.2

(a) Perform the contraction $L_{\mu\nu}W^{\mu\nu}$ for inclusive inelastic electron–proton scattering (remember $q^\mu L_{\mu\nu} = q^\nu L_{\mu\nu} = 0$). Hence verify that the inclusive differential cross section in terms of 'laboratory' variables, and neglecting the electron mass, has the form

$$\frac{\mathrm{d}^2\sigma}{\mathrm{d}\Omega \mathrm{d}k'} = \frac{\alpha^2}{4k^2\sin^4(\theta/2)}[W_2\cos^2(\theta/2) + W_1 2\sin^2(\theta/2)].$$

(b) By calculating the Jacobian

$$J = \begin{vmatrix} \partial u/\partial x & \partial u/\partial y \\ \partial v/\partial x & \partial v/\partial y \end{vmatrix}$$

for a change of variables $(x,y) \to (u,v)$

$$\mathrm{d}u\,\mathrm{d}v = |J|\mathrm{d}x\,\mathrm{d}y$$

find expressions for $\mathrm{d}^2\sigma/\mathrm{d}Q^2\,\mathrm{d}\nu$ and $\mathrm{d}^2\sigma/\mathrm{d}x\,\mathrm{d}y$, where Q^2 and ν have their usual significance, and x is the scaling variable $Q^2/2M\nu$ and $y = \nu/k$.

9.3 Consider the description of inelastic electron–proton scattering in terms of virtual photon cross sections:

(a) In the 'laboratory' frame with

$$p^\mu = (M,0,0,0) \qquad \text{and} \qquad q^\mu = (q^0,0,0,q^3)$$

evaluate the transverse spin sum

$$\tfrac{1}{2}\sum_{\lambda=\pm1}\epsilon_\mu(\lambda)\epsilon^*_\nu(\lambda)W^{\mu\nu}.$$

Hence show that the 'Hand' cross section for transverse virtual photons is

$$\sigma_{\rm T} = (4\pi^2\alpha/K)W_1.$$

(b) Using the definition

$$\epsilon^\mu_{\rm S} = (1/\sqrt{Q^2})(q^3, 0, 0, q^0)$$

and rewriting this in terms of the 'laboratory' 4-vectors p^μ and q^μ, evaluate the longitudinal/scalar virtual photon cross section. Hence show that

$$W_2 = \frac{K}{4\pi^2\alpha}\frac{Q^2}{Q^2+\nu^2}(\sigma_{\rm S}+\sigma_{\rm T}).$$

9.4 In this problem, we consider the representation of the 4×4 Dirac matrices in which (see (3.40))

$$\boldsymbol{\alpha} = \begin{pmatrix} \boldsymbol{\sigma} & \mathbf{0} \\ \mathbf{0} & -\boldsymbol{\sigma} \end{pmatrix} \qquad \beta = \begin{pmatrix} \mathbf{0} & \mathbf{1} \\ \mathbf{1} & \mathbf{0} \end{pmatrix}.$$

Define also the 4×4 matrix $\gamma_5 = \begin{pmatrix} \mathbf{1} & \mathbf{0} \\ \mathbf{0} & -\mathbf{1} \end{pmatrix}$ and the Dirac four-component spinor $u = \begin{pmatrix} \phi \\ \chi \end{pmatrix}$. Then the two-component spinors ϕ, χ satisfy

$$\begin{aligned} \boldsymbol{\sigma}\cdot\boldsymbol{p}\phi &= E\phi - m\chi \\ \boldsymbol{\sigma}\cdot\boldsymbol{p}\chi &= -E\chi + m\phi. \end{aligned}$$

(a) Show that for a massless Dirac particle, ϕ and χ become helicity eigenstates (see section 3.3) with positive and negative helicity respectively.

(b) Defining

$$P_{\rm R} = \frac{1+\gamma_5}{2} \qquad P_{\rm L} = \frac{1-\gamma_5}{2}$$

show that $P_{\rm R}^2 = P_{\rm L}^2 = 1$, $P_{\rm R}P_{\rm L} = 0 = P_{\rm L}P_{\rm R}$, and that $P_{\rm R} + P_{\rm L} = 1$. Show also that

$$P_{\rm R}\begin{pmatrix}\phi\\ \chi\end{pmatrix} = \begin{pmatrix}\phi\\ 0\end{pmatrix} \qquad P_{\rm L}\begin{pmatrix}\phi\\ \chi\end{pmatrix} = \begin{pmatrix}0\\ \chi\end{pmatrix}$$

and hence that $P_{\rm R}$ and $P_{\rm L}$ are projection operators for massless Dirac particles, onto states of definite helicity. Discuss what happens when $m \neq 0$.

(c) The general massless spinor u can be written

$$u = (P_{\rm L} + P_{\rm R})u \equiv u_{\rm L} + u_{\rm R}$$

where $u_{\rm L}$, $u_{\rm R}$ have the indicated helicities. Show that

$$\bar{u}\gamma^{\mu}u = \bar{u}_{\rm L}\gamma^{\mu}u_{\rm L} + \bar{u}_{\rm R}\gamma^{\mu}u_{\rm R}$$

where $\bar{u}_{\rm L} = u_{\rm L}^{\dagger}\gamma^0$, $\bar{u}_{\rm R} = u_{\rm R}^{\dagger}\gamma^0$; and deduce that in electromagnetic interactions of massless fermions helicity is conserved.

(d) In weak interactions an *axial vector current* $\bar{u}\gamma^{\mu}\gamma_5 u$ also enters. Is helicity still conserved?

(e) Show that the 'Dirac' mass term $m\bar{\hat{\psi}}\hat{\psi}$ may be written as $m(\bar{\hat{\psi}}_{\rm L}\hat{\psi}_{\rm R} + \bar{\hat{\psi}}_{\rm R}\hat{\psi}_{\rm L})$.

9.5 In the HERA colliding beam machine, positrons of total energy 27.5 GeV collide head on with protons of total energy 820 GeV. Neglecting both the positron and the proton rest masses, calculate the centre-of-mass energy in such a collision process.

Some theories have predicted the existence of 'leptoquarks', which could be produced at HERA as a resonance state formed from the incident positron and the struck quark. How would a distribution of such events look, if plotted versus the variable x?

9.6

(a) By the expedient of inserting a δ-function, the differential cross section for Drell–Yan production of a lepton pair of mass $\sqrt{q^2}$ may be written as

$$\frac{{\rm d}\sigma}{{\rm d}q^2} = \int {\rm d}x_1\,{\rm d}x_2\,\frac{{\rm d}^2\sigma}{{\rm d}x_1\,{\rm d}x_2}\delta(q^2 - sx_1x_2).$$

Show that this is equivalent to the form

$$\begin{aligned}\frac{{\rm d}\sigma}{{\rm d}q^2} &= \frac{4\pi\alpha^2}{9q^4}\int {\rm d}x_1\,{\rm d}x_2\,x_1x_2\delta(x_1x_2 - \tau) \\ &\quad \times \sum_a e_a^2[q_a(x_1)\bar{q}_a(x_2) + \bar{q}_a(x_1)q_a(x_2)]\end{aligned}$$

which, since $q^2 = s\tau$, exhibits a scaling law of the form

$$s^2{\rm d}\sigma/{\rm d}q^2 = F(\tau).$$

(b) Introduce the Feynman scaling variable

$$x_{\mathrm{F}} = x_1 - x_2$$

with

$$q^2 = s x_1 x_2$$

and show that

$$\mathrm{d}q^2\, \mathrm{d}x_{\mathrm{F}} = (x_1 + x_2) s \mathrm{d}x_1\, \mathrm{d}x_2.$$

Hence show that the Drell–Yan formula can be rewritten as

$$\frac{\mathrm{d}^2\sigma}{\mathrm{d}q^2\, \mathrm{d}x_{\mathrm{F}}} = \frac{4\pi\alpha^2}{9q^4} \frac{\tau}{(x_{\mathrm{F}}^2 + 4\tau)^{1/2}} \sum_a e_a^2 [q_a(x_1)\bar{q}_a(x_2) + \bar{q}_a(x_1) q_a(x_2)].$$

9.7 Verify that if the quarks participating in the Drell–Yan subprocess $\mathrm{q\bar{q}} \to \gamma \to \mu\bar{\mu}$ had spin-0, the CM angular distribution of the final $\mu^+\mu^-$pair would be proportional to $(1 - \cos^2\theta)$.

Part IV

Loops and Renormalization


Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

10

Loops and Renormalization I: The ABC Theory

We have seen how Feynman diagrams represent terms in a perturbation theory expansion of physical amplitudes, namely the Dyson expansion of section 6.2. Terms of a given order all involve the same power of a 'coupling constant', which is the multiplicative constant appearing in the interaction Hamiltonian – for example, 'g' in the ABC theory, or the charge 'e' in electrodynamics. In practice, it often turns out that the relevant parameter is actually the square of the coupling constant, and factors of 4π have a habit of appearing on a regular basis; so, for QED, the perturbation series is conveniently ordered according to powers of the *fine structure constant* $\alpha = e^2/4\pi \approx 1/137$.

Equivalently, this is an expansion in terms of the number of vertices appearing in the diagrams, since one power of the coupling constant is associated with each vertex. For a given physical process, the lowest-order diagrams (the ones with the fewest vertices) are those in which each vertex is connected to every other vertex by just one internal line; these are called *tree diagrams*. The Yukawa (u-channel) exchange process of figure 6.4, and the s-channel process of figure 6.5, are both examples of tree diagrams, and indeed all of our calculations so far have not gone further than this lowest-order ('tree') level. Admittedly, since α is after all pretty small, tree diagrams in QED are likely to give us a good approximation to compare with experiment. Nevertheless, a long history of beautiful and ingenious experiments has resulted in observables in QED being determined to an accuracy far better than the $O(1\%)$ represented by the leading (tree) terms. More generally, precision experiments at LEP and other laboratories have an accuracy sensitive to higher-order corrections in the Standard Model. Hence, some understanding of the physics beyond the tree approximation is now essential for phenomenology.

All higher-order processes beyond the tree approximation involve *loops*, a concept easier to recognize visually than to define in words. In section 6.3.5 we already met (figure 6.8) one example of an $O(g^4)$ correction to the $O(g^2)$ C-exchange tree diagram of figure 6.4, which contains one loop. The crucial point is that whereas a tree diagram can be cut into two separate pieces by severing just one internal line, to cut a loop diagram into two separate pieces requires the severing of at least two internal lines.

In these last two chapters of volume 1, we aim to provide an introduction to higher-order processes, confining ourselves to 'one-loop' order. In the

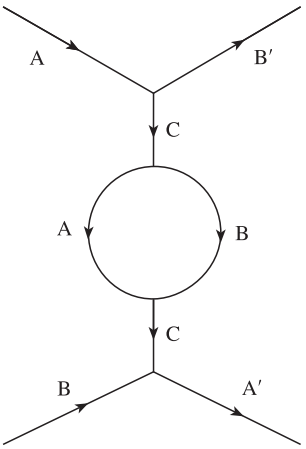


FIGURE 10.1
$O(g^4)$ contribution to the process $\mathrm{A}+\mathrm{B}\to\mathrm{A}+\mathrm{B}$, involving the modification of the C propagator by the insertion of a loop.

present chapter we shall concentrate mainly on the particular loop appearing in figure 6.8. This will lead us into the physics of *renormalization* for the ABC theory, which – as a Yukawa-like theory – is a good theoretical laboratory for studying 'one-loop physics', without the complications of spinor and gauge fields. In the following chapter, we shall discuss one-loop diagrams in QED, emphasizing some important physical consequences, such as corrections to Coulomb's law, anomalous magnetic moments and the running coupling constant.

10.1 The propagator correction in ABC theory

10.1.1 The $O(g^2)$ self-energy $\Pi_{\mathrm{C}}^{[2]}(q^2)$

We consider figure 6.8, reproduced here again as figure 10.1. In section 6.3.5, we gave the extra rule ('(iii)') needed to write down the invariant amplitude for this process. We first show how this rule arises in the special case of figure 10.1.

Clearly, figure 10.1 is a fourth-order process, so it must emerge from the term

$$\begin{aligned}\frac{(-\mathrm{i}g)^4}{4!}\int\!\!\int\!\!\int\!\!\int \mathrm{d}^4x_1\,\mathrm{d}^4x_2\,\mathrm{d}^4x_3\,\mathrm{d}^4x_4\,\langle 0|\hat{a}_{\mathrm{A}}(p'_{\mathrm{A}})\hat{a}_{\mathrm{B}}(p'_{\mathrm{B}})\\ \times T\{\hat{\phi}_{\mathrm{A}}(x_1)\hat{\phi}_{\mathrm{B}}(x_1)\hat{\phi}_{\mathrm{C}}(x_1)\ldots\hat{\phi}_{\mathrm{A}}(x_4)\hat{\phi}_{\mathrm{B}}(x_4)\hat{\phi}_{\mathrm{C}}(x_4)\}\\ \times \hat{a}^{\dagger}_{\mathrm{A}}(p_{\mathrm{A}})\hat{a}^{\dagger}_{\mathrm{B}}(p_{\mathrm{B}})|0\rangle(16E_{\mathrm{A}}E_{\mathrm{B}}E'_{\mathrm{A}}E'_{\mathrm{B}})^{1/2}\end{aligned}\tag{10.1}$$

of the Dyson expansion. Since it is basically a u-channel exchange process ($u = (p_{\rm A} - p'_{\rm B})^2 = (p'_{\rm A} - p_{\rm B})^2$), the vev's involving the external creation and annihilation operators must appear as they do in equation (6.89) ('ingoing A, outgoing B$'$ at one point x_2; ingoing B, outgoing A$'$ at another point x_1') rather than as in equation (6.88) ('ingoing A and B at x_2; outgoing A$'$ and B$'$ at x_1'). In (10.1), however, we unfortunately have *four* space–time points to choose from, rather than merely the two in (6.74). Figuring out exactly which choices are in fact equivalent and which are not is best left to private struggle, especially since we are not seriously interested in the numerical value of our fourth-order corrections in this case. Let us simply consider one choice, analogous to (6.89). This yields the amplitude (cf (6.91))

$$(-{\rm i}g)^4 \iiiint {\rm d}^4x_1\,{\rm d}^4x_2\,{\rm d}^4x_3\,{\rm d}^4x_4\,{\rm e}^{{\rm i}(p'_{\rm A}-p_{\rm B})\cdot x_1}{\rm e}^{{\rm i}(p'_{\rm B}-p_{\rm A})\cdot x_2} \times \langle 0|T\{\hat\phi_{\rm C}(x_1)\hat\phi_{\rm C}(x_2)\hat\phi_{\rm A}(x_3)\hat\phi_{\rm B}(x_3)\hat\phi_{\rm C}(x_3)\hat\phi_{\rm A}(x_4)\hat\phi_{\rm B}(x_4)\hat\phi_{\rm C}(x_4)\}|0\rangle \tag{10.2}$$

and we have discarded the numerical factor 1/4!. Once again, there are many terms in the expansion of the vev of the eight operators in (10.2). But, with an eye on the structure of the Feynman amplitude at which we are aiming (figure 10.1), let us consider again just a single contribution

$$\begin{aligned}(-{\rm i}g)^4 \iiiint {\rm d}^4x_1\,{\rm d}^4x_2\,{\rm d}^4x_3\,{\rm d}^4x_4\,{\rm e}^{{\rm i}(p'_{\rm A}-p_{\rm B})\cdot x_1}{\rm e}^{{\rm i}(p'_{\rm B}-p_{\rm A})\cdot x_2} \\ \times \langle 0|T(\hat\phi_{\rm C}(x_1)\hat\phi_{\rm C}(x_3))|0\rangle\langle 0|T(\hat\phi_{\rm C}(x_2)\hat\phi_{\rm C}(x_4))|0\rangle \\ \times \langle 0|T(\hat\phi_{\rm A}(x_3)\hat\phi_{\rm A}(x_4))|0\rangle\langle 0|T(\hat\phi_{\rm B}(x_3)\hat\phi_{\rm B}(x_4))|0\rangle\end{aligned} \tag{10.3}$$

which contains four propagators connected as in figure 10.2.

As we saw in section 6.3.2, each of these propagators is a function only of the difference of the two space–time points involved. Introducing relative coordinates $x = x_1 - x_3$, $y = x_2 - x_4$, $z = x_3 - x_4$ and the CM coordinate $X = \frac{1}{4}(x_1 + x_2 + x_3 + x_4)$, we find (problem 10.1) that (10.3) becomes

$$\begin{aligned}(-{\rm i}g)^4 \iiiint {\rm d}^4X\,{\rm d}^4x\,{\rm d}^4y\,{\rm d}^4z\,{\rm e}^{{\rm i}(p'_{\rm A}+p'_{\rm B}-p_{\rm A}-p_{\rm B})\cdot X}{\rm e}^{{\rm i}(p'_{\rm A}-p_{\rm B})\cdot(3x-y+2z)/4} \\ \times {\rm e}^{{\rm i}(p'_{\rm B}-p_{\rm A})\cdot(-x+3y-2z)/4}D_{\rm C}(x)D_{\rm C}(y)D_{\rm A}(z)D_{\rm B}(z)\end{aligned} \tag{10.4}$$

where D_i is the position–space propagator for type-i particles ($i = {\rm A}, {\rm B}, {\rm C}$), defined as in (6.98). The integral over X gives the expected overall 4-momentum conservation factor, $(2\pi)^4\delta^4(p'_{\rm A}+p'_{\rm B}-p_{\rm A}-p_{\rm B})$. Setting $q = p_{\rm A}-p'_{\rm B} = p'_{\rm A}-p_{\rm B}$ (where 4-momentum conservation has been used), (10.4) becomes

$$\begin{aligned}(-{\rm i}g)^4(2\pi)^4\delta^4(p'_{\rm A}+p'_{\rm B}-p_{\rm A}-p_{\rm B}) \iiint {\rm d}^4x\,{\rm d}^4y\,{\rm d}^4z\,{\rm e}^{{\rm i}q\cdot x}D_{\rm C}(x) \\ \times {\rm e}^{-{\rm i}q\cdot y}D_{\rm C}(y){\rm e}^{{\rm i}q\cdot z}D_{\rm A}(z)D_{\rm B}(z).\end{aligned} \tag{10.5}$$

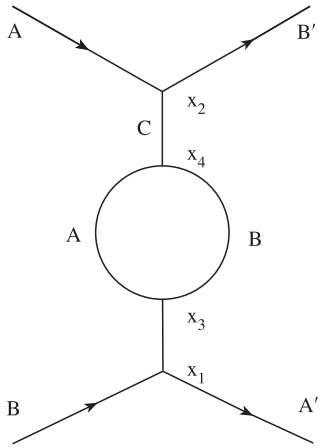


FIGURE 10.2
The space–time structure of the integrand in (10.3).

The integrals over x and y separate out completely, each being just the Fourier transform of a C propagator – that is, the momentum–space propagator $\tilde{D}_{\rm C}(q)$. Since the latter is a function of q^2 only, we end up with two factors of $\mathrm{i}/(q^2 - m_{\rm C}^2 + \mathrm{i}\epsilon)$, corresponding to the two C propagators in the momentum–space Feynman diagram of figure 10.1. Note that the Mandelstam u-variable is defined by $u = (p_{\rm A} - p'_{\rm B})^2$ and is thus equal to q^2; we shall, however, continue to use q^2 rather than u in what follows.

The remaining factor represents the loop. Including $(-\mathrm{i}g)^2$ for the two vertices in the loop, it is given by

$$(-\mathrm{i}g)^2 \int \mathrm{d}^4 z\, \mathrm{e}^{\mathrm{i}q\cdot z} D_{\rm A}(z) D_{\rm B}(z) \tag{10.6}$$

which is the main result of our calculation so far. Since we want to end up finally with a momentum–space amplitude, let us introduce the A and B propagators in momentum space, and write (10.6) as (cf (6.99))

$$\begin{aligned}
&(-\mathrm{i}g)^2 \int \mathrm{d}^4 z\, \mathrm{e}^{\mathrm{i}q\cdot z} \int \frac{\mathrm{d}^4 k_1}{(2\pi)^4} \mathrm{e}^{-\mathrm{i}k_1\cdot z} \frac{\mathrm{i}}{k_1^2 - m_{\rm A}^2 + \mathrm{i}\epsilon} \int \frac{\mathrm{d}^4 k_2}{(2\pi)^4} \mathrm{e}^{-\mathrm{i}k_2\cdot z} \frac{\mathrm{i}}{k_2^2 - m_{\rm B}^2 + \mathrm{i}\epsilon} \\
&\quad = (-\mathrm{i}g)^2 \iint \frac{\mathrm{d}^4 k_1}{(2\pi)^4} \frac{\mathrm{d}^4 k_2}{(2\pi)^4} \frac{\mathrm{i}}{k_1^2 - m_{\rm A}^2 + \mathrm{i}\epsilon} \frac{\mathrm{i}}{k_2^2 - m_{\rm B}^2 + \mathrm{i}\epsilon} \\
&\qquad \times (2\pi)^4 \delta^4(k_1 + k_2 - q) \\
&\quad = (-\mathrm{i}g)^2 \int \frac{\mathrm{d}^4 k}{(2\pi)^4} \frac{\mathrm{i}}{k^2 - m_{\rm A}^2 + \mathrm{i}\epsilon} \frac{\mathrm{i}}{(q-k)^2 - m_{\rm B}^2 + \mathrm{i}\epsilon} \qquad (10.7) \\
&\quad \equiv -\mathrm{i}\Pi_C^{[2]}(q^2), \qquad (10.8)
\end{aligned}$$

where we have defined the function $-\mathrm{i}\Pi_{\mathrm{C}}^{[2]}(q^2)$ as the loop (or 'bubble') amplitude appearing in figure 10.1. It is a function of q^2, as follows from Lorentz invariance. The $^{[2]}$ refers to the two powers of g, as will be explained shortly, after (10.15).

Careful consideration of the equivalences among the various contractions shows that the amplitude corresponding to figure 10.1 is, in fact, just the simple expression

$$(-\mathrm{i}g)^2(2\pi)^4\delta^4(p'_{\mathrm{A}}+p'_{\mathrm{B}}-p_{\mathrm{A}}-p_{\mathrm{B}})\frac{\mathrm{i}}{q^2-m_{\mathrm{C}}^2+\mathrm{i}\epsilon}(-\mathrm{i}\Pi_{\mathrm{C}}^{[2]}(q^2))\frac{\mathrm{i}}{q^2-m_{\mathrm{C}}^2+\mathrm{i}\epsilon} \tag{10.9}$$

where $\Pi_{\mathrm{C}}^{[2]}(q^2)$ is given in (10.8). We see that whereas the 'single-particle' pieces, involving one C-exchange, do not involve any integral in momentum–space, the loop (which involves both A and B particles) does involve a momentum integral. This can be simply understood in terms of 4-momentum conservation, which holds at every vertex of a Feynman graph. At the top (or bottom) vertex of figure 10.1, the 4-momentum q of the C-particle is fully determined by that of the incoming and outgoing particles ($q = p_{\mathrm{A}} - p'_{\mathrm{B}} = p'_{\mathrm{A}} - p_{\mathrm{B}}$). This same 4-momentum q flows in (and out) of the loop in figure 10.1, but nothing determines how it is to be shared between the A- and B-particles; all that can be said is that if the 4-momentum of A is k (as in (10.7)) then that of B is $q-k$, so that their sum is q. The 'free' variable k then has to be integrated over, and this is the physical origin of rule (iii) of section 6.3.5.

We have devoted some time to the steps leading to expression (10.7), not only in order to follow the emergence of rule (iii) mathematically, but so as to lend some plausibility to a very important statement: the Feynman rules for associating factors with vertices and propagators, which we learned for tree graphs in chapters 6 and 8, *also* work, with the addition of rule (iii), for all more complicated graphs as well! Having seen most of just one fairly short calculation of a higher-order amplitude, the reader may perhaps now begin to appreciate just how powerful is the precise correspondence between 'diagrams and amplitudes', given by the Feynman rules.

Having arrived at the expression for our first one-loop graph, we must at once draw the reader's attention to the *bad news*: the integral in (10.7) is divergent at large values of k. We shall postpone a more detailed mathematical analysis until section 10.3.1, but the divergence can be plausibly inferred just from a simple counting of powers: there are four powers of k in the numerator and four in the denominator, and the likelihood is that the integral diverges as $\int_0^\Lambda k^3\mathrm{d}k/k^4 \sim \ln\Lambda$, as $\Lambda \to \infty$. This is plainly a disaster: a quantity which was supposed to be a small correction in perturbation theory is actually infinite! Such divergences, occurring as loop momenta go to infinity, are called 'ultraviolet divergences', and they are ubiquitous in quantum field theory. Only after a long struggle with these infinities was it understood how to obtain physically sensible results from such perturbation expansions. Depending on the type of field theory involved, the infinities can often be 'tamed' through a

procedure known as *renormalization*, to which we shall provide an introduction in this and the following chapter.

The physical ideas behind renormalization are, however, just as relevant in cases – such as condensed matter physics – where the analogous higher-order (loop) corrections are not infinite, though possibly large. In quantum mechanics, infinite momentum corresponds to zero distance, and our fields are certainly 'point-like'. But in condensed matter physics there is generally a natural non-zero smallest distance – the lattice size, or an atomic diameter, for example. In quantum field theory, such a 'shortest distance' would correspond to a 'highest momentum', meaning that the magnitudes of loop momenta would run from zero up to some finite limit Λ, say, rather than infinity. Such a Λ is called a (momentum) 'cut-off'. With such a cut-off in place, our loop integrals are of course finite – but it would seem that we have then maltreated our field theory in some way. However, we might well ask whether we seriously believe that any of our quantum field theories is literally valid for arbitrarily high energies (or arbitrarily small distances). The answer is surely no: we are virtually certain that 'new physics' will come into play at some stage, which is not contained in – say – the QED, or even the Standard Model, Lagrangian. At what scale this new physics will enter (the Planck energy? 1 TeV?) we do not know, but surely the current models will break down at some point. We should not be too alarmed, therefore, by formal divergences as $\Lambda \to \infty$. Rather, it may be sensible to regard a cut-off Λ as standing for some 'new physics' scale, accepting some such manoeuvre as physically realistic as well as mathematically prudent.

At the same time, however, we would not want our physical predictions, made using quantum field theories, to depend sensitively on Λ – i.e. on the unknown short-distance physics, in this interpretation. Indeed, theories exist (for example, those in the Standard Model and the ABC theory) which can be reformulated in such a way that all dependence on Λ disappears, as $\Lambda \to \infty$; these are, precisely, *renormalizable* quantum field theories. Roughly speaking, a renormalizable quantum field theory is one such that, when formulae are expressed in terms of certain 'physical' parameters taken from experiment, rather than in terms of the original parameters appearing in the Lagrangian, calculated quantities will be finite and independent of Λ as $\Lambda \to \infty$.

Solid state physics provides a close analogy. There, the usefulness of a description of, say, electrons in a metal in terms of their 'effective charge' and 'effective mass', rather than their free-space values, is well established. In this analogy, the free-space quantities correspond to our Lagrangian values, while the effective parameters correspond to our 'physical' ones. In both cases, the interactions are causing changes to the parameters.

It is clear that we need to understand more precisely just what our 'physical parameters' might be and how they might be defined. This is what we aim to do in the remainder of the present section, and in the next one, before returning in section 10.3 to the mathematical details associated with evaluating (10.7), and indicating how renormalization works for the self-energy. Having

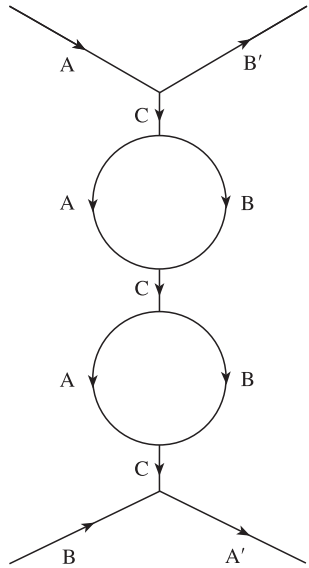


FIGURE 10.3
$O(g^6)$ term in $\mathrm{A} + \mathrm{B} \to \mathrm{A} + \mathrm{B}$, involving the insertion of two loops in the C propagator.

thus prepared the ground, we shall introduce a more powerful approach in section 10.4, and offer a few preliminary remarks about 'renormalizability' in section 10.5, returning to that topic at the end of the following chapter. Although usually not explicitly indicated, loop corrections considered in this and the following section will be understood to be defined with a cut-off Λ, so that they are finite.

To begin the discussion of the physical significance of our $O(g^4)$ correction, (10.9), it is convenient to consider both the $O(g^2)$ term (6.100) and the $O(g^4)$ correction together, obtaining

$$(-\mathrm{i}g)^2(2\pi)^4\delta^4(p'_{\mathrm{A}} + p'_{\mathrm{B}} - p_{\mathrm{A}} - p_{\mathrm{B}}) \times \left\{ \frac{\mathrm{i}}{q^2 - m_{\mathrm{C}}^2} + \frac{\mathrm{i}}{q^2 - m_{\mathrm{C}}^2}(-\mathrm{i}\Pi_{\mathrm{C}}^{[2]}(q^2))\frac{\mathrm{i}}{q^2 - m_{\mathrm{C}}^2} \right\} \qquad (10.10)$$

where the $\mathrm{i}\epsilon$ in the C propagators does not need to be retained. Both the form of (10.10), and inspection of figure 10.1, suggest that the $O(g^4)$ term we have calculated can be regarded as an $O(g^2)$ correction to the propagator for the C-particle. Indeed, we can easily imagine adding in the $O(g^6)$ term shown in figure 10.3, and in fact the whole infinite series of such 'bubbles' connected by simple C propagators. The infinite geometric series for the

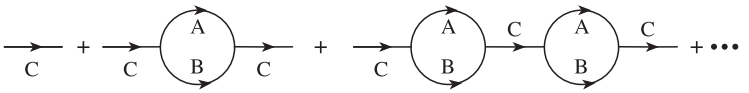


FIGURE 10.4
Series of one-loop (or 'bubble') insertions in the C propagator.

corrected propagator shown in figure 10.4 has the form

$$\frac{\mathrm{i}}{q^2-m_{\mathrm{C}}^2}+\frac{\mathrm{i}}{q^2-m_{\mathrm{C}}^2}(-\mathrm{i}\Pi_{\mathrm{C}}^{[2]}(q^2))\frac{\mathrm{i}}{q^2-m_{\mathrm{C}}^2} +\frac{\mathrm{i}}{q^2-m_{\mathrm{C}}^2}(-\mathrm{i}\Pi_{\mathrm{C}}^{[2]}(q^2))\frac{\mathrm{i}}{q^2-m_{\mathrm{C}}^2}(-\mathrm{i}\Pi_{\mathrm{C}}^{[2]}(q^2))\frac{\mathrm{i}}{q^2-m_{\mathrm{C}}^2}+\cdots \tag{10.11}$$

$$=\frac{\mathrm{i}}{q^2-m_{\mathrm{C}}^2}(1+r+r^2+\cdots) \tag{10.12}$$

where

$$r=\Pi_{\mathrm{C}}^{[2]}(q^2)/(q^2-m_{\mathrm{C}}^2). \tag{10.13}$$

The geometric series in (10.12) may be summed, at least formally[1], to give $(1-r)^{-1}$ so that (10.12) becomes

$$\frac{\mathrm{i}}{q^2-m_{\mathrm{C}}^2}\frac{1}{1-\Pi_{\mathrm{C}}^{[2]}(q^2)/(q^2-m_{\mathrm{C}}^2)}=\frac{\mathrm{i}}{q^2-m_{\mathrm{C}}^2-\Pi_{\mathrm{C}}^{[2]}(q^2)}. \tag{10.14}$$

In this form it is particularly clear that we are dealing with corrections to the simple C propagator $\mathrm{i}/(q^2-m_{\mathrm{C}}^2)$. $\Pi_{\mathrm{C}}^{[2]}$ is called the $O(g^2)$ *self-energy*.

Before proceeding with the analysis of (10.14), we note that it is a special case of the more general expression

$$\tilde{D}_{\mathrm{C}}^{'}(q^2)=\frac{\mathrm{i}}{q^2-m_{\mathrm{C}}^2-\Pi_{\mathrm{C}}(q^2)} \tag{10.15}$$

where $\tilde{D}_{\mathrm{C}}^{'}(q^2)$ is the *complete* (including all corrections) C propagator, and $\Pi_{\mathrm{C}}(q^2)$ is the sum of all 'insertions' in the C line, *excluding* those which can be cut into two separate bits by severing a single line: $\Pi_{\mathrm{C}}(q^2)$ is the *one-particle irreducible self-energy* and we must exclude all one-particle bits from it as they are already included in the geometric series summation (cf (10.11)). The amplitude $\Pi_{\mathrm{C}}^{[2]}$ which we have calculated is simply the lowest-order ($O(g^2)$) contribution to $\Pi_{\mathrm{C}}(q^2)$; an $O(g^4)$ contribution to $\Pi_{\mathrm{C}}(q^2)$ is shown in figure 10.5.

[1]Properly speaking this is valid only for $|r|<1$, yet we know that $\Pi_{\mathrm{C}}^{[2]}(q^2)$ actually diverges! As we shall see, however, renormalization will be carried out after making such quantities finite by 'regularization' (section 10.3.2), and then working systematically at a given order in g (section 10.4).

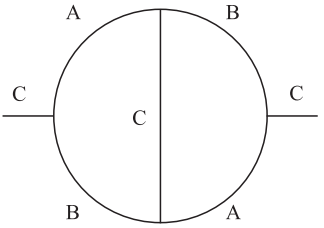


FIGURE 10.5
$O(g^4)$ contribution to $\Pi_{\rm C}(q^2)$.

10.1.2 Mass shift

We return to the expression (10.14) which includes the effect of all the iterated $O(g^2)$ bubbles in the C propagator, where $\Pi_{\rm C}^{[2]}(q^2)$ is given by

$$-\mathrm{i}\Pi_{\rm C}^{[2]}(q^2) = (-\mathrm{i}g)^2 \int \frac{\mathrm{d}^4k}{(2\pi)^4} \frac{\mathrm{i}}{k^2 - m_{\rm A}^2 + \mathrm{i}\epsilon} \frac{\mathrm{i}}{(q-k)^2 - m_{\rm B}^2 + \mathrm{i}\epsilon}. \qquad (10.16)$$

Postponing the evaluation of (10.16) (and in particular the treatment of its divergence) until section 10.3, we proceed to discuss the further implications of (10.14).

First, suppose $\Pi_{\rm C}^{[2]}$ were simply a constant, $\delta m_{\rm C}^2$ say. In the absence of this correction, we know (cf section 6.3.3) that the vanishing of the denominator of the C propagator would correspond to the 'mass-shell condition' $q^2 = m_{\rm C}^2$ appropriate to a free particle of momentum q and energy $q_0 = (\boldsymbol{q}^2 + m_{\rm C}^2)^{1/2}$, where $m_{\rm C}$ is the mass of a C particle. It seems very plausible, therefore, to interpret the constant $\delta m_{\rm C}^2$ as a shift in the $(\text{mass})^2$ of the C particle, the denominator of (10.14) now vanishing at $q_0 = (\boldsymbol{q}^2 + m_{\rm C}^2 + \delta m_{\rm C}^2)^{1/2}$, if $\Pi_{\rm C}^{[2]} \simeq \delta m_{\rm C}^2$. The idea that the mass of a particle can be changed from its 'free space' value by the presence of interactions with its 'environment' is a familiar one in condensed matter physics, as noted above. In the case of electrons in a metal, for example, it is not surprising that the presence of the lattice ions, and the attendant band structure, affect the response of conduction electrons to external fields, so that their apparent inertia changes. In the present case, the 'environment' is, in fact, the *vacuum*. The process described by the bubble $\Pi_{\rm C}^{[2]}(q^2)$ is one in which a C particle dissociates virtually into an A–B pair, which then recombine into the C particle, no other 'external' source being present. As in earlier uses of the word, by 'virtual' here is meant a process in which the participating particles leave their mass-shells. Thus, in particular, in the expression (10.16) for $\Pi_{\rm C}^{[2]}$, it will in general be the case that $k^2 \neq m_{\rm A}^2$, and $(q-k)^2 \neq m_{\rm B}^2$.

In the case of the electron in a metal, both the 'free' and the 'effective' masses are measurable quantities. But we cannot get outside the vacuum!

This strongly suggests that what we must mean by 'the physical (mass)^2' of a particle in our ABC theory is *not* the 'free' (Lagrangian) value m_i^2, which is unmeasurable, but the effective (mass)^2 which includes all vacuum interactions. This 'physical (mass)^2' may be *defined* to be that value of q^2 for which

$$q^2 - m_i^2 - \Pi_i(q^2) = 0 \tag{10.17}$$

where $\Pi_i(q^2)$ is the complete one-particle irreducible self-energy for particle type 'i'. If we call the physical mass $m_{\mathrm{ph},i}$, then, we will have $q^2 - m_i^2 - \Pi_i(q^2) = 0$ when $q^2 = m_{\mathrm{ph},i}^2$.

What we are dealing with in (10.14) is just the lowest-order contribution to $\Pi_{\mathrm{C}}(q^2)$, namely $\Pi_{\mathrm{C}}^{[2]}(q^2)$, so that in our case $m_{\mathrm{ph,C}}^2$ is determined by the condition

$$q^2 - m_{\mathrm{C}}^2 - \Pi_{\mathrm{C}}^{[2]}(q^2) = 0 \qquad \text{when } q^2 = m_{\mathrm{ph,C}}^2, \tag{10.18}$$

which (to this order) is

$$m_{\mathrm{ph,C}}^2 = m_{\mathrm{C}}^2 + \Pi_{\mathrm{C}}^{[2]}(m_{\mathrm{ph,C}}^2). \tag{10.19}$$

Once we have calculated $\Pi_{\mathrm{C}}^{[2]}$ (see section 10.3), equation (10.19) could be regarded as an equation to determine $m_{\mathrm{ph,C}}^2$ in terms of the parameter m_{C}^2, which appeared in the original ABC Lagrangian. This might, indeed, be the way such an equation would be viewed in condensed matter physics, where we should know the values of the parameters in the Lagrangian. But in the field-theory case m_{C}^2 is unobservable, so that such an equation has no predictive value. Instead, we may regard it as an equation determining (up to $O(g^2)$) m_{C}^2 in terms of $m_{\mathrm{ph,C}}^2$, thus enabling us to eliminate – to this order in g – all occurrences of the unobservable parameter m_{C}^2 from our amplitudes in favour of the physical parameter $m_{\mathrm{ph,C}}^2$. Note that $\Pi_{\mathrm{C}}^{[2]}$ contains two powers of g, so that in the spirit of systematic perturbation theory, the mass shift represented by (10.19) is a second-order correction.

The crucial point here is that $\Pi_{\mathrm{C}}^{[2]}$ depends on the cut-off Λ, whereas the physical mass $m_{\mathrm{ph,C}}^2$ clearly does not. But there is nothing to stop us supposing that the unknown and unobservable Lagrangian parameter m_{C}^2 depends on Λ in just such a way as to cancel the Λ-dependence of $\Pi_{\mathrm{C}}^{[2]}$, leaving $m_{\mathrm{ph,C}}^2$ independent of Λ. This is the beginning of the 'renormalization procedure' in quantum field theory.

10.1.3 Field strength renormalization

We now need to consider the more realistic case in which $\Pi_{\mathrm{C}}^{[2]}(q^2)$ is not a constant. Let us expand it about the point $q^2 = m_{\mathrm{ph,C}}^2$, writing

$$\Pi_{\mathrm{C}}^{[2]}(q^2) \approx \Pi_{\mathrm{C}}^{[2]}(m_{\mathrm{ph,C}}^2) + (q^2 - m_{\mathrm{ph,C}}^2)\frac{\mathrm{d}\Pi_{\mathrm{C}}^{[2]}}{\mathrm{d}q^2}\bigg|_{q^2=m_{\mathrm{ph,C}}^2} + \cdots. \tag{10.20}$$

The corrected propagator (10.14) then becomes

$$\frac{\mathrm{i}}{q^2 - m_{\mathrm{C}}^2 - \Pi_{\mathrm{C}}^{[2]}(m_{\mathrm{ph,C}}^2) - (q^2 - m_{\mathrm{ph,C}}^2)\dfrac{\mathrm{d}\Pi_{\mathrm{C}}^{[2]}}{\mathrm{d}q^2}\Big|_{q^2=m_{\mathrm{ph,C}}^2} + \cdots} \tag{10.21}$$

$$= \frac{\mathrm{i}}{(q^2 - m_{\mathrm{ph,C}}^2)\left[1 - \dfrac{\mathrm{d}\Pi_{\mathrm{C}}^{[2]}}{\mathrm{d}q^2}\Big|_{q^2=m_{\mathrm{ph,C}}^2}\right] + O(q^2 - m_{\mathrm{ph,C}}^2)^2}. \tag{10.22}$$

The expression (10.22) has indeed the expected form for a 'physical C' propagator, having the simple behaviour $\sim 1/(q^2 - m_{\mathrm{ph,C}}^2)$ for $q^2 \approx m_{\mathrm{ph,C}}^2$. However, the *normalization* of this (corrected) propagator is different from that of the 'free' one, $\mathrm{i}/(q^2 - m_{\mathrm{C}}^2)$, because of the extra factor

$$\left[1 - \frac{\mathrm{d}\Pi_{\mathrm{C}}^{[2]}}{\mathrm{d}q^2}\Big|_{q^2=m_{\mathrm{ph,C}}^2}\right]^{-1}.$$

To the order at which we are working ($O(g^2)$), it is consistent to replace this expression by

$$1 + \frac{\mathrm{d}\Pi_{\mathrm{C}}^{[2]}}{\mathrm{d}q^2}\Big|_{q^2=m_{\mathrm{ph,C}}^2}.$$

Let us see how this factor may be understood.

Our $O(g^2)$ corrected propagator is an approximation to the exact propagator which we may write as $\langle\Omega|T(\hat{\phi}_{\mathrm{C}}(x_1)\hat{\phi}_{\mathrm{C}}(x_2))|\Omega\rangle$, in coordinate space, where $|\Omega\rangle$ is the exact vacuum. The free propagator, however, is $\langle 0|T(\hat{\phi}_{\mathrm{C}}(x_1)\hat{\phi}_{\mathrm{C}}(x_2))|0\rangle$ as calculated in section 6.3.2. Consider one term in the latter, $\theta(t_1 - t_2)\times \langle 0|\hat{\phi}_{\mathrm{C}}(x_1)\hat{\phi}_{\mathrm{C}}(x_2)|0\rangle$, and insert a complete set of free-particle states '$1 = \sum_n |n\rangle\langle n|$' between the two free fields, obtaining

$$\theta(t_1 - t_2)\sum_n \langle 0|\hat{\phi}_{\mathrm{C}}(x_1)|n\rangle\langle n|\hat{\phi}_{\mathrm{C}}(x_2)|0\rangle. \tag{10.23}$$

The only free particle state $|n\rangle$ having a non-zero matrix element of the free field $\hat{\phi}_{\mathrm{C}}$ to the vacuum is the $1-\mathrm{C}$ state, for which $\langle 0|\hat{\phi}_{\mathrm{C}}(x)|\mathrm{C}, k\rangle = \mathrm{e}^{-\mathrm{i}k\cdot x}$ as we learned in chapters 5 and 6. Thus (10.23) becomes (cf equation (6.92))

$$\theta(t_1 - t_2)\int \frac{\mathrm{d}^3\boldsymbol{k}}{(2\pi)^3 2\omega_k}\mathrm{e}^{-\mathrm{i}\omega_k(t_1-t_2)+\mathrm{i}\boldsymbol{k}\cdot(\boldsymbol{x}_1-\boldsymbol{x}_2)} \tag{10.24}$$

which is exactly the first term of equation (6.92). Consider now carrying out a similar manipulation for the corresponding term of the interacting propagator, obtaining

$$\theta(t_1 - t_2)\sum_n \langle\Omega|\hat{\phi}_{\mathrm{C}}(x_1)\overline{|n\rangle}\,\overline{\langle n|}\hat{\phi}_{\mathrm{C}}(x_2)|\Omega\rangle \tag{10.25}$$

where the states $\overline{|n\rangle}$ are now the exact eigenstates of the full Hamiltonian. The crucial difference between (10.23) and (10.25) is that in (10.25), multi-particle states can appear in the states $\overline{|n\rangle}$. For example, the state $|\mathrm{A},\mathrm{B}\rangle$ consisting of an A particle and a B particle will enter, because the interaction couples this state to the 1-C states created and destroyed in $\hat{\phi}_{\mathrm{C}}$: indeed, just such an A+B state is present in $\Pi_{\mathrm{C}}^{[2]}$! This means that, whereas in the free case the 'content' of the state $\langle 0|\hat{\phi}_{\mathrm{C}}(x)$ was fully exhausted by the $1-\mathrm{C}$ state $|\mathrm{C},k\rangle$ (in the sense that all overlaps with other states $|n\rangle$ were zero), this is not so in the interacting case. The 'content' of $\langle\Omega|\hat{\phi}_{\mathrm{C}}(x)$ is not fully exhausted by the state $\overline{|\mathrm{C},k\rangle}$: rather, it has overlaps with many other states. Now the sum total of all these overlaps (in the sense of '$\sum_n \overline{|n\rangle}\,\overline{\langle n|}$') must be unity. Thus it seems clear that the 'strength' of the single matrix element $\langle\Omega|\hat{\phi}_{\mathrm{C}}(x)\overline{|\mathrm{C},k\rangle}$ in the interacting case cannot be the same as the free case (where the single state exhausted the completeness sum). However, we expect it to be true that $\langle\Omega|\hat{\phi}_{\mathrm{C}}(x)\overline{|\mathrm{C},k\rangle}$ is still basically the wavefunction for the C-particle. Hence we shall write

$$\langle\Omega|\hat{\phi}_{\mathrm{C}}(x)\overline{|\mathrm{C},k\rangle} = \sqrt{Z_{\mathrm{C}}}\mathrm{e}^{-\mathrm{i}k\cdot x} \tag{10.26}$$

where $\sqrt{Z_{\mathrm{C}}}$ is a constant to take account of the change in normalization – the *renormalization*, in fact – required by the altered 'strength' of the matrix element.

If (10.26) is accepted, we can now imagine repeating the steps leading from equation (6.92) to equation (6.98) but this time for $\langle\Omega|T(\hat{\phi}_{\mathrm{C}}(x_1)\hat{\phi}_{\mathrm{C}}(x_2))|\Omega\rangle$, retaining explicitly only the single-particle state $\overline{|\mathrm{C},k\rangle}$ in (10.25), and using the physical $(\mathrm{mass})^2$, $m^2_{\mathrm{ph,C}}$. We should then arrive at a propagator in the interacting case which has the form

$$\begin{aligned}\langle\Omega|T(\hat{\phi}_{\mathrm{C}}(x_1)\hat{\phi}_{\mathrm{C}}(x_2))|\Omega\rangle \quad &= \quad \int \frac{\mathrm{d}^4k}{(2\pi)^4}\mathrm{e}^{-\mathrm{i}k\cdot(x_1-x_2)}\Bigg\{\frac{\mathrm{i}Z_{\mathrm{C}}}{k^2-m^2_{\mathrm{ph,C}}+\mathrm{i}\epsilon} \\ &\qquad + \text{multiparticle contributions}\Bigg\}. \end{aligned} \tag{10.27}$$

The single-particle contribution in (10.27) – after undoing the Fourier transform – has exactly the same form as the one we found in (10.22), if we identify the *field strength renormalization constant* Z_{C} with the proportionality factor in (10.22), to this order:

$$Z_{\mathrm{C}} \approx Z_{\mathrm{C}}^{[2]} = 1 + \left.\frac{\mathrm{d}\Pi_{\mathrm{C}}^{[2]}}{\mathrm{d}q^2}\right|_{q^2=m^2_{\mathrm{ph,C}}}. \tag{10.28}$$

This is how the change in normalization in (10.22) is to be interpreted.

It may be helpful to sketch briefly an analogy between this 'renormalization' and a very similar one in ordinary quantum mechanical perturbation theory. Suppose we have a Hamiltonian $H = H_0 + V$ and that the $|n\rangle$ are

a complete set of orthonormal states such that $H_0|n\rangle = E_n^{(0)}|n\rangle$. The exact eigenstates $\overline{|n\rangle}$ satisfy

$$(H_0 + V)\overline{|n\rangle} = E_n\overline{|n\rangle}. \tag{10.29}$$

To obtain $\overline{|n\rangle}$ and E_n in perturbation theory, we write

$$\overline{|n\rangle} = \sqrt{N_n}|n\rangle + \sum_{i\neq n} c_{i,n}|i\rangle \tag{10.30}$$

where, if $|n\rangle$ is also normalized, we have

$$1 = N_n + \sum_{i\neq n} |c_{i,n}|^2. \tag{10.31}$$

N_n cannot be unity, since non-zero amounts of the states $|i\rangle$ $(i \neq n)$ have been 'mixed in' by the perturbation- just as the A + B state was introduced into the summation '$\sum_n \overline{|n\rangle}\,\overline{\langle n|}$', in addition to the 1 − C state. Inserting (10.30) into (10.29) and taking the bracket with $\langle j|$ yields

$$c_{j,n} = -\frac{\langle j|V\overline{|n\rangle}}{E_j^{(0)} - E_n} \tag{10.32}$$

which is still an exact expression. The lowest non-trivial approximation to $c_{j,n}$ is to take $\overline{|n\rangle} \approx \sqrt{N_n}|n\rangle$ and $E_n \approx E_n^{(0)}$ in (10.32), giving

$$c_{j,n} \approx -\sqrt{N_n}\frac{\langle j|V|n\rangle}{E_j^{(0)} - E_n^{(0)}} \equiv -\sqrt{N_n}\frac{V_{jn}}{E_j^{(0)} - E_n^{(0)}}. \tag{10.33}$$

Equation (10.31) then gives N_n as

$$N_n \approx 1\Big/\left(1 + \sum_j |V_{jn}|^2/(E_j^{(0)} - E_n^{(0)})^2\right) \approx 1 - \sum_j |V_{jn}|^2/(E_j^{(0)} - E_n^{(0)})^2 \tag{10.34}$$

to second order in V_{jn}. The reader may ponder on the analogy between (10.34) and (10.28).

10.2 The vertex correction

At the same order (g^4) of perturbation theory, we should also include, for consistency, the processes shown in figures 10.6(a) and (b). Figure 10.6(a), for example, has the general form

$$-\mathrm{i}g\frac{\mathrm{i}}{q^2 - m_{\mathrm{C}}^2}(-\mathrm{i}gG^{[2]}(p_{\mathrm{A}}, p'_{\mathrm{B}})) \tag{10.35}$$

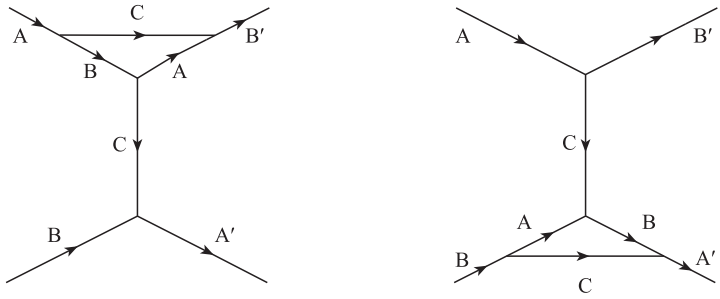


FIGURE 10.6
$O(g^4)$ contributions to A + B → A + B, involving corrections to the ABC vertices in figure 6.4.

where $-\mathrm{i}gG^{[2]}$ is the 'triangle' loop, given by an expression similar to (10.16) but with a factor $(-\mathrm{i}g)^3$ and three propagators. The 'vertex correction' $G^{[2]}$ depends on just two of its external 4-momenta because the third is determined by 4-momentum conservation, as usual. Thus, the addition of figure 10.6(a) and the $O(g^2)$ C-exchange tree diagram gives

$$-\mathrm{i}g\frac{\mathrm{i}}{q^2-m_{\mathrm{C}}^2}\{-\mathrm{i}g+(-\mathrm{i}gG^{[2]}(p_{\mathrm{A}},p'_{\mathrm{B}}))\} \tag{10.36}$$

from which it seems plausible that $G^{[2]}$ will contribute – among other effects – to a change in g. This change will be of order g^2, since we may write the $\{\ldots\}$ bracket in (10.36) as

$$-\mathrm{i}g\{1+G^{[2]}(p_{\mathrm{A}},p'_{\mathrm{B}})\} \tag{10.37}$$

where $G^{[2]}$ is dimensionless and contains a g^2 factor – hence the superscript [2].

Once again, the effect of interactions with the environment (i.e. vacuum fluctuations) has been to alter the value of a Lagrangian parameter away from the 'free' value. In the case of g the change is analogous to that in which an electron in a metal acquires an 'effective charge'. How we define the 'physical g' is less clear than in the case of the physical mass and we shall not pursue this point here, since we shall discuss it again in the more interesting case of the charge 'e' in QED, in the following chapter. At all events, some suitable definition of 'g_{ph}' can be given, so that it can be related to g after the relevant amplitudes have been computed.

Let us briefly recapitulate progress. We are studying higher-order (one-loop) corrections to tree graph amplitudes in the ABC model, which has the Lagrangian density:

$$\hat{\mathcal{L}}=\sum_i\{\tfrac{1}{2}\partial_\mu\hat{\phi}_i\partial^\mu\hat{\phi}_i-\tfrac{1}{2}m_i^2\hat{\phi}_i^2\}-g\hat{\phi}_{\mathrm{A}}\hat{\phi}_{\mathrm{B}}\hat{\phi}_{\mathrm{C}}. \tag{10.38}$$

FIGURE 10.7
Elementary one-loop amplitudes: (a) self-energy; (b) vertex correction.

We have found that the loops considered so far, namely those in figures 10.1 and 10.5, have the following qualitative effects:

(i) the position of the single-particle mass-shell condition becomes shifted away from the 'Lagrangian' value m_i^2 to a 'physical' value $m_{\mathrm{ph},i}^2$ given by the vanishing of an expression such as (10.17);

(ii) the vacuum-to-one-particle matrix elements of the fields $\hat{\phi}_i$ have to be 'renormalized' by a factor $\sqrt{Z_i}$, given by (10.28) to $O(g^2)$ for i=C, and these factors have to be included in S-matrix elements;

(iii) the propagators contain some contribution from two-particle states (e.g. ' A + B ' for the C propagator);

(iv) the Lagrangian coupling g is shifted by the interactions to a 'physical' value g_{ph}.

Responsible for these effects were two 'elementary' loops, that for $-\mathrm{i}\Pi^{[2]}$ shown in figure 10.7(a) and that for $-\mathrm{i}gG^{[2]}$ shown in figure 10.7(b). It is noteworthy that the effects (i), (ii) and (iv) all relate to changes (renormalizations, shifts) in the fields and parameters of the original Lagrangian. We say, collectively, that the 'fields, masses and coupling have been renormalized' – i.e. generically altered from their 'free' values, by the virtual interactions represented generically by figures 10.7(a) and (b). However, whereas in condensed matter physics one might well have the ambition to calculate such effects from first principles, in the field-theory case that makes no sense. Rather, by rewriting all calculated expressions (at a given order of perturbation theory) in terms of 'renormalized' quantities, we aim to eliminate the 'unknown physics scale', Λ, from the theory. Let us now see how this works in more mathematical detail.

10.3 Dealing with the bad news: a simple example

10.3.1 Evaluating $\Pi_{\rm C}^{[2]}(q^2)$

We turn our attention to the actual evaluation of a one-loop amplitude, beginning with the simplest, which is $-\mathrm{i}\Pi_{\rm C}^{[2]}(q^2)$:

$$-\mathrm{i}\Pi_{\rm C}^{[2]}(q^2) = (-\mathrm{i}g)^2 \int \frac{\mathrm{d}^4k}{(2\pi)^4} \frac{\mathrm{i}}{k^2 - m_{\rm A}^2 + \mathrm{i}\epsilon} \frac{\mathrm{i}}{(q-k)^2 - m_{\rm B}^2 + \mathrm{i}\epsilon}; \qquad (10.39)$$

in particular, we want to know the precise mathematical form of the divergence which arises when the momentum integral in (10.39) is not cut off at an upper limit Λ. This will necessitate the introduction of a few modest tricks from a large armoury (mostly due to Feynman) for dealing with such integrals.

The first move in evaluating (10.39) is to 'combine the denominators' using the identity (problem 10.2)

$$\frac{1}{AB} = \int_0^1 \frac{\mathrm{d}x}{[(1-x)A + xB]^2} \qquad (10.40)$$

(similar 'Feynman identities' exist for combining three or more denominator factors). Applying (10.40) to (10.39) we obtain

$$\begin{aligned} -\mathrm{i}\Pi_{\rm C}^{[2]}(q^2) &= g^2 \int_0^1 \mathrm{d}x \int \frac{\mathrm{d}^4k}{(2\pi)^4} \\ &\quad \times \frac{1}{[(1-x)(k^2 - m_{\rm A}^2 + \mathrm{i}\epsilon) + x((q-k)^2 - m_{\rm B}^2 + \mathrm{i}\epsilon)]^2} \end{aligned} \qquad (10.41)$$

Collecting up terms inside the $[\ldots]$ bracket and changing the integration variable to $k' = k - xq$ leads to (problem 10.3)

$$-\mathrm{i}\Pi_{\rm C}^{[2]}(q^2) = g^2 \int_0^1 \mathrm{d}x \int \frac{\mathrm{d}^4k'}{(2\pi)^4} \frac{1}{(k'^2 - \Delta + \mathrm{i}\epsilon)^2} \qquad (10.42)$$

where

$$\Delta = -x(1-x)q^2 + xm_{\rm B}^2 + (1-x)m_{\rm A}^2. \qquad (10.43)$$

The d^4k' integral means $\mathrm{d}k'^0\,\mathrm{d}^3\boldsymbol{k}'$, and $k'^2 = (k'^0)^2 - \boldsymbol{k}'^2$.

We now perform the k'^0 integration in (10.42) for which we will need the contour integration techniques explained in appendix F. The integral we want to calculate is

$$\int_{-\infty}^{\infty} \frac{\mathrm{d}k'^0}{[(k'^0)^2 - A]^2} = \frac{\partial}{\partial A} \int_{-\infty}^{\infty} \frac{\mathrm{d}k'^0}{[(k'^0)^2 - A]} \equiv \frac{\partial}{\partial A} I(A) \qquad (10.44)$$

where $A = \boldsymbol{k}'^2 + \Delta - \mathrm{i}\epsilon$. We rewrite $I(A)$ as

$$I(A) = \lim_{R\to\infty} \int_{\mathcal{C}_R} \frac{\mathrm{d}z}{[z^2 - A]} \qquad (10.45)$$

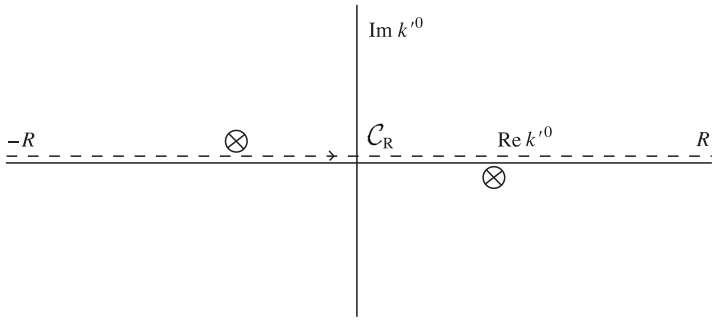


FIGURE 10.8
Location of the poles of (10.42) in the complex k'^0-plane.

where the contour $\mathcal{C}_R$ is the real axis from $-R$ to R. Next, we identify the points where the integrand $[z^2 - A]^{-1}$ ceases to be analytic (called 'poles'), which are at $z = \pm\sqrt{A} = \pm(\boldsymbol{k}'^2 + \Delta - \mathrm{i}\epsilon)^{1/2}$. Figure 10.8 shows the location of these points in the complex $z(k'^0)$-plane: note that the 'iϵ' determines in which half-plane each point lies (compare the similar role of the 'iϵ' in $(z+\mathrm{i}\epsilon)^{-1}$, in the proof in appendix F of the representation (6.93) for the θ-function). We must now 'close the contour' in order to be able to use Cauchy's integral formula of (F.19). We may do this by means of a large semicircle in either the upper ($\mathcal{C}_+$) or lower ($\mathcal{C}_-$) half-plane (again compare the discussion in appendix F). The contribution from either such semicircle vanishes as $R \to \infty$, since on either we have $z = R\mathrm{e}^{\mathrm{i}\theta}$, and

$$\int_{\mathcal{C}_+ \text{ or } \mathcal{C}_-} \frac{\mathrm{d}z}{z^2 - A} = \int \frac{R\mathrm{e}^{\mathrm{i}\theta}\mathrm{i}\,\mathrm{d}\theta}{R^2\mathrm{e}^{2\mathrm{i}\theta} - A} \to 0 \qquad \text{as } R \to \infty. \tag{10.46}$$

For definiteness, let us choose to close the contour in the upper half-plane. Then we are evaluating

$$I(A) = \lim_{R\to\infty} \oint_{\mathcal{C}=\mathcal{C}_R+\mathcal{C}_+} \frac{\mathrm{d}z}{(z - \sqrt{A})(z + \sqrt{A})} \tag{10.47}$$

around the closed contour $\mathcal{C}$ shown in figure 10.9, which encloses the single non-analytic point at $z = -\sqrt{A}$. Applying Cauchy's integral formula (F.19) with $a = -\sqrt{A}$ and $f(z) = (z - \sqrt{A})^{-1}$, we find

$$I(A) = 2\pi\mathrm{i}\frac{1}{-2\sqrt{A}} \tag{10.48}$$

and thus

$$\int_{-\infty}^{\infty} \frac{\mathrm{d}k'^0}{[(k'^0)^2 - A]^2} = \frac{\pi\mathrm{i}}{2A^{3/2}}. \tag{10.49}$$

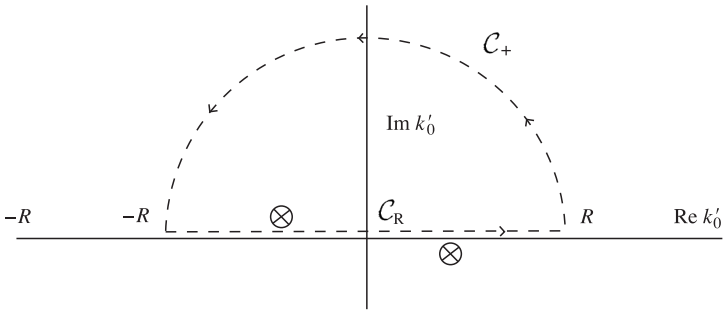


FIGURE 10.9
The closed contour $\mathcal{C}$ used in the integral (10.47).

The reader may like to try taking the other choice ($\mathcal{C}_-$) of closing contour, and check that the answer is the same. Reinstating the remaining integrals in (10.42) we have finally (as $\epsilon \to 0$)

$$-\mathrm{i}\Pi_{\mathrm{C}}^{[2]}(q^2) = \frac{\mathrm{i}}{8\pi^2} g^2 \int_0^1 \mathrm{d}x \int_0^\infty \frac{u^2\,\mathrm{d}u}{(u^2+\Delta)^{3/2}} \tag{10.50}$$

where $u = |\boldsymbol{k}'|$ and the integration over the angles of $\boldsymbol{k}'$ has yielded a factor of 4π. We see that the u-integral behaves as $\int \mathrm{d}u/u$ for large u, which is logarithmically divergent, as expected from the start.

10.3.2 Regularization and renormalization

Faced with results which are infinite, one can either try to go back to the very beginnings of the theory and see if a totally new start can avoid the infinities or one can see if they can somehow be 'lived with'. The first approach may yet, ultimately, turn out to be correct: perhaps a future theory will be altogether free of divergences (such theories do in fact exist, but none as yet successfully describes the pattern of particles and forces we actually seem to have in Nature). For the moment, it is the second approach which has been pursued – indeed with great success as we shall see in the next chapter and in volume 2.

Accepting the general framework of quantum field theory, then, the first thing we must obviously do is to modify the theory in some way so that integrals such as (10.50) do not actually diverge, so that we can at least discuss finite rather than infinite quantities. This step is called 'regularization' of the theory. There are many ways to do this but for our present purposes a simple one will do well enough, which is to cut off the u-integration in (10.50) at some finite value Λ (remember u is $|\boldsymbol{k}'|$, so Λ here will have dimensions of energy, or mass); such a step was given some physical motivation in section 10.1.1.

Then we can evaluate the integral straightforwardly and move on to the next stage.

With the upper limit in (10.50) replaced by Λ, we can evaluate the u-integral, obtaining (problem 10.4)

$$\Pi_{\rm C}^{[2]}(q^2,\Lambda^2) = \frac{-g^2}{8\pi^2}\int_0^1 {\rm d}x\,\left\{\ln\left(\frac{\Lambda+(\Lambda^2+\Delta)^{1/2}}{\Delta^{1/2}}\right) - \frac{\Lambda}{(\Lambda^2+\Delta)^{1/2}}\right\} \quad (10.51)$$

where from (10.43)

$$\Delta = -x(1-x)q^2 + xm_{\rm B}^2 + (1-x)m_{\rm A}^2. \quad (10.52)$$

Note that $\Delta > 0$ for $q^2 < 0$.

Inspection of (10.51) shows that as $\Lambda \to \infty$, $\Pi_{\rm C}^{[2]}(q^2,\Lambda^2)$ contains a divergent part proportional to $\ln\Lambda$. It is useful to isolate this divergent part, as follows. For large Λ, we can expand the terms in (10.51) in powers of Δ/Λ^2, writing

$$\Lambda + (\Lambda^2+\Delta)^{1/2} = 2\Lambda(1+\frac{\Delta}{4\Lambda^2}+\ldots) \quad (10.53)$$

and

$$\frac{\Lambda}{(\Lambda^2+\Delta)^{1/2}} = 1 - \frac{\Delta}{2\Lambda^2} + \ldots \quad (10.54)$$

It follows that

$$\Pi_{\rm C}^{[2]}(q^2,\Lambda^2) = \frac{-g^2}{8\pi^2}\int_0^1 {\rm d}x\,\left\{\ln\Lambda + (\ln 2 - 1) - \frac{1}{2}\ln\Delta\right\} \quad (10.55)$$

where terms that go to zero as $\Lambda \to \infty$ have been omitted.

Relation (10.19) then becomes

$$m_{\rm C}^2(\Lambda^2) = m_{\rm ph,C}^2 - \Pi_{\rm C}^{[2]}(q^2 = m_{\rm ph,C}^2, \Lambda^2) \quad (10.56)$$

and there will be similar relations for the A and B masses. As noted previously, after (10.19), the shift represented by (10.56) is in an $O(g^2)$ perturbative correction (because $\Pi_{\rm C}^{[2]}$ contains a factor g^2), so that – again in the spirit of systematic perturbation theory – it will be adequate to this order in g^2 to replace the Lagrangian masses $m_{\rm A}^2$, $m_{\rm B}^2$, and $m_{\rm C}^2$ *inside* the expressions for $\Pi_{\rm A}^{[2]}$, $\Pi_{\rm B}^{[2]}$ and $\Pi_{\rm C}^{[2]}$ by their physical counterparts. In this way the relations (10.56) and the two similar ones give us the prescription for rewriting the m_i^2 in terms of the $m_{{\rm ph},i}^2$ and Λ^2. Of course, when this is done in the propagators, the result is just to produce the desired form $\sim(q^2 - m_{{\rm ph},i}^2)^{-1}$, to this order.

So, for the propagator at this one-loop order, the effect of such mass shifts is essentially trivial: the large Λ behaviour is simply absorbed into m_i^2. What about $Z_{\rm C}$? This was defined via (10.28) in terms of the quantity

$$\left.\frac{{\rm d}\Pi_{\rm C}^{[2]}}{{\rm d}q^2}\right|_{q^2=m_{\rm ph,C}^2}. \quad (10.57)$$

However, equation (10.55) shows that the divergent part of $\Pi_{\rm C}^{[2]}$ is independent of q^2, or equivalently that the quantity (10.57) is finite. It follows that $Z_{\rm C}$ is finite in this theory. In other theories, quantities analogous to (10.55) might contain a q^2-dependent divergence, which would be formally absorbed in the rescaling represented by $Z_{\rm C}$.

We may also analyse the vertex correction $G^{[2]}$ of figure 10.6, and conclude that it too is finite, because there are now three propagators giving six powers of k in the denominator, with still only a four-dimensional d^4k integration. Once again, the analogous vertex correction in QED is divergent, as we shall see in chapter 11; there too this divergence can be absorbed into a redefinition of the physical charge. The ABC theory is, in fact, a 'super-renormalizable' one, meaning (loosely) that it has fewer divergences than might be expected. We shall come back to the classification of theories (renormalizable, non-renormalizable and super-renormalizable) at the end of the following chapter.

While it is not our purpose to present a full discussion of one-loop renormalization in the ABC theory (because it is not of any direct physical interest) we will use it to introduce one more important idea before turning, in the next chapter, to one-loop QED.

10.4 Bare and renormalized perturbation theory

10.4.1 Reorganizing perturbation theory

We have seen that, of the one-loop effects listed at the end of section 10.2, the mass shifts given by equations such as (10.14) do involve formal divergences as $\Lambda \to \infty$, but the vertex correction and field strength renormalization are finite in the ABC theory. We shall find that in QED the corresponding quantities are all divergent, so that the perturbative replacement of all Lagrangian parameters by their 'physical' counterparts, together with field strength renormalizations, is mandatory in QED in order to get rid of $\ln\Lambda$ terms. However, this process – of evaluating the connections between the two sets of parameters, and then inserting them into all the calculated amplitudes – is likely to be very cumbersome. In this section, we shall introduce an alternative formulation, which has both calculational and conceptual advantages.

By way of motivation, consider the QED analogue of the divergent part of equation (10.7), which contributes a correction to the bare electron mass of the form $\alpha m \ln(\Lambda/m)$ where m is the electron mass. At $\Lambda = 100$ GeV the magnitude of this is about 0.04 MeV (if we take m to have the physical value), which is a shift of some 10%. The application of perturbation theory would seem more plausible if this kind of correction were to be included from the start, so that the 'free' part of the Hamiltonian (or Lagrangian) involved the physical fields and parameters, rather than the (unobserved) ones appearing

in the original theory. Then the main effects, in some sense, would already be included by the use of these (empirical) physical quantities, and corrections would be 'more plausibly' small. This is indeed the main reason for the usefulness of such 'effective' parameters in the analogous case of condensed matter physics. Actually, of course, in quantum field theory the corrections will be just as infinite (if we send Λ to infinity) in this approach also, since whichever way we set the calculation up, we shall get loops, which are divergent. All the same, this kind of 'reorganization' does offer a more systematic approach to renormalization.

To illustrate the idea, consider again our ABC Lagrangian

$$\hat{\mathcal{L}} = \hat{\mathcal{L}}_{0,\mathrm{A}} + \hat{\mathcal{L}}_{0,\mathrm{B}} + \hat{\mathcal{L}}_{0,\mathrm{C}} + \hat{\mathcal{L}}_{\mathrm{int}} \tag{10.58}$$

where

$$\hat{\mathcal{L}}_{0,\mathrm{C}} = \tfrac{1}{2}\partial_\mu\hat{\phi}_\mathrm{C}\partial^\mu\hat{\phi}_\mathrm{C} - \tfrac{1}{2}m_\mathrm{C}^2\hat{\phi}_\mathrm{C}^2 \tag{10.59}$$

and similarly for $\hat{\mathcal{L}}_{0,\mathrm{A}}$, $\hat{\mathcal{L}}_{0,\mathrm{B}}$; and where

$$\hat{\mathcal{L}}_{\mathrm{int}} = -g\hat{\phi}_\mathrm{A}\hat{\phi}_\mathrm{B}\hat{\phi}_\mathrm{C}. \tag{10.60}$$

There are two obvious moves to make: (i) introduce the rescaled (renormalized) fields by

$$\hat{\phi}_{\mathrm{ph},i}(x) = Z_i^{-1/2}\hat{\phi}_i(x) \tag{10.61}$$

in order to get rid of the $\sqrt{Z_i}$ factors in the S-matrix elements; and (ii) introduce the physical masses $m_{\mathrm{ph},i}^2$. Consider first the non-interacting parts of $\hat{\mathcal{L}}$, namely

$$\hat{\mathcal{L}}_0 = \hat{\mathcal{L}}_{0,\mathrm{A}} + \hat{\mathcal{L}}_{0,\mathrm{B}} + \hat{\mathcal{L}}_{0,\mathrm{C}}. \tag{10.62}$$

Singling out the C-parameters for definiteness, $\hat{\mathcal{L}}_0$ can then be written as

$$\begin{aligned}
\hat{\mathcal{L}}_0 &= \tfrac{1}{2}Z_\mathrm{C}\partial_\mu\hat{\phi}_{\mathrm{ph,C}}\partial^\mu\hat{\phi}_{\mathrm{ph,C}} - \tfrac{1}{2}m_\mathrm{C}^2Z_\mathrm{C}\hat{\phi}_{\mathrm{ph,C}}^2 + \cdots \\
&= \tfrac{1}{2}\partial_\mu\hat{\phi}_{\mathrm{ph,C}}\partial^\mu\hat{\phi}_{\mathrm{ph,C}} - \tfrac{1}{2}m_{\mathrm{ph,C}}^2\hat{\phi}_{\mathrm{ph,C}}^2 \\
&\quad + \tfrac{1}{2}(Z_\mathrm{C}-1)\partial_\mu\hat{\phi}_{\mathrm{ph,C}}\partial^\mu\hat{\phi}_{\mathrm{ph,C}} - \tfrac{1}{2}(m_\mathrm{C}^2Z_\mathrm{C} - m_{\mathrm{ph,C}}^2)\hat{\phi}_{\mathrm{ph,C}}^2 + \cdots \qquad (10.63) \\
&\equiv \hat{\mathcal{L}}_{0\mathrm{ph,C}} + \{\tfrac{1}{2}\delta Z_\mathrm{C}\partial_\mu\hat{\phi}_{\mathrm{ph,C}}\partial^\mu\hat{\phi}_{\mathrm{ph,C}} \\
&\qquad - \tfrac{1}{2}(\delta Z_\mathrm{C}m_{\mathrm{ph,C}}^2 + \delta m_\mathrm{C}^2Z_\mathrm{C})\hat{\phi}_{\mathrm{ph,C}}^2\} + \cdots \qquad (10.64)
\end{aligned}$$

where $\hat{\mathcal{L}}_{0\mathrm{ph,C}}$ is the standard free-C Lagrangian in terms of the physical field and mass, which leads to a Feynman propagator $\mathrm{i}/(k^2 - m_{\mathrm{ph,C}}^2 + \mathrm{i}\epsilon)$ in the usual way; also, $\delta Z_\mathrm{C} = Z_\mathrm{C} - 1$ and $\delta m_\mathrm{C}^2 = m_\mathrm{C}^2 - m_{\mathrm{ph,C}}^2$. In (10.64) the dots signify similar rearrangements of $\hat{\mathcal{L}}_{0,\mathrm{A}}$ and $\hat{\mathcal{L}}_{0,\mathrm{B}}$. Note that Z_C and m_C^2 are understood to depend on Λ, as usual, although this has not been indicated explicitly.

We now regard '$\hat{\mathcal{L}}_{0\mathrm{ph,A}} + \hat{\mathcal{L}}_{0\mathrm{ph,B}} + \hat{\mathcal{L}}_{0\mathrm{ph,C}}$' as the 'unperturbed' part of $\hat{\mathcal{L}}$, and all the remainder of (10.64) as perturbations additional to the original $\hat{\mathcal{L}}_{\mathrm{int}}$

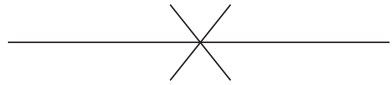


FIGURE 10.10
Counter term corresponding to the terms in braces in (10.64).

(much of theoretical physics consists of exploiting the identity '$a+b=(a+c)+(b-c)$'). The effect of this rearrangement is to introduce new perturbations, namely $\frac{1}{2}\delta Z_{\rm C}\partial_\mu\hat{\phi}_{\rm ph,C}\partial^\mu\hat{\phi}_{\rm ph,C}$ and the $\hat{\phi}^2_{\rm ph,C}$ term in (10.64), together with similar terms for the A and B fields. Such additional perturbations are called '*counter terms*' and they must be included in our new perturbation theory based on the $\hat{\mathcal{L}}_{0{\rm ph},i}$ pieces. As usual, this is conveniently implemented in terms of associated Feynman diagrams. Since both of these counter terms involve just the square of the field, it should be clear that they only have non-zero matrix elements between one-particle states, so that the associated diagram has the form shown in figure 10.10, which includes both these C-contributions. Problem 10.5 shows that the Feynman rule for figure 10.10 is that it contributes $\mathrm{i}[\delta Z_{\rm C}k^2-(\delta Z_{\rm C}m^2_{\rm ph,C}+\delta m^2_{\rm C}Z_{\rm C})]$ to the 1 C $\to$ 1 C amplitude.

The original interaction term $\hat{\mathcal{L}}_{\rm int}$ may also be rewritten in terms of the physical fields and a physical (renormalized) coupling constant $g_{\rm ph}$:

$$\begin{aligned} -g\hat{\phi}_{\rm A}\hat{\phi}_{\rm B}\hat{\phi}_{\rm C} &= -g(Z_{\rm A}Z_{\rm B}Z_{\rm C})^{1/2}\hat{\phi}_{\rm ph,A}\hat{\phi}_{\rm ph,B}\hat{\phi}_{\rm ph,C} \\ &= -g_{\rm ph}\hat{\phi}_{\rm ph,A}\hat{\phi}_{\rm ph,B}\hat{\phi}_{\rm ph,C}-(Z_{\rm V}-1)g_{\rm ph}\hat{\phi}_{\rm ph,A}\hat{\phi}_{\rm ph,B}\hat{\phi}_{\rm ph,C} \end{aligned} \tag{10.65}$$

where

$$Z_{\rm V}g_{\rm ph}=g(Z_{\rm A}Z_{\rm B}Z_{\rm C})^{1/2}. \tag{10.66}$$

The interpretation of (10.66) is clearly that '$g_{\rm ph}$' is the coupling constant describing the interactions among the $\hat{\phi}_{{\rm ph},i}$ fields, while the '$(Z_{\rm V}-1)$' term is another counter term, having the structure shown in figure 10.11.

In summary, we have reorganized $\hat{\mathcal{L}}$ so as to base perturbation theory on a part describing the free renormalized fields (rather than the fields in the original Lagrangian); in this formulation we find that, in addition to the (renormalized) ABC-interaction term, further terms have appeared which are interpreted as additional perturbations, called counter terms. These counter terms are determined, at each order in this (renormalized) perturbation theory, by what are basically self-consistency conditions – such as, for example, the requirement that the propagators really do reduce to the physical ones at the 'mass-shell' points. We shall now illustrate this procedure for the C propagator.

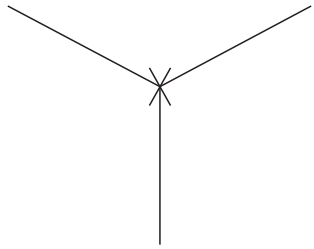


FIGURE 10.11
Counter term corresponding to the '$(Z_{\rm V}-1)$' term in (10.66).

10.4.2 The $O(g_{\rm ph}^2)$ renormalized self-energy revisited: how counter terms are determined by renormalization conditions

Let us return to the calculation of the C propagator, following the same procedure as in section 10.1, but this time 'perturbing' away from $\hat{\mathcal{L}}_{0{\rm ph},i}$ and including the contribution from the counter term of figure 10.10, in addition to the $O(g_{\rm ph}^2)$ self energy. The expression (10.14) will now be replaced by

$$\frac{\mathrm{i}}{q^2-m_{\rm ph,C}^2+q^2\delta Z_{\rm C}-\delta Z_{\rm C}m_{\rm ph,C}^2-\delta m_{\rm C}^2 Z_{\rm C}-\Pi_{\rm ph,C}^{[2]}(q^2,\Lambda^2)} \tag{10.67}$$

where

$$-\mathrm{i}\Pi_{\rm ph,C}^{[2]}(q^2,\Lambda^2)=(-\mathrm{i}g_{\rm ph})^2\int\frac{\mathrm{d}^4k}{(2\pi)^4}\frac{\mathrm{i}}{k^2-m_{\rm ph,A}^2+\mathrm{i}\epsilon}\cdot\frac{\mathrm{i}}{(q-k)^2-m_{\rm ph,B}^2+\mathrm{i}\epsilon} \tag{10.68}$$

and where we have indicated the cut-off dependence on the left-hand side, leaving it understood on the right. Comparing (10.68) with (10.39) we see that they are exactly the same, except that $\Pi_{\rm ph,C}^{[2]}$ involves the 'physical' coupling constant $g_{\rm ph}$ and the physical masses, as expected in this renormalized perturbation theory. In particular, $\Pi_{\rm ph,C}^{[2]}$ will be divergent in exactly the same way as $\Pi_{\rm C}^{[2]}$, as the cut-off Λ goes to infinity.

The essence of this 'reorganized' perturbation theory is that we now determine $\delta Z_{\rm C}$ and $\delta m_{\rm C}^2$ from the condition that as $q^2\to m_{\rm ph,C}^2$, the propagator (10.67) reduces to $\mathrm{i}/(q^2-m_{\rm ph,C}^2)$, i.e. it correctly represents the physical C propagator at the mass-shell point, with standard normalization. Expanding $\Pi_{\rm ph,C}^{[2]}(q^2)$ about $q^2=m_{\rm ph,C}^2$ then, we reach the approximate form of (10.67),

valid for $q^2 \approx m^2_{\rm ph,C}$:

$$\frac{\rm i}{(q^2-m^2_{\rm ph,C})Z_{\rm C}-\delta m^2_{\rm C}Z_{\rm C}-\Pi^{[2]}_{\rm ph,C}(m^2_{\rm ph,C},\Lambda^2)-(q^2-m^2_{\rm ph,C})\dfrac{{\rm d}\Pi^{[2]}_{\rm ph,C}}{{\rm d}q^2}\Big|_{q^2=m^2_{\rm ph,C}}}. \tag{10.69}$$

Requiring that this has the form ${\rm i}/(q^2-m^2_{\rm ph,C})$ gives

$$\begin{aligned} &\text{condition (a)} \qquad \delta m^2_{\rm C} = -Z_{\rm C}^{-1}\Pi^{[2]}_{\rm ph,C}(m^2_{\rm ph,C},\Lambda^2) \\ &\text{condition (b)} \qquad Z_{\rm C} = 1+\frac{{\rm d}\Pi^{[2]}_{\rm ph,C}}{{\rm d}q^2}\Big|_{q^2=m^2_{\rm ph,C}}. \end{aligned} \tag{10.70}$$

Looking first at condition (b), we see that our renormalization constant $Z_{\rm C}$ has, in this approach, been determined up to $O(g^2_{\rm ph})$ by an equation that is, in fact, very similar to (10.28), but it is expressed in terms of physical parameters. As regards (a), since $Z_{\rm C} = 1+O(g^2_{\rm ph})$, it is sufficient to replace it by 1 on the right-hand side of (a), so that, to this order, $\delta m^2_{\rm C} \approx -\Pi^{[2]}_{\rm ph,C}(m^2_{\rm ph,C},\Lambda^2)$. Once again, this is similar to (10.56), but written in terms of the physical quantities from the outset. We indicate that these evaluations of $Z_{\rm C}$ and $\delta m^2_{\rm C}$ are correct to second order by adding a superscript, as in $Z^{[2]}_{\rm C}$.

Of course, we have *not* avoided the infinities (in the limit $\Lambda\to\infty$) in this approach! It is still true that the loop integral in $\Pi^{[2]}_{\rm ph,C}$ diverges logarithmically and so the mass shift $(\delta m^{[2]}_{\rm C})^2$ is infinite as $\Lambda\to\infty$. Nevertheless, this is a conceptually cleaner way to do the business. It is called 'renormalized perturbation theory', as opposed to our first approach which is called 'bare perturbation theory'. What we there called the 'Lagrangian fields and parameters' are usually called the 'bare' ones; the 'renormalized' quantities are 'clothed' by the interactions.

We may now return to our propagator (10.67), and insert the results (10.70) to obtain the final important expression for the C propagator containing the one-loop $O(g^2_{\rm ph})$ renormalized self-energy:

$$\boxed{\frac{\rm i}{q^2-m^2_{\rm ph,C}-\overline{\Pi}^{[2]}_{\rm ph,C}(q^2)}} \tag{10.71}$$

where

$$\overline{\Pi}^{[2]}_{\rm ph,C}(q^2) = \Pi^{[2]}_{\rm ph,C}(q^2,\Lambda^2)-\Pi^{[2]}_{\rm ph,C}(m^2_{\rm ph,C},\Lambda^2)-(q^2-m^2_{\rm ph,C})\frac{{\rm d}\Pi^{[2]}_{\rm ph,C}}{{\rm d}q^2}\Big|_{q^2=m^2_{\rm ph,C}}. \tag{10.72}$$

We remind the reader that $\Pi^{[2]}_{\rm ph,C}(q^2,\Lambda^2)$ has exactly the same form as $\Pi^{[2]}_{\rm C}(q^2,\Lambda^2)$

except that g^2 and m_i^2 are replaced by $g^2_{\rm ph}$ and $m^2_{{\rm ph},i}$. From (10.55) it then follows that, as $\Lambda \to \infty$,

$$\Pi^{[2]}_{\rm ph,C}(q^2,\Lambda^2) = -\frac{g^2_{\rm ph}}{8\pi^2}\ln\Lambda - \frac{g^2_{\rm ph}}{8\pi^2}(\ln 2 - 1) + \frac{g^2_{\rm ph}}{16\pi^2}\int_0^1 {\rm d}x \ln\Delta(x,q^2), \quad (10.73)$$

and hence

$$\Pi^{[2]}_{\rm ph,C}(q^2,\Lambda^2) - \Pi^{[2]}_{\rm ph,C}(m^2_{\rm ph,C},\Lambda^2) = \frac{g^2_{\rm ph}}{16\pi^2}\int_0^1 {\rm d}x \ln\left\{\frac{\Delta(x,q^2)}{\Delta(x,m^2_{\rm ph,C})}\right\} \quad (10.74)$$

which is finite as $\Lambda \to \infty$. It is also clear from (10.73) that ${\rm d}\Pi^{[2]}_{\rm ph,C}/{\rm d}q^2$ is finite as $\Lambda \to \infty$. Thus the quantity $\overline{\Pi}^{[2]}_{\rm ph,C}(q^2)$ is finite as $\Lambda \to \infty$, and is understood to be evaluated in that limit; the *subtraction* in (10.74) has removed the infinity. The additional subtraction in (10.72) would in fact have removed a logarithmic divergence in $Z_{\rm C}$, had there been one. Note that the form of (10.72) guarantees that the leading behaviour of $\overline{\Pi}^{[2]}_{\rm ph,C}(q^2)$ near $q^2 = m^2_{\rm ph,C}$ is $(q^2 - m^2_{\rm ph,C})^2$, so that the behaviour of (10.71) near the mass-shell point is indeed ${\rm i}/(q^2 - m^2_{\rm ph,C})$ as desired.

A succinct way of summarizing our final renormalized result (10.71), with the definition (10.72), is to say that the C propagator may be defined by (10.71) where the $O(g^2_{\rm ph})$ renormalized self-energy $\overline{\Pi}^{[2]}_{\rm ph,C}$ satisfies the *renormalization conditions*

$$\overline{\Pi}^{[2]}_{\rm ph,C}(q^2 = m^2_{\rm ph,C}) = 0 \qquad \frac{\rm d}{{\rm d}q^2}\overline{\Pi}^{[2]}_{\rm ph,C}(q^2)\bigg|_{q^2=m^2_{\rm ph,C}} = 0. \quad (10.75)$$

Relations analogous to (10.75) clearly hold for the A and B self-energies also. In this definition, the explicit introduction and cancellation of large-Λ terms has disappeared from sight, and *all that remains is the importation of one constant from experiment*, $m^2_{\rm ph,C}$, *and a (hidden) rescaling of the fields.* It is useful to bear this viewpoint in mind when considering more general theories, including ones that are 'non-renormalizable' (see section 11.8 of the following chapter).

There is a lot of good physics in the expression (10.71), which we shall elucidate in the realistic case of QED in the next chapter. For the moment, we just whet the reader's appetite by pointing out that (10.71) must amount to the prediction of a finite, calculable correction to the Yukawa 1 − C exchange potential, which after all is given by the Fourier transform of the (static form of) the propagator, as we learned long ago. In the case of QED, this will amount to a calculable correction to Coulomb's law, due to radiative corrections, as we shall discuss in section 11.5.1.

There is an important technical implication we may draw from (10.75). Consider the Feynman diagram of figure 10.12 in which a propagator correction has been inserted in an *external* line. This diagram is of order $g^4_{\rm ph}$, and

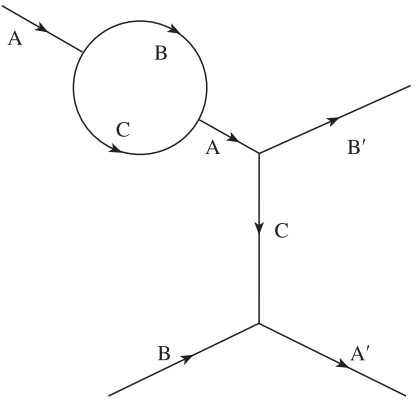


FIGURE 10.12
$O(g^4)$ contribution to A + B $\rightarrow$ A + B, involving a propagator correction inserted in an external line.

should presumably be included along with the others at this order. However, the conditions (10.75) – in this case written for $\overline{\Pi}^{[2]}_{\mathrm{ph,A}}$ – imply that it vanishes. Omitting irrelevant factors, the amplitude for figure 10.12 is

$$\overline{\Pi}^{[2]}_{\mathrm{ph,A}}(p_{\mathrm{A}})\frac{1}{p^2_{\mathrm{A}} - m^2_{\mathrm{ph,A}}}\frac{1}{q^2 - m^2_{\mathrm{ph,C}}} \tag{10.76}$$

and we need to take the limit $p^2_{\mathrm{A}} \rightarrow m^2_{\mathrm{ph,A}}$ since the external A particle is on-shell. Expanding $\overline{\Pi}^{[2]}_{\mathrm{ph,A}}$ about the point $p^2_{\mathrm{A}} = m^2_{\mathrm{ph,A}}$ and using conditions (10.75) for C $\rightarrow$ A we see that (10.76) vanishes. Thus with this definition, propagator corrections do not need to be applied to external lines.

10.5 Renormalizability

We have seen how divergences present in self-energy loops like figure 10.7(a) can be eliminated by supposing that the 'bare' masses in the original Lagrangian depend on the cut-off in just such a way as to cancel the divergences, leaving a finite value for the physical masses. The latter are, however, parameters to be taken from experiment: they are not calculable. Alternatively, we may rephrase perturbation theory in terms of renormalized quantities from the outset, in which case the loop divergence is cancelled by appropriate counter terms; but again the physical masses have to be taken from experiment. We pointed out that, in the ABC theory, neither the field strength renormalizations Z_i nor the vertex diagrams of figure 10.5 were divergent, but we shall see

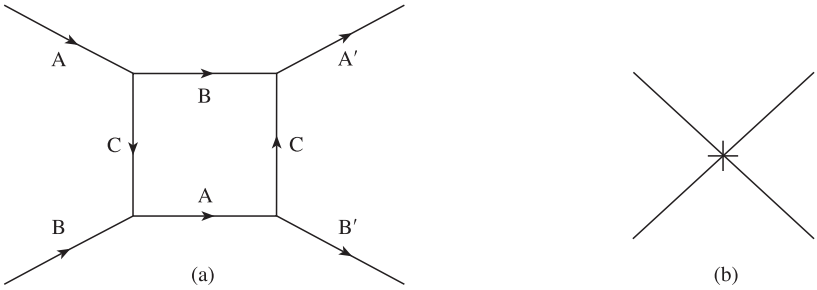


FIGURE 10.13
(a) $O(g^4)$ one-loop contribution to A + B → A + B; (b) counter term that would be required if (a) were divergent.

in the next chapter that the analogous quantities in QED are divergent. These divergences too can be absorbed into redefinitions of the 'physical' fields and a 'physical' coupling constant (the latter again to be taken from experiment). Or, again, such divergences can be cancelled by appropriate counter terms in the renormalized perturbation theory approach.

In general, a theory will have various divergences at the one-loop level, and new divergences will enter as we go up in order of perturbation theory (or number of loops). Typically, therefore, quantum field theories betray sensitivity to unknown short-distance physics by the presence of formal divergences in loops, as a cut-off $\Lambda \to \infty$. In a *renormalizable* theory, this sensitivity can be systematically removed by accepting that a finite number of parameters are uncalculable, and must be taken from experiment. These are the suitably defined 'physical' values of the masses and coupling constants appearing in the Lagrangian. Once these parameters are given, all other quantities are finite and calculable, to any desired order in perturbation theory – assuming, of course, that terms in successive orders diminish sensibly in size.

Alternatively, we may say that a renormalizable theory is one in which a finite number of counter terms can be so chosen as to cancel all divergences order by order in renormalized perturbation theory. Note, now, that the only available counter terms are the ones which arise in the process of 'reorganizing' the original theory in terms of renormalized quantities plus extra bits (the counter terms). All the counter terms must correspond to masses, interactions, etc which are present in the original (or 'bare') Lagrangian – which is, in fact, *the* theory we are trying to make sense of! We are not allowed to add in any old kind of counter term – if we did, we would be redefining the theory.

We can illustrate this point by considering, for example, a one-loop ($O(g^4)$) contribution to AB → AB scattering, as shown in figure 10.13(a). If this graph is divergent, we will need a counter term with the structure shown in figure 10.13(b) to cancel the divergence – but there is no such 'contact' AB → AB

interaction in the original theory (it would have the form $\lambda\hat{\phi}_{\mathrm{A}}^2(x)\hat{\phi}_{\mathrm{B}}^2(x)$). In fact, the graph is convergent, as indicated by the usual power-counting (four powers of k in the numerator, eight in the denominator from the four propagators). And indeed, the ABC theory is renormalizable – or rather, as noted earlier, 'super-renormalizable'.

We shall have something more to say about renormalizability and non-renormalizability (is it fatal?), at the end of the following chapter. The first and main business, however, will be to apply what we have learned here to QED.

Problems

10.1 Carry out the indicated change of variables so as to obtain (10.4) from (10.3).

10.2 Verify the Feynman identity (10.40).

10.3 Obtain (10.42) from (10.41).

10.4 Obtain (10.51) from (10.50), having replaced the upper limit of the u-integral by Λ.

10.5 Obtain the Feynman rule quoted in the text for the sum of the counter terms appearing in (10.64).

11

Loops and Renormalization II: QED

> The present electrodynamics is certainly incomplete, but is no longer certainly incorrect.
>
> —F. J. Dyson (1949b)

We now turn to the analysis of loop corrections in QED. As we might expect, a theory with fermionic and gauge fields proves to be a tougher opponent than one with only spinless particles, even though we restrict ourselves to one-loop diagrams only.

At the outset we must make one important disclaimer. In QED many loop diagrams diverge not only as the loop momentum goes to infinity ('ultraviolet divergence') but also as it goes to zero ('infrared divergence'). This phenomenon can only arise when there are massless particles in the theory – for otherwise the propagator factors $\approx(k^2 - M^2)^{-1}$ will always prevent any infinity at low k. Of course, in a gauge theory we do have just such massless quanta. Our main purpose here is to demonstrate how the *ultraviolet* divergences can be tamed and we must refer the reader to Weinberg (1995, chapter 13), or to Peskin and Schroeder (1995, section 6.5), for instruction in dealing with the infrared problem. The remedy lies, essentially, in a careful consideration of the contribution, to physical cross sections, of amplitudes involving the *real* emission of very low frequency photons, along with infrared divergent virtual photon processes. It is a 'technical' problem, having to do with massless particles (of which there are not that many), whereas ultraviolet divergences are generic.

11.1 Counter terms

We shall consider the simplest case of a single fermion of bare mass m_0 and bare charge e_0 ($e_0 > 0$) interacting with the Maxwell field, for which the bare (i.e. actual!) Lagrangian is

$$\hat{\mathcal{L}} = \bar{\hat{\psi}}_0(\mathrm{i}\hat{\not\partial} - m_0)\hat{\psi}_0 - e_0\bar{\hat{\psi}}_0\gamma^\mu\hat{\psi}_0\hat{A}_{0\mu} - \frac{1}{4}\hat{F}_{0\mu\nu}\hat{F}_0^{\mu\nu} - \frac{1}{2\xi_0}(\partial\cdot\hat{A}_0)^2 \qquad (11.1)$$

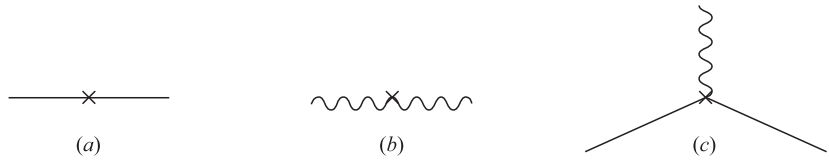


FIGURE 11.1
Counter terms in QED: (a) electron mass and wavefunction; (b) photon wavefunction; (c) vertex part.

according to chapter 7. We shall adopt the 'renormalized perturbation theory' approach and begin by introducing field strength renormalizations via

$$\hat{\psi} = Z_2^{-1/2}\hat{\psi}_0 \tag{11.2}$$

$$\hat{A}^\mu = Z_3^{-1/2}\hat{A}_0^\mu \tag{11.3}$$

where the 'physical' fields and parameters will now simply have no '0' subscript. This will lead to a rewriting of the free and gauge-fixing part of (11.1):

$$\begin{aligned}
&\bar{\hat{\psi}}_0(\mathrm{i}\partial\!\!\!/ - m_0)\hat{\psi}_0 - \frac{1}{4}\hat{F}_{0\mu\nu}\hat{F}_0^{\mu\nu} - \frac{1}{2\xi_0}(\partial\cdot\hat{A}_0)^2 \\
&\quad = \bar{\hat{\psi}}(\mathrm{i}\partial\!\!\!/ - m)\hat{\psi} - \frac{1}{4}\hat{F}_{\mu\nu}\hat{F}^{\mu\nu} - \frac{1}{2\xi}(\partial\cdot\hat{A})^2 \\
&\qquad + [(Z_2-1)\bar{\hat{\psi}}\mathrm{i}\partial\!\!\!/\hat{\psi} - \delta m\bar{\hat{\psi}}\hat{\psi}] - \tfrac{1}{4}(Z_3-1)\hat{F}_{\mu\nu}\hat{F}^{\mu\nu}
\end{aligned} \tag{11.4}$$

where $\xi = \xi_0/Z_3$ and $\delta m = m_0 Z_2 - m$ (compare (10.64)). We see the emergence of the expected '$\bar{\hat{\psi}}\ldots\hat{\psi}$' and '$\hat{F}\cdot\hat{F}$' counter terms in (11.4), affecting both the fermion and the gauge-field propagators. Next, we write the interaction in terms of a physical e, and the physical fields, together with a compensating third counter term:

$$-e_0\bar{\hat{\psi}}_0\gamma^\mu\hat{\psi}_0\hat{A}_{0\mu} = -e\bar{\hat{\psi}}\gamma^\mu\hat{\psi}\hat{A}_\mu - (Z_1-1)e\bar{\hat{\psi}}\gamma^\mu\hat{\psi}\hat{A}_\mu \tag{11.5}$$

where, with the aid of (11.2) and (11.3),

$$Z_1 e = e_0 Z_2 Z_3^{1/2}. \tag{11.6}$$

The three counter terms are represented diagrammatically as shown in figures 11.1(a), (b) and (c), for which the Feynman rules are, respectively,

$$\begin{aligned}
&\text{(a):} \quad \mathrm{i}[k\!\!\!/(Z_2-1) - \delta m] \\
&\text{(b):} \quad -\mathrm{i}(g^{\mu\nu}k^2 - k^\mu k^\nu)(Z_3-1) \\
&\text{(c):} \quad -\mathrm{i}e\gamma^\mu(Z_1-1).
\end{aligned} \tag{11.7}$$

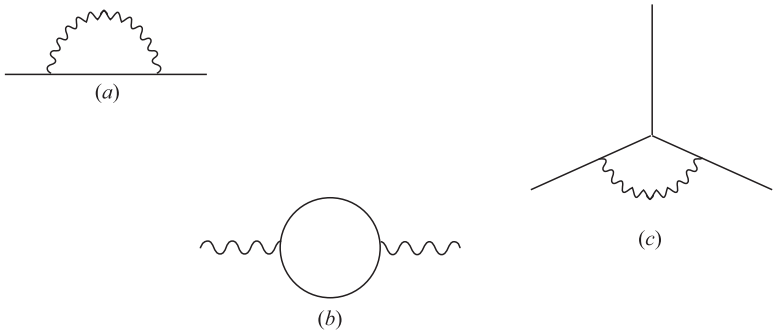


FIGURE 11.2
Elementary one-loop divergent diagrams in QED.

These counter terms will compensate for the ultraviolet divergences of the three elementary loop diagrams of figure 11.2, and in fact they are sufficient to eliminate all such divergences in all QED loops.

Before proceeding further we remark that we already have a first indication that renormalizing a gauge theory presents some new features. Consider the two counter terms involving $Z_2 - 1$ and $Z_1 - 1$; their sum gives

$$\bar{\hat{\psi}}[\mathrm{i}(Z_2 - 1)\partial\!\!\!/ - e(Z_1 - 1)\hat{A\!\!\!/}]\hat{\psi} \tag{11.8}$$

which is *not* of the 'gauge principle' form '$\mathrm{i}\partial\!\!\!/ - eA\!\!\!/$'! Unless, of course, $Z_1 = Z_2$. This relation between the two quite different renormalization constants is, in fact, true to all orders in perturbation theory, as a consequence of a *Ward identity* (Ward 1950), which is itself a consequence of gauge invariance. We shall discuss the Ward identity and $Z_1 = Z_2$ at the one loop level in section 11.6.

11.2 The $O(e^2)$ fermion self-energy

In analogy with $-\mathrm{i}\Pi_{\mathrm{C}}^{[2]}$, the amplitude corresponding to figure 11.2(a) is the fermion self-energy $-\mathrm{i}\Sigma^{[2]}$ where

$$-\mathrm{i}\Sigma^{[2]}(p) = (-\mathrm{i}e)^2 \int \gamma^\nu \frac{-\mathrm{i}g_{\mu\nu}}{k^2} \frac{\mathrm{i}}{p\!\!\!/ - k\!\!\!/ - m} \gamma^\mu \frac{\mathrm{d}^4 k}{(2\pi)^4} \tag{11.9}$$

and we have now chosen the gauge $\xi = 1$. As expected, the d^4k integral in (11.9) diverges for large k – this time more seriously than the integral in $\Pi_{\mathrm{C}}^{[2]}$, because there are only three powers of k in the denominator of (11.9) as opposed to four in (10.7). Once again, we need to choose some form of

regularization to make (11.9) ultraviolet finite. We shall not be specific (as yet) about what choice we are making, since whatever it may be the outcome will be qualitatively similar to the $\Pi_{\rm C}^{[2]}$ case.

There is, however, one interesting new feature in this (fermion) case. As previously indicated, power-counting in the integral of (11.9) might lead us to expect that – if we adopt a simple cut-off – the leading ultraviolet divergence of $\Sigma^{[2]}$ would be proportional to Λ rather than $\ln\Lambda$. This is because we have that one extra power of k in the numerator and $\Sigma^{[2]}$ has dimensions of mass. However, this is not so. The leading p-independent divergence is, in fact, proportional to $m\ln(\Lambda/m)$. The reason for this is important and it has interesting generalizations. Suppose that m in (11.4) were set equal to zero. Then, as we saw in problem 9.4, the two helicity components $\hat{\psi}_{\rm L}$ and $\hat{\psi}_{\rm R}$ of the electron field will not be coupled by the QED interaction. It follows that no terms of the form $\bar{\hat{\psi}}_{\rm L}\hat{\psi}_{\rm R}$ or $\bar{\hat{\psi}}_{\rm R}\hat{\psi}_{\rm L}$ can be generated, and hence no perturbatively induced mass term, if $m=0$. The perturbative mass shift must be proportional to m and therefore, on dimensional grounds, only logarithmically divergent.

There is also a p-dependent divergence of the self-energy, of which warning was given in section 10.3.2. As in the scalar case, this will be associated with the field strength renormalization factor Z_2. It is proportional to $\not{p}\ln(\Lambda/m)$ (Z_2 is the coefficient of $\bar{\partial}\!\!\!/$ in (11.8), which leads to $\not{p}$ in momentum space). The upshot is that the fermion propagator, including the one-loop renormalized self-energy, is given by

$$\frac{\mathrm{i}}{\not{p}-m-\bar{\Sigma}^{[2]}(p)} \tag{11.10}$$

where (cf (10.74))

$$\bar{\Sigma}^{[2]}(p)=\Sigma^{[2]}(p)-\Sigma^{[2]}(\not{p}=m)-(\not{p}-m)\frac{\mathrm{d}\Sigma^{[2]}}{\mathrm{d}\not{p}}\bigg|_{\not{p}=m}. \tag{11.11}$$

Whatever form of regularization is used, the twice-subtracted $\bar{\Sigma}^{[2]}$ will be finite and independent of the regulator when it is removed. In terms of the 'compensating' quantities Z_2 and m_0-m, we find (problem 11.1, cf (10.70))

$$Z_2=1+\frac{\mathrm{d}\Sigma^{[2]}}{\mathrm{d}\not{p}}\bigg|_{\not{p}=m} \qquad m_0-m=-Z_2^{-1}\Sigma^{[2]}(\not{p}=m). \tag{11.12}$$

Note that, as in the case of $\bar{\Pi}_C^{[2]}$, the definition (11.11) of $\bar{\Sigma}^{[2]}$ implies that propagator corrections vanish for external (on-shell) fermions. The quantities Z_2 and m_0 determined by (11.12) now carry a superscript '[2]' to indicate that they are correct at $O(e^2)$.

We must now remind the reader that, although we have indeed eliminated the ultraviolet divergences in $\bar{\Sigma}^{[2]}$ by the subtractions of (11.11), there remains an untreated infrared divergence in $\mathrm{d}\Sigma^{[2]}/\mathrm{d}\not{p}$. To show how this is dealt with

would take us beyond our intended scope, as explained at the start of the chapter. Suffice it to say that by the introduction of a 'regulating' photon mass μ^2, and consideration of relevant real photon processes along with virtual ones, these infrared problems can be controlled (Weinberg 1995, Peskin and Schroeder 1995).

11.3 The $O(e^2)$ photon self-energy

The amplitude corresponding to figure 11.2(b) is $\mathrm{i}\Pi^{[2]}_{\mu\nu}(q)$ where

$$\mathrm{i}\Pi^{[2]}_{\mu\nu}(q) = (-1)(-\mathrm{i}e)^2\mathrm{Tr}\int\frac{\mathrm{d}^4k}{(2\pi)^4}\frac{\mathrm{i}}{\not q+\not k-m}\gamma_\mu\frac{\mathrm{i}}{\not k-m}\gamma_\nu \quad (11.13)$$

$$= -e^2\int\frac{\mathrm{d}^4k}{(2\pi)^4}\frac{\mathrm{Tr}[(\not q+\not k+m)\gamma_\mu(\not k+m)\gamma_\nu]}{[(q+k)^2-m^2][k^2-m^2]}. \quad (11.14)$$

Once again, this *photon self-energy* is analogous to the scalar particle self-energy of chapter 10. There are two new features to be commented on in (11.14). The first is the overall '-1' factor, which occurs whenever there is a *closed fermion loop*. The keen reader may like to pursue this via problem 11.2. The second feature is the appearance of the trace symbol 'Tr': this is plausible as the amplitude is basically a $1\gamma\to1\gamma$ one with no spinor indices, but again the reader can follow that through in problem 11.3.

We now want to go some way into the calculation of $\Pi^{[2]}_{\mu\nu}$ because it will, in the end, contain important physics – for example, corrections to Coulomb's law. The first step is to evaluate the numerator trace factor using the theorems of section 8.2.3. We find (problem 11.4)

$$\mathrm{Tr}[(\not q+\not k+m)\gamma_\mu(\not k+m)\gamma_\nu] = 4\{(q_\mu+k_\mu)k_\nu+(q_\nu+k_\nu)k_\mu - g_{\mu\nu}((q\cdot k)+k^2-m^2)\}. \quad (11.15)$$

We then use the Feynman identity (10.40) to combine the denominators, yielding

$$\frac{1}{[(q+k)^2-m^2][k^2-m^2]} = \int_0^1\mathrm{d}x\frac{1}{[k'^2-\Delta_\gamma+\mathrm{i}\epsilon]^2} \quad (11.16)$$

where $k'=k+xq$, $\Delta_\gamma=-x(1-x)q^2+m^2$ (note that Δ_γ is precisely the same as Δ of (10.43) with $m_A=m_B=m$) and we have reinstated the implied '$\mathrm{i}\epsilon$'. Making the shift to the variable k' in the numerator factor (11.15) produces a revised numerator which is

$$4\{2k'_\mu k'_\nu-g_{\mu\nu}(k'^2-\Delta_\gamma)-2x(1-x)(q_\mu q_\nu-g_{\mu\nu}q^2)+\text{terms linear in } k'\} \quad (11.17)$$

where the terms linear in k' will vanish by symmetry when integrated over k' in (11.14). Our result so far is therefore

$$\mathrm{i}\Pi^{[2]}_{\mu\nu}(\mathrm{q}^2) = -4e^2\int_0^1 \mathrm{d}x\left\{\int\frac{\mathrm{d}^4k'}{(2\pi)^4}\left[\frac{2k'_\mu k'_\nu}{(k'^2-\Delta_\gamma+\mathrm{i}\epsilon)^2}-\frac{g_{\mu\nu}}{(k'^2-\Delta_\gamma+\mathrm{i}\epsilon)}\right]\right\} + 8e^2(q_\mu q_\nu - g_{\mu\nu}q^2)\int_0^1\mathrm{d}x\int\frac{\mathrm{d}^4k'}{(2\pi)^4}\frac{x(1-x)}{(k'^2-\Delta_\gamma+\mathrm{i}\epsilon)^2}. \quad (11.18)$$

Consider now the ultraviolet divergences of (11.18), adopting a simple cut-off as a regularization. The terms in the first line are both apparently quadratically divergent, while the integral in the second line is logarithmically divergent. What counter terms do we have to cancel these divergences? The answer is that the '(Z_3-1)' counter term of figure 11.1(b) is of exactly the right form to cancel the logarithmic divergence in the second line of (11.18), but we have no counter term proportional to the $g_{\mu\nu}$ term in the first line. Note, incidentally, that we can argue from Lorentz covariance (see appendix D) that

$$\int\frac{\mathrm{d}^4k'}{(2\pi)^4}\frac{k'_\mu k'_\nu}{(k'^2-\Delta_\gamma+\mathrm{i}\epsilon)^2} = f(\Delta_\gamma)g_{\mu\nu} \quad (11.19)$$

so that taking the dot product of both sides with $g^{\mu\nu}$ we deduce that

$$\int\frac{\mathrm{d}^4k'}{(2\pi)^4}\frac{2k'_\mu k'_\nu}{(k'^2-\Delta_\gamma+\mathrm{i}\epsilon)^2} = \frac{1}{2}\int\frac{\mathrm{d}^4k'}{(2\pi)^4}\frac{k'^2 g_{\mu\nu}}{(k'^2-\Delta_\gamma+\mathrm{i}\epsilon)^2}. \quad (11.20)$$

It follows that both the terms in the first line of (11.18) produce a divergence of the form $\sim\Lambda^2 g_{\mu\nu}$, and they do not cancel, at least in our simple cut-off regularization.

A term proportional to $g_{\mu\nu}$ is, in fact, a photon mass term. A Lagrangian mass term for the photon would have the form $\frac{1}{2}m^2_{\gamma_0}g_{\mu\nu}\hat{A}^\mu_0\hat{A}^\nu_0$, which after introducing the rescaled $\hat{A}_\mu$ will generate a counter term proportional to $g_{\mu\nu}\hat{A}^\mu\hat{A}^\nu$, and an associated Feynman amplitude proportional to $g_{\mu\nu}$. But such a term $m^2_{\gamma_0}$ violates gauge invariance! (It is plainly not invariant under (7.69).) Evidently the simple momentum cut-off that we have adopted as a regularization procedure does not respect gauge invariance. We saw in section 8.6.2 that gauge invariance implied the condition

$$q^\mu T_\mu = 0 \quad (11.21)$$

where q is the 4-momentum of a photon entering a one-photon amplitude T_μ. Our discussion of (11.21) was limited in section 8.6.2 to the case of a real external photon, whereas the photon lines in $\mathrm{i}\Pi^{[2]}_{\mu\nu}$ are internal and virtual; nevertheless it is still true that gauge invariance implies (Peskin and Schroeder 1995, section 7.4)

$$q^\mu\Pi^{[2]}_{\mu\nu} = q^\nu\Pi^{[2]}_{\mu\nu} = 0. \quad (11.22)$$

Condition (11.22) is guaranteed by the tensor structure $(q_\mu q_\nu - g_{\mu\nu}q^2)$ of the *second* line in (11.18), provided the divergence is regularized. As previously implied, a simple cut-off Λ suffices for this term, since it does not alter the

tensor structure, and the Λ-dependence can be compensated by the '$Z_3 - 1$' counter term which has the same tensor structure (cf figure 11.2(b)). But what about the first line of (11.18)? Various gauge-invariant regularizations have been used, the effect of all of which is to cause the first line of (11.18) to vanish. The most widely used, since the 1970s, is the *dimensional regularization* technique introduced by 't Hooft and Veltman (1972), which involves the 'continuation' of the number of space–time dimensions from four to d (< 4). As d is reduced, the integrals tend to diverge less, and the divergences can be isolated via the terms which diverge as $d \to 4$. Using gauge-invariant dimensional regularization, the two terms in the first line of (11.18) are found to cancel each other exactly, leaving just the manifestly gauge invariant second line (see appendix O of volume 2).

We proceed to the next step, renormalizing the gauge-invariant part of $\mathrm{i}\Pi^{[2]}_{\mu\nu}(q^2)$.

11.4 The $O(e^2)$ renormalized photon self-energy

The surviving (gauge-invariant) term of $\Pi^{[2]}_{\mu\nu}$ is

$$\begin{aligned} \mathrm{i}\Pi^{[2]}_{\mu\nu}(q^2) &= 8e^2(q_\mu q_\nu - q^2 g_{\mu\nu}) \int_0^1 \mathrm{d}x \int \frac{\mathrm{d}^4 k'}{(2\pi)^4} \frac{x(1-x)}{(k'^2 - \Delta_\gamma + \mathrm{i}\epsilon)^2} \qquad (11.23) \\ &\equiv \mathrm{i}(q^2 g_{\mu\nu} - q_\mu q_\nu)\Pi^{[2]}_\gamma(q^2). \qquad (11.24) \end{aligned}$$

The $\mathrm{d}^4 k'$ integral in (11.23) is exactly the same as the one in (10.42), with Δ replaced by Δ_γ. It contains a logarithmic divergence, which we regulate as before by a simple cut-off Λ, so that we are dealing with the gauge-invariant quantity $\Pi^{[2]}_\gamma(q^2, \Lambda^2)$. The calculation leading to (10.55) then tells us that, as $\Lambda \to \infty$,

$$\Pi^{[2]}_\gamma(q^2, \Lambda^2) = -\frac{e^2}{\pi^2} \int_0^1 \mathrm{d}x \left\{ \ln \Lambda + (\ln 2 - 1) - \frac{1}{2} \ln \Delta_\gamma \right\}. \qquad (11.25)$$

The analogue of (10.11) is then (in the gauge $\xi = 1$)

$$\begin{aligned} &\frac{-\mathrm{i}g_{\mu\nu}}{q^2} + \frac{-\mathrm{i}g_{\mu\rho}}{q^2} \cdot \mathrm{i}(q^2 g^{\rho\sigma} - q^\rho q^\sigma)\Pi^{[2]}_\gamma(q^2, \Lambda^2) \cdot \frac{-\mathrm{i}g_{\sigma\nu}}{q^2} \\ &\quad + \frac{-\mathrm{i}g_{\mu\rho}}{q^2} \cdot \mathrm{i}(q^2 g^{\rho\sigma} - q^\rho q^\sigma)\Pi^{[2]}_\gamma(q^2, \Lambda^2) \cdot \frac{-\mathrm{i}g_{\sigma\tau}}{q^2} \\ &\quad \cdot \mathrm{i}(q^2 g^{\tau\eta} - q^\tau q^\eta) \cdot \Pi^{[2]}_\gamma(q^2, \Lambda^2) \cdot \frac{-\mathrm{i}g_{\eta\nu}}{q^2} + \cdots \\ &= \frac{-\mathrm{i}g_{\mu\nu}}{q^2} + \frac{-\mathrm{i}g_{\mu\rho}}{q^2} P^\rho_\nu \Pi^{[2]}_\gamma(q^2, \Lambda^2) + \frac{-\mathrm{i}g_{\mu\rho}}{q^2} P^\rho_\tau P^\tau_\nu (\Pi^{[2]}_\gamma(q^2, \Lambda^2)^2 + \cdots \end{aligned} \qquad (11.26)$$

where

$$P^\rho_\nu = g^\rho_\nu - \frac{q^\rho q_\nu}{q^2}$$

and

$$g^\rho_\nu = \delta^\rho_\nu$$

(i.e. the 4×4 unit matrix). It is easy to check (problem 10.5) that $P^\rho_\tau P^\tau_\nu = P^\rho_\nu$. Hence the series (11.26) becomes

$$\begin{aligned}
&\frac{-\mathrm{i}g_{\mu\nu}}{q^2} + \frac{-\mathrm{i}g_{\mu\rho}}{q^2}P^\rho_\nu[\Pi^{[2]}_\gamma(q^2,\Lambda^2) + (\Pi^{[2]}_\gamma(q^2,\Lambda^2))^2 + \cdots] \\
&\quad = \frac{-\mathrm{i}g_{\mu\nu}}{q^2} + \frac{-\mathrm{i}g_{\mu\rho}}{q^2}P^\rho_\nu[1 + \Pi^{[2]}_\gamma(q^2,\Lambda^2) + (\Pi^{[2]}_\gamma(q^2,\Lambda^2))^2 + \cdots] + \frac{\mathrm{i}g_{\mu\rho}}{q^2}P^\rho_\nu \\
&\quad = \frac{-\mathrm{i}(g_{\mu\nu} - q_\mu q_\nu/q^2)}{q^2(1-\Pi^{[2]}_\gamma(q^2,\Lambda^2))} - \frac{\mathrm{i}}{q^2}\left(\frac{q_\mu q_\nu}{q^2}\right) \qquad (11.27)
\end{aligned}$$

after summing the geometric series, exactly as in (10.11)–(10.14).

But we have forgotten the counter term of figure 11.1(b), which contributes an amplitude $-\mathrm{i}(g^{\mu\nu}q^2 - q^\mu q^\nu)(Z_3 - 1)$. This has the effect of replacing $\Pi^{[2]}_\gamma$ in (11.27) by $\Pi^{[2]}_\gamma - (Z_3 - 1)$ and we arrive at the form

$$\frac{-\mathrm{i}(g_{\mu\nu} - q_\mu q_\nu/q^2)}{q^2(Z_3 - \Pi^{[2]}_\gamma(q^2,\Lambda^2))} - \frac{\mathrm{i}}{q^2}\frac{q_\mu q_\nu}{q^2}. \qquad (11.28)$$

Now in any S-matrix element, at least one end of this corrected propagator will connect to an external charged particle line via a vertex of the form $j^\mu_{\mathrm{a}}(p,p')$ (cf (8.98) and (8.99) for example), as in figure 11.3. But, as we have seen in (8.100), current conservation implies

$$q_\mu j^\mu_{\mathrm{a}}(p,p') = 0. \qquad (11.29)$$

Hence the parts of (11.28) with $q_\mu q_\nu$ factors will not contribute to physical scattering amplitudes, and our $O(e^2)$ corrected photon propagator effectively takes the simple form

$$\frac{-\mathrm{i}g_{\mu\nu}}{q^2(Z_3 - \Pi^{[2]}_\gamma(q^2,\Lambda^2))}. \qquad (11.30)$$

We must now determine Z_3 from the condition (just as for the C propagator) that (11.30) has the form $-\mathrm{i}g_{\mu\nu}/q^2$ as $q^2 \to 0$ (the mass-shell condition). This gives

$$Z^{[2]}_3 = 1 + \Pi^{[2]}_\gamma(0,\Lambda^2) \qquad (11.31)$$

the superscript on Z_3 indicating as usual that it is an $O(e^2)$ calculation as evidenced by the e^2 factor in (11.18). We note from equation (11.25) that $\Pi^{[2]}_\gamma(0,\Lambda^2)$ contains a $\ln\Lambda$ part, so that this time the field renormalization constant Z_3 diverges when the cut-off is removed.

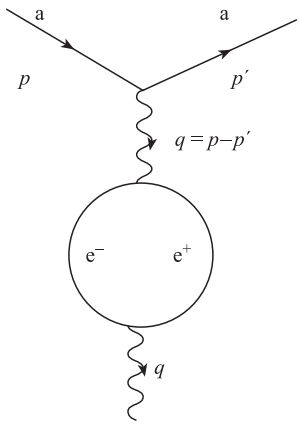


FIGURE 11.3
One-loop corrected photon propagator connected to a charged particle vertex.

Inserting (11.31) into (11.30) we obtain the final important expression for the γ-propagator including the one-loop renormalized self-energy (cf (10.71)):

$$\boxed{\frac{-\mathrm{i}g_{\mu\nu}}{q^2(1-\bar{\Pi}_\gamma^{[2]}(q^2))}} \tag{11.32}$$

where

$$\bar{\Pi}_\gamma^{[2]}(q^2) = \Pi_\gamma^{[2]}(q^2,\Lambda^2) - \Pi_\gamma^{[2]}(0,\Lambda^2). \tag{11.33}$$

Equation (11.25) then leads to the result

$$\bar{\Pi}_\gamma^{[2]}(q^2) = -\frac{2\alpha}{\pi}\int_0^1 \mathrm{d}x\, x(1-x)\ln\left[\frac{m^2}{m^2-q^2x(1-x)}\right], \tag{11.34}$$

which was first given by Schwinger (1949a). This 'once-subtracted' $\bar{\Pi}_\gamma^{[2]}$ is *finite* as $\Lambda\to\infty$, and tends to zero as $q^2\to 0$.

The generalization of (11.32) to all orders will be given by

$$\frac{-\mathrm{i}g_{\mu\nu}}{q^2(1-\bar{\Pi}_\gamma(q^2))} \tag{11.35}$$

where $\bar{\Pi}_\gamma(q^2)$ is the all-orders analogue of $\bar{\Pi}_\gamma^{[2]}$ in (11.32), and is similarly related to the 1-γ irreducible photon self-energy $\bar{\Pi}_{\mu\nu}$ via the analogue of (11.24):

$$\mathrm{i}\bar{\Pi}_{\mu\nu}(q^2) = \mathrm{i}(q^2 g_{\mu\nu} - q_\mu q_\nu)\bar{\Pi}_\gamma(q^2). \tag{11.36}$$

Because $\bar{\Pi}_{\mu\nu}$, and hence $\bar{\Pi}_\gamma$, has no 1–γ intermediate states, it is expected to

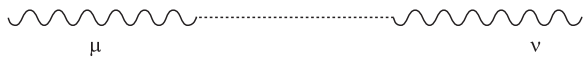


FIGURE 11.4
The contribution of a massless particle to the photon self-energy.

have no contribution of the form A^2/q^2. If such a contribution were present, (11.35) shows that it would result in a photon propagator having the form

$$\frac{-\mathrm{i}g_{\mu\nu}}{q^2 - A^2} \tag{11.37}$$

which is, of course, that of a *massive* particle. Thus, provided no such contribution is present, the photon mass will remain zero through all radiative corrections. It is important to note, though, that gauge invariance is fully satisfied by the general form (11.36) relating $\bar{\Pi}_{\mu\nu}$ to $\bar{\Pi}_\gamma$; it does not prevent the occurrence of such an 'A^2/q^2' piece in $\bar{\Pi}_\gamma$. Remarkably, therefore, it seems possible, after all, to have a massive photon while respecting gauge invariance! This loophole in the argument 'gauge invariance implies $m_\gamma = 0$' was first pointed out by Schwinger (1962).

Such a $1/q^2$ contribution in $\bar{\Pi}_\gamma$ must, of course, correspond to a massless single particle intermediate state, via a diagram of the form shown in figure 11.4. Thus if the theory contains a massless particle, not the photon (since 1–γ states are omitted from $\bar{\Pi}_{\mu\nu}$) but coupling to it, the photon can acquire mass. This is one way of understanding the 'Higgs mechanism' for generating a mass for a gauge-field quantum while still respecting the gauge symmetry (Englert and Brout 1964, Higgs 1964, Guralnik *et al.* 1964). The massless particle involved is called a 'Goldstone boson'. As we shall see in volume 2, just such a photon mass is generated in a superconductor, and a similar mechanism is invoked in the Standard Model to give masses to the $\mathrm{W}^\pm$ and Z^0 gauge bosons, which mediate the weak interactions.

11.5 The physics of $\bar{\Pi}_\gamma^{[2]}(q^2)$

We now consider some immediate physical consequences of the formulae (11.32) and (11.34).

11.5.1 Modified Coulomb's law

In section 1.3.3 we saw how, in the static limit, a propagator of the form $-g_{\mathrm{N}}^2(\boldsymbol{q}^2 + m_{\mathrm{U}}^2)^{-1}$ could be interpreted (via a Fourier transform) in terms of a

Yukawa potential

$$\frac{-g_{\rm N}^2}{4\pi}\frac{{\rm e}^{-r/a}}{r}$$

where $a = m_{\rm U}^{-1}$ (in units $\hbar = c = 1$). As $m_{\rm U} \to 0$ we arrive at the Coulomb potential, associated with the propagator $\sim 1/\boldsymbol{q}^2$ in the static ($q_0 = 0$) limit. It follows that the corrected propagator (11.32) must represent a correction to the $1/r$ Coulomb potential.

To see what it is, we expand the denominator of (11.32) so as to write (11.32) as

$$\frac{-{\rm i}g_{\mu\nu}}{q^2}(1 + \bar{\Pi}_\gamma^{[2]}(q^2)) \tag{11.38}$$

which is in fact the perturbative $O(\alpha)$ correction to the propagator (we shall return to (11.32) in a moment). At low energies, and in the static limit, $q^2 = -\boldsymbol{q}^2$ will be small compared to the fermion (mass)^2 in (11.34), and we may expand the logarithm in powers of $\boldsymbol{q}^2/m^2$, with the result that the static propagator becomes (problem 11.6)

$$\frac{{\rm i}g_{\mu\nu}}{\boldsymbol{q}^2}\left(1 + \frac{\alpha}{15\pi}\boldsymbol{q}^2/m^2\right) \tag{11.39}$$

$$= \frac{{\rm i}g_{\mu\nu}}{\boldsymbol{q}^2} + {\rm i}g_{\mu\nu}\frac{\alpha}{15\pi}\frac{1}{m^2}. \tag{11.40}$$

The Fourier transform of the first term in (11.40) is proportional to the familiar coulombic $1/r$ potential (see appendix G, for example), while the Fourier transform of the *constant* ($\boldsymbol{q}^2$-independent) second term is a δ-function:

$$\int {\rm e}^{{\rm i}\boldsymbol{q}\cdot\boldsymbol{r}}\frac{{\rm d}^3\boldsymbol{q}}{(2\pi)^3} = \delta^3(\boldsymbol{r}). \tag{11.41}$$

When (11.40) is used in any scattering process between two charged particles, each charged particle vertex will carry a charge e (or $-e$) and so the total effective potential will be (in the attractive case)

$$-\left\{\frac{\alpha}{r} + \frac{4\alpha^2}{15m^2}\delta^3(\boldsymbol{r})\right\}. \tag{11.42}$$

The second term in (11.42) may be treated as a perturbation in hydrogenic atoms, taking m to be the electron mass. Application of first-order perturbation theory yields an energy shift

$$\begin{aligned}\Delta E_n^{(1)} &= -\frac{4\alpha^2}{15m^2}\int \psi_n^*(\boldsymbol{r})\delta^3(\boldsymbol{r})\psi_n(\boldsymbol{r})\,{\rm d}^3\boldsymbol{r} \\ &= -\frac{4\alpha^2}{15m^2}|\psi_n(0)|^2. \end{aligned}\tag{11.43}$$

Only s-state wavefunctions are non-vanishing at the origin, where they take the value (in hydrogen)

$$\psi_n(0) = \frac{1}{\sqrt{\pi}} \left(\frac{\alpha m}{n}\right)^{3/2} \tag{11.44}$$

where n is the principal quantum number. Hence for this case

$$\Delta E_n^{(1)} = -\frac{4\alpha^5 m}{15\pi n^3}. \tag{11.45}$$

For example, in the 2s state the energy shift is -1.122×10^{-7} eV. Although we did not discuss the Coulomb spectrum predicted by the Dirac equation in chapter 3, it turns out that the $2^2\mathrm{S}_{\frac{1}{2}}$ and $2^2P_{\frac{1}{2}}$ levels are degenerate if no radiative corrections (such as the previous one) are applied. In fact, the levels are found experimentally to be split apart by the famous 'Lamb shift', which amounts to $\Delta E/2\pi\hbar = 1058$ MHz in frequency units. The shift we have calculated, for the $2s$ level, is -27.13 MHz in these units, so it is a small – but still perfectly measurable – contribution to the entire shift. This particular contribution was first calculated by Uehling (1935).

While small in hydrogen and ordinary atoms, the 'Uehling effect' dominates the radiative corrections in muonic atoms, where the 'm' in (11.44) becomes the muon mass m_μ. This means that the result (11.45) becomes

$$-\frac{4\alpha^5}{15\pi n^3} \left(\frac{m_\mu}{m}\right)^2 m_\mu.$$

Since the unperturbed energy levels are (in this case) proportional to m_μ, this represents a relative enhancement of $\sim(m_\mu/m)^2 \sim (210)^2$. This calculation cannot be trusted in detail, however, as the muonic atom radius is itself $\sim 1/210$ times smaller than the electron radius in hydrogen, so that the approximation $|\boldsymbol{q}| \sim 1/r \ll m$, which led to (11.42), is no longer accurate enough. Nevertheless the order of magnitude is correct.

11.5.2 Radiatively induced charge form factor

This leads us to consider (11.38) more generally, without making the low $\boldsymbol{q}^2$ expansion. In chapter 8 we learned how the static Coulomb potential became modified by a form factor $F(\boldsymbol{q}^2)$ if the scattering centre was not point-like, and we also saw how the idea could be extended to covariant form factors for spin-0 and spin-$\frac{1}{2}$ particles. Referring to the case of $\mathrm{e}^-\mu^-$ scattering for definiteness (section 8.7), we may consider the effect of inserting (11.38) into (8.182). The result is

$$e^2 \bar{u}_{k'} \gamma_\mu u_k \left\{ \frac{g^{\mu\nu}}{q^2} (1 + \bar{\Pi}_\gamma^{[2]}(q^2)) \right\} \bar{u}_{p'} \gamma_\nu u_p. \tag{11.46}$$

Referring now to the discussion of form factors for charged spin-$\frac{1}{2}$ particles in section 8.8, we can share the correction (11.46) equally between the e^- and the μ^- vertices and write

$$e\bar{u}_{k'}\gamma_\mu u_k \to e\bar{u}_{k'}\gamma_\mu u_k(1+\bar{\Pi}_\gamma^{[2]}(q^2))^{1/2} \approx e\bar{u}_{k'}\gamma_\mu u_k(1+\tfrac{1}{2}\bar{\Pi}_\gamma^{[2]}(q^2)) \quad (11.47)$$

for the electron, and similarly for the muon. From (8.208) this means that our 'radiative correction' has generated some effective extension of the charge, as given by a charge form factor $\mathcal{F}_1(q^2) = 1+\frac{1}{2}\bar{\Pi}_\gamma^{[2]}(q^2)$. Note that the condition $\mathcal{F}_1(0) = 1$ is satisfied since $\bar{\Pi}_\gamma^{[2]}(0) = 0$.

In the static case, or for scattering of equal mass particles in the CM system, we have $q^2 = -\boldsymbol{q}^2$ and we may consider the Fourier transform of the function $\mathcal{F}_1(-\boldsymbol{q}^2)$, to obtain the charge distribution. The integral is discussed in Weinberg (1995, section 10.2) and in Peskin and Schroeder (1995, section 7.5). The latter authors show that the approximate radial distribution of charge is $\sim e^{-2mr}/(mr)^{3/2}$, indicating that it has a range $\sim\frac{1}{2m}$. This is precisely the mass of the fermion–antifermion intermediate state in the loop which yields $\bar{\Pi}_\gamma^{[2]}$, so this result represents a plausible qualitative extension of Yukawa's relationship (1.20) to the case of *two*-particle exchange. In any case, the range represented by $\bar{\Pi}_\gamma^{[2]}$ is of order of the fermion Compton wavelength $1/m$, which is an important insight; this is why we need to do better than the point-like approximation (11.42) in the case of muonic atoms.

11.5.3 The running coupling constant

There is yet another way of interpreting (11.38). Referring to (11.46), we may regard

$$e^2(q^2) = e^2[1+\bar{\Pi}_\gamma^{[2]}(q^2)] \quad (11.48)$$

as a 'q^2-dependent effective charge'. In fact, it is usually written as a 'q^2-dependent fine structure constant'

$$\alpha(q^2) = \alpha[1+\bar{\Pi}_\gamma^{[2]}(q^2)]. \quad (11.49)$$

The concept of a q^2-dependent charge may be startling but the related one of a spatially dependent charge is, in fact, familiar from the theory of dielectrics. Consider a test charge q in a polarizable dielectric medium, such as water. If we introduce another test charge $-q$ into the medium, the electric field between the two test charges will line up the water molecules (which have a permanent electric dipole moment) as shown in figure 11.5. There will be an induced dipole moment $\boldsymbol{P}$ per unit volume, and the effect of $\boldsymbol{P}$ on the resultant field is (from elementary electrostatics) the same as that produced by a volume charge equal to $-\operatorname{div}\boldsymbol{P}$. If, as is usual, $\boldsymbol{P}$ is taken to be proportional to $\boldsymbol{E}$, so that $\boldsymbol{P} = \chi\epsilon_0\boldsymbol{E}$, Gauss' law will be modified from

$$\operatorname{div}\boldsymbol{E} = \rho_{\text{free}}/\epsilon_0 \quad (11.50)$$

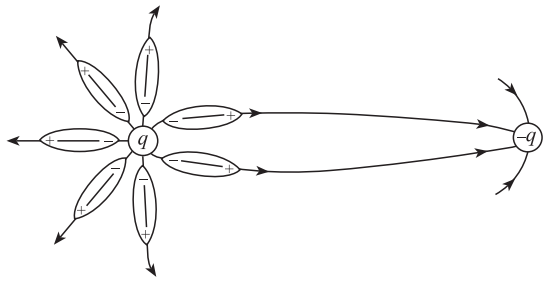


FIGURE 11.5
Screening of charge in a dipolar medium (from Aitchison 1985).

to

$$\operatorname{div} \boldsymbol{E} = (\rho_{\text{free}} - \operatorname{div} \boldsymbol{P})/\epsilon_0 = \rho_{\text{free}}/\epsilon_0 - \operatorname{div}(\chi \boldsymbol{E}) \tag{11.51}$$

where ρ_{free} refers to the test charges introduced into the dielectric. If χ is slowly varying as compared to $\boldsymbol{E}$, it may be taken as approximately constant in (11.51), which may then be written as

$$\operatorname{div} \boldsymbol{E} = \rho_{\text{free}}/\epsilon \tag{11.52}$$

where $\epsilon = (1 + \chi)\epsilon_0$ is the dielectric constant of the medium, ϵ_0 being that of the vacuum. Thus the field is effectively reduced by the factor $(1+\chi)^{-1} = \epsilon_0/\epsilon$.

This is all familiar ground. Note, however, that this treatment is essentially macroscopic, the molecules being replaced by a continuous distribution of charge density $-$ div $\boldsymbol{P}$. When the distance between the two test charges is as small as, roughly, the molecular diameter, this reduction – or *screening* effect – must cease and the field between them has the full unscreened value. In general, the electrostatic potential between two test charges q_1 and q_2 in a dielectric can be represented phenomenologically by

$$V(r) = q_1 q_2/4\pi\epsilon(r)r \tag{11.53}$$

where $\epsilon(r)$ is assumed to vary slowly from the value ϵ for $r \gg d$ to the value ϵ_0 for $r \ll d$, where d is the diameter of the polarized molecules. The situation may be described in terms of an effective charge

$$q' = q/[\epsilon(r)]^{1/2} \tag{11.54}$$

for each of the test charges. Thus we have an effective charge which depends on the interparticle separation, as shown in figure 11.6.

Now consider the application of this idea to QED, replacing the polarizable medium by the *vacuum*. The important idea is that, in the vicinity of a test charge *in vacuo*, charged pairs can be created. Pairs of particles of mass m can exist for a time of the order of $\Delta t \sim \hbar/mc^2$. They can spread apart

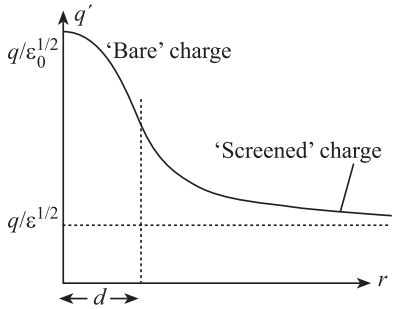


FIGURE 11.6
Effective (screened) charge versus separation between charges (from Aitchison 1985).

a distance of order $c\Delta t$ in this time, i.e. a distance of approximately $\hbar/mc$, which is the Compton wavelength $\lambda\!\!\!^{-}_c$. This distance gives a measure of the 'molecular diameter' we are talking about, since it is the polarized virtual pairs which now provide a *vacuum* screening effect around the original charged particle. The largest 'diameter' will be associated with the smallest mass m, in this case the electron mass. Not coincidentally, this estimate of the range of the 'spreading' of the charge 'cloud' is just what we found in section 11.5.2: namely, the fermion Compton wavelength. The longest-range part of the cloud will be that associated with the lightest charged fermion, the electron.

In this analogy the bare vacuum (no virtual pairs) corresponds to the 'vacuum' used in the previous macroscopic analysis and the physical vacuum (virtual pairs) to the polarizable dielectric. We cannot, of course, get outside the physical vacuum, so that we are really always dealing with effective charges that depend on r. What, then, do we mean by the familiar symbol e? This is simply the effective charge as $r \to \infty$ or $q^2 \to 0$; or, in practice, the charge relevant for distances much larger than the particles' Compton wavelength. This is how our $q^2 \to 0$ definition is to be understood.

Let us consider, then, how $\alpha(q^2)$ varies when q^2 moves to large space-like values, such that $-q^2$ is much greater than m^2 (i.e. to distances well within the 'cloud'). For $|q^2| \gg m^2$ we find (problem 11.7) from (11.34) that

$$\bar{\Pi}_\gamma^{[2]}(q^2) = \frac{\alpha}{3\pi}\left[\ln\left(\frac{|q^2|}{m^2}\right) - \frac{5}{3} + O(m^2/|q^2|)\right] \tag{11.55}$$

so that our q^2-dependent fine structure constant, to leading order in α is

$$\alpha(q^2) \approx \alpha\left[1 + \frac{\alpha}{3\pi}\ln\left(\frac{|q^2|}{Am^2}\right)\right] \tag{11.56}$$

for large values of $|q^2|/m^2$, where $A = \exp 5/3$.

Equation (11.56) shows that the effective strength $\alpha(q^2)$ tends to increase at large $|q^2|$ (short distances). This is, after all, physically reasonable: the reduction in the effective charge caused by the dielectric constant associated with the *polarization of the vacuum* disappears (the charge increases) as we pass inside some typical dipole length. In the present case, that length is m^{-1} (in our standard units $\hbar = c = 1$), the fermion Compton wavelength, a typical distance over which the fluctuating pairs extend.

The foregoing is the reason why this whole phenomenon is called *vacuum polarization*, and why the original diagram which gave $\Pi_\gamma^{[2]}$ is called a vacuum polarization diagram.

Equation (11.56) is the lowest-order correction to α, in a form valid for $|q^2| \gg m^2$. It turns out that, in this limit, the dominant vacuum polarization contributions (for a theory with one charged fermion) can be isolated in each order of perturbation theory and summed explicitly. The result of summing these 'leading logarithms' is

$$\alpha(Q^2) = \frac{\alpha}{[1 - (\alpha/3\pi)\ln(Q^2/Am^2)]} \qquad \text{for } Q^2 \gg m^2 \tag{11.57}$$

where we now introduce $Q^2 = -q^2$, a positive quantity when q is a momentum transfer. The justification for (11.57) – which of course amounts to the very plausible return to (11.32) instead of (11.38) – is subtle, and depends upon ideas grouped under the heading of the 'renormalization group'. This is beyond the scope of the present volume, but will be taken up again in volume 2.

Equation (11.57) presents some interesting features. First, note that for typical large $Q^2 \sim (50\ \text{GeV})^2$, say, the change in the effective α predicted by (11.57) is quite measurable. Let us write

$$\alpha(Q^2) = \frac{\alpha}{1 - \Delta\alpha(Q^2)} \tag{11.58}$$

in general, where $\Delta\alpha(Q^2)$ includes the contributions from all charged fermions with mass m such that $m^2 \ll Q^2$. The contribution from the charged leptons is then straightforward, being given by

$$\Delta\alpha_{\text{leptons}} = \frac{\alpha}{3\pi} \sum_l \ln(Q^2/Am_l^2) \tag{11.59}$$

where m_l is the lepton mass. Including the e, μ and τ one finds (problem 11.8)

$$\Delta\alpha_{\text{leptons}}(Q^2 = (50\ \text{GeV})^2) \approx 0.03. \tag{11.60}$$

However, the corresponding quark loop contributions are subject to strong interaction corrections, and are not straightforward to calculate. We shall not pursue this in detail here, noting just that the total contribution from the five quarks u, d, s, c and b has a value very similar to (11.60) for the leptons (see,

for example, Altarelli *et al.* 1989). Including both the leptonic and hadronic contributions then yields the estimate

$$\alpha(Q^2 = (50\ \text{GeV})^2) \approx \tfrac{1}{137} \times \tfrac{1}{0.94} \approx \tfrac{1}{129}. \tag{11.61}$$

The predicted increase of $\alpha(Q^2)$ at large Q^2 has been tested by measuring the differential cross section for Bhabha scattering,

$$\mathrm{e}^-\mathrm{e}^+ \to \mathrm{e}^-\mathrm{e}^+. \tag{11.62}$$

We are interested in the contribution from one-photon exchange in the t-channel, which will contain the factor $\alpha(Q^2)$. To favour this contribution, the CM energy should be well beyond the Z^0 peak in the s-channel (cf figure 9.16). This was the case at the highest LEP energy, $\sqrt{s} = 198$ GeV, which also allowed large Q^2 values to be probed. The L3 experiment covered the region $1800\ \text{GeV}^2 < Q^2 < 21600\ \text{GeV}^2$ (Achard *et al.* 2005). These results, and earlier data from L3 (Acciari *et al.* 2000) and OPAL (Abbiendi *et al.* 2000), clearly show the expected rise in $\alpha(Q^2)$ as Q^2 increases, and are in good quantitative agreement with the theoretical prediction of QED (Burkhardt and Pietrzyk 2001).

The notion of a q^2-dependent coupling constant is, in fact, quite general – for example, we could just as well interpret (10.71) in terms of a q^2-dependent $g^2_{\mathrm{ph}}(q^2)$. Such 'varying constants' are called *running coupling constants*. Until 1973 it was generally believed that they would all behave in essentially the same way as (11.57) – namely, a logarithmic rise as Q^2 increases. Many people (in particular Landau 1955) noted that if equation (11.57) is taken at face value for arbitrarily large Q^2, then $\alpha(Q^2)$ itself will diverge at $Q^2 = Am^2 \exp(3\pi/\alpha)$. Taking m to be the mass of an electron, this is of course an absurdly high energy. Besides, as such energies are reached, approximations made in arriving at (11.57) will break down; all we can really say is that perturbation theory will fail as we approach such energies.

While this may be an academic point in QED, it turns out that there is one part of the Standard Model where it may be relevant. This is the 'Higgs sector' involving a complex scalar field, as will be discussed in volume 2. In this case, the 'running' of the Higgs coupling constant can be invoked to suggest a useful upper bound on the Higgs mass (Maiani 1991).

The significance of the 1973 date is that it was in that year that one of the most important discoveries in 'post-QED' quantum field theory was made, by Politzer (1973) and by Gross and Wilczek (1973). They performed a similar one-loop calculation in the more complicated case of QCD, which is a 'non-Abelian gauge theory' (as is the theory of the weak interactions in the electroweak theory). They found that the QCD analogue of (11.57) was

$$\alpha_{\mathrm{s}}(Q^2) = \frac{\alpha_{\mathrm{s}}(\mu^2)}{[1 + \frac{\alpha_{\mathrm{s}}}{12\pi}(33 - 2f)\ln(Q^2/\mu^2)]} \tag{11.63}$$

where f is the number of fermion–antifermion loops considered, and μ is a

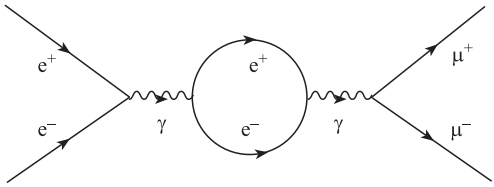


FIGURE 11.7
Vacuum polarization insertion in the virtual one-photon annihilation amplitude in $e^+e^- \to \mu^+\mu^-$.

reference mass scale. The crucial difference from (11.57) is the large positive contribution '+33', which is related to the contributions from the gluonic self-interactions (non-existent among photons). The quantity $\alpha_s(Q^2)$ now tends to *decrease* at large Q^2 (provided $f \leq 16$), tending ultimately to zero. This property is called 'asymptotic freedom' and is highly relevant to understanding the success of the parton model of chapter 9, in which the quarks and gluons are taken to be essentially free at large values of Q^2. This can be qualitatively understood in terms of $\alpha_s(Q^2) \to 0$ for high momentum transfers ('deep scattering'). The non-Abelian parts of the Standard Model will be considered in volume 2, where we shall return again to $\alpha_s(Q^2)$.

11.5.4 $\bar{\Pi}_\gamma^{[2]}$ in the *s*-channel

We have still not exhausted the riches of $\bar{\Pi}_\gamma^{[2]}(q^2)$. Hitherto we have concentrated on regarding our corrected propagator as appearing in a t-channel exchange process, where $q^2 < 0$. But of course it could also perfectly well enter an s-channel process such as $e^+e^- \to \mu^+\mu^-$ (see problem 8.18), as in figure 11.7. In this case, the 4-momentum carried by the photon is $q = p_{e^+} + p_{e^-} = p_{\mu^+} + p_{\mu^-}$, so that q^2 is precisely the usual invariant variable 's' (cf section 6.3.3), which in turn is the square of the CM energy and is therefore positive. In fact, the process of figure 11.7 occurs physically only for $q^2 = s > 4m_\mu^2$, where m_μ is the muon mass.

Consider, therefore, our formula (11.34) for $q^2 > 0$, that is, in the time-like rather than the space-like ($q^2 < 0$) region. The crucial new point is that the argument $[m^2 - q^2x(1-x)]$ of the logarithm can now become negative, so that $\bar{\Pi}_\gamma^{[2]}$ must develop an imaginary part. The smallest q^2 for which this can happen will correspond to the largest possible value of the product $x(1-x)$, for $0 < x < 1$. This value is $\frac{1}{4}$, and so $\bar{\Pi}_\gamma^{[2]}$ becomes imaginary for $q^2 > 4m^2$, which is the threshold for real creation of an e^+e^- pair.

This is the first time that we have encountered an imaginary part in a Feynman amplitude which, for figure 11.7 and omitting all the spinor factors,

is once again

$$\frac{1}{q^2(1-\bar{\Pi}_\gamma^{[2]}(q^2))} \tag{11.64}$$

but now $q^2 > 4m_\mu^2$, which is greater than $4m^2$ so that $\bar{\Pi}_\gamma^{[2]}(q^2)$ in (11.64) has an imaginary part. There is a good physical reason for this, which has to do with *unitarity*. This was introduced in section 6.2.2 in terms of the relation $SS^\dagger = I$ for the S-matrix. The invariant amplitude $\mathcal{M}$ is related to S by $S_{\rm fi} = 1 + {\rm i}(2\pi)^4\delta^4(p_{\rm i} - p_{\rm f})\mathcal{M}_{\rm fi}$ (cf (6.102)). Inserting this into $SS^\dagger = I$ leads to an equation of the form (for help see Peskin and Schroeder (1995, section 7.3))

$$2{\rm Im}\mathcal{M}_{\rm fi} = \sum_{\rm k} \mathcal{M}^*_{\rm kf}\mathcal{M}_{\rm ki}(2\pi)^4\delta\Big(p_{\rm i} - \sum q_{\rm k}\Big) \tag{11.65}$$

where '$\sum_{\rm k}$' stands for the phase space integral involving momenta $q_1, q_2, \ldots$ over the states allowed by energy–momentum conservation. This implies that as the energy crosses each threshold for production of a newly allowed state, there will be a new contribution to the imaginary part of $\mathcal{M}$. This is exactly what we are seeing here, at the e^+e^- threshold.

It is interesting, incidentally, that (11.65) can be used to derive the relativistic generalization of the optical theorem given in appendix H (note that the right-hand side of (11.65) is clearly related to the total cross section for ${\rm i} \to {\rm k}$, if ${\rm i} = {\rm f}$).

As regards the real part of $\bar{\Pi}_\gamma^{[2]}(q^2)$ in the time-like region, it will be given by (11.57) with Q^2 replaced by q^2, or s, for large values of q^2. Again, measurements have verified the predicted variation of $\alpha(q^2)$ in the time-like region (Miyabayashi *et al.* 1995, Ackerstaff *et al.* 1998, Abbiendi *et al.* 1999, 2000).

There is one more 'elementary' loop that we must analyse – the vertex correction shown in figure 11.8, which we now discuss. We will see how the important relation $Z_1 = Z_2$ emerges, and introduce some of the physics contained in the renormalized vertex.

11.6 The $O(e^2)$ vertex correction, and $Z_1 = Z_2$

The amplitude corresponding to figure 11.8 is

$$\begin{aligned} -{\rm i}e\bar{u}(p')\Gamma_\mu^{[2]}(p,p')u(p) &= \bar{u}(p')\int(-{\rm i}e\gamma^\nu)\frac{-{\rm i}g_{\lambda\nu}}{k^2}\frac{\rm i}{\not p' - \not k - m} \\ &\quad\times(-{\rm i}e\gamma_\mu)\frac{\rm i}{\not p - \not k - m}(-{\rm i}e\gamma^\lambda)\frac{{\rm d}^4k}{(2\pi)^4}u(p) \end{aligned} \tag{11.66}$$

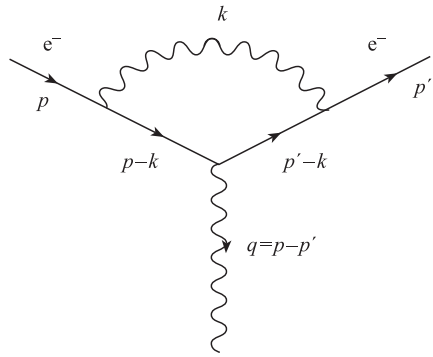


FIGURE 11.8
One-loop vertex correction.

where $\gamma_\mu = g_{\mu\sigma}\gamma^\sigma$, and $\Gamma_\mu^{[2]}$ represents the correction to the standard vertex and again $\xi = 1$. We find

$$\Gamma_\mu^{[2]}(p,p') = -\mathrm{i}e^2 \int \frac{1}{k^2}\gamma^\lambda \frac{1}{\not{p}' - \not{k} - m}\gamma_\mu \frac{1}{\not{p} - \not{k} - m}\gamma_\lambda \frac{\mathrm{d}^4k}{(2\pi)^4}. \tag{11.67}$$

The integral is logarithmically divergent at large k, by power counting, and the divergence will be cancelled by the Z_1 counter term of figure 11.1(c). It turns out to be infrared divergent also, as was $\mathrm{d}\Sigma^{[2]}/\mathrm{d}\not{p}$. As in the latter case, we leave the infrared problem aside, concentrating on the removal of ultraviolet divergences.

Z_1 is determined by the requirement that the total amplitude at $q = p - p' = 0$, for on-shell fermions, is just $-\mathrm{i}e\bar{u}(p)\gamma_\mu u(p)$, this being our definition of 'e'. Hence we have (at $O(e^2)$)

$$-\mathrm{i}e\bar{u}(p)\Gamma_\mu^{[2]}(p,p)u(p) - \mathrm{i}e\bar{u}(p)\gamma_\mu(Z_1^{[2]} - 1)u(p) = 0 \tag{11.68}$$

and so

$$\Gamma_\mu^{[2]}(p,p) + \gamma_\mu(Z_1^{[2]} - 1) = 0. \tag{11.69}$$

The renormalized vertex correction $\bar{\Gamma}_\mu^{[2]}$ may then be defined as

$$\bar{\Gamma}_\mu^{[2]}(p,p') = \Gamma_\mu^{[2]}(p,p') + (Z_1^{[2]} - 1)\gamma_\mu = \Gamma_\mu^{[2]}(p,p') - \Gamma_\mu^{[2]}(p,p) \tag{11.70}$$

and in this 'once-subtracted' form it is finite, and equal to zero at $q = 0$.

We shall consider some physical consequences of $\bar{\Gamma}_\mu^{[2]}$ in a moment, but first we show that (at $O(e^2)$) $Z_1^{[2]} = Z_2^{[2]}$, and explain the significance of this important relation. It is, after all, at first sight a rather surprising equality between two apparently unrelated quantities, one associated with the fermion

self-energy, the other with the vertex part. From (11.9) we have, for the fermion self-energy,

$$\Sigma^{[2]}(p) = -\mathrm{i}e^2 \int \frac{1}{k^2}\gamma^\lambda \frac{1}{\not p - \not k - m}\gamma_\lambda \frac{\mathrm{d}^4 k}{(2\pi)^4}. \tag{11.71}$$

One can discern some kind of similarity between (11.71) and (11.67), which can be elucidated with the help of a little algebra.

Consider differentiating the identity $(\not p - m)(\not p - m)^{-1} = 1$ with respect to p^μ:

$$\begin{aligned} 0 &= \frac{\partial}{\partial p^\mu}[(\not p - m)(\not p - m)^{-1}] \\ &= \left[\frac{\partial}{\partial p^\mu}(\not p - m)\right](\not p - m)^{-1} + (\not p - m)\frac{\partial}{\partial p^\mu}(\not p - m)^{-1} \\ &= \gamma_\mu(\not p - m)^{-1} + (\not p - m)\frac{\partial}{\partial p^\mu}(\not p - m)^{-1}. \end{aligned} \tag{11.72}$$

It follows that

$$\frac{\partial}{\partial p^\mu}(\not p - m)^{-1} = -(\not p - m)^{-1}\gamma_\mu(\not p - m)^{-1} \tag{11.73}$$

from which the *Ward identity* (Ward 1950) follows immediately:

$$-\frac{\partial \Sigma^{[2]}}{\partial p^\mu} = \Gamma_\mu^{[2]}(p, p' = p). \tag{11.74}$$

Derived here to one-loop order, the identity is, in fact, true to all orders, *provided* that a gauge-invariant regularization is adopted. Note that the identity deals with $\Gamma_\mu^{[2]}$ at zero momentum transfer ($q = p - p' = 0$), which is the value at which e is defined. Note also that consistently with (11.74), each of $\partial\Sigma^{[2]}/\partial\not p$ and $\Gamma_\mu^{[2]}$ are both infrared and ultraviolet divergent, though we shall only be concerned with the latter.

The quantities $\Sigma^{[2]}$ and $\Gamma_\mu^{[2]}$ are both $O(e^2)$, and contain ultraviolet divergences which are cancelled by the $O(e^2)$ counter terms. From (11.11) and (11.12) we have

$$\Sigma^{[2]} = \bar\Sigma^{[2]} - Z_2^{[2]}(m_0 - m) + (\not p - m)(Z_2^{[2]} - 1) \tag{11.75}$$

where $\bar\Sigma^{[2]}$ is finite, and from (11.70) we have

$$\Gamma_\mu^{[2]}(p, p') = \bar\Gamma_\mu^{[2]}(p, p') - (Z_1^{[2]} - 1)\gamma_\mu \tag{11.76}$$

where $\bar\Gamma_\mu^{[2]}$ is finite. Inserting (11.75) and (11.76) into (11.74) and equating the infinite parts gives

$$Z_1^{[2]} = Z_2^{[2]}. \tag{11.77}$$

This relation is true to all orders ($Z_1 = Z_2$), provided a gauge-invariant regularization is used. It is a very significant relation, as already indicated after (11.8). It shows, first, that the gauge principle survives renormalization provided the regularization is gauge invariant. More physically, it tells us that the bare and renormalized charges are related simply by (cf (11.6))

$$e = e_0 Z_3^{1/2}. \tag{11.78}$$

In other words, the interaction-dependent rescaling of the bare charge is due solely to vacuum polarization effects in the photon propagator, which are the same for *all* charged particles interacting with the photon. By contrast, both Z_1 and Z_2 do depend on the specific type of the interacting charged particle, since these quantities involve the particle masses. The ratio of bare to renormalized charge is independent of particle type. Hence if a set of bare charges are all equal (or 'universal'), the renormalized ones will be too. But we saw in section 2.6 how just such a notion of universality was present in theories constructed according to the (electromagnetic) gauge principle. We now see how the universality survives renormalization. In volume 2 we shall find that a similar universality holds, empirically, in the case of the weak interaction, giving a strong indication that this force too should be described by a renormalizable gauge theory.

11.7 The anomalous magnetic moment and tests of QED

Returning now to $\Gamma_\mu^{[2]}$, just as in section 11.5.2 we regarded the vacuum polarization correction $1 + \frac{1}{2}\bar{\Pi}_\gamma^{[2]}$ as a contribution to the fermion's charge form factor $\mathcal{F}_1(q^2)$, so we may expect that the vertex correction will also contribute to the form factor. Indeed, let us recall the general form of the electromagnetic vertex for a spin-$\frac{1}{2}$ particle (cf (8.208)):

$$-\mathrm{i}e\bar{u}(p',s')\left[\mathcal{F}_1(q^2)\gamma_\mu + \mathrm{i}\kappa\frac{\mathcal{F}_2(q^2)}{2m}\sigma_{\mu\nu}q^\nu\right]u(p,s) \tag{11.79}$$

where κ is the 'anomalous' part of the magnetic moment, i.e. the magnetic moment is $(e\hbar/2m)(1+\kappa)$, the '1' being the Dirac value calculated in section 3.5. In (11.79), $\mathcal{F}_1$ and $\mathcal{F}_2$ are each normalized to 1 at $q^2 = 0$. Our vertex $\Gamma_\mu^{[2]}$ contributes to both the charge and the magnetic moment form factors; let us call the contributions $\mathcal{F}_1^{[2]}$ and $\kappa\mathcal{F}_2^{[2]}$. Now the Z_1 counter term multiplies γ_μ, and therefore clearly cancels a divergence in $\mathcal{F}_1^{[2]}$. Is there also, we may ask, a divergence in $\kappa\mathcal{F}_2^{[2]}$?

Actually, $\kappa\mathcal{F}_2^{[2]}$ is convergent, and this is highly significant to the physics of renormalization. Had it been divergent, we would either have had to abandon

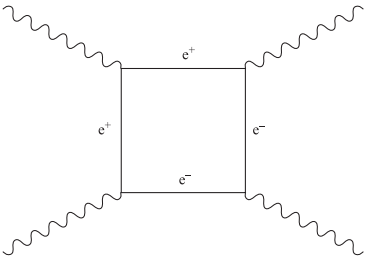


FIGURE 11.9
Contribution (which is finite) to $\gamma\gamma \to \gamma\gamma$.

the theory or introduce a *new* counter term to cancel the divergence. This counter term would have the general form

$$\frac{K}{m}\bar{\hat{\psi}}\sigma_{\mu\nu}\hat{\psi}\hat{F}^{\mu\nu}; \tag{11.80}$$

it is, indeed, an 'anomalous magnetic moment' interaction. But no such term exists in the original QED Lagrangian (11.1)! Its appearance does not seem to follow from the gauge principle argument, even though it is, in fact, gauge invariant. Part of the meaning of the renormalizability of QED (or any theory) is that all infinities can be cancelled by counter terms of the same form as the terms appearing in the original Lagrangian. This means, in other words, that all infinities can be cancelled by assuming an appropriate cut-off dependence for the fields and parameters in the bare Lagrangian. The interaction (11.80) is certainly gauge invariant – but it is *non-renormalizable* – as we shall discuss further later. The message is that, in a renormalizable theory, amplitudes which do not have counterparts in the interactions present in the bare Lagrangian must be *finite*. Figure 11.9 shows another example of an amplitude which turns out to be finite: there is no '$\hat{A}^4$' type of interaction in QED (cf figure 10.13 (a) and the attendant comment in section 10.5).

The calculation of the renormalized $\bar{\mathcal{F}}_1(q^2)$ and of $\kappa\mathcal{F}_2(q^2)$ is quite laborious, not least because three denominators are involved in the $\Gamma_\mu^{[2]}$ integral (11.67). The dedicated reader can follow the story in section 6.3 of Peskin and Schroeder (1995). The most important result is the value obtained for κ, the QED-induced anomalous magnetic moment of the fermion, first calculated by Schwinger (1948a). He obtained

$$\kappa = \frac{\alpha}{2\pi} \approx 0.001\,1614 \tag{11.81}$$

which means a g-factor corrected from the $g = 2$ Dirac value to

$$g = 2 + \frac{\alpha}{\pi} \tag{11.82}$$

or, equivalently,

$$[(g-2)/2]_{\text{Schwinger}} = \frac{\alpha}{2\pi} \approx 0.0011614. \tag{11.83}$$

Note that since κ is a dimensionless quantity, it cannot depend on the mass m of the internal fermion in (11.66). Contributions from two-loop (and higher) diagrams can involve different leptons in internal lines, and hence can depend on lepton mass ratios.

The prediction (11.83) may be compared with the experimental values which are, for the electron (Hanneke *et al.* 2008)

$$a_{\text{e,expt}} \equiv [(g_{\text{e}}-2)/2]_{\text{expt}} = 115\,965\,218\,0.73\,(0.28)\times 10^{-12}\ [0.24\ \text{ppb}] \tag{11.84}$$

and for the muon (Bennett *et al.* 2006)

$$a_{\mu,\text{expt}} \equiv [(g_{\mu}-2)/2]_{\text{expt}} = 116\,592\,080\,(63)\ \times 10^{-11}\ [0.54\ \text{ppm}], \tag{11.85}$$

where the bracketed figures are the quoted uncertainties (statistical and systematic combined in quadrature). Of course, in Schwinger's day the experimental accuracy was far different, but there was a real discrepancy (Kusch and Foley 1947) with the Dirac value ($a = 0$). Schwinger's one-loop calculation provided a fundamental early confirmation of QED, and was the start of a long confrontation between theory and experiment which still continues. The interested reader is referred to the extensive review by Jegerlehner and Nyffeler (2009), upon which we shall draw in the following.

The extraordinarily precise values in (11.84) and (11.85) represent the result of ever more sophisticated and imaginative experimentation. The measurement of $a_{\text{e,expt}}$ is some 2250 times more accurate than that of $a_{\mu,\text{exp}}$. Yet the latter is capable of probing the Standard Model more deeply, for an interesting reason. Consider expanding the vacuum polarization formula (11.18) in powers of m/Λ, having done the momentum integrals as in (10.51) and removed the $\ln\Lambda$ divergence by the subtraction (11.33). The resulting expression will be finite as $\Lambda \to \infty$, but for finite Λ it will contain Λ-dependent terms, the first being of order (m^2/Λ^2). This suggests that the contribution of a 'beyond QED physics' scale to $a_{\mu,\text{theory}}$ (modelled crudely by our cut-off) would be enhanced by a factor $(m_\mu/m_{\text{e}})^2 \approx 43\,000$ relative to its contribution to $a_{\text{e,theory}}$.[1] This outweighs by a factor of 19 the greater experimental accuracy in $a_{\text{e,exp}}$.

This is both good news and bad news. We may distinguish three distinct contributions to 'beyond QED physics' in $a_{\text{e,theory}}$ and $a_{\mu,\text{theory}}$: (i) SM weak interactions; (ii) SM strong (or hadronic) interactions; (iii) beyond the SM physics. Representative diagrams contributing to (i) and (ii) are shown in figure 11.10 (a) and (b) respectively. Sensitivity of $a_{\ell,\text{theory}}$ to effects under (i) is welcome, since they are calculable, and in principle may provide precision

[1] The sensitivity would be even greater for a_τ of course, but the very short lifetime of the τ precludes an accurate measurement of its magnetic moment, at present.

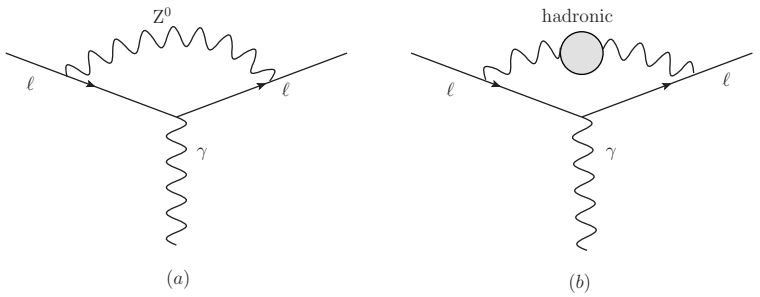


FIGURE 11.10
'Beyond QED' contributions to $a_{\ell,\text{theory}}$ ($\ell = \text{e}, \mu$) due to (a) weak and (b) strong interaction corrections.

tests of the theory. Effects under (ii), however, are difficult to control, and may limit the precision of the theoretical prediction – and hence the capacity to discern the appearance of 'beyond the SM physics'.

In the case of $a_{\text{e,theory}}$, it turns out that the sensitivity to effects under (i) and (ii) is very small. This allows for an essentially pure QED high precision prediction of a_e. The accuracy of the experimental number requires calculation of QED corrections up to 8th order – i.e. terms proportional to $(\alpha/\pi)^4$, which contain 4 loops; there are 891 such diagrams. Their contribution has been calculated by numerical methods by Kinoshita and collaborators (Aoyama *et al.* 2007, 2008; Kinoshita and Nio 2006), who have also estimated the 10th order (5-loop) contributions. To compare with experiment, a value of the fine structure constant α is required. The most accurate value currently quoted is (Bouchendira *et al.* 2011)

$$\alpha^{-1} = 137.035\ 999\ 037\ (91)\ [0.66\ \text{ppb}]. \tag{11.86}$$

With this α the theoretical (QED) prediction of a_e is

$$a_{\text{e,theory}}^{\text{QED}} = 115\ 965\ 218\ 1.13\ (0.11)\ (0.37)\ (0.77) \times 10^{-12} \tag{11.87}$$

where the first, second, and third uncertainties come from the calculated 8th order terms, the 10th order estimate, and the fine structure constant (11.86). The theory is thus in good agreement with experiment, at an extraordinary level of precision:

$$a_{\text{e,expt}} - a_{\text{e,theory}}^{\text{QED}} = -0.40\ (0.88) \times 10^{-12}. \tag{11.88}$$

The QED part of the Standard Model is indeed the paradigm quantum field theory. Further progress will depend on the evaluation of the 10th order (5-loop) terms.

Turning now to $a_{\mu,\text{theory}}$, the 'pure QED' part has been evaluated up to 4 loops and estimated at the 5-loop level, with the result (Jegerlehner and Nyffeler 2009)

$$a_{\mu,\text{theory}}^{\text{QED}} = 116\ 584\ 718.1\ (0.2) \times 10^{-11} \tag{11.89}$$

where the error results from the uncertainties in the lepton mass ratios, the numerical error in the α^4 terms, the estimated uncertainty in the α^5 terms, and the uncertainty in the value of α, which in (11.89) is determined from $a_{\text{e,expt}}$. There are also electroweak and hadronic contributions, $a_{\mu,\text{theory}}^{\text{E-W}}$ and $a_{\mu,\text{theory}}^{\text{had.}}$. The first of these has been evaluated up to 2 loops, and the 3-loop effects are negligible; the result is (Jegerlehner and Nyffeler 2009)

$$a_{\mu,\text{theory}}^{\text{E-W}} = 153.2\ (1.8) \times 10^{-11}. \tag{11.90}$$

$a_{\mu,\text{theory}}^{\text{had.}}$ is considerably larger, and has larger uncertainties. Its value is the subject of intensive ongoing theoretical effort, and is likely to be regularly updated. Here we give the value arrived at by Jegerlehner and Nyffeler (2009), namely

$$a_{\mu,\text{theory}}^{\text{had.}} = 6918.8\ (65) \times 10^{-11}. \tag{11.91}$$

Adding together (11.89), (11.90) and (11.91) gives the Standard Model prediction

$$a_{\mu,\text{theory}}^{\text{SM}} = 116\ 591\ 790.1\ (65) \times 10^{-11}. \tag{11.92}$$

It is worth stressing that *all* of the Standard Model (electromagnetic, weak and strong theories) is needed for the result (11.92); it is also interesting that the theoretical error is essentially the same as the experimental one, at this stage.

Comparison of (11.92) and (11.85) yields

$$a_{\mu,\text{expt}} - a_{\mu,\text{theory}}^{\text{SM}} = 290\ (90) \times 10^{-11}. \tag{11.93}$$

Equation (11.93) represents a discrepancy of some 3 standard deviations. This discrepancy between experiment and the SM prediction has persisted now for a number of years, and is one of the very few significant (at this level) such discrepancies. While it may be premature to conclude that a_μ can definitely not be understood without some 'beyond the SM' physics, many such possibilities are reviewed by Jegerlehner and Nyffeler (2009). No doubt this epic confrontation between theory and experiment will continue to be pursued: it is a classic example of the way in which a very high-precision measurement in a thoroughly 'low-energy' area of physics (a magnetic moment) can have profound impact on the 'high-energy' frontier – a circumstance we may be increasingly dependent upon.

One conclusion we can certainly draw is that renormalizable quantum field theories are the most predictive theories we have. We end this volume with some general reflections on renormalizable, and non-renormalizable, theories.

11.8 Which theories are renormalizable – and does it matter?

In the course of our travels thus far, we have met theories which exhibit three different types of ultraviolet behaviour. In the ABC theory at one-loop order, we found that both the field strength renormalizations and the vertex correction were finite; only the mass shifts diverged as $\Lambda \to \infty$. The theory was called 'super-renormalizable'. In QED, we needed divergent renormalization constants Z_i as well as an infinite mass shift – but (although we did not attempt to explain why) these counter terms were enough to cure divergences systematically to all orders and the theory was renormalizable. Finally, we asserted that the anomalous coupling (11.80) was non-renormalizable. In the final section of this volume we shall try to shed more light on these distinctions and their significance.

Is there some way of telling which of these ultraviolet behaviours a given Lagrangian is going to exhibit, without going through the calculations? The answer is yes (nearly), and the test is surprisingly simple. It has to do with the *dimensionality of a theory's coupling constant*. We have seen (section 6.3.1) that the dimensionality of 'g' in the ABC theory is M^1 (using mass as the remaining dimension when $\hbar = c = 1$), that of e in QED is M^0 (section 7.4) and that of the coefficient of the anomalous coupling $\bar{\hat{\psi}}\sigma_{\mu\nu}\hat{\psi}\hat{F}^{\mu\nu}$ in (11.80) is M^{-1}. These couplings have positive, zero and negative mass dimension, respectively. It is no accident that the three theories, with different dimensions for their couplings, have different ultraviolet behaviour and hence different renormalizability.

That coupling constant dimensionality and ultraviolet behaviour are related can be understood by simple dimensional considerations. Compare, for example, the vertex corrections in the ABC theory (figure 10.6) and in QED (figure 11.8). These amplitudes behave essentially as

$$G^{[2]} \sim g_{\text{ph}}^2 \int \frac{\mathrm{d}^4 k}{k^2 k^2 k^2} \tag{11.94}$$

and

$$\Gamma^{[2]} \sim e^2 \int \frac{\mathrm{d}^4 k}{k^2 \not{k} \not{k}} \tag{11.95}$$

respectively, for large k. Both are dimensionless: but in (11.94) the positive (mass)^2 dimension of g_{ph}^2 is compensated by two additional factors of k^2 in the denominator of the integral, as compared with (11.95), with the result that (11.94) is ultraviolet convergent but (11.95) is not. The analysis can be extended to higher-order diagrams: for the ABC theory, the more powers of g_{ph} which are involved, the more denominator factors are necessary, and hence the better the convergence is. Indeed, in this kind of 'super-renormalizable'

theory, only a finite number of diagrams are ultraviolet divergent, to all orders in perturbation theory.

It is clear that some kind of opposite situation must obtain when the coupling constant dimensionality is negative; for then, as the order of the perturbation theory increases, the negative powers of M in the coupling constant factors must be compensated by positive powers of k in the numerators of loop integrals. Hence the divergence will tend to get worse at each successive order. A famous example of such a theory is Fermi's original theory of β-decay (Fermi 1934a, b), referred to in section 1.3.5, in which the interaction density has the 'four-fermion' form

$$G_{\mathrm{F}}\bar{\hat{\psi}}_{\mathrm{p}}(x)\hat{\psi}_{\mathrm{n}}(x)\bar{\hat{\psi}}_{\mathrm{e}}(x)\hat{\psi}_{\nu_{\mathrm{e}}}(x) \tag{11.96}$$

where G_{F} is the 'Fermi constant'. To find the dimensionality of G_{F}, we first establish that of the fermion field by considering a mass term $m\bar{\hat{\psi}}\hat{\psi}$, for example. The integral of this over $\mathrm{d}^3\boldsymbol{x}$ gives one term in the Hamiltonian, which has dimension M. We deduce that $[\hat{\psi}] = \frac{3}{2}$, since $[\mathrm{d}^3\boldsymbol{x}] = -3$. Hence $[\bar{\hat{\psi}}\hat{\psi}\bar{\hat{\psi}}\hat{\psi}] = 6$, and so $[G_{\mathrm{F}}] = -2$. The coupling constant G_{F} in (11.96) therefore has a negative mass dimension, just like the coefficient K/m in (11.80). Indeed, the four-fermion theory is also non-renormalizable.

Must such a theory be rejected? Let us briefly sketch the consequences of an interaction of the form (11.96), but slightly simpler, namely

$$G_{\mathrm{F}}\bar{\hat{\psi}}_{\mathrm{n}}(x)\hat{\psi}_{\mathrm{n}}(x)\bar{\hat{\psi}}_{\nu_{\mathrm{e}}}(x)\hat{\psi}_{\nu_{\mathrm{e}}}(x) \tag{11.97}$$

where, for the present purposes, the neutron is regarded as point-like. Consider, for example, the scattering process $\nu_{\mathrm{e}} + \mathrm{n} \to \nu_{\mathrm{e}} + \mathrm{n}$. To lowest order in G_{F}, this is given by the tree diagram – or 'contact term' – of figure 11.11, which contributes a constant $-\mathrm{i}G_{\mathrm{F}}$ to the invariant amplitude for the process, disregarding the spinor factors for the moment. A one-loop $O(G_{\mathrm{F}}^2)$ correction is shown in figure 11.12. Inspection of figure 11.12 shows that this is an s-channel process (recall section 6.3.3): let us call the amplitude $-\mathrm{i}G_{\mathrm{F}}G_l^{[2]}(s)$, where one G_{F} factor has been extracted, so that the correction can be compared with the tree amplitude and $G_l^{[2]}(s)$ is dimensionless. Then $G_l^{[2]}(s)$ is given by

$$G_l^{[2]}(s) = -\mathrm{i}G_{\mathrm{F}}\int\frac{\mathrm{d}^4k}{(2\pi)^4}\frac{\mathrm{i}}{\not{k} - m_{\nu_{\mathrm{e}}}}\frac{\mathrm{i}}{(p_{\nu_{\mathrm{e}}} + p_{\mathrm{n}} - \not{k}) - m_{\mathrm{n}}}. \tag{11.98}$$

As expected, the negative mass dimension of G_{F} leaves fewer k-factors in the denominator of the loop integral. Indeed, manipulations exactly like those we used in the case of $\Sigma^{[2]}$ shows that $G_l^{[2]}(s)$ has a quadratic divergence, and that $\mathrm{d}G_l^{[2]}/\mathrm{d}s$ has a logarithmic divergence. The extra denominators associated with second and higher derivatives of $G_l^{[2]}(s)$ are sufficient to make these integrals finite.

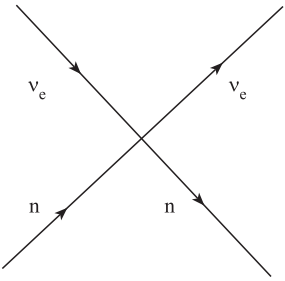


FIGURE 11.11
Lowest order contribution to $\nu_{\rm e} + {\rm n} \to \nu_{\rm e} + {\rm n}$ in the model defined by the interaction (11.97).

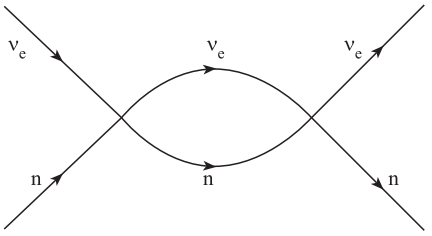


FIGURE 11.12
Second-order (one-loop) contribution to $\nu_{\rm e} + {\rm n} \to \nu_{\rm e} + {\rm n}$.

The standard procedure would now be to cancel these divergences with counter terms. There will certainly be one counter term arising naturally from writing the bare version of (11.97) as (cf (11.5)):

$$G_{0\rm F}\bar{\hat{\psi}}_{0\rm n}\hat{\psi}_{0\rm n}\bar{\hat{\psi}}_{0\nu_{\rm e}}\hat{\psi}_{0\nu_{\rm e}} = G_{\rm F}\bar{\hat{\psi}}_{\rm n}\hat{\psi}_{\rm n}\bar{\hat{\psi}}_{\nu_{\rm e}}\hat{\psi}_{\nu_{\rm e}} + (Z_4 - 1)G_{\rm F}\bar{\hat{\psi}}_{\rm n}\hat{\psi}_{\rm n}\bar{\hat{\psi}}_{\nu_{\rm e}}\hat{\psi}_{\nu_{\rm e}} \qquad (11.99)$$

where $Z_4 G_{\rm F} = G_{0\rm F} Z_{2,{\rm n}} Z_{2,\nu_{\rm e}}$ and the Z_2's are the field strength renormalization constants for the n and $\nu_{\rm e}$ fields. Including the tree graph of figure 11.11, the amplitude of figure 11.12, and the counter term, the total amplitude to $O(G_{\rm F}^2)$ is given by

$$\mathrm{i}\mathcal{M} = -\mathrm{i}G_{\rm F} - \mathrm{i}G_{\rm F}G_l^{[2]}(s) - \mathrm{i}G_{\rm F}(Z_4 - 1). \qquad (11.100)$$

As in our earlier examples, Z_4 will be determined from a *renormalization condition*. In this case, we might demand, for example, that the amplitude $\mathcal{M}$ reduces to $G_{\rm F}$ at the threshold value $s = s_0$, where $s_0 = (m_{\rm n} + m_{\nu_{\rm e}})^2$. Then to $O(G_{\rm F}^2)$ we find

$$Z_4^{[2]} = 1 - G_l^{[2]}(s_0) \qquad (11.101)$$

and our amplitude (11.100) is, in fact,

$$-\mathrm{i}G_\mathrm{F} - \mathrm{i}G_\mathrm{F}[G_l^{[2]}(s) - G_l^{[2]}(s_0)]. \tag{11.102}$$

In (11.102), we see the familiar outcome of such renormalization – the appearance of subtractions of the divergent amplitude (cf (10.74), (11.11), (11.33) and (11.70)). In fact, because $\mathrm{d}G_l^{[2]}/\mathrm{d}s$ is also divergent, we need a second subtraction – and correspondingly, a *new* counter term, not present in the original Lagrangian, of the form

$$G_d \bar{\hat{\psi}}_\mathrm{n} \not{\partial} \hat{\psi}_\mathrm{n} \bar{\hat{\psi}}_{\nu_\mathrm{e}} \not{\partial} \hat{\psi}_{\nu_\mathrm{e}}$$

for example; there will also be others, but we are concerned only with the general idea. The occurrence of such a new counter term is characteristic of a non-renormalizable theory, but at this stage of the proceedings the only penalty we pay is the need to import another constant from experiment, namely the value D of $\mathrm{d}G_l^{[2]}/\mathrm{d}s$ at some fixed s, say $s = s_0$; D will be related to the renormalized value of G_d. We will then write our *renormalized amplitude*, up to $0(G_\mathrm{F}^2)$, as

$$-\mathrm{i}G_\mathrm{F}[1 + D(s - s_0) + \bar{G}_l^{[2]}(s)] \tag{11.103}$$

where $\bar{G}_l^{[2]}(s)$ is finite, and vanishes along with its first derivative at $s = s_0$; that is, $\bar{G}_l^{[2]}(s)$ contributes calculable terms of order $(s - s_0)^2$ if expanded about $s = s_0$.

The moral of the story so far, then, is that we can perform a one-loop renormalization of this theory, at the cost of taking additional parameters from experiments and introducing new terms in the Lagrangian. What about the next order? Figure 11.13 shows a two-loop diagram in our theory, which is of order G_F^3. Writing the amplitude as $-\mathrm{i}G_\mathrm{F}G_l^{[3]}(s)$, the ultraviolet behaviour of $G_l^{[3]}(s)$ is given by

$$(-\mathrm{i}G_\mathrm{F})^2 \int \frac{\mathrm{d}^4k_1 \mathrm{d}^4k_2}{k^4} \tag{11.104}$$

where k is a linear function of k_1 and k_2. This has a leading ultraviolet divergence $\sim \Lambda^4$, even worse than that of $G_l^{[2]}$. As suggested earlier, it is indeed the case that, the higher we go in perturbation theory in this model, the worse the divergences become. We can, of course, eliminate this divergence in $G_l^{[3]}$ by performing a further subtraction, requiring the provision of more parameters from experiment. By now the pattern should be becoming clear: new counter terms will have to be introduced at each order of perturbation theory, and ultimately we shall need an infinite number of them, and hence an infinite number of parameters determined from experiment – and we shall have zero predictive capacity.

Does this imply that the theory is useless? We have learned that $\bar{G}_l^{[2]}(s)$ produces a calculable term of order $G_\mathrm{F}^2(s - s_0)^2$ when expanded about $s = s_0$;

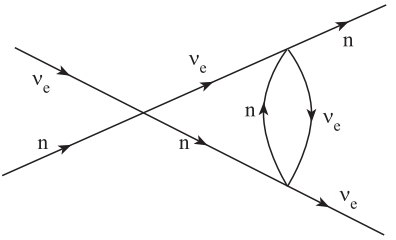


FIGURE 11.13
A two-loop contribution to $\nu_e + n \to \nu_e + n$ in the model defined by (11.97).

and that $\bar{G}_l^{[3]}$ will produce a calculable term of order $G_F^3(s-s_0)^3$, and so on. Now, from the discussion after (11.96), G_F itself is a dimensionless number divided by the square of some mass. As we saw in section 1.3.5 (and will return to in more detail in volume 2), in the case of the physical weak interaction this mass in G_F is the W-mass, and $G_F \sim \alpha/M_W^2$. Hence our loop corrections have the form $\alpha^2(s-s_0)^2/M_W^4, \alpha^3(s-s_0)^3/M_W^6 \ldots$. We now see that for low enough energy close to threshold, where $(s-s_0) \ll M_W^2$, it will be a good approximation to stop at the one-loop level. As we go up in energy, we will need to include higher-order loops, and correspondingly more parameters will have to be drawn from experiment. But only when we begin to approach an energy $\sqrt{s} \sim M_W/\sqrt{\alpha} \sim G_F^{-1/2} \sim 300$ GeV will this theory be terminally sick. This was pointed out by Heisenberg (1939). For this argument to work, it is important that the ultraviolet divergences at a given order in perturbation theory (i.e. a given number of loops) should have been removed by renormalization, otherwise factors of Λ^2 will enter – in place of the $(s-s_0)$ factors, for example.

We have seen that a non-renormalizable theory can be useful at energies well below the 'natural' scale specified by its coupling constant. Let us look at this in a slightly different way, by considering the two four-fermion interaction terms introduced at one loop,

$$G_F \bar{\hat{\psi}}_n \hat{\psi}_n \bar{\hat{\psi}}_{\nu_e} \hat{\psi}_{\nu_e} \qquad \text{and} \qquad G_d \bar{\hat{\psi}}_n \not{\partial} \hat{\psi}_n \bar{\hat{\psi}}_{\nu_e} \not{\partial} \hat{\psi}_{\nu_e}. \tag{11.105}$$

We know that $G_F \sim M_W^{-2}$, and similarly $G_d \sim M_W^{-4}$ (from dimensional counting, or from the association of the G_d term with the $O(G_F^2)$ counter term). From dimensional analysis, or by referring to (11.103) and remembering that D is of order G_F for consistency, we see that the second term in (11.105), when evaluated at tree level, is of order $(s-s_0)/M_W^2$ times the first. It follows that *higher derivative interactions, and in general terms with successively larger negative mass dimension, are increasingly suppressed at low energies.*

Where, then, do *renormalizable* theories fit into this? Those with couplings having positive mass dimension ('super-renormalizable') have, as we have seen, a limited number of infinities and can be quickly renormalized. The 'merely renormalizable' theories have dimensionless coupling constants,

such as e (or α). In this case, since there are no mass factors (for good or ill) to be associated with powers of α, as we go up in order of perturbation theory it would seem plausible that the divergences get essentially no worse, and can be cured by the counter terms which compensated those simplest divergences which we examined in earlier sections – though for QED the proof is difficult, and took many years to perfect.

Given any renormalizable theory, such as QED, it is always possible to suppose that the 'true' theory contains additional non-renormalizable terms, provided their mass scale is very much larger than the energy scale at which the theory has been tested. For example, a term of the form (11.80) with 'K/m' replaced by some very large inverse mass M^{-1} would be possible, and would contribute an amount of order $4e/M$ to a lepton magnetic moment. The present level of agreement between theory and experiment in the case of the electron's moment implies that $M \geq 4 \times 10^9$ GeV.

From this perspective, then, it may be less of a mystery why renormalizable theories are generally the *relevant* ones at presently posed energies. Returning to the line of thought introduced in section 10.1.1, we may imagine that a 'true' theory exists at some enormously high energy Λ (the Planck scale?) which, though not itself a local quantum field theory, can be written in terms of all possible fields and their couplings, as allowed by certain symmetry principles. Our particular renormalizable subset of these theories then emerges as a *low-energy effective theory*, due to the strong suppression of the non-renormalizable terms. Of course, for this point of view to hold, we must assume that the latter interactions do not have 'unnaturally large' couplings, when expressed in terms of Λ.

This interpretation, if correct, deals rather neatly with what was, for many physicists, an awkward aspect of renormalizable theories. On the one hand, it was certainly an achievement to have rendered all perturbative calculations finite as the cut-off went to infinity; but on the other, it was surely unreasonable to expect any such theory, established by confrontation with experiments in currently accessible energy regimes, really to describe physics at arbitrarily high energies. On the 'low-energy effective field theory' interpretation, we can enjoy the calculational advantages of renormalizable field theories, while acknowledging – with no contradiction – the likelihood that at some scale 'new physics' will enter.

Having thus argued that renormalizable theories emerge 'naturally' as low-energy theories, we now seem to be faced with another puzzle: why were weak interactions successfully describable, for many years, in terms of the non-renormalizable four-fermion theory? The answer is that non-renormalizable theories may be physically detectable at low energies if they contribute to processes that would otherwise be forbidden. For example, the fact that (as far as we know) neutrinos have neither electromagnetic nor strong interactions, but only weak interactions, allowed the four-fermion theory to be detected – but amplitudes were suppressed by powers of s/M_{W}^2 (relative to comparable electromagnetic ones) and this was, indeed, why it was called 'weak'!

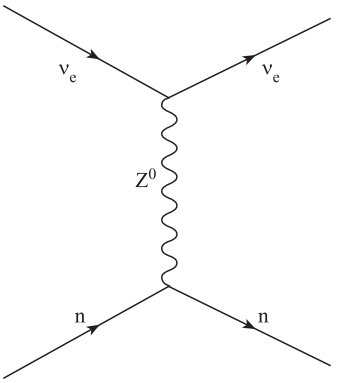


FIGURE 11.14
One-Z (Yukawa-type) exchange process in $\nu_e + n \to \nu_e + n$.

In the case of the weak interaction, the reader may perhaps wonder why – if it was understood that the four-fermion theory could after all be handled up to energies of order 10 GeV – so much effort went in to creating a renormalizable theory of weak interactions, as it undoubtedly did. Part of the answer is that the utility of non-renormalizable interactions was a rather late realization (see, for example, Weinberg 1979). But surely the prospect of having a theory with the predictive power of QED was a determining factor. At all events, the preceding argument for the 'naturalness' of renormalizable theories as low-energy effective theories provides strong expectation that such a description of weak interactions should exist.

We shall discuss the construction of the currently accepted renormalizable theory of electroweak interactions in volume 2. We can already anticipate that the first step will be to replace the 'negative-mass-dimensioned' constant G_F by a dimensionless one. The most obvious way to do this is to envisage a Yukawa-type theory of weak interactions mediated by a massive quantum (as, of course, Yukawa himself did – see section 1.3.5). The four-fermion process of figure 11.11 would then be replaced by that of figure 11.14, with amplitude (omitting spinors) $\sim g_Z^2/(q^2 - m_Z^2)$ where g_Z is dimensionless. For small $q^2 \ll m_Z^2$, this reduces to the contact four-fermion form of figure 11.11, with an effective $G_F \sim g_Z^2/m_Z^2$, showing the origin of the negative mass dimensions of G_F. It is clear that even if the new theory were to be renormalizable, many low-energy processes would be well described by an *effective* non-renormalizable four-fermion theory, as was indeed the case historically.

Unfortunately, we shall see in volume 2 that the application of this simple idea to the charge-changing weak interactions does not, after all, lead to a renormalizable theory. This teaches us an important lesson: a dimensionless coupling does not necessarily guarantee renormalizability.

To arrive at a renormalizable theory of the weak interactions it seems to be necessary to describe them in terms of a gauge theory (recall the 'universality' hints mentioned in section 11.6). Yet the mediating gauge field quanta have mass, which appears to contradict gauge invariance. The remarkable story of how gauge field quanta can acquire mass while preserving gauge invariance is reserved for volume 2.

A number of other non-renormalizable interactions are worth mentioning. Perhaps the most famous of all is gravity, characterized by Newton's constant $G_{\rm N}$, which has the value $(1.2\times10^{19}\ {\rm GeV})^{-2}$. The detection of gravity at energies so far below 10^{19} GeV is due, of course, to the fact that the gravitational fields of all the particles in a macroscopic piece of matter add up coherently. At the level of the individual particles, its effect is still entirely negligible. Another example may be provided by baryon and/or lepton violating interactions, mediated by highly suppressed non-renormalizable terms.[2] Such things are frequently found when the low-energy limit is taken of theories defined (for example) at energies of order 10^{16} GeV or higher.

The stage is now set for the discussion, in volume 2, of the renormalizable non-Abelian gauge field theories which describe the weak and strong sectors of the Standard Model.

Problems

11.1 Establish the values of the counter terms given in (11.12).

11.2 Convince yourself of the rule 'each closed fermion loop carries an additional factor -1'.

11.3 Explain why the trace is taken in (11.14).

11.4 Verify (11.15).

11.5 Verify the quoted relation $P^\rho_\tau P^\tau_\nu = P^\rho_\nu$ where $P^\rho_\nu = g^\rho_\nu - q^\rho q_\nu/q^2$ (cf (11.26)).

11.6 Verify (11.39) for $\boldsymbol{q}^2 \ll m^2$.

11.7 Verify (11.55) for $-q^2 \gg m^2$.

11.8 Check the estimate (11.60).

11.9 Find the dimensionality of 'E' in an interaction of the form $E(\hat{F}_{\mu\nu}\hat{F}^{\mu\nu})^2$. Express this interaction in terms of the $\hat{\boldsymbol{E}}$ and $\hat{\boldsymbol{B}}$ fields. Is such a term finite or infinite in QED? How might it be measured?

[2] The most general renormalizable Lagrangian with the field content, and the gauge symmetries, of the Standard Model automatically conserves baryon and lepton number (Weinberg 1996, pp 316-7).

A

Non-relativistic Quantum Mechanics

This appendix is intended as a very terse 'revision' summary of those aspects of non-relativistic quantum mechanics that are particularly relevant for this book. A fuller account may be found in Mandl (1992), for example.

Natural units $\hbar = c = 1$ (see appendix B).

Fundamental postulate of quantum mechanics:

$$[\hat{p}_i, \hat{x}_j] = -\mathrm{i}\delta_{ij}. \tag{A.1}$$

Coordinate representation:

$$\hat{\boldsymbol{p}} = -\mathrm{i}\boldsymbol{\nabla} \tag{A.2}$$

$$\hat{H}\psi(\boldsymbol{x}, t) = \mathrm{i}\frac{\partial\psi(\boldsymbol{x}, t)}{\partial t}. \tag{A.3}$$

Schrödinger equation for a spinless particle:

$$\hat{H} = \frac{\hat{\boldsymbol{p}}^2}{2m} + \hat{V} \tag{A.4}$$

and so

$$\left(-\frac{1}{2m}\boldsymbol{\nabla}^2 + \hat{V}(\boldsymbol{x}, t)\right)\psi(\boldsymbol{x}, t) = \mathrm{i}\frac{\partial\psi(\boldsymbol{x}, t)}{\partial t}. \tag{A.5}$$

Probability density and current (see problem 3.1 (a)):

$$\rho = \psi^*\psi = |\psi|^2 \geq 0 \tag{A.6}$$

$$\boldsymbol{j} = \frac{1}{2m\mathrm{i}}[\psi^*(\boldsymbol{\nabla}\psi) - (\boldsymbol{\nabla}\psi^*)\psi] \tag{A.7}$$

with

$$\frac{\partial\rho}{\partial t} + \boldsymbol{\nabla}\cdot\boldsymbol{j} = 0. \tag{A.8}$$

Free-particle solutions:

$$\phi(\boldsymbol{x}, t) = u(\boldsymbol{x})\mathrm{e}^{-\mathrm{i}Et} \tag{A.9}$$

$$\hat{H}_0 u = Eu \tag{A.10}$$

where

$$\hat{H}_0 = \hat{H}(\hat{V} = 0). \tag{A.11}$$

Box normalization:

$$\int_V u^*(\boldsymbol{x})u(\boldsymbol{x})\,\mathrm{d}^3\boldsymbol{x} = 1. \tag{A.12}$$

Angular momentum: Three Hermitian operators $(\hat{J}_x, \hat{J}_y, \hat{J}_z)$ satisfying

$$[\hat{J}_x, \hat{J}_y] = \mathrm{i}\hbar\hat{J}_z$$

and corresponding relations obtained by rotating the x–y–z subscripts. The result $[\hat{\boldsymbol{J}}^2, \hat{J}_z] = 0$ implies complete sets of states exist with definite values of $\hat{\boldsymbol{J}}^2$ and $\hat{J}_z$. Eigenvalues of $\hat{\boldsymbol{J}}^2$ are (with $\hbar = 1$) $j(j+1)$ where $j = 0, \frac{1}{2}, 1, \ldots$; eigenvalues of $\hat{J}_z$ are m where $-j \le m \le j$, for given j. For orbital angular momentum, $\hat{\boldsymbol{J}} \to \hat{\boldsymbol{L}} = \boldsymbol{r} \times \hat{\boldsymbol{p}}$ and eigenfunctions are spherical harmonics $Y_{\ell m}(\theta, \phi)$, for which eigenvalues of $\hat{\boldsymbol{L}}^2$ and $\hat{L}_z$ are $l(l+1)$ and m where $-l \le m \le l$. For spin-$\frac{1}{2}$ angular momentum, $\hat{\boldsymbol{J}} \to \frac{1}{2}\boldsymbol{\sigma}$ where the Pauli matrices $\boldsymbol{\sigma} = (\sigma_x, \sigma_y, \sigma_z)$ are

$$\sigma_x = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \qquad \sigma_y = \begin{pmatrix} 0 & -\mathrm{i} \\ \mathrm{i} & 0 \end{pmatrix} \qquad \sigma_z = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}. \tag{A.13}$$

Eigenvectors of s_z are $\begin{pmatrix} 1 \\ 0 \end{pmatrix}$ (eigenvalue $+\frac{1}{2}$), and $\begin{pmatrix} 0 \\ 1 \end{pmatrix}$ (eigenvalue $-\frac{1}{2}$).

Interaction with electromagnetic field: Particle of charge q in electromagnetic vector potential $\boldsymbol{A}$

$$\boxed{\hat{\boldsymbol{p}} \to \hat{\boldsymbol{p}} - q\boldsymbol{A}}. \tag{A.14}$$

Thus

$$\frac{1}{2m}(\hat{\boldsymbol{p}} - q\boldsymbol{A})^2\psi = \mathrm{i}\frac{\partial\psi}{\partial t} \tag{A.15}$$

and so

$$-\frac{1}{2m}\boldsymbol{\nabla}^2\psi + \mathrm{i}\frac{q}{m}\boldsymbol{A}\cdot\boldsymbol{\nabla}\psi + \frac{q^2}{2m}\boldsymbol{A}^2\psi = \mathrm{i}\frac{\partial\psi}{\partial t}. \tag{A.16}$$

Note: (i) chosen gauge $\boldsymbol{\nabla}\cdot\boldsymbol{A} = 0$; (ii) q^2 term is usually neglected.

Example: Magnetic field along z-axis, possible $\boldsymbol{A}$ consistent with $\boldsymbol{\nabla}\cdot\boldsymbol{A} = 0$ is $\boldsymbol{A} = \frac{1}{2}B(-y, x, 0)$ such that $\boldsymbol{\nabla}\times\boldsymbol{A} = (0, 0, B)$. Inserting this into the second term on left-hand side of (A.16) gives

$$\frac{\mathrm{i}qB}{2m}\left(-y\frac{\partial}{\partial x} + x\frac{\partial}{\partial y}\right)\psi = -\frac{qB}{2m}\hat{L}_z\psi \tag{A.17}$$

which generalizes to the standard orbital magnetic moment interaction $-\hat{\boldsymbol{\mu}}\cdot\boldsymbol{B}\psi$ where

$$\hat{\boldsymbol{\mu}} = \frac{qB}{2m}\hat{\boldsymbol{L}}. \tag{A.18}$$

Time-dependent perturbation theory:

$$\hat{H} = \hat{H}_0 + \hat{V} \tag{A.19}$$

$$\hat{H}\psi = \mathrm{i}\frac{\partial \psi}{\partial t}. \tag{A.20}$$

Unperturbed problem:

$$\hat{H}_0 u_n = E_n u_n. \tag{A.21}$$

Completeness:

$$\psi(\boldsymbol{x}, t) = \sum_n a_n(t) u_n(\boldsymbol{x}) \mathrm{e}^{-\mathrm{i}E_n t}. \tag{A.22}$$

First-order perturbation theory:

$$\boxed{a_{\mathrm{fi}} = -\mathrm{i}\int\!\!\int \mathrm{d}^3\boldsymbol{x}\,\mathrm{d}t\, u_{\mathrm{f}}^*(\boldsymbol{x})\mathrm{e}^{+\mathrm{i}E_{\mathrm{f}}t}\hat{V}(\boldsymbol{x}, t)u_{\mathrm{i}}(\boldsymbol{x})\mathrm{e}^{-\mathrm{i}E_{\mathrm{i}}t}} \tag{A.23}$$

which has the form

$$a_{\mathrm{fi}} = -\mathrm{i}\int (\text{volume element})(\text{final state})^*(\text{perturbing potential})(\text{initial state}) \tag{A.24}$$

Important examples:

(i) $\hat{V}$ independent of t:

$$a_{\mathrm{fi}} = -\mathrm{i}V_{\mathrm{fi}} 2\pi\delta(E_{\mathrm{f}} - E_{\mathrm{i}}) \tag{A.25}$$

where

$$V_{\mathrm{fi}} = \int \mathrm{d}^3\boldsymbol{x}\, u_{\mathrm{f}}^*(\boldsymbol{x})\hat{V}(\boldsymbol{x})u_{\mathrm{i}}(\boldsymbol{x}). \tag{A.26}$$

(ii) Oscillating time-dependent potential:

(a) if $\hat{V} \sim \mathrm{e}^{-\mathrm{i}\omega t}$, time integral of a_{fi} is

$$\int \mathrm{d}t\, \mathrm{e}^{+\mathrm{i}E_{\mathrm{f}}t}\mathrm{e}^{-\mathrm{i}\omega t}\mathrm{e}^{-\mathrm{i}E_{\mathrm{i}}t} = 2\pi\delta(E_{\mathrm{f}} - E_{\mathrm{i}} - \omega) \tag{A.27}$$

i.e. the system has absorbed energy from potential;

(b) if $\hat{V} \sim \mathrm{e}^{+\mathrm{i}\omega t}$, time integral of a_{fi} is

$$\int \mathrm{d}t\, \mathrm{e}^{+\mathrm{i}E_{\mathrm{f}}t}\mathrm{e}^{+\mathrm{i}\omega t}\mathrm{e}^{-\mathrm{i}E_{\mathrm{i}}t} = 2\pi\delta(E_{\mathrm{f}} + \omega - E_{\mathrm{i}}) \tag{A.28}$$

i.e. the potential has absorbed energy from system.

Absorption and emission of photons: For electromagnetic radiation, far from its sources, the vector potential satisfies the wave equation

$$\boldsymbol{\nabla}^2 \boldsymbol{A} - \frac{\partial^2 \boldsymbol{A}}{\partial t^2} = 0. \tag{A.29}$$

Solution:

$$\boldsymbol{A}(\boldsymbol{x}, t) = \boldsymbol{A}_0 \exp(-\mathrm{i}\omega t + \mathrm{i}\boldsymbol{k} \cdot \boldsymbol{x}) + \boldsymbol{A}_0^* \exp(+\mathrm{i}\omega t - \mathrm{i}\boldsymbol{k} \cdot \boldsymbol{x}). \tag{A.30}$$

With gauge condition $\boldsymbol{\nabla} \cdot \boldsymbol{A} = 0$ we have

$$\boldsymbol{k} \cdot \boldsymbol{A}_0 = 0 \tag{A.31}$$

and there are two independent polarization vectors for photons.

Treat the interaction in first-order perturbation theory:

$$\hat{V}(\boldsymbol{x}, t) = (\mathrm{i}q/m)\boldsymbol{A}(\boldsymbol{x}, t) \cdot \boldsymbol{\nabla}. \tag{A.32}$$

Thus

$$\begin{aligned} \boldsymbol{A}_0 \exp(-\mathrm{i}\omega t + \mathrm{i}\boldsymbol{k} \cdot \boldsymbol{x}) &\equiv \text{absorption of photon of energy } \omega \\ \boldsymbol{A}_0^* \exp(+\mathrm{i}\omega t + \mathrm{i}\boldsymbol{k} \cdot \boldsymbol{x}) &\equiv \text{emission of photon of energy } \omega. \end{aligned} \tag{A.33}$$

B

Natural Units

In particle physics, a widely adopted convention is to work in a system of units, called natural units, in which

$$\boxed{\hbar = c = 1.} \tag{B.1}$$

This avoids having to keep track of untidy factors of $\hbar$ and c throughout a calculation; only at the end is it necessary to convert back to more usual units. Let us spell out the implications of this choice of c and $\hbar$.

(i) $c = 1$. In conventional MKS units c has the value

$$c = 3 \times 10^8 \text{ m s}^{-1}. \tag{B.2}$$

By choosing units such that

$$c = 1 \tag{B.3}$$

since a velocity has the dimensions

$$[c] = [\text{L}][\text{T}]^{-1} \tag{B.4}$$

we are implying that our unit of length is numerically equal to our unit of time. In this sense, length and time are equivalent dimensions:

$$[\text{L}] = [\text{T}]. \tag{B.5}$$

Similarly, from the energy–momentum relation of special relativity

$$E^2 = \boldsymbol{p}^2 c^2 + m^2 c^4 \tag{B.6}$$

we see that the choice of $c = 1$ also implies that energy, mass and momentum all have equivalent dimensions. In fact, it is customary to refer to momenta in units of 'MeV/c' or 'GeV/c'; these all become 'MeV' or 'GeV' when $c = 1$.

(ii) $\hbar = 1$. The numerical value of Planck's constant is

$$\hbar = 6.6 \times 10^{-22} \text{ MeV s} \tag{B.7}$$

and $\hbar$ has dimensions of energy multiplied by time so that

$$[\hbar] = [\text{M}][\text{L}]^2[\text{T}]^{-1}. \tag{B.8}$$

Setting $\hbar = 1$ therefore relates our units of [M], [L] and [T]. Since [L] and

[T] are equivalent by our choice of $c = 1$, we can choose [M] as the single independent dimension for our natural units:

$$\boxed{[\mathrm{M}] = [\mathrm{L}]^{-1} = [\mathrm{T}]^{-1}.} \tag{B.9}$$

An example: the pion Compton wavelength How do we convert from natural units to more conventional units? Consider the pion Compton wavelength

$$\lambda_\pi = \hbar / M_\pi c \tag{B.10}$$

evaluated in both natural and conventional units. In natural units

$$\lambda_\pi = 1/M_\pi \tag{B.11}$$

where $M_\pi \simeq 140$ MeV$/c^2$. In conventional units, using $M_\pi, \hbar$ (B.7) and c (B.2), we have the familiar result

$$\lambda_\pi = 1.41 \text{ fm} \tag{B.12}$$

where the 'fermi' or femtometre, fm, is defined as

$$1 \text{ fm} = 10^{-15} \text{ m}.$$

We therefore have the correspondence

$$\boxed{\lambda_\pi = 1/M_\pi = 1.41 \text{ fm}.} \tag{B.13}$$

Practical cross section calculations: An easy-to-remember relation may be derived from the result

$$\boxed{\hbar c \simeq 200 \text{ MeV fm}} \tag{B.14}$$

obtained directly from (B.2) and (B.7). Hence, in natural units, we have the relation

$$\boxed{1 \text{ fm} \simeq \frac{1}{200 \text{ MeV}} = 5 \text{ (GeV)}^{-1}.} \tag{B.15}$$

Cross sections are calculated without $\hbar$'s and c's and all masses, energies and momenta typically in MeV or GeV. To convert the result to an area, we merely remember the dimensions of a cross section:

$$[\sigma] = [\mathrm{L}]^2 = [\mathrm{M}]^{-2}. \tag{B.16}$$

If masses, momenta and energies have been specified in GeV, from (B.15) we derive the useful result (from the more precise relation $\hbar c = 197.328$ MeV fm)

$$\boxed{\left(\frac{1}{1 \text{ GeV}}\right)^2 = 1 \text{ (GeV)}^{-2} = 0.389\,39 \text{ mb}} \tag{B.17}$$

where a millibarn, mb, is defined to be

$$1 \text{ mb} = 10^{-31} \text{ m}^2.$$

Note that a 'typical' hadronic cross section corresponds to an area of about λ_π^2 where

$$\lambda_\pi^2 = 1/M_\pi^2 = 20 \text{ mb}.$$

Electromagnetic cross sections are an order of magnitude smaller: specifically for lowest order $e^+e \to \mu^+\mu^-$

$$\sigma \approx \frac{86.8}{s} \text{ nb} \tag{B.18}$$

where s is in $(\text{GeV})^2$ (see problem 8.18(d) in chapter 8).


Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

C

Maxwell's Equations: Choice of Units

In high-energy physics, it is not the convention to use the rationalized MKS system of units when treating Maxwell's equations. Since the discussion is always limited to field equations *in vacuo*, it is usually felt desirable to adopt a system of units in which these equations take their simplest possible form – in particular, one such that the constants ϵ_0 and μ_0, employed in the MKS system, do not appear. These two constants enter, of course, via the force laws of Coulomb and Ampère, respectively. These laws relate a mechanical quantity (force) to electrical ones (charge and current). The introduction of ϵ_0 in Coulomb's law

$$\boldsymbol{F} = \frac{q_1 q_2 \boldsymbol{r}}{4\pi\epsilon_0 r^3} \tag{C.1}$$

enables one to choose arbitrarily one of the electrical units and assign to it a dimension independent of those entering into mechanics (mass, length and time). If, for example, we use the coulomb as the basic electrical quantity (as in the MKS system), ϵ_0 has dimension (coulomb)^2 $[\mathrm{T}]^2/[\mathrm{M}][\mathrm{L}]^3$. Thus the common practical units (volt, ampère, coulomb, etc) can be employed in applications to both fields and circuits. However, for our purposes this advantage is irrelevant, since we are only concerned with the field equations, not with practical circuits. In our case, we prefer to define the electrical units in terms of mechanical ones in such a way as to reduce the field equations to their simplest form. The field equation corresponding to (C.1) is

$$\boldsymbol{\nabla} \cdot \boldsymbol{E} = \rho/\epsilon_0 \qquad \text{(Gauss' law: MKS)} \tag{C.2}$$

and this may obviously be simplified if we choose the unit of charge such that ϵ_0 becomes unity. Such a system, in which CGS units are used for the mechanical quantities, is a variant of the electrostatic part of the 'Gaussian CGS' system. The original Gaussian system set $\epsilon_0 \to 1/4\pi$, thereby simplifying the force law (C.1), but introducing a compensating 4π into the field equation (C.2). The field equation is, in fact, primary, and the 4π is a geometrical factor appropriate only to the specific case of three dimensions, so that it should not appear in a field equation of general validity. The system in which ϵ_0 in (C.2) may be replaced by unity is called the 'rationalized Gaussian CGS' or 'Heaviside–Lorentz' system:

$$\boldsymbol{\nabla} \cdot \boldsymbol{E} = \rho \qquad \text{(Gauss' law; Heaviside–Lorentz).} \tag{C.3}$$

Generally, systems in which the 4π factors appear in the force equations rather than the field equations are called 'rationalized'.

Of course, (C.3) is only the first of the Maxwell equations in Heaviside–Lorentz units. In the Gaussian system, μ_0 in Ampère's force law

$$\boldsymbol{F} = \frac{\mu_0}{4\pi} \iint \frac{\boldsymbol{j}_1 \times (\boldsymbol{j}_2 \times \boldsymbol{r}_{12})}{r_{12}^3} \, \mathrm{d}^3\boldsymbol{r}_1 \, \mathrm{d}^3\boldsymbol{r}_2 \tag{C.4}$$

was set equal to 4π, thereby defining a unit of current (the electromagnetic unit or Biot (Bi emu)). The unit of charge (the electrostatic unit or Franklin (Fr esu)) has already been defined by the (Gaussian) choice $\epsilon_0 = 1/4\pi$ and currents via $\mu_0 \to 4\pi$, and c appears explicitly in the equations. In the rationalized (Heaviside–Lorentz) form of this system, $\epsilon_0 \to 1$ and $\mu_0 \to 1$, and the remaining Maxwell equations are

$$\boldsymbol{\nabla} \times \boldsymbol{E} = -\frac{1}{c}\frac{\partial \boldsymbol{B}}{\partial t} \tag{C.5}$$

$$\boldsymbol{\nabla} \cdot \boldsymbol{B} = 0 \tag{C.6}$$

$$\boldsymbol{\nabla} \times \boldsymbol{B} = \boldsymbol{j} + \frac{1}{c}\frac{\partial \boldsymbol{E}}{\partial t}. \tag{C.7}$$

A further discussion of units in electromagnetic theory is given in Panofsky and Phillips (1962, appendix I).

Finally, throughout this book we have used a particular choice of units for mass, length and time such that $\hbar = c = 1$ (see appendix B). In that case, the Maxwell equations we use are as in (C.3), (C.5)–(C.7), but with c replaced by unity.

As an example of the relation between MKS and the system employed in this book (and universally in high-energy physics), we remark that the fine structure constant is written as

$$\alpha = \frac{e^2}{4\pi\epsilon_0\hbar c} \qquad \text{in MKS units} \tag{C.8}$$

or as

$$\alpha = \frac{e^2}{4\pi} \qquad \text{in Heaviside–Lorentz units with } \hbar = c = 1. \tag{C.9}$$

Clearly the value of $\alpha(\simeq 1/137)$ is the same in both cases, but the numerical values of 'e' in (C.8) and in (C.9) are, of course, different.

The choice of rationalized MKS units for Maxwell's equations is a part of the SI system of units. In this system of units the numerical values of μ_0 and ϵ_0 are

$$\mu_0 = 4\pi \times 10^{-7} \qquad (\mathrm{kg\ m\ C^{-2} = H\ m^{-1}})$$

and, since $\mu_0\epsilon_0 = 1/c^2$,

$$\epsilon_0 = \frac{10^7}{4\pi c^2} = \frac{1}{36\pi \times 10^9} \qquad (\mathrm{C^2\ s^2\ kg^{-1}\ m^{-3} = F\ m^{-1}}).$$

D

Special Relativity: Invariance and Covariance

The co-ordinate 4-vector x^μ is defined by

$$x^\mu = (x^0, x^1, x^2, x^3)$$

where $x^0 = t$ (with $c = 1$) and $(x^1, x^2, x^3) = \boldsymbol{x}$. Under a Lorentz transformation along the x^1-axis with velocity v, x^μ transforms to

$$\begin{aligned} x^{0\prime} &= \gamma(x^0 - vx^1) \\ x^{1\prime} &= \gamma(-vx^0 + x^1) \\ x^{2\prime} &= x^2 \\ x^{3\prime} &= x^3 \end{aligned} \tag{D.1}$$

where $\gamma = (1 - v^2)^{-1/2}$.

A general 'contravariant 4-vector' is defined to be any set of four quantities $A^\mu = (A^0, A^1, A^2, A^3) \equiv (A^0, \boldsymbol{A})$ which transform under Lorentz transformations exactly as the corresponding components of the coordinate 4-vector x^μ. Note that the definition is phrased in terms of the *transformation property* (under Lorentz transformations) of the object being defined. An important example is the energy–momentum 4-vector $p^\mu = (E, \boldsymbol{p})$, where for a particle of rest mass m, $E = (\boldsymbol{p}^2 + m^2)^{1/2}$. Another example is the 4-gradient $\partial^\mu = (\partial^0, -\boldsymbol{\nabla})$ (see problem 2.1) where

$$\partial^0 = \frac{\partial}{\partial t} \qquad \boldsymbol{\nabla} = \left(\frac{\partial}{\partial x^1}, \frac{\partial}{\partial x^2}, \frac{\partial}{\partial x^3}\right). \tag{D.2}$$

Lorentz transformations leave the expression $A^{0\,2} - \boldsymbol{A}^2$ invariant for a general 4-vector A^μ. For example, $E^2 - \boldsymbol{p}^2 = m^2$ is invariant, implying that the rest mass m is invariant under Lorentz transformations. Another example is the four-dimensional invariant differential operator analogous to $\boldsymbol{\nabla}^2$, namely

$$\Box = \partial^{0\,2} - \boldsymbol{\nabla}^2$$

which is precisely the operator appearing in the massless wave equation

$$\Box\phi = \partial^{0\,2}\phi - \boldsymbol{\nabla}^2\phi = 0.$$

The expression $A^{0\,2} - \boldsymbol{A}^2$ may be regarded as the scalar product of A^μ with a related 'covariant vector' $A_\mu = (A^0, -\boldsymbol{A})$. Then

$$A^{0\,2} - \boldsymbol{A}^2 = \sum_\mu A^\mu A_\mu$$

where, in practice, the summation sign on repeated 'upstairs' and 'downstairs' indices is always omitted. We shall often shorten the expression '$A^\mu A_\mu$' even further, to 'A^2'; thus $p^2 = E^2 - \boldsymbol{p}^2 = m^2$. The 'downstairs' version of ∂^μ is $\partial_\mu = (\partial^0, \boldsymbol{\nabla})$. Then $\partial_\mu \partial^\mu = \partial^2 = \Box$. 'Lowering' and 'raising' indices is effected by the *metric tensor* $g^{\mu\nu}$ or $g_{\mu\nu}$, where $g^{00} = g_{00} = 1$, $g^{11} = g^{22} = g^{33} = g_{11} = g_{22} = g_{33} = -1$, all other components vanishing. Thus if $A_\mu = g_{\mu\nu} A^\nu$ then $A_0 = A^0$, $A_1 = -A^1$, etc.

In the same way, the scalar product $A \cdot B$ of two 4-vectors is

$$A \cdot B = A^\mu B_\mu = A^0 B^0 - \boldsymbol{A} \cdot \boldsymbol{B} \tag{D.3}$$

and this is also invariant under Lorentz transformations. For example, the invariant four-dimensional divergence of a 4-vector $j^\mu = (\rho, \boldsymbol{j})$ is

$$\partial^\mu j_\mu = \partial^0 \rho - (-\boldsymbol{\nabla}) \cdot \boldsymbol{j} = \partial^0 \rho + \boldsymbol{\nabla} \cdot \boldsymbol{j} = \partial_\mu j^\mu \tag{D.4}$$

since the spatial part of ∂^μ is $-\boldsymbol{\nabla}$.

Because the Lorentz transformation is linear, it immediately follows that the sum (or difference) of two 4-vectors is also a 4-vector. In a reaction of the type '$1 + 2 \to 3 + 4 + \cdots N$' we express the conservation of both energy and momentum as one '4-momentum conservation equation':

$$p_1^\mu + p_2^\mu = p_3^\mu + p_4^\mu + \cdots p_N^\mu. \tag{D.5}$$

In practice, the 4-vector index on all the p's is conventionally omitted in conservation equations such as (D.5), but it is nevertheless important to remember, in that case, that it is actually *four* equations, one for the energy components and a further three for the momentum components. Further, it follows that quantities such as $(p_1 + p_2)^2, (p_1 - p_3)^2$ are invariant under Lorentz transformations.

We may also consider products of the form $A^\mu B^\nu$, where A and B are 4-vectors. As μ and ν each run over their four possible values $(0, 1, 2, 3)$ 16 different 'components' are generated $(A^0 B^0, A^0 B^1, \ldots, A^3 B^3)$. Under a Lorentz transformation, the components of A and B will transform into definite linear combinations of themselves, as in the particular case of (D.1). It follows that the 16 components of $A^\mu B^\nu$ will also transform into well-defined linear combinations of themselves (try it for $A^0 B^1$ and (D.1)). Thus we have constructed a new object whose 16 components transform by a well-defined linear transformation law under a Lorentz transformation, as did the components of a 4-vector. This new quantity, defined by its transformation law, is called a *tensor* – or more precisely a 'contravariant second-rank tensor', the 'contravariant' referring to the fact that both indices are upstairs, the 'second rank' meaning that it has two indices. An important example of such a tensor is provided by $\partial^\mu A^\nu(x) - \partial^\nu A^\mu(x)$, which is the *electromagnetic field strength tensor* $F^{\mu\nu}$, introduced in chapter 2. More generally we can consider tensors $B^{\mu\nu}$ which are not literally formed by 'multiplying' two vectors together, but which transform in just the same way; and we can introduce third- and

higher-rank tensors similarly, which can also be 'mixed', with some upstairs and some downstairs indices.

We now state a very useful and important fact. Suppose we 'dot' a downstairs 4-vector A_μ into a contravariant second-rank tensor $B^{\mu\nu}$, via the operation $A_\mu B^{\mu\nu}$, where as always a sum on the repeated index μ is understood. Then this quantity transforms as a 4-vector, via its 'loose' index ν. This is obvious if $B^{\mu\nu}$ is actually a product such as $B^{\mu\nu} = C^\mu D^\nu$, since then we have $A_\mu B^{\mu\nu} = (A \cdot C) D^\nu$, and $(A \cdot C)$ is an invariant, which leaves the 4-vector D^μ as the only 'transforming' object left. But even if $B^{\mu\nu}$ is not such a product, it transforms under Lorentz transformations in exactly the same way as if it were, and this leads to the same result. An example is provided by the quantity $\partial_\mu F^{\mu\nu}$ which enters on the left-hand side of the Maxwell equations in the form (2.18).

This example brings us conveniently to the remaining concept we need to introduce here, which is the important one of '*covariance*'. Referring to (2.18), we note that it has the form of an equality between two quantities ($\partial_\mu F^{\mu\nu}$ on the left, $j^\nu_{\rm em}$ on the right) *each of which transforms in the same way under Lorentz transformations* – namely as a contravariant 4-vector. One says that (2.18) is 'Lorentz covariant', the word 'covariant' here meaning precisely that both sides transform in the same way (i.e. consistently) under Lorentz transformations. Confusingly enough, this use of the word 'covariant' is evidently quite different from the one encountered previously in an expression such as 'a covariant 4-vector', where it just meant a 4-vector with a downstairs index. This new meaning of 'covariant' is actually much better captured by an alternative name for the same thing, which is 'form invariant', as we will shortly see.

Why is this idea so important? Consider the (special) relativity principle, which states that the laws of physics should be the same in all inertial frames. The way in which this physical requirement is implemented mathematically is precisely via the notion of *covariance under Lorentz transformations*. For, consider how a law will typically be expressed. Relative to one inertial frame, we set up a coordinate system and describe the phenomena in question in terms of suitable coordinates, and such other quantities (forces, fields, etc) as may be necessary. We write the relevant law mathematically as equations relating these quantities, all referred to our chosen frame and coordinate system. What the relativity principle requires is that these relationships – these equations – must *have the same form* when the quantities in them are referred to a different inertial frame. Note that we must say 'have the same form', rather than 'be identical to', since we know very well that coordinates, at least, are not identical in two different inertial frames (cf (D.1)). This is why the term 'form invariant' is a more helpful one than 'covariant' in this context, but the latter is more commonly used.

A more elementary example may be helpful. Consider Newton's law in the simple form $\boldsymbol{F} = m\ddot{\boldsymbol{r}}$. This equation is 'covariant under rotations', meaning that it preserves the same form under a rotation of the coordinate system –

and this in turn means that the physics it expresses is independent of the orientation of our coordinate axes. The 'same form' in this case is of course just $\boldsymbol{F}' = m\ddot{\boldsymbol{r}}'$. We emphasize again that the components of $\boldsymbol{F}'$ are *not* the same as those of $\boldsymbol{F}$, nor are the components of $\ddot{\boldsymbol{r}}'$ the same as those of $\ddot{\boldsymbol{r}}$; but the *relationship* between $\boldsymbol{F}'$ and $\ddot{\boldsymbol{r}}'$ is exactly the same as the relationship between $\boldsymbol{F}$ and $\ddot{\boldsymbol{r}}$, and that is what is required.

It is important to understand why this deceptively simple result ('$\boldsymbol{F}' = m\ddot{\boldsymbol{r}}'$') has been obtained. The reason is that we have assumed (or asserted) that 'force' is in fact to be represented mathematically as a 3-vector quantity. Once we have said that, the rest follows. More formally, the transformation law of the components of $\boldsymbol{r}$ is $r'_i = R_{ij}r_j$ (sum on j understood), where the matrix of transformation coefficients $\boldsymbol{R}$ is 'orthogonal' ($\boldsymbol{R}\boldsymbol{R}^{\mathrm{T}} = \boldsymbol{R}^{\mathrm{T}}\boldsymbol{R} = \boldsymbol{I}$), which ensures that the length (squared) of $\boldsymbol{r}$ is invariant , $\boldsymbol{r}^2 = \boldsymbol{r}'^2$. To say that 'force is a 3-vector' then implies that the components of $\boldsymbol{F}$ transform by the same set of coefficients R_{ij}: $F'_i = R_{ij}F_j$. Thus starting from the law $F_j = m\ddot{r}_j$ which relates the components in one frame, by multiplying both sides of the equation by R_{ij} and summing over j we arrive at $F'_i = m\ddot{r}'_i$, which states precisely that the components in the primed frame bear the same relationship to each other as the components in the unprimed frame did. This is the property of covariance under rotations, and it ensures that the physics embodied in the law is the same for all systems which differ from one another only by a rotation.

In just the same way, if we can write equations of physics as equalities between quantities which transform in the same way (i.e. 'are covariant') under Lorentz transformations, we will guarantee that these laws obey the relativity principle. This is indeed the case in the *Lorentz covariant formulation* of Maxwell's equations, given in (2.18), which we now repeat here: $\partial_\mu F^{\mu\nu} = j^\nu_{\mathrm{em}}$. To check covariance, we follow essentially the same steps as in the case of Newton's equations, except that the transformations being considered are Lorentz transformations. Inserting the expression (2.19) for $F^{\mu\nu}$, the equation can be written as $(\partial_\mu\partial^\mu)A^\nu - \partial^\nu(\partial_\mu A^\mu) = j^\nu_{\mathrm{em}}$. The two quantities enclosed in parentheses are actually *in*variants, as was mentioned earlier. This means that $\partial_\mu\partial^\mu$ is equal to $\partial_\mu{}'\partial'^\mu$, and similarly $\partial_\mu A^\mu = \partial_\mu{}' A'^\mu$, so that we can write the equation as $(\partial'_\mu\partial'^\mu)A^\nu - \partial^\nu(\partial'_\mu A'^\mu) = j^\nu_{\mathrm{em}}$. It is now clear that if we apply a Lorentz transformation to both sides, A^ν and ∂^ν will become A'^ν and ∂'^ν respectively, while j^ν_{em} will become j'^ν_{em}, since all these quantities are 4-vectors, transforming the same way (as the 3-vectors did in the Newton case). Thus we obtain just the same form of equation, written in terms of the 'primed frame' quantities, and this is the essence of (Lorentz transformation) covariance.

Actually, the detailed 'check' that we have just performed is really unnecessary. All that is required for covariance is that (once again!) both sides of equations transform the same way. That this is true of (2.18) can be seen 'by inspection', once we understand the significance (for instance) of the fact that the μ indices are 'dotted' so as to form an invariant. This example should

convince the reader of the power of the 4-vector notation for this purpose: compare the 'by inspection' covariance of (2.18) with the job of verifying Lorentz covariance starting from the original Maxwell equations (2.1), (2.2), (2.3) and (2.8)! The latter involves establishing the rather complicated transformation law for the fields $\boldsymbol{E}$ and $\boldsymbol{B}$ (which, of course, form parts of the tensor $F^{\mu\nu}$). One can indeed show in this way that the Maxwell equations are covariant under Lorentz transformations, but they are not *manifestly* (i.e. without doing any work) so, whereas in the form (2.18) they are.


Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

E

Dirac δ-Function

Consider approximating an integral by a sum over strips Δx wide as shown in figure E.1:

$$\int_{x_1}^{x_2} f(x)\,\mathrm{d}x \simeq \sum_i f(x_i)\Delta x. \tag{E.1}$$

Consider the function $\delta(x - x_j)$ shown in figure E.2,

$$\delta(x - x_j) = \begin{cases} 1/\Delta x & \text{in the } j\text{th interval} \\ 0 & \text{all others} \end{cases} \tag{E.2}$$

Clearly this function has the properties

$$\sum_i f(x_i)\delta(x_i - x_j)\Delta x = f(x_j) \tag{E.3}$$

and

$$\sum_i \delta(x_i - x_j)\Delta x = 1. \tag{E.4}$$

In the limit as we pass to an integral form, we might expect (applying (E.1) to the left-hand sides) that these equations reduce to

$$\int_{x_1}^{x_2} f(x)\delta(x - x_j)\,\mathrm{d}x = f(x_j) \tag{E.5}$$

and

$$\int_{x_1}^{x_2} \delta(x - x_j)\,\mathrm{d}x = 1 \tag{E.6}$$

provided that $x_1 < x_j < x_2$. Clearly such 'δ-functions' can easily be generalized to more dimensions, e.g. three dimensions:

$$\mathrm{d}V = \mathrm{d}x\,\mathrm{d}y\,\mathrm{d}z \equiv \mathrm{d}^3\boldsymbol{r} \qquad \delta(\boldsymbol{r} - \boldsymbol{r}_j) \equiv \delta(x - x_j)\delta(y - y_j)\delta(z - z_j). \tag{E.7}$$

Informally, therefore, we can think of the δ-function as a function that is zero everywhere except where its argument vanishes – at which point it is infinite in such a way that its integral has unit area, and equations (E.5) and (E.6) hold. Do such amazing functions exist? In fact, the informal idea just given does not define a respectable mathematical function. More properly the use of the 'δ-function' can be justified by introducing the notion of 'distributions'

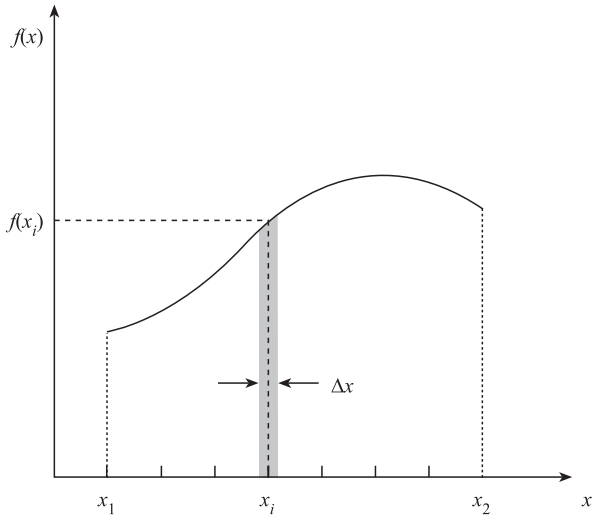


FIGURE E.1
Approximate evaluation of integral.

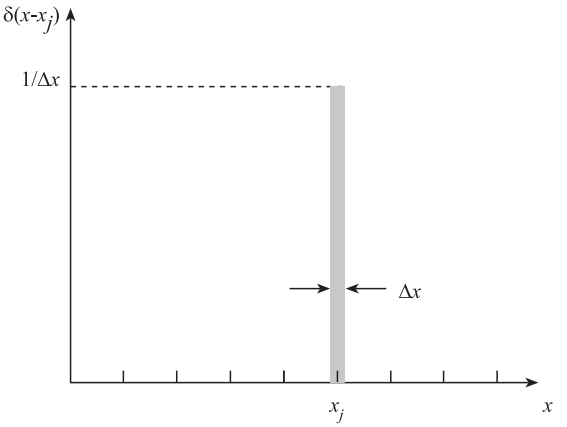


FIGURE E.2
The function $\delta(x - x_j)$.

or 'generalized functions'. Roughly speaking, this means we can think of the 'δ-function' as the limit of a sequence of functions, whose properties converge to those given here. The following useful expressions all approximate the

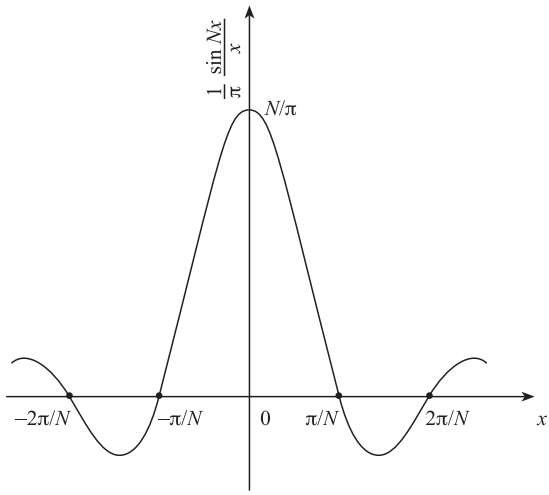


FIGURE E.3
The function (E.10) for finite N.

δ-function in this sense:

$$\delta(x) = \begin{cases} \lim_{\epsilon\to 0} \frac{1}{\epsilon} & \text{for } -\epsilon/2 \le x \le \epsilon/2 \\ 0 & \text{for } |x| > \epsilon/2 \end{cases} \tag{E.8}$$

$$\delta(x) = \lim_{\epsilon\to 0} \frac{1}{\pi} \frac{\epsilon}{x^2+\epsilon^2} \tag{E.9}$$

$$\delta(x) = \lim_{N\to\infty} \frac{1}{\pi} \frac{\sin(Nx)}{x}. \tag{E.10}$$

The first of these is essentially the same as (E.2), and the second is a 'smoother' version of the first. The third is sketched in figure E.3: as N tends to infinity, the peak becomes infinitely high and narrow, but it still preserves unit area.

Usually, under integral signs, δ-functions can be manipulated with no danger of obtaining a mathematically incorrect result. However, care must be taken when products of two such generalized functions are encountered.

Resumé of Fourier series and Fourier transforms

Fourier's theorem asserts that any suitably well-behaved periodic function with period L can be expanded as follows:

$$f(x) = \sum_{n=-\infty}^{\infty} a_n \mathrm{e}^{\mathrm{i}2n\pi x/L}. \tag{E.11}$$

Using the orthonormality relation

$$\frac{1}{L}\int_{-L/2}^{L/2} \mathrm{e}^{-2\pi \mathrm{i}mx/L}\mathrm{e}^{2\pi \mathrm{i}nx/L}\,\mathrm{d}x = \delta_{mn} \tag{E.12}$$

with the Krönecker δ-symbol defined by

$$\delta_{mn} = \begin{cases} 1 & \text{if } m = n \\ 0 & \text{if } m \neq n \end{cases} \tag{E.13}$$

the coefficients in the expansion may be determined:

$$a_m = \frac{1}{L}\int_{-L/2}^{L/2} f(x)\mathrm{e}^{-2\pi \mathrm{i}mx/L}\,\mathrm{d}x. \tag{E.14}$$

Consider the limit of these expressions as $L \to \infty$. We may write

$$f(x) = \sum_{n=-\infty}^{\infty} F_n \Delta n \tag{E.15}$$

with

$$F_n = a_n \mathrm{e}^{2\pi \mathrm{i}nx/L} \tag{E.16}$$

and the interval $\Delta n = 1$. Defining

$$2\pi n/L = k \tag{E.17}$$

and

$$La_n = g(k) \tag{E.18}$$

we can take the limit $L \to \infty$ to obtain

$$\begin{aligned} f(x) &= \int_{-\infty}^{\infty} F_n\,\mathrm{d}n \\ &= \int_{-\infty}^{\infty} \frac{g(k)\mathrm{e}^{\mathrm{i}kx}}{L}\frac{L\mathrm{d}k}{2\pi}. \end{aligned} \tag{E.19}$$

Thus

$$f(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} g(k)\mathrm{e}^{\mathrm{i}kx}\,\mathrm{d}k \tag{E.20}$$

and similarly from (E.14)

$$g(k) = \int_{-\infty}^{\infty} f(x) \mathrm{e}^{-\mathrm{i}kx} \, \mathrm{d}x. \tag{E.21}$$

These are the Fourier transform relations, and they lead us to an important representation of the Dirac δ-function.

Substitute $g(k)$ from (E.21) into (E.20) to obtain

$$f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \mathrm{d}k \, \mathrm{e}^{\mathrm{i}kx} \int_{-\infty}^{\infty} \mathrm{d}x' \, \mathrm{e}^{-\mathrm{i}kx'} f(x'). \tag{E.22}$$

Reordering the integrals, we arrive at the result

$$f(x) = \int_{-\infty}^{\infty} \mathrm{d}x' f(x') \left(\frac{1}{2\pi} \int_{-\infty}^{\infty} \mathrm{e}^{\mathrm{i}k(x-x')} \, \mathrm{d}k \right) \tag{E.23}$$

valid for any function $f(x)$. Thus the expression

$$\frac{1}{2\pi} \int_{-\infty}^{\infty} \mathrm{e}^{\mathrm{i}k(x-x')} \, \mathrm{d}k \tag{E.24}$$

has the remarkable property of vanishing everywhere except at $x = x'$, and its integral with respect to x' over any interval including x is unity (set $f = 1$ in (E.23)). In other words, (E.24) provides us with a new representation of the Dirac δ-function:

$$\boxed{\delta(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \mathrm{e}^{\mathrm{i}kx} \, \mathrm{d}k.} \tag{E.25}$$

Equation (E.25) is very important. It is the representation of the δ-function which is most commonly used, and it occurs throughout this book. Note that if we replace the upper and lower limits of integration in (E.25) by N and $-N$, and consider the limit $N \to \infty$, we obtain exactly (E.10).

The integral in (E.25) represents the superposition, with identical uniform weight $(2\pi)^{-1}$, of plane waves of all wavenumbers. Physically it may be thought of (cf (E.20)) as the Fourier transform of unity. Equation (E.25) asserts that the contributions from all these waves cancel completely, unless the phase parameter x is zero – in which case the integral manifestly diverges and '$\delta(0)$ is infinity' as expected. The fact that the Fourier transform of a constant is a δ-function is an extreme case of the bandwidth theorem from Fourier transform theory, which states that if the (suitably defined) 'spread' in a function $g(k)$ is Δk, and that of its transform $f(x)$ is Δx, then $\Delta x \Delta k \geq \frac{1}{2}$. In the present case Δk is tending to infinity and Δx to zero.

One very common use of (E.25) refers to the normalization of plane-wave states. If we rewrite it in the form

$$\delta(k' - k) = \int_{-\infty}^{\infty} \frac{\mathrm{e}^{-\mathrm{i}k'x}}{(2\pi)^{1/2}} \frac{\mathrm{e}^{\mathrm{i}kx}}{(2\pi)^{1/2}} \, \mathrm{d}x \tag{E.26}$$

we can interpret it to mean that the wavefunctions $e^{ikx}/(2\pi)^{1/2}$ and $e^{ik'x}/(2\pi)^{1/2}$ are orthogonal on the real axis $-\infty \leq x \leq \infty$ for $k \neq k'$ (since the left-hand side is zero), while for $k = k'$ their overlap is infinite, in such a way that the *integral* of this overlap is unity. This is the continuum analogue of orthonormality for wavefunctions labelled by a discrete index, as in (E.12). We say that the plane waves in (E.26) are 'normalized to a δ-function'. There is, however, a problem with this: plane waves are not square integrable and thus do not strictly belong to a Hilbert space. Mathematical physicists concerned with such matters have managed to deal with this by introducing 'rigged' Hilbert spaces in which such a normalization is legitimate. Although we often, in the text, appear to be using 'box normalization' (i.e. restricting space to a finite volume V), in practice when we evaluate integrals over plane waves the limits will be extended to infinity, and results like (E.26) will be used repeatedly.

Important three- and four-dimensional generalizations of (E.25) are:

$$\int e^{i\boldsymbol{k}\cdot\boldsymbol{x}}\, d^3\boldsymbol{k} = (2\pi)^3\delta(\boldsymbol{x}) \tag{E.27}$$

and

$$\int e^{ik\cdot x}\, d^4k = (2\pi)^4\delta(x) \tag{E.28}$$

where $k \cdot x = k^0x^0 - \boldsymbol{k}\cdot\boldsymbol{x}$ (see appendix D) and $\delta(x) = \delta(x^0)\delta(\boldsymbol{x})$.

Properties of the δ-function

The basic properties of the δ-function are exemplified by the equations (see (E.5) and (E.6))

$$\int_{-\infty}^{\infty} \delta(x-a)\, dx = 1, \quad \delta(x-a) = 0 \ \text{ for } x \neq a, \tag{E.29}$$

where a is any real number; and

$$\int_{-\infty}^{\infty} f(x)\ \delta(x-a)\, dx = f(a), \tag{E.30}$$

where $f(x)$ is any continuous function of x. Other useful properties follow:

(i)
$$\delta(ax) = \frac{1}{|a|}\delta(x). \tag{E.31}$$

Proof

For $a > 0$,

$$\int_{-\infty}^{\infty} \delta(ax)\, dx = \int_{-\infty}^{\infty} \delta(y)\frac{dy}{a} = \frac{1}{a}; \tag{E.32}$$

for $a < 0$,

$$\int_{-\infty}^{\infty} \delta(ax)\, dx = \int_{\infty}^{-\infty} \delta(y)\frac{dy}{a} = \int_{-\infty}^{\infty} \delta(y)\frac{dy}{|a|} = \frac{1}{|a|}. \tag{E.33}$$

(ii) $$\delta(x) = \delta(-x) \qquad \text{i.e. an even function.} \tag{E.34}$$

Proof

$$f(0) = \int \delta(x) f(x)\,\mathrm{d}x. \tag{E.35}$$

If $f(x)$ is an odd function, $f(0) = 0$. Thus $\delta(x)$ must be an even function.

(iii) $$\delta(f(x)) = \sum_i \frac{1}{|\mathrm{d}f/\mathrm{d}x|_{x=a_i}} \delta(x - a_i) \tag{E.36}$$

where a_i are the roots of $f(x) = 0$.

Proof

The δ-function is only non-zero when its argument vanishes. Thus we are concerned with the roots of $f(x) = 0$. In the vicinity of a root

$$f(a_i) = 0 \tag{E.37}$$

we can make a Taylor expansion

$$f(x) = \cancel{f(a_i)} + (x - a_i) \left(\frac{\mathrm{d}f}{\mathrm{d}x}\right)_{x=a_i} + \cdots. \tag{E.38}$$

Thus the δ-function has non-zero contributions from each of the roots a_i of the form

$$\delta(f(x)) = \sum_i \delta \left[(x - a_i) \left(\frac{\mathrm{d}f}{\mathrm{d}x}\right)_{x=a_i} \right]. \tag{E.39}$$

Hence (using property (i)) we have

$$\delta(f(x)) = \sum_i \frac{1}{|\mathrm{d}f/\mathrm{d}x|_{x=a_i}} \delta(x - a_i). \tag{E.40}$$

Consider the example

$$\delta(x^2 - a^2). \tag{E.41}$$

Thus

$$f(x) = x^2 - a^2 = (x - a)(x + a) \tag{E.42}$$

with two roots $x = \pm a$ $(a > 0)$, and $\mathrm{d}f/\mathrm{d}x = 2x$. Hence

$$\delta(x^2 - a^2) = \frac{1}{2a}[\delta(x - a) + \delta(x + a)]. \tag{E.43}$$

(iv) $$x\delta(x) = 0. \tag{E.44}$$

This is to be understood as always occurring under an integral. It is obvious from the definition or from property (ii).

(v)
$$\int_{-\infty}^{\infty} f(x)\delta'(x)\,\mathrm{d}x = -f'(0) \tag{E.45}$$

where
$$\delta'(x) = \frac{\mathrm{d}}{\mathrm{d}x}\delta(x). \tag{E.46}$$

Proof

$$\begin{aligned}\int_{-\infty}^{\infty} f(x)\delta'(x)\,\mathrm{d}x &= -\int_{-\infty}^{\infty} f'(x)\delta(x)\,\mathrm{d}x + [f(x)\delta(x)]_{-\infty}^{\infty} \\ &= -f'(0)\end{aligned} \tag{E.47}$$

since the second term vanishes.

(vi)
$$\int_{-\infty}^{x} \delta(x'-a)\,\mathrm{d}x' = \theta(x-a) \tag{E.48}$$

where
$$\theta(x) = \begin{cases} 0 & \text{for } x < 0 \\ 1 & \text{for } x > 0 \end{cases} \tag{E.49}$$

is the so-called 'θ-function'.

Proof

For $x > a$,
$$\int_{-\infty}^{x} \delta(x'-a)\,\mathrm{d}x' = 1; \tag{E.50}$$

for $x < a$,
$$\int_{-\infty}^{x} \delta(x'-a)\,\mathrm{d}x' = 0. \tag{E.51}$$

By a simple extension it is easy to prove the result
$$\int_{x_1}^{x_2} \delta(x-a)\,\mathrm{d}x = \theta(x_2-a) - \theta(x_1-a). \tag{E.52}$$

(vii)
$$\delta(x-y)\ \delta(x-z) = \delta(x-y)\ \delta(y-z). \tag{E.53}$$

Proof

Take any continuous function of z, $f(z)$. Then
$$\int_{-\infty}^{\infty} f(z)\ \mathrm{d}z\{\delta(x-y)\ \delta(x-z)\} = f(x)\ \delta(x-y) \tag{E.54}$$
$$= f(y)\ \delta(x-y) = \int_{-\infty}^{\infty} f(z)\mathrm{d}z\{\delta(x-y)\ \delta(y-z)\}. \tag{E.55}$$

Thus the two sides of (vii) are equivalent as factors in an integrand with z as the integration variable.

Exercise

Use property (iii) plus the definition of the θ-function to perform the p^0 integration and prove the useful phase space formula

$$\int \mathrm{d}^4p\,\delta(p^2 - m^2)\theta(p^0) = \int \mathrm{d}^3\boldsymbol{p}/2E \tag{E.56}$$

where

$$p^2 = (p^0)^2 - \boldsymbol{p}^2 \tag{E.57}$$

and

$$E = +(\boldsymbol{p}^2 + m^2)^{1\,2}. \tag{E.58}$$

The relation (E.51) shows that the expression $\mathrm{d}^3\boldsymbol{p}/2E$ is Lorentz invariant: on the left-hand side, d^4p and $\delta(p^2 - m^2)$ are invariant, while $\theta(p^0)$ depends only on the sign of p^0, which cannot be changed by a 'proper' Lorentz transformation – that is, one that does not reverse the sense of time.


Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

F

Contour Integration

We begin by recalling some relevant results from the calculus of real functions of two real variables x and y, which we shall phrase in 'physical' terms. Consider a particle moving in the xy-plane subject to a force $\boldsymbol{F} = (P(x,y), Q(x,y))$ whose x- and y-components P and Q vary throughout the plane. Suppose the particle moves, under the action of the force, around a closed path $\mathcal{C}$ in the xy-plane. Then the total work done by the force on the particle, $W_{\mathcal{C}}$, will be given by the integral

$$W_{\mathcal{C}} = \oint_{\mathcal{C}} \boldsymbol{F} \cdot \mathrm{d}\boldsymbol{r} = \oint_{\mathcal{C}} P\,\mathrm{d}x + Q\,\mathrm{d}y \tag{F.1}$$

where the $\oint$ sign means that the integration path is closed. Using Stokes' theorem, we can rewrite (F.1) as a surface integral

$$W_{\mathcal{C}} = \iint_{\mathcal{S}} \operatorname{curl} \boldsymbol{F} \cdot \mathrm{d}\boldsymbol{S} \tag{F.2}$$

where $\mathcal{S}$ is any surface bounded by $\mathcal{C}$ (as a butterfly net is bounded by the rim). Taking $\mathcal{S}$ to be the area in the xy-plane enclosed by $\mathcal{C}$, we have $\mathrm{d}\boldsymbol{S} = \mathrm{d}x\,\mathrm{d}y\,\boldsymbol{k}$ and

$$W_{\mathcal{C}} = \iint_{\mathcal{S}} \left(\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y} \right) \mathrm{d}x\,\mathrm{d}y. \tag{F.3}$$

A mathematically special, but physically common, case is that in which $\boldsymbol{F}$ is a 'conservative force', derivable from a potential function $V(x,y)$ (in this two-dimensional example) such that

$$P(x,y) = -\frac{\partial V}{\partial x} \qquad \text{and} \qquad Q(x,y) = -\frac{\partial V}{\partial y} \tag{F.4}$$

the minus signs being the usual convention. In that case, it is clear that

$$\frac{\partial P}{\partial y} = \frac{\partial Q}{\partial x} \tag{F.5}$$

and hence $W_{\mathcal{C}}$ in (F.3) is *zero*. The condition (F.5) is, in fact, both necessary and sufficient for $W_{\mathcal{C}} = 0$.

There can, however, be surprises. Consider, for example, the potential

$$V(x,y) = -\tan^{-1} y/x. \tag{F.6}$$

In this case the components of the associated force are

$$P = -\frac{\partial V}{\partial x} = \frac{-y}{x^2+y^2} \qquad \text{and} \qquad Q = -\frac{\partial V}{\partial y} = \frac{x}{x^2+y^2}. \tag{F.7}$$

Let us calculate the work done by this force in the case that $\mathcal{C}$ is the circle of unit radius centred on the origin, traversed in the anticlockwise sense. We may parametrize a point on this circle by $(x = \cos\theta, y = \sin\theta)$, so that (F.1) becomes

$$W_{\mathcal{C}} = \oint_{\mathcal{C}} -\sin\theta(-\sin\theta\,\mathrm{d}\theta) + \cos\theta(\cos\theta\,\mathrm{d}\theta) = \oint_{\mathcal{C}} \mathrm{d}\theta = 2\pi \tag{F.8}$$

a result which is plainly different from zero. The reason is that although this force is (minus) the gradient of a potential, the latter is not single-valued, in the sense that it does not return to its original value after a circuit round the origin. Indeed, the V of (F.6) is just $-\theta$, which changes by -2π on such a circuit, exactly as calculated in (F.8) allowing for the minus signs in (F.4). Alternatively, we may suspect that the trouble has to do with the 'blow up' of the integrand of (F.7) at the point $x = y = 0$, which is also true.

Much of the foregoing has direct parallels within the theory of functions of a complex variable $z = x + \mathrm{i}y$, to which we now give a brief and informal introduction, limiting ourselves to the minimum required in the text[1]. The crucial property, to which all the results we need are related, is *analyticity*. A function $f(z)$ is *analytic* in a region $\mathcal{R}$ of the complex plane if it has a unique derivative at every point of $\mathcal{R}$. The derivative at a point z is defined by the natural generalization of the real variable definition:

$$\frac{\mathrm{d}f}{\mathrm{d}z} = \lim_{\Delta z \to 0} \left\{ \frac{f(z+\Delta z) - f(z)}{\Delta z} \right\}. \tag{F.9}$$

The crucial new feature in the complex case, however, is that 'Δz' is actually an (infinitesimal) *vector*, in the xy (Argand) plane. Thus we may immediately ask: along which of the infinitely many possible *directions* of Δz are we supposed to approach the point z in (F.9)? The answer is: along any! This is the force of the word 'unique' in the definition of analyticity, and it is a very powerful requirement.

Let $f(z)$ be an analytic function of z in some region $\mathcal{R}$, and let u and v be the real and imaginary parts of f: $f = u + \mathrm{i}v$, where u and v are each functions of x and y. Let us evaluate $\mathrm{d}f/\mathrm{d}z$ at the point $z = x + \mathrm{i}y$ in two different ways, which must be equivalent.

(i) By considering $\Delta z = \Delta x$ (i.e. $\Delta y = 0$). In this case

$$\begin{aligned} \frac{\mathrm{d}f}{\mathrm{d}z} &= \lim_{\Delta x \to 0} \left\{ \frac{u(x+\Delta x, y) - u(x,y) + \mathrm{i}v(x+\Delta x, y) - \mathrm{i}v(x,y)}{\Delta x} \right\} \\ &= \frac{\partial u}{\partial x} + \mathrm{i}\frac{\partial v}{\partial x} \end{aligned} \tag{F.10}$$

from the definition of a partial derivative.

[1] For a fuller introduction, see for example Boas (1983, chapter 14).

(ii) By considering $\Delta z = \mathrm{i}\Delta y$ (i.e. $\Delta x = 0$). In this case

$$\begin{aligned}\frac{\mathrm{d}f}{\mathrm{d}z} &= \lim_{\Delta y \to 0}\left\{\frac{u(x,y+\Delta y)-u(x,y)+\mathrm{i}v(x,y+\Delta y)-\mathrm{i}v(x,y)}{\mathrm{i}\Delta y}\right\} \\ &= \frac{\partial v}{\partial y} - \mathrm{i}\frac{\partial u}{\partial y}. \end{aligned} \tag{F.11}$$

Equating (F.10) and (F.11) we obtain the *Cauchy–Reimann (CR) relations*

$$\frac{\partial u}{\partial x} = \frac{\partial v}{\partial y} \qquad \frac{\partial u}{\partial y} = -\frac{\partial v}{\partial x} \tag{F.12}$$

which are the necessary and sufficient conditions for f to be analytic.

Consider now an integral of the form

$$I = \oint_{\mathcal{C}} f(z)\,\mathrm{d}z \tag{F.13}$$

where again the symbol $\oint$ means that the integration path (or *contour*) in the complex plane in closed. Inserting $f = u + \mathrm{i}v$ and $z = x + \mathrm{i}y$, we may write (F.13) as

$$I = \oint (u\,\mathrm{d}x - v\,\mathrm{d}y) + \mathrm{i}\oint (v\,\mathrm{d}x + u\,\mathrm{d}y). \tag{F.14}$$

Thus the single complex integral (F.13) is equivalent to the two real-plane integrals (F.14); one is the real part of I, the other is the imaginary part, and each is of the form (F.1). In the first, we have $P = u, Q = -v$. Hence the condition (F.5) for the integral to vanish is $\partial u/\partial y = -\partial v/\partial x$, which is precisely the second CR relation! Similarly, in the second integral in (F.14) we have $P = v$ and $Q = u$ so that condition (F.5) becomes $\partial v/\partial y = \partial u/\partial x$, which is the first CR relation. It follows that if $f(z)$ is analytic inside and on $\mathcal{C}$, then

$$\oint_{\mathcal{C}} f(z)\,\mathrm{d}z = 0, \tag{F.15}$$

a result known as *Cauchy's theorem*, the foundation of complex integral calculus.

Now let us consider a simple case in which (as in (F.7)) the result of integrating a complex function around a closed curve is *not* zero – namely the integral

$$\oint_{\mathcal{C}} \frac{\mathrm{d}z}{z} \tag{F.16}$$

where $\mathcal{C}$ is the circle of radius ρ enclosing the origin. On this circle, $z = \rho \mathrm{e}^{\mathrm{i}\theta}$ where ρ is fixed and $0 \le \theta \le 2\pi$, so

$$\oint_{\mathcal{C}} \frac{\mathrm{d}z}{z} = \oint_{\mathcal{C}} \frac{\rho \mathrm{i}\mathrm{e}^{\mathrm{i}\theta}\mathrm{d}\theta}{\rho \mathrm{e}^{\mathrm{i}\theta}} = \mathrm{i}\oint \mathrm{d}\theta = 2\pi\mathrm{i}. \tag{F.17}$$

Cauchy's theorem does not apply in this case because the function being integrated (z^{-1}) is not analytic at $z = 0$. Writing $\mathrm{d}z/z$ in terms of x and y we have

$$\begin{aligned}\frac{\mathrm{d}z}{z} &= \frac{\mathrm{d}x + \mathrm{i}\,\mathrm{d}y}{x + \mathrm{i}y} = \frac{(x - \mathrm{i}y)}{x^2 + y^2}(\mathrm{d}x + \mathrm{i}\,\mathrm{d}y) \\ &= \left(\frac{x\,\mathrm{d}x + y\,\mathrm{d}y}{x^2 + y^2}\right) + \mathrm{i}\left(\frac{-y\,\mathrm{d}x + x\,\mathrm{d}y}{x^2 + y^2}\right). \end{aligned} \tag{F.18}$$

The reader will recognize the imaginary part of (F.18) as involving precisely the functions (F.7) studied earlier, and may like to find the real potential function appropriate to the real part of (F.18).

We note that the result (F.17) is independent of the circle's radius ρ. This means that we can shrink or expand the circle how we like, without affecting the answer. The reader may like to show that the circle can, in fact, be distorted into a simple closed loop of any shape, enclosing $z = 0$, and the answer will still be $2\pi\mathrm{i}$. In general, a contour may be freely distorted in any region in which the integrand is analytic.

We are now in a position to prove the main integration formula we need, which is *Cauchy's integral formula*: let $f(z)$ be analytic inside and on a simple closed curve $\mathcal{C}$ which encloses the point $z = a$; then

$$\boxed{\oint_{\mathcal{C}} \frac{f(z)}{z - a}\,\mathrm{d}z = 2\pi\mathrm{i}f(a)} \tag{F.19}$$

where it is understood that $\mathcal{C}$ is traversed in an anticlockwise sense around $z = a$. The proof follows. The integrand in (F.19) is analytic inside and on $\mathcal{C}$, except at $z = a$; we may therefore distort the contour $\mathcal{C}$ by shrinking it into a very small circle of fixed radius ρ around the point $z = a$. On this circle, z is given by $z = a + \rho\mathrm{e}^{\mathrm{i}\theta}$, and

$$\oint_{\mathcal{C}} \frac{f(z)}{z - a}\,\mathrm{d}z = \int_0^{2\pi} \frac{f(a + \rho\mathrm{e}^{\mathrm{i}\theta})\rho\mathrm{i}\mathrm{e}^{\mathrm{i}\theta}}{\rho\mathrm{e}^{\mathrm{i}\theta}}\,\mathrm{d}\theta = \int_0^{2\pi} f(a + \rho\mathrm{e}^{\mathrm{i}\theta})\mathrm{i}\,\mathrm{d}\theta. \tag{F.20}$$

Now, since f is analytic at $z = a$, it has a unique derivative there, and is consequently continuous at $z = a$. We may then take the limit $\rho \to 0$ in (F.20), obtaining $\lim_{\rho\to 0} f(a + \rho\mathrm{e}^{\mathrm{i}\theta}) = f(a)$, and hence

$$\oint_{\mathcal{C}} \frac{f(z)}{z - a}\,\mathrm{d}z = f(a)\int_0^{2\pi} \mathrm{i}\,\mathrm{d}\theta = 2\pi\mathrm{i}f(a) \tag{F.21}$$

as stated.

We now use these results to establish the representation of the θ-function (see (E.47)) quoted in section 6.3.2. Consider the function $F(t)$ of the real variable t defined by

$$F(t) = \frac{\mathrm{i}}{2\pi}\oint_{\mathcal{C}=\mathcal{C}_1+\mathcal{C}_2} \frac{\mathrm{e}^{-\mathrm{i}zt}}{z + \mathrm{i}\epsilon}\,\mathrm{d}z \tag{F.22}$$

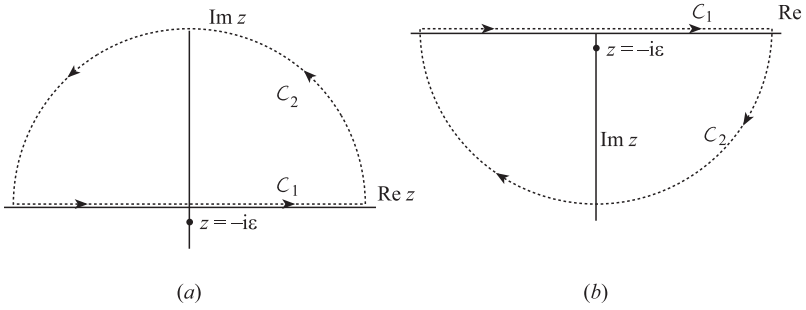


FIGURE F.1
Contours for $F(t)$: (a) $t < 0$; (b) $t > 0$.

where ϵ is an infinitesimally small *positive* number (i.e. it will tend to zero through positive values). The closed contour $\mathcal{C}$ is made up of $\mathcal{C}_1$ which is the real axis from $-R$ to R (we shall let $R \to \infty$ at the end), and of $\mathcal{C}_2$ which is a large semicircle of radius R with diameter the real axis, in either the upper or lower half-plane, the choice being determined by the sign of t, as we shall now explain (see figure F.1). Suppose first that $t < 0$, and let z on $\mathcal{C}_2$ be parametrized as $z = Re^{\mathrm{i}\theta} = R\cos\theta + \mathrm{i}R\sin\theta$. Then

$$\mathrm{e}^{-\mathrm{i}zt} = \mathrm{e}^{\mathrm{i}z|t|} = \mathrm{e}^{-R\sin\theta|t|}e^{\mathrm{i}R\cos\theta|t|} \tag{F.23}$$

from which it follows that the contribution to (F.22) from $\mathcal{C}_2$ will vanish exponentially as $R \to \infty$ *provided that* $\theta > 0$, i.e. we choose $\mathcal{C}_2$ to be in the upper half-plane (figure F.1(a)). In that case the integrand of (F.22) is analytic inside and on $\mathcal{C}$ (the only non-analytic point is outside $\mathcal{C}$ at $z = -\mathrm{i}\epsilon$) and so

$$F(t) = 0 \qquad \text{for } t < 0. \tag{F.24}$$

However, suppose $t > 0$. Then

$$\mathrm{e}^{-\mathrm{i}zt} = \mathrm{e}^{R\sin\theta t}\mathrm{e}^{-\mathrm{i}R\cos\theta t} \tag{F.25}$$

and in this case we must choose the 'contour-closing' $\mathcal{C}_2$ to be in the lower half-plane ($\theta < 0$) or else (F.25) will diverge exponentially as $R \to \infty$. With this choice the $\mathcal{C}_2$ contribution will again go to zero as $R \to \infty$. However, this time the whole closed contour $\mathcal{C}$ does enclose the point $z = -\mathrm{i}\epsilon$ (see figure F.1(b)), and we may apply Cauchy's integral formula to get, for $t > 0$,

$$F(t) = -2\pi\mathrm{i}\frac{\mathrm{i}}{2\pi}\mathrm{e}^{-\epsilon t}, \tag{F.26}$$

the minus sign at the front arising from the fact (see figure F.1(b)) that $\mathcal{C}$ is

now being traversed in a *clockwise* sense around $z = -\mathrm{i}\epsilon$ (this just inverts the limits in (F.21)). Thus as $\epsilon \to 0$,

$$F(t) \to 1 \qquad \text{for } t > 0. \tag{F.27}$$

Summarizing these manoeuvres, for $t < 0$ we chose $\mathcal{C}_2$ in (F.22) in the upper half-plane (figure F.1(a)), and its contribution vanished as $R \to \infty$. In this case we have, as $R \to \infty$,

$$F(t) \to \frac{\mathrm{i}}{2\pi} \int_{-\infty}^{\infty} \frac{\mathrm{e}^{-\mathrm{i}zt}}{z + \mathrm{i}\epsilon}\, \mathrm{d}z = 0 \qquad \text{for } t < 0. \tag{F.28}$$

For $t > 0$ we chose $\mathcal{C}_2$ in the lower half-plane (figure F.1(b)), when again its contribution vanished as $R \to \infty$. However, in this case F does not vanish, but instead we have, as $R \to \infty$,

$$F(t) \to \frac{\mathrm{i}}{2\pi} \int_{-\infty}^{\infty} \frac{\mathrm{e}^{-\mathrm{i}zt}}{z + \mathrm{i}\epsilon}\, \mathrm{d}z = 1 \qquad \text{for } t > 0. \tag{F.29}$$

Equations (F.28) and (F.29) show that we may indeed write

$$\theta(t) = \lim_{\epsilon \to 0} \frac{\mathrm{i}}{2\pi} \int_{-\infty}^{\infty} \frac{\mathrm{e}^{-\mathrm{i}zt}}{z + \mathrm{i}\epsilon}\, \mathrm{d}z \tag{F.30}$$

as claimed in section 6.3, equation (6.93).

G

Green Functions

Let us start with a simple but important example. We seek the solution $G_0(\boldsymbol{r})$ of the equation

$$\boldsymbol{\nabla}^2 G_0(\boldsymbol{r}) = \delta(\boldsymbol{r}). \tag{G.1}$$

There is a 'physical' way to look at this equation which will give us the answer straightaway. Recall that Gauss' law in electrostatics (appendix C) is

$$\boldsymbol{\nabla} \cdot \boldsymbol{E} = \rho/\epsilon_0 \tag{G.2}$$

and that $\boldsymbol{E}$ is expressed in terms of the electrostatic potential V as $\boldsymbol{E} = -\boldsymbol{\nabla}V$. Then (G.2) becomes

$$\boldsymbol{\nabla}^2 V = -\rho/\epsilon_0 \tag{G.3}$$

which is known as Poisson's equation. Comparing (G.3) and (G.1), we see that $(-G_0(\boldsymbol{r})/\epsilon_0)$ can be regarded as the 'potential' due to a source ρ which is concentrated entirely at the origin, and whose total 'charge' is unity, since (see appendix E)

$$\int \delta(\boldsymbol{r})\, \mathrm{d}^3\boldsymbol{r} = 1. \tag{G.4}$$

In other words, $(-G_0/\epsilon_0)$ is effectively the potential due to a unit point charge at the origin. But we know exactly what this potential is from Coulomb's law, namely

$$\frac{-G_0(\boldsymbol{r})}{\epsilon_0} = \frac{1}{4\pi\epsilon_0 r} \tag{G.5}$$

whence

$$G_0(\boldsymbol{r}) = -\frac{1}{4\pi r}. \tag{G.6}$$

We may also check this result mathematically as follows. Using (G.6), equation (G.1) is equivalent to

$$\boldsymbol{\nabla}^2 \frac{1}{r} = -4\pi\delta(\boldsymbol{r}). \tag{G.7}$$

Let us consider the integral of both sides of this equation over a spherical volume of arbitrary radius R surrounding the origin. The integral of the left-hand side becomes, using Gauss' divergence theorem,

$$\int_V \left(\boldsymbol{\nabla}^2 \frac{1}{r}\right) \mathrm{d}^3\boldsymbol{r} = \int_V \boldsymbol{\nabla} \cdot \left(\boldsymbol{\nabla}\frac{1}{r}\right) \mathrm{d}^3\boldsymbol{r} = \int_{S \text{ bounding } V} \boldsymbol{\nabla}\left(\frac{1}{r}\right) \cdot \hat{\boldsymbol{n}}\, \mathrm{d}S. \tag{G.8}$$

Now

$$\boldsymbol{\nabla}\left(\frac{1}{r}\right) = -\frac{1}{r^2}\hat{\boldsymbol{r}} = -\frac{1}{R^2}\hat{\boldsymbol{r}}$$

on the surface S, while $\hat{\boldsymbol{n}} = \hat{\boldsymbol{r}}$ and $\mathrm{d}S = R^2\,\mathrm{d}\Omega$ with $\mathrm{d}\Omega$ the element of solid angle on the sphere. So

$$\int_V \boldsymbol{\nabla}^2\left(\frac{1}{r}\right)\mathrm{d}^3\boldsymbol{r} = -\int_S \mathrm{d}\Omega = -4\pi \tag{G.9}$$

which using (G.4) is precisely the integral of the right-hand side of (G.7), as required.

Consider now the solutions of

$$(\boldsymbol{\nabla}^2 + \boldsymbol{k}^2)G_k(\boldsymbol{r}) = \delta(\boldsymbol{r}). \tag{G.10}$$

We are interested in rotationally invariant solutions, for which G_k is a function of $r = |\boldsymbol{r}|$ alone. For $r \neq 0$, equation (G.10) is easy to solve. Setting $G_k(r) = f(r)/r$, and using

$$\boldsymbol{\nabla}^2 = \frac{1}{r^2}\frac{\partial}{\partial r}r^2\frac{\partial}{\partial r} + \text{parts depending on } \frac{\partial}{\partial\theta} \text{ and } \frac{\partial}{\partial\phi}$$

we find that $f(r)$ satisfies

$$\frac{\mathrm{d}^2 f}{\mathrm{d}r^2} + \boldsymbol{k}^2 f = 0$$

the general solution to which is ($k = |\boldsymbol{k}|$)

$$f(r) = A\mathrm{e}^{\mathrm{i}kr} + B\mathrm{e}^{-\mathrm{i}kr},$$

leading to

$$G_k(r) = A\frac{\mathrm{e}^{\mathrm{i}kr}}{r} + B\frac{\mathrm{e}^{-\mathrm{i}kr}}{r} \tag{G.11}$$

for $r \neq 0$. In the application to scattering problems (appendix H) we shall want G_k to contain purely outgoing waves, so we will pick the 'A'-type solution in (G.11).

Consider therefore the expression

$$(\boldsymbol{\nabla}^2 + \boldsymbol{k}^2)\left(\frac{A\mathrm{e}^{\mathrm{i}kr}}{r}\right) \tag{G.12}$$

where r is now allowed to take the value zero. Making use of the vector operator result

$$\boldsymbol{\nabla}^2(fg) = (\boldsymbol{\nabla}^2 f)g + 2\boldsymbol{\nabla}f\cdot\boldsymbol{\nabla}g + f(\boldsymbol{\nabla}^2 g)$$

with 'f' $= \mathrm{e}^{\mathrm{i}kr}$ and 'g' $= 1/r$, together with

$$\boldsymbol{\nabla}^2\mathrm{e}^{\mathrm{i}kr} = \frac{2\mathrm{i}k\mathrm{e}^{\mathrm{i}kr}}{r} - \boldsymbol{k}^2\mathrm{e}^{\mathrm{i}kr} \qquad \boldsymbol{\nabla}\mathrm{e}^{\mathrm{i}kr} = \frac{\mathrm{i}k\boldsymbol{r}\mathrm{e}^{\mathrm{i}kr}}{r} \qquad \boldsymbol{\nabla}\frac{1}{r} = -\frac{\boldsymbol{r}}{r^3}$$

we find

$$\begin{aligned}(\boldsymbol{\nabla}^2 + \boldsymbol{k}^2)\left(\frac{A\mathrm{e}^{\mathrm{i}kr}}{r}\right) &= A\mathrm{e}^{\mathrm{i}kr}\boldsymbol{\nabla}^2\left(\frac{1}{r}\right) \\ &= -4\pi A\mathrm{e}^{\mathrm{i}kr}\delta(\boldsymbol{r}) \\ &= -4\pi A\delta(\boldsymbol{r}) \end{aligned} \tag{G.13}$$

where we have replaced r by zero in the exponent of the last term of the last line in (G.13), since the δ-function ensures that only this point need be considered for this term. By choosing the constant $A = -1/4\pi$, we find that the (outgoing wave) solution of (G.10) is

$$G_k^{(+)}(r) = -\frac{\mathrm{e}^{\mathrm{i}kr}}{4\pi r}. \tag{G.14}$$

We are also interested in spherically symmetric solutions of (restoring c and $\hbar$ explicitly for the moment)

$$\left(\boldsymbol{\nabla}^2 - \frac{m^2c^2}{\hbar^2}\right)\phi(\boldsymbol{r}) = \delta(\boldsymbol{r}) \tag{G.15}$$

which is the equation analogous to (G.1) for a static classical scalar potential of a field whose quanta have mass m. The solutions to (G.15) are easily found from the previous work by letting $k \to \mathrm{i}mc/\hbar$. Retaining now the solution which goes to zero as $r \to \infty$, we find

$$\phi(\boldsymbol{r}) = -\frac{1}{4\pi}\frac{\mathrm{e}^{-r/a}}{r} \tag{G.16}$$

where $a = \hbar/mc$, the Compton wavelength of the quantum, with mass m. The potential (G.16) is (up to numerical constants) the famous Yukawa potential, in which the quantity 'a' is called the *range*: as r gets greater than a, $\phi(\boldsymbol{r})$ becomes exponentially small. Thus, just as the Coulomb potential is the solution of Poisson's equation (G.3) corresponding to a point source at the origin, so the Yukawa potential is the solution of the analogous equation (G.15), also with a point source at the origin. Note that as $a \to \infty$, $\phi(\boldsymbol{r}) \to G_0(\boldsymbol{r})$.

Functions such as G_k, G_0 and ϕ, which generically satisfy equations of the form

$$\Omega_r G(\boldsymbol{r}) = \delta(\boldsymbol{r}) \tag{G.17}$$

where Ω_r is some linear differential operator, are said to be *Green functions* of the operator Ω_r. From the examples already treated, it is clear that $G(\boldsymbol{r})$ in (G.17) has the general interpretation of a 'potential' due to a point source at the origin, when Ω_r is the appropriate operator for the field theory in question.

Green functions play an important role in the solution of differential equations of the type

$$\Omega_r \psi(\boldsymbol{r}) = s(\boldsymbol{r}) \tag{G.18}$$

where $s(\boldsymbol{r})$ is a known 'source function' (e.g. the charge density in (G.3)). The solution of (G.18) may be written as

$$\psi(\boldsymbol{r}) = u(\boldsymbol{r}) + \int G(\boldsymbol{r}-\boldsymbol{r}')s(\boldsymbol{r}')\,\mathrm{d}^3\boldsymbol{r}' \tag{G.19}$$

where $u(\boldsymbol{r})$ is a solution of $\Omega_r u(\boldsymbol{r}) = 0$. Thus once we know G, we have the solution via (G.19).

Equation (G.19) has a simple physical interpretation. We know that $G(\boldsymbol{r})$ is the solution of (G.18) with $s(\boldsymbol{r})$ replaced by $\delta(\boldsymbol{r})$. But by writing

$$s(\boldsymbol{r}) = \int \delta(\boldsymbol{r}-\boldsymbol{r}')s(\boldsymbol{r}')\,\mathrm{d}^3\boldsymbol{r}' \tag{G.20}$$

we can formally regard $s(\boldsymbol{r})$ as being made up of a *superposition* of point sources, distributed at points $\boldsymbol{r}'$ with a weighting function $s(\boldsymbol{r}')$. Then, since the operator Ω_r is (by assumption) linear, the solution for such a superposition of point sources must be just the same superposition of the point source solutions, namely the integral on the right-hand side of (G.19). This integral term is, in fact, the 'particular integral' of the differential equation (G.18), while the $u(\boldsymbol{r})$ is the 'complementary function'.

Equation (G.19) can also be checked analytically. First note that it is generally the case that the operator Ω_r is translationally invariant, so that

$$\Omega_r = \Omega_{r-r'}; \tag{G.21}$$

the right-hand side of (G.21) amounts to shifting the origin to the point $\boldsymbol{r}'$. Applying Ω_r to both sides of (G.19), we find

$$\begin{aligned}
\Omega_r\psi(\boldsymbol{r}) &= \Omega_r u(\boldsymbol{r}) + \int \Omega_r G(\boldsymbol{r}-\boldsymbol{r}')s(\boldsymbol{r}')\,\mathrm{d}^3\boldsymbol{r}' \\
&= 0 + \int \Omega_{r-r'} G(\boldsymbol{r}-\boldsymbol{r}')s(\boldsymbol{r}')\,\mathrm{d}^3\boldsymbol{r}' = \int \delta(\boldsymbol{r}-\boldsymbol{r}')s(\boldsymbol{r}')\,\mathrm{d}^3\boldsymbol{r}' \\
&= s(\boldsymbol{r})
\end{aligned}$$

as required in (G.18).

Finally, consider the Fourier transform of equation (G.10), defined as

$$\int \mathrm{e}^{-\mathrm{i}\boldsymbol{q}\cdot\boldsymbol{r}}(\boldsymbol{\nabla}^2 + \boldsymbol{k}^2)G_k(\boldsymbol{r})\,\mathrm{d}^3\boldsymbol{r} = \int \mathrm{e}^{-\mathrm{i}\boldsymbol{q}\cdot\boldsymbol{r}}\delta(\boldsymbol{r})\,\mathrm{d}^3\boldsymbol{r}.$$

The right-hand side is unity, by equation (G.4). On the left-hand side we may use the result

$$\int u(r)\boldsymbol{\nabla}^2 v(\boldsymbol{r})\,\mathrm{d}^3\boldsymbol{r} = \int (\boldsymbol{\nabla}^2 u(\boldsymbol{r}))v(\boldsymbol{r})\,\mathrm{d}^3\boldsymbol{r}$$

(proved by integrating by parts, assuming u and v go to zero sufficiently fast

at the boundaries of the integral) to obtain

$$\begin{aligned}\int \mathrm{e}^{-\mathrm{i}\boldsymbol{q}\cdot\boldsymbol{r}}(\boldsymbol{\nabla}^2+\boldsymbol{k}^2)G_k(\boldsymbol{r})\,\mathrm{d}^3\boldsymbol{r} &= \int\{(\boldsymbol{\nabla}^2\mathrm{e}^{-\mathrm{i}\boldsymbol{q}\cdot\boldsymbol{r}})+\boldsymbol{k}^2\mathrm{e}^{-\mathrm{i}\boldsymbol{q}\cdot\boldsymbol{r}}\}G_k(\boldsymbol{r})\,\mathrm{d}^3\boldsymbol{r}\\ &= \int(-\boldsymbol{q}^2+\boldsymbol{k}^2)\mathrm{e}^{-\mathrm{i}\boldsymbol{q}\cdot\boldsymbol{r}}G_k(\boldsymbol{r})\,\mathrm{d}^3\boldsymbol{r}\\ &= (-\boldsymbol{q}^2+\boldsymbol{k}^2)\tilde{G}_k(\boldsymbol{q})\end{aligned}$$

where $\tilde{G}_k(\boldsymbol{q})$ is the Fourier transform of $G_k(\boldsymbol{r})$. Since this expression has to equal unity, we have

$$\tilde{G}_k(\boldsymbol{q})=\frac{1}{\boldsymbol{k}^2-\boldsymbol{q}^2}. \tag{G.22}$$

There is, however, a problem with (G.22) as it stands, which is that it is undefined when the variable $\boldsymbol{q}^2$ takes the value equal to the parameter $\boldsymbol{k}^2$ in the original equation. Indeed, various definitions are possible, corresponding to the type of solution in $\boldsymbol{r}$-space for $G_k(\boldsymbol{r})$ (i.e. ingoing, outgoing or standing wave). It turns out (see the exercise at the end of this appendix) that the specification which is equivalent to the solution $G_k^{(+)}(\boldsymbol{r})$ in (G.14) is to add an infinitesimally small imaginary part in the denominator of (G.22):

$$\tilde{G}_k^{(+)}(\boldsymbol{q})=\frac{1}{\boldsymbol{k}^2-\boldsymbol{q}^2+\mathrm{i}\epsilon}. \tag{G.23}$$

In exactly the same way, the Fourier transform of $\phi(\boldsymbol{r})$ satisfying (G.15) is

$$\tilde{\phi}(\boldsymbol{q})=\frac{-1}{\boldsymbol{q}^2+m^2}, \tag{G.24}$$

where we have reverted to units such that $\hbar=c=1$.

The relativistic generalization of this result is straightforward. Consider the equation

$$(\Box+m^2)G(x)=-\delta(x) \tag{G.25}$$

where x is the coordinate 4-vector and $\delta(x)$ is the four-dimensional δ-function, $\delta(x^0)\delta(\boldsymbol{x})$; the sign in (G.25) has been chosen to be consistent with (G.15) in the static case. Taking the four-dimensional Fourier transform, and making suitable assumptions about the vanishing of G at the boundary of space–time, we obtain

$$(-q^2+m^2)\tilde{G}(q)=-1 \tag{G.26}$$

where

$$\tilde{G}(q)=\int \mathrm{e}^{\mathrm{i}q\cdot x}G(x)\,\mathrm{d}^4x$$

and so

$$\tilde{G}(q)=\frac{1}{q^2-m^2}. \tag{G.27}$$

As we have seen in detail in chapter 6, the Feynman prescription for selecting

the physically desired solution amounts to adding an 'iϵ' term in the denominator of (G.27):

$$\tilde{G}^{(+)}(q) = \frac{1}{q^2 - m^2 + \mathrm{i}\epsilon}. \tag{G.28}$$

Exercise

Verify the 'iϵ' specification in (G.23), using the methods of appendix F. [*Hint*: You need to show that the Fourier transform of (G.23), defined by

$$\hat{G}_k^{(+)}(\boldsymbol{r}) = \frac{1}{(2\pi)^3} \int \mathrm{e}^{\mathrm{i}\boldsymbol{q}\cdot\boldsymbol{r}} \tilde{G}_k^{(+)}(\boldsymbol{q})\, \mathrm{d}^3\boldsymbol{q}, \tag{G.29}$$

is equal to $G_k^{(+)}(\boldsymbol{r})$ of (G.14). Do the integration over the polar angles of $\boldsymbol{q}$, taking the direction of $\boldsymbol{r}$ as the polar axis. This gives

$$\hat{G}_k^{(+)}(\boldsymbol{r}) = \frac{-1}{8\pi^2} \int_{-\infty}^{\infty} \left(\frac{\mathrm{e}^{\mathrm{i}qr} - \mathrm{e}^{-\mathrm{i}qr}}{\mathrm{i}r} \right) \frac{q\, \mathrm{d}q}{q^2 - k^2 - \mathrm{i}\epsilon} \tag{G.30}$$

where $q = |\boldsymbol{q}|$, $r = |\boldsymbol{r}|$, and we have used the fact that the integrand is an even function of q to extend the lower limit to $-\infty$, with an overall factor of $1/2$. Now convert q to the complex variable z. Locate the poles of $(z^2 - k^2 - \mathrm{i}\epsilon)^{-1}$ (compare the similar calculation in section 10.3.1, and in appendix F). Apply Cauchy's integral formula (F.17), closing the $\mathrm{e}^{\mathrm{i}zr}$ part in the upper half z-plane, and the $\mathrm{e}^{-\mathrm{i}zr}$ part in the lower half z-plane.]

H

Elements of Non-relativistic Scattering Theory

H.1 Time-independent formulation and differential cross section

We consider the scattering of a particle of mass m by a fixed spherically symmetric potential $V(\boldsymbol{r})$; we shall retain $\hbar$ explicitly in what follows. The potential is assumed to go to zero rapidly as $r \to \infty$, as for the Yukawa potential (G.16); it will turn out that the important Coulomb case can be treated as the $a \to \infty$ limit of (G.16). We shall treat the problem here as a *stationary state* one, in which the Schrödinger wavefunction $\psi(\boldsymbol{r}, t)$ has the form

$$\psi(\boldsymbol{r}, t) = \phi(\boldsymbol{r}) \mathrm{e}^{-\mathrm{i}Et\hbar} \tag{H.1}$$

where E is the particle's energy, and where $\phi(\boldsymbol{r})$ satisfies the equation

$$\left[\frac{-\hbar^2}{2m}\boldsymbol{\nabla}^2 + V(\boldsymbol{r})\right]\phi(\boldsymbol{r}) = E\phi(\boldsymbol{r}). \tag{H.2}$$

We shall take V to be spherically symmetric, so that $V(\boldsymbol{r}) = V(r)$ where $r = |\boldsymbol{r}|$. In this approach to scattering, we suppose the potential to be 'bathed' in a steady flux of incident particles, all of energy E. The wavefunction for the incident beam, far from the region near the origin where V is appreciably non-zero, is then just a plane wave of the form $\phi_{\mathrm{inc}} = \mathrm{e}^{\mathrm{i}kz}$, where the z-axis has been chosen along the propagation direction, and where $E = \hbar^2\boldsymbol{k}^2/2m$ with $\boldsymbol{k} = (0, 0, k)$. This plane wave is normalized to one particle per unit volume, and yields a steady-state flux of

$$\begin{aligned} \boldsymbol{j}_{\mathrm{inc}} &= \frac{\hbar}{2m\mathrm{i}}[\phi^*_{\mathrm{inc}}\boldsymbol{\nabla}\phi_{\mathrm{inc}} - \phi_{\mathrm{inc}}\boldsymbol{\nabla}\phi^*_{\mathrm{inc}}] \\ &= \hbar\boldsymbol{k}/m = \boldsymbol{p}/m \end{aligned} \tag{H.3}$$

where the momentum is $\boldsymbol{p} = \hbar\boldsymbol{k}$. As expected, the incident flux is given by the velocity $\boldsymbol{v}$ per unit volume.

Though we have represented the incident beam as a plane wave, it will, in practice, be collimated. We could, of course, superpose such plane waves,

with different $\boldsymbol{k}$'s, to make a wave-packet of any desired localization. But the dimensions of practical beams are so much greater than the de Broglie wavelength $\lambda = h/p$ of our particles, that our plane wave will be a very good approximation to a realistic packet.

The form of the complete solution to (H.2), even in the region where V is essentially zero, is not simply the incident plane wave, however. The presence of the potential gives rise also to a *scattered wave*, whose form as $r \to \infty$ is

$$\phi_{\rm sc} = f(\theta,\phi)\frac{{\rm e}^{{\rm i}kr}}{r}. \tag{H.4}$$

We shall actually derive this later, but its physical interpretation is simply that it is an outgoing ($\sim {\rm e}^{{\rm i}kr}$ rather than ${\rm e}^{-{\rm i}kr}$) 'spherical wave', with a factor $f(\theta,\phi)$ called the *scattering amplitude* that allows for the fact that even though $V(r)$ is spherically symmetric, the solution, in general, will not be (recall the bound-state solutions of the Coulomb potential in the hydrogen atom). Calculating the radial component of the flux corresponding to (H.4) yields

$$\begin{aligned} j_{r,{\rm sc}} &= \frac{\hbar}{2m{\rm i}}\left[\phi^*_{\rm sc}\frac{\partial}{\partial r}\phi_{\rm sc} - \phi_{\rm sc}\frac{\partial}{\partial r}\phi^*_{\rm sc}\right] \\ &= \frac{\hbar k}{m}|f(\theta,\phi)|^2/r^2. \end{aligned} \tag{H.5}$$

The flux in the two non-radial directions will contain an extra power of r in the denominator – recall that

$$\boldsymbol{\nabla} = \hat{r}\frac{\partial}{\partial r} + \hat{\theta}\frac{1}{r}\frac{\partial}{\partial\theta} + \hat{\phi}\frac{1}{r\sin\theta}\frac{\partial}{\partial\phi}$$

and so (H.5) represents the correct asymptotic form of the scattered flux.

The *cross section* is now easily found. The differential cross section, ${\rm d}\sigma$, for scattering into the element of solid angle ${\rm d}\Omega$ is defined by

$${\rm d}\sigma = j_{r,{\rm sc}}\,{\rm d}S/|\boldsymbol{j}_{\rm inc}| \tag{H.6}$$

where ${\rm d}S = r^2\,{\rm d}\Omega$, so that from (H.3) and (H.5)

$$\frac{{\rm d}\sigma}{{\rm d}\Omega} = |f(\theta,\phi)|^2. \tag{H.7}$$

The total cross section is then just

$$\sigma = \int |f(\theta,\phi)|^2\,{\rm d}\Omega. \tag{H.8}$$

It is important to realize that the complete asymptotic form of the solution to (H.2) is the superposition of $\phi_{\rm inc}$ and $\phi_{\rm sc}$:

$$\phi(\boldsymbol{r}) \overset{r\to\infty}{\to} {\rm e}^{{\rm i}kz} + f(\theta,\phi)\frac{{\rm e}^{{\rm i}kr}}{r}. \tag{H.9}$$

Note that in the 'forward direction' (i.e. within a region close to the z-axis, as determined by the collimation), the incident and scattered waves will interfere. Careful analysis reveals a depletion of the incident beam in the forward direction (the 'shadow' of the scattering centre), which corresponds exactly to the total flux scattered into all angles (Gottfried 1966, section 12.3). This is expressed in the *optical theorem*:

$$\operatorname{Im} f(0) = \frac{k}{4\pi}\sigma. \tag{H.10}$$

H.2 Expression for the scattering amplitude: Born approximation

We begin by rewriting (H.2) as

$$(\boldsymbol{\nabla}^2 + \boldsymbol{k}^2)\phi(\boldsymbol{r}) = \frac{2m}{\hbar^2}V(\boldsymbol{r})\phi(\boldsymbol{r}). \tag{H.11}$$

This equation is of exactly the form discussed in appendix G, e.g. equation (G.18) with $\Omega_r = \boldsymbol{\nabla}^2 + \boldsymbol{k}^2$. Further, we know that the Green function for this Ω_r, corresponding to the desired outgoing wave solution, is given by (G.14). Using then (G.19) and (G.14), we can immediately write the 'formal solution' of (H.11) as

$$\phi(\boldsymbol{r}) = \mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{r}} + \frac{2m}{\hbar^2}\int -\frac{1}{4\pi}\,\frac{\mathrm{e}^{\mathrm{i}k|\boldsymbol{r}-\boldsymbol{r}'|}}{|\boldsymbol{r}-\boldsymbol{r}'|}V(\boldsymbol{r}')\phi(\boldsymbol{r}')\,\mathrm{d}^3\boldsymbol{r}' \tag{H.12}$$

where we have chosen '$u(\boldsymbol{r})$' in (G.19) to be the incident plane wave ϕ_{inc}, and have used $\boldsymbol{k}\cdot\boldsymbol{r} = kz$. We say 'formal' because of course the unknown $\phi(\boldsymbol{r}')$ still appears on the right-hand side of (H.12).

It may therefore seem that we have made no progress – but in fact (H.12) leads to a very useful expression for $f(\theta,\phi)$, which is the quantity we need to calculate. This can be found by considering the asymptotic ($r\to\infty$) limit of the integral term in (H.12). We have

$$\begin{aligned} |\boldsymbol{r}-\boldsymbol{r}'| &= (\boldsymbol{r}^2 + \boldsymbol{r}'^2 - 2\boldsymbol{r}\cdot\boldsymbol{r}')^{1/2} \\ &\sim r - \boldsymbol{r}\cdot\boldsymbol{r}'/r + O\left(\frac{1}{r}\right)\text{terms}. \end{aligned} \tag{H.13}$$

Thus in the exponent we may write

$$\mathrm{e}^{\mathrm{i}k|\boldsymbol{r}-\boldsymbol{r}'|} \approx \mathrm{e}^{\mathrm{i}k(r-\boldsymbol{r}\cdot\boldsymbol{r}'/r)} = \mathrm{e}^{\mathrm{i}kr}\mathrm{e}^{-\mathrm{i}\boldsymbol{k}'\cdot\boldsymbol{r}'}$$

where $\boldsymbol{k}' = k\hat{\boldsymbol{r}}$ is the *outgoing wavevector*, pointing along the direction of the

outgoing scattered wave which enters $\mathrm{d}S$. In the denominator factor we may simply say $|\boldsymbol{r}-\boldsymbol{r}'|^{-1} \approx r^{-1}$ since the next term in (H.13) will produce a correction of order r^{-2}. Putting this together, we have

$$\phi(\boldsymbol{r}) \overset{r\to\infty}{\to} \mathrm{e}^{\mathrm{i}kz} - \frac{m}{2\pi\hbar^2}\frac{\mathrm{e}^{\mathrm{i}kr}}{r}\int \mathrm{e}^{-\mathrm{i}\boldsymbol{k}'\cdot\boldsymbol{r}'} V(\boldsymbol{r}')\phi(\boldsymbol{r}')\,\mathrm{d}^3\boldsymbol{r}' \tag{H.14}$$

from which follows the formula for $f(\theta,\phi)$:

$$f(\theta,\phi) = -\frac{m}{2\pi\hbar^2}\int \mathrm{e}^{-\mathrm{i}\boldsymbol{k}'\cdot\boldsymbol{r}'} V(\boldsymbol{r}')\phi(\boldsymbol{r}')\,\mathrm{d}^3\boldsymbol{r}'. \tag{H.15}$$

No approximations have been made thus far, in deriving (H.15) – but of course it still involves the unknown $\phi(\boldsymbol{r}')$ inside the integral. However, it is in a form which is very convenient for setting up a *systematic approximation scheme* – a kind of perturbation theory – in powers of V. If the potential is relatively 'weak', its effect will be such as to produce only a slight distortion of the incident wave, and so $\phi(\boldsymbol{r}) \approx \mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{r}}$+'small correction'. This suggests that it may be a good approximation to replace $\phi(\boldsymbol{r}')$ in (H.15) by the undistorted incident wave $\mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{r}'}$, giving the *approximate* scattering amplitude

$$f_{\mathrm{BA}}(\theta,\phi) = -\frac{m}{2\pi\hbar^2}\int \mathrm{e}^{\mathrm{i}\boldsymbol{q}\cdot\boldsymbol{r}'} V(\boldsymbol{r}')\,\mathrm{d}^3\boldsymbol{r}' \tag{H.16}$$

where the wave vector transfer $\boldsymbol{q}$ is given by

$$\boldsymbol{q} = \boldsymbol{k} - \boldsymbol{k}'. \tag{H.17}$$

This is called the 'Born approximation to the scattering amplitude'. The criteria for the validity of the Born approximation are discussed in many standard quantum mechanics texts.

The approximation can be improved by returning to (H.12) for $\phi(\boldsymbol{r})$, and replacing $\phi(\boldsymbol{r}')$ inside the integral by $\mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{r}'}$ just as we did in (H.16); this will give us a formula for the first-order (in V) correction to $\phi(\boldsymbol{r})$. We can now insert *this* expression for $\phi(\boldsymbol{r}')$ (i.e. $\phi(\boldsymbol{r}') = \mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{r}'} + O(V)$ correction) into (H.15), which will give us f_{BA} again as the first term, but also another term, of order V^2 (since V appears in the integral in (H.15)). By iterating the process indefinitely, the *Born series* can be set up, to all orders in V.

H.3 Time-dependent approach

In this approach we consider the potential $V(\boldsymbol{r})$ as causing transitions between states describing the incident and scattered particles. From standard time-dependent perturbation theory in quantum mechanics, the transition

probability per unit time for going from state $|\mathrm{i}\rangle$ to state $|\mathrm{f}\rangle$, to first order in V, is given by

$$\dot{P}_{\mathrm{fi}} = \frac{2\pi}{\hbar}|\langle \mathrm{f}|V|\mathrm{i}\rangle|^2 \rho(E_{\mathrm{f}})|_{E_{\mathrm{f}}=E_{\mathrm{i}}} \tag{H.18}$$

where $\rho(E_{\mathrm{f}})\mathrm{d}E_{\mathrm{f}}$ is the number of final states in the energy range $\mathrm{d}E_{\mathrm{f}}$ around the energy-conserving point $E_{\mathrm{i}} = E_{\mathrm{f}}$. Equation (H.18) is often known as the 'Golden Rule'. In the present case, if we adopt the same normalization as in the previous section, the initial and final states are represented by the wavefunction $\mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{r}}$ and $\mathrm{e}^{-\mathrm{i}\boldsymbol{k}'\cdot\boldsymbol{r}}$, so that

$$\langle \mathrm{f}|V|\mathrm{i}\rangle = \int \mathrm{e}^{\mathrm{i}\boldsymbol{q}\cdot\boldsymbol{r}} V(\boldsymbol{r})\,\mathrm{d}^3\boldsymbol{r} \equiv \tilde{V}(\boldsymbol{q}). \tag{H.19}$$

Also, the number of such states in a volume element $\mathrm{d}^3\boldsymbol{p}'$ of momentum space ($\boldsymbol{p}' = \hbar\boldsymbol{k}'$) is $\mathrm{d}^3\boldsymbol{p}'/(2\pi\hbar)^3$.

In spherical polar coordinates, with $\mathrm{d}\Omega$ standing for the element of solid angle around the direction $(\theta, /\phi)$ of $\boldsymbol{p}'$, we have

$$\mathrm{d}^3\boldsymbol{p}' = \boldsymbol{p}'^2\,\mathrm{d}|\boldsymbol{p}'|\,\mathrm{d}\Omega = m|\boldsymbol{p}'|\,\mathrm{d}E'\,\mathrm{d}\Omega \tag{H.20}$$

where we have used $E' = \boldsymbol{p}'^2/2m$. It follows that

$$\rho(E')\,\mathrm{d}E' = \frac{\mathrm{d}^3\boldsymbol{p}'}{(2\pi\hbar)^3} = \frac{m}{(2\pi\hbar)^3}|\boldsymbol{p}'|\,\mathrm{d}\Omega\,\mathrm{d}E' \tag{H.21}$$

and so

$$\rho(E') = \frac{m}{(2\pi\hbar)^3}|\boldsymbol{p}'|\,\mathrm{d}\Omega. \tag{H.22}$$

Inserting (H.19) and (H.22) into (H.18) we obtain, for this case,

$$\dot{P}_{\mathrm{fi}} = \frac{2\pi}{\hbar}|\tilde{V}(\boldsymbol{q})|^2 \frac{m}{(2\pi\hbar)^3}|\boldsymbol{p}|\,\mathrm{d}\Omega. \tag{H.23}$$

To get the cross section, we need to divide this expression by the incident flux, which is $|\boldsymbol{p}|/m$ as in (H.3). Thus the differential cross section for scattering into the element of solid angle $d\Omega$ in the direction (θ, ϕ) is

$$\mathrm{d}\sigma = \left(\frac{m}{2\pi\hbar^2}\right)^2 |\tilde{V}(\boldsymbol{q})|^2\,\mathrm{d}\Omega. \tag{H.24}$$

Comparing (H.24) with (H.7) and (H.16), we see that this application of the Golden Rule (first-order time-dependent perturbation theory) is exactly equivalent to the Born approximation in the time-independent approach. It is, however, the time-dependent approach which is much closer to the corresponding quantum field theory formulation we introduce in chapter 6.


Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

I

The Schrödinger and Heisenberg Pictures

The standard introductory formalism of quantum mechanics is that of Schrödinger, in which the dynamical variables (such as $\boldsymbol{x}$ and $\hat{\boldsymbol{p}} = -\mathrm{i}\boldsymbol{\nabla}$) are independent of time, while the wavefunction ψ changes with time according to the general equation

$$\hat{H}\psi(\boldsymbol{x}, t) = \mathrm{i}\frac{\partial\psi(\boldsymbol{x}, t)}{\partial t} \tag{I.1}$$

where $\hat{H}$ is the Hamiltonian. Matrix elements of operators $\hat{A}$ depending on $\boldsymbol{x}, \hat{\boldsymbol{p}} \ldots$ then have the form

$$\langle\phi|\hat{A}|\psi\rangle = \int \phi^*(\boldsymbol{x}, t)\hat{A}\psi(\boldsymbol{x}, t)\,\mathrm{d}^3\boldsymbol{x} \tag{I.2}$$

and will, in general, depend on time via the time dependences of ϕ and ψ. Although used almost universally in introductory courses on quantum mechanics, this formulation is not the only possible one, nor is it always the most convenient.

We may, for example, wish to bring out similarities (and differences) between the general dynamical frameworks of quantum and classical mechanics. The formulation here does not seem to be well adapted to this purpose, since in the classical case the dynamical variables depend on time ($\boldsymbol{x}(t), \boldsymbol{p}(t) \ldots$) and obey equations of motion, while the quantum variables $\hat{A}$ are time-independent and the 'equation of motion' (I.1) is for the wavefunction ψ, which has no classical counterpart. In quantum mechanics, however, it is always possible to make unitary transformations of the state vector or wavefunctions. We can make use of this possibility to obtain an alternative formulation of quantum mechanics, which is in some ways closer to the spirit of classical mechanics, as follows.

Equation (I.1) can be formally solved to give

$$\psi(\boldsymbol{x}, t) = \mathrm{e}^{-\mathrm{i}\hat{H}t}\psi(\boldsymbol{x}, 0) \tag{I.3}$$

where the exponential (of an operator!) can be defined by the corresponding power series, for example:

$$\mathrm{e}^{-\mathrm{i}\hat{H}t} = 1 - \mathrm{i}\hat{H}t + \frac{1}{2!}(-\mathrm{i}\hat{H}t)^2 + \cdots. \tag{I.4}$$

It is simple to check that (I.3) as defined by (I.4), does satisfy (I.1) and that

the operator $\hat{U} = \exp(-\mathrm{i}\hat{H}t)$ is unitary:

$$U^\dagger = [\exp(-\mathrm{i}\hat{H}t)]^\dagger = \exp(\mathrm{i}\hat{H}^\dagger t) = \exp(\mathrm{i}\hat{H}t) = U^{-1} \tag{I.5}$$

where the Hermitian property $\hat{H}^\dagger = \hat{H}$ has been used. Thus (I.3) can be viewed as a unitary transformation from the time-dependent wavefunction $\psi(\boldsymbol{x}, t)$ to the time-independent one $\psi(\boldsymbol{x}, 0)$. Correspondingly the matrix element (I.2) is then

$$\langle\phi|\hat{A}|\psi\rangle = \int \phi^*(\boldsymbol{x}, 0)\mathrm{e}^{\mathrm{i}\hat{H}t}\hat{A}\mathrm{e}^{-\mathrm{i}\hat{H}t}\psi(\boldsymbol{x}, 0)\,\mathrm{d}^3\boldsymbol{x} \tag{I.6}$$

which can be regarded as the matrix element of the *time-dependent* operator

$$\hat{A}(t) = \mathrm{e}^{\mathrm{i}\hat{H}t}\hat{A}\mathrm{e}^{-\mathrm{i}\hat{H}t} \tag{I.7}$$

between time-independent wavefunctions $\phi^*(\boldsymbol{x}, 0), \psi(\boldsymbol{x}, 0)$.

Since (I.6) is perfectly general, it is clear that we can calculate amplitudes in quantum mechanics in either of the two ways outlined: (i) by using time-dependent ψ's and time-independent $\hat{A}$'s, which is called the 'Schrödinger picture': or (ii) by using time-independent ψ's and time-dependent $\hat{A}$'s, which is called the 'Heisenberg picture'. The wavefunctions and operators in the two pictures are related by (I.3) and (I.7). We note that the pictures coincide at the (conventionally chosen) time $t = 0$.

Since $\hat{A}(t)$ is now time-dependent, we can ask for its equation of motion. Differentiating (I.7) carefully, we find (if $\hat{A}$ does not depend explicitly on t) that

$$\frac{\mathrm{d}\hat{A}(t)}{\mathrm{d}t} = -\mathrm{i}[\hat{A}(t), \hat{H}] \tag{I.8}$$

which is called the Heisenberg equation of motion for $\hat{A}(t)$. On the right-hand side of (I.8), $\hat{H}$ is the Schrödinger operator; however, if $\hat{H}$ is substituted for $\hat{A}$ in (I.7), one finds $\hat{H}(t) = \hat{H}$, so $\hat{H}$ can equally well be interpreted as the Heisenberg operator. For simple Hamiltonians $\hat{H}$, (I.8) leads to operator equations quite analogous to classical equations of motion, which can sometimes be solved explicitly (see section 5.2.2 of chapter 5).

The foregoing ideas apply equally well to the operators and state vectors of quantum field theory.

J

Dirac Algebra and Trace Identities

J.1 Dirac algebra

J.1.1 γ matrices

The fundamental anticommutator

$$\{\gamma^\mu, \gamma^\nu\} = 2g^{\mu\nu} \tag{J.1}$$

may be used to prove the following results.

$$\begin{aligned}
\gamma_\mu \gamma^\mu &= 4 && \text{(J.2)}\\
\gamma_\mu \not{a} \gamma^\mu &= -2\not{a} && \text{(J.3)}\\
\gamma_\mu \not{a} \not{b} \gamma^\mu &= 4a \cdot b && \text{(J.4)}\\
\gamma_\mu \not{a} \not{b} \not{c} \gamma^\mu &= -2\not{c}\not{b}\not{a} && \text{(J.5)}\\
\not{a}\not{b} &= -\not{b}\not{a} + 2a \cdot b. && \text{(J.6)}
\end{aligned}$$

As an example, we prove this last result:

$$\begin{aligned}
\not{a}\not{b} &= a_\mu b_\nu \gamma^\mu \gamma^\nu \\
&= a_\mu b_\nu (-\gamma^\nu \gamma^\mu + 2g^{\mu\nu}) \\
&= -\not{b}\not{a} + 2a \cdot b.
\end{aligned}$$

J.1.2 γ_5 identities

Define

$$\gamma_5 = \mathrm{i}\gamma^0\gamma^1\gamma^2\gamma^3. \tag{J.7}$$

In the usual representation with

$$\gamma^0 = \begin{pmatrix} \mathbf{1} & \mathbf{0} \\ \mathbf{0} & -\mathbf{1} \end{pmatrix} \qquad \text{and} \qquad \boldsymbol{\gamma} = \begin{pmatrix} \mathbf{0} & \boldsymbol{\sigma} \\ -\boldsymbol{\sigma} & \mathbf{0} \end{pmatrix} \tag{J.8}$$

γ_5 is the matrix

$$\gamma_5 = \begin{pmatrix} \mathbf{0} & \mathbf{1} \\ \mathbf{1} & \mathbf{0} \end{pmatrix}. \tag{J.9}$$

Either from the definition or using this explicit form it is easy to prove that

$$\gamma_5^2 = 1 \tag{J.10}$$

and

$$\{\gamma_5, \gamma^\mu\} = 0 \tag{J.11}$$

i.e. γ_5 anticommutes with the other γ-matrices. Defining the totally antisymmetric tensor

$$\epsilon_{\mu\nu\rho\sigma} = \begin{cases} +1 & \text{for an even permutation of 0, 1, 2, 3} \\ -1 & \text{for an odd permutation of 0, 1, 2, 3} \\ 0 & \text{if two or more indices are the same} \end{cases} \tag{J.12}$$

we may write

$$\gamma_5 = \frac{\mathrm{i}}{4!}\epsilon_{\mu\nu\rho\sigma}\gamma^\mu\gamma^\nu\gamma^\rho\gamma^\sigma. \tag{J.13}$$

With this form it is possible to prove

$$\gamma_5\gamma_\sigma = \frac{\mathrm{i}}{3!}\epsilon_{\mu\nu\rho\sigma}\gamma^\mu\gamma^\nu\gamma^\rho \tag{J.14}$$

and the identity

$$\gamma^\mu\gamma^\nu\gamma^\rho = g^{\mu\nu}\gamma^\rho - g^{\mu\rho}\gamma^\nu + g^{\nu\rho}\gamma^\mu + \mathrm{i}\gamma_5\epsilon^{\mu\nu\rho\sigma}\gamma_\sigma. \tag{J.15}$$

J.1.3 Hermitian conjugate of spinor matrix elements

$$[\bar{u}(p', s')\Gamma u(p, s)]^\dagger = \bar{u}(p, s)\bar{\Gamma}u(p', s') \tag{J.16}$$

where Γ is any collection of γ matrices and

$$\bar{\Gamma} \equiv \gamma^0\Gamma^\dagger\gamma^0. \tag{J.17}$$

For example

$$\overline{\gamma^\mu} = \gamma^\mu \tag{J.18}$$

and

$$\overline{\gamma^\mu\gamma_5} = \gamma^\mu\gamma_5. \tag{J.19}$$

J.1.4 Spin sums and projection operators

Positive-energy projection operator:

$$[\Lambda_+(p)]_{\alpha\beta} \equiv \sum_s u_\alpha(p, s)\bar{u}_\beta(p, s) = (\not{p} + m)_{\alpha\beta}. \tag{J.20}$$

Negative-energy projection operator:

$$[\Lambda_-(p)]_{\alpha\beta} \equiv -\sum_s v_\alpha(p, s)\bar{v}_\beta(p, s) = (-\not{p} + m)_{\alpha\beta}. \tag{J.21}$$

Note that these forms are specific to the normalizations

$$\bar{u}u = 2m \qquad \bar{v}v = -2m \tag{J.22}$$

for the spinors.

J.2 Trace theorems

$$\begin{aligned} \mathrm{Tr}\mathbf{1} &= 4 \quad \text{(theorem 1)} && \text{(J.23)} \\ \mathrm{Tr}\gamma_5 &= 0 \quad \text{(theorem 2)} && \text{(J.24)} \\ \mathrm{Tr}(\text{odd number of } \gamma\text{'s}) &= 0 \quad \text{(theorem 3).} && \text{(J.25)} \end{aligned}$$

Proof

Consider

$$T \equiv \mathrm{Tr}(\not{a}_1\not{a}_2\ldots\not{a}_n) \tag{J.26}$$

where n is odd. Now insert $1 = (\gamma_5)^2$ into T, so that

$$T = \mathrm{Tr}(\not{a}_1\not{a}_2\ldots\not{a}_n\gamma_5\gamma_5). \tag{J.27}$$

Move the first γ_5 to the front of T by repeatedly using the result

$$\not{a}\gamma_5 = -\gamma_5\not{a}. \tag{J.28}$$

We therefore pick up n minus signs:

$$\begin{aligned} T &= \mathrm{Tr}(\not{a}_1\ldots\not{a}_n) = (-1)^n\mathrm{Tr}(\gamma_5\not{a}_1\ldots\not{a}_n\gamma_5) \\ &= (-1)^n\mathrm{Tr}(\not{a}_1\ldots\not{a}_n\gamma_5\gamma_5) \qquad \text{(cyclic property of trace)} \\ &= -\mathrm{Tr}(\not{a}_1\ldots\not{a}_n) \qquad \text{for } n \text{ odd.} \end{aligned} \tag{J.29}$$

Thus, for n odd, T must vanish:

$$\mathrm{Tr}(\not{a}\not{b}) = 4a\cdot b \qquad \text{(theorem 4).} \tag{J.30}$$

Proof

$$\begin{aligned} \mathrm{Tr}(\not{a}\not{b}) &= \tfrac{1}{2}\mathrm{Tr}(\not{a}\not{b} + \not{b}\not{a}) \\ &= \tfrac{1}{2}a_\mu b_\nu \mathrm{Tr}(\mathbf{1}.2g^{\mu\nu}) \\ &= 4a\cdot b. \\ \mathrm{Tr}(\not{a}\not{b}\not{c}\not{d}) &= 4[(a\cdot b)(c\cdot d) + (a\cdot d)(b\cdot c) - (a\cdot c)(b\cdot d)]. \qquad \text{(theorem 5)} \end{aligned} \tag{J.31}$$

Proof

$$\mathrm{Tr}(\not{a}\not{b}\not{c}\not{d}) = 2(a\cdot b)\mathrm{Tr}(\not{c}\not{d}) - \mathrm{Tr}(\not{b}\not{a}\not{c}\not{d}) \tag{J.32}$$

using the result of (J.6). We continue taking $\not{a}$ through the trace in this manner and use (J.30) to obtain

$$\begin{aligned}\mathrm{Tr}(\not{a}\not{b}\not{c}\not{d}) &= 2(a\cdot b)4(c\cdot d) - 2(a\cdot c)\mathrm{Tr}(\not{b}\not{d}) + \mathrm{Tr}(\not{b}\not{c}\not{a}\not{d}) \\ &= 8(a\cdot b)(c\cdot d) - 8(a\cdot c)(b\cdot d) + 8(b\cdot c)(a\cdot d) - \mathrm{Tr}(\not{b}\not{c}\not{d}\not{a})\end{aligned} \tag{J.33}$$

and, since we can bring $\not{a}$ to the front of the trace, we have proved the theorem.

$$\mathrm{Tr}[\gamma_5\not{a}] = 0. \qquad \text{(theorem 6)} \tag{J.34}$$

This is a special case of theorem 3 since γ_5 contains four γ matrices.

$$\mathrm{Tr}[\gamma_5\not{a}\not{b}] = 0. \qquad \text{(theorem 7)} \tag{J.35}$$

This is not so obvious; it may be proved by writing out all the possible products of γ matrices that arise.

$$\mathrm{Tr}[\gamma_5\not{a}\not{b}\not{c}] = 0. \qquad \text{(theorem 8)} \tag{J.36}$$

Again this is a special case of theorem 3.

$$\mathrm{Tr}[\gamma_5\not{a}\not{b}\not{c}\not{d}] = 4\mathrm{i}\epsilon_{\alpha\beta\gamma\delta}a^\alpha b^\beta c^\gamma d^\delta. \qquad \text{(theorem 9)} \tag{J.37}$$

This theorem follows by looking at components: the ϵ tensor just gives the correct sign of the permutation.

The ϵ tensor is the four-dimensional generalization of the three-dimensional antisymmetric tensor ϵ_{ijk}. In the three-dimensional case we have the well-known results

$$(\boldsymbol{b}\times\boldsymbol{c})_i = \epsilon_{ijk}b_jc_k \tag{J.38}$$

and

$$\boldsymbol{a}\cdot(\boldsymbol{b}\times\boldsymbol{c}) = \epsilon_{ijk}a_ib_jc_k \tag{J.39}$$

for the triple scalar product.

K

Example of a Cross Section Calculation

In this appendix we outline in more detail the calculation of the e^-s^+ elastic scattering cross section in section 8.3.2. The standard factors for the unpolarized cross section lead to the expression

$$\begin{aligned} d\bar{\sigma} &= \frac{1}{4E\omega|\boldsymbol{v}|}\frac{1}{2}\sum_{ss'}|\mathcal{M}_{e^-s^+}(s,s')|^2 \mathrm{dLips}(s;k',p') && \text{(K.1)} \\ &= \frac{1}{4[(k.p)^2-m^2M^2]^{1/2}}\frac{1}{2}\sum_{ss'}|\mathcal{M}_{e^-s^+}(s,s')|^2 \mathrm{dLips}(s;k',p') && \text{(K.2)} \end{aligned}$$

using the result of problem 6.9, and the definition of Lorentz-invariant phase space:

$$\mathrm{dLips}(s;k',p') \equiv (2\pi)^4\delta^4(k'+p'-k-p)\frac{d^3\boldsymbol{p'}}{(2\pi)^3 2E'}\frac{d^3\boldsymbol{k'}}{(2\pi)^3 2\omega'}. \qquad \text{(K.3)}$$

Instead of evaluating the matrix element and phase space integral in the CM frame, or writing the result in invariant form, we shall perform the calculation entirely in the 'laboratory' frame, defined as the frame in which the target (i.e. the s-particle) is at rest:

$$p^\mu = (M,\mathbf{0}) \qquad \text{(K.4)}$$

where M is the s-particle mass. Let us look in some detail at the 'laboratory' frame kinematics for elastic scattering (figure K.1). Conservation of energy and momentum in the form

$$p'^2 = (p+q)^2 \qquad \text{(K.5)}$$

allows us to eliminate p' to obtain the elastic scattering condition

$$2p\cdot q + q^2 = 0 \qquad \text{(K.6)}$$

or

$$\boxed{2p\cdot q = Q^2} \qquad \text{(K.7)}$$

if we introduce the positive quantity

$$Q^2 = -q^2 \qquad \text{(K.8)}$$

for a scattering process.

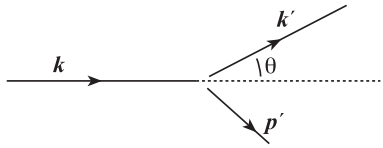


FIGURE K.1
Laboratory frame kinematics.

In all the applications with which we are concerned it will be a good approximation to neglect electron mass effects for high-energy electrons. We therefore set

$$k^2 = k'^2 \simeq 0 \tag{K.9}$$

so that

$$s + t + u \simeq 2M^2 \tag{K.10}$$

where

$$s = (k+p)^2 = (k'+p')^2 \tag{K.11}$$
$$t = (k-k')^2 = (p'-p)^2 = q^2 \tag{K.12}$$
$$u = (k-p')^2 = (k'-p)^2 \tag{K.13}$$

are the usual Mandelstam variables. For the electron 4-vectors

$$k^\mu = (\omega, \boldsymbol{k}) \tag{K.14}$$
$$k'^\mu = (\omega', \boldsymbol{k}') \tag{K.15}$$

we can neglect the difference between the magnitude of the 3-momentum and the energy,

$$\omega \simeq |\boldsymbol{k}| \equiv k \tag{K.16}$$
$$\omega' \simeq |\boldsymbol{k}'| \equiv k' \tag{K.17}$$

and in this approximation

$$q^2 = -2kk'(1 - \cos\theta) \tag{K.18}$$

or

$$\boxed{q^2 = -4kk' \sin^2(\theta/2).} \tag{K.19}$$

The elastic scattering condition (K.7) gives the following relation between k, k' and θ:

$$(k/k') = 1 + (2k/M)\sin^2(\theta/2). \tag{K.20}$$

It is important to realize that this relation is *only* true for elastic scattering: for inclusive inelastic electron scattering k, k' and θ are independent variables.

The first element of the cross section, the flux factor, is easy to evaluate:

$$4[(k\cdot p)^2 - m^2M^2]^{\frac{1}{2}} \simeq 4Mk \tag{K.21}$$

in the approximation of neglecting the electron mass m. We now consider the calculation of the spin-averaged matrix element and the phase space integral in turn.

K.1 The spin-averaged squared matrix element

The Feynman rules for es scattering enable us to write the spin sum in the form

$$\frac{1}{2}\sum_{s,s'} |\mathcal{M}_{\mathrm{e}^-\mathrm{s}^+}(s,s')|^2 = \left(\frac{4\pi\alpha}{q^2}\right)^2 L_{\mu\nu}T^{\mu\nu} \tag{K.22}$$

where $L_{\mu\nu}$ is the lepton tensor, $T^{\mu\nu}$ the s-particle tensor and the one-photon exchange approximation has been assumed. From problem 8.12 we find the result

$$L_{\mu\nu}T^{\mu\nu} = 8[2(k\cdot p)(k'\cdot p) + (q^2/2)M^2]. \tag{K.23}$$

In the 'laboratory' frame, neglecting the electron mass, this becomes

$$L_{\mu\nu}T^{\mu\nu} = 16M^2kk'\cos^2(\theta/2). \tag{K.24}$$

K.2 Evaluation of two-body Lorentz-invariant phase space in 'laboratory' variables

We must evaluate

$$\mathrm{dLips}(s;k',p') \equiv \frac{1}{(4\pi)^2}\delta^4(k'+p'-k-p)\frac{\mathrm{d}^3\boldsymbol{p}'}{E'}\frac{\mathrm{d}^3\boldsymbol{k}'}{\omega'} \tag{K.25}$$

in terms of 'laboratory' variables. This is in fact rather tricky and requires some care. There are several ways it can be done:

(i) Use CM variables, put the cross section into invariant form, and then translate to the 'laboratory' frame. This involves relating $\mathrm{d}q^2$ to $\mathrm{d}(\cos\theta)$ which we shall do as an exercise at the end of this appendix.

(ii) Alternatively, we can work directly in terms of 'laboratory' variables and write

$$\mathrm{d}^3\boldsymbol{p}'/2E' = \mathrm{d}^4p'\,\delta(p'^2 - M^2)\theta(p'^0). \tag{K.26}$$

The four-dimensional δ-function then removes the integration over d^4p' leaving us only with an integration over the single δ-function $\delta(p'^2 - M^2)$, in which p' is understood to be replaced by $k+p-k'$. For details of this last integration, see Bjorken and Drell (1964, p 114).

(iii) We shall evaluate the phase space integral in a more direct manner. We begin by performing the integral over $\mathrm{d}^3\boldsymbol{p}'$ using the three-dimensional δ-function from $\delta^4(k' + p' - k - p)$. In the 'laboratory' frame $\boldsymbol{p} = 0$, so we have

$$\int \mathrm{d}^3\boldsymbol{p}'\,\delta^3(\boldsymbol{k}' + \boldsymbol{p}' - \boldsymbol{k})f(\boldsymbol{p}', \boldsymbol{k}', \boldsymbol{k}) = f(\boldsymbol{p}', \boldsymbol{k}', \boldsymbol{k})|_{\boldsymbol{p}'=\boldsymbol{k}-\boldsymbol{k}'}. \quad \text{(K.27)}$$

In the particular function $f(\boldsymbol{p}', \boldsymbol{k}', \boldsymbol{k})$ that we require, $\boldsymbol{p}'$ only appears via E', since

$$E'^2 = \boldsymbol{p}'^2 + M^2 \quad \text{(K.28)}$$

and

$$\boldsymbol{p}'^2 = k^2 + k'^2 - 2kk'\cos\theta \quad \text{(K.29)}$$

(setting the electron mass m to zero). We now change $\mathrm{d}^3\boldsymbol{k}'$ to angular variables:

$$\mathrm{d}^3\boldsymbol{k}'/\omega' \simeq k'\mathrm{d}k'\mathrm{d}\Omega \quad \text{(K.30)}$$

leading to

$$\mathrm{dLips}(s; k', p') = \frac{1}{(4\pi)^2}\,\mathrm{d}\Omega\,\mathrm{d}k'\,\frac{k'}{E'}\delta(E' + k' - k - M). \quad \text{(K.31)}$$

Since E' is a function of k' and θ for a given k (cf (K.28) and (K.29)), the δ-function relates k' and θ as required for elastic scattering (cf (K.20)), but *until* the δ function integration is performed they must be regarded as independent variables. We have the integral

$$\frac{1}{(4\pi)^2}\int \mathrm{d}\Omega\,\mathrm{d}k'\,\frac{k'}{E'}\delta(f(k', \cos\theta)) \quad \text{(K.32)}$$

where

$$f(k', \cos\theta) = [(k^2 + k'^2 - 2kk'\cos\theta) + M^2]^{\frac{1}{2}} + k' - k - M \quad \text{(K.33)}$$

remaining to be evaluated. In order to obtain a differential cross section, we wish to integrate over k'; for this k' integration we must regard $\cos\theta$ in $f(k', \cos\theta)$ as a constant, and use the result (E.36):

$$\delta(f(x)) = \frac{1}{|f'(x)|_{x=x_0}}\delta(x - x_0) \quad \text{(K.34)}$$

where $f(x_0) = 0$. The required derivative is

$$\left.\frac{\mathrm{d}f}{\mathrm{d}k'}\right|_{\text{constant }\cos\theta} = \frac{1}{E'}(E' + k' - k\cos\theta) \quad \text{(K.35)}$$

and the δ-function requires that k' is determined from k and θ by the elastic scattering condition

$$k' = \frac{k}{1 + (2k/M)\sin^2(\theta/2)} \equiv k'(\cos\theta). \tag{K.36}$$

The integral (K.32) becomes

$$\frac{1}{(4\pi)^2}\int \mathrm{d}\Omega\,\mathrm{d}k'\,\frac{k'}{E'}\frac{1}{|\mathrm{d}f/\mathrm{d}k'|_{k'=k'(\cos\theta)}}\,\delta[k' - k'(\cos\theta)] \tag{K.37}$$

and, after some juggling, $\mathrm{d}f/\mathrm{d}k'$ evaluated at $k' = k'(\cos\theta)$ may be written as

$$\left.\frac{\mathrm{d}f}{\mathrm{d}k'}\right|_{k'=k'(\cos\theta)} = \frac{Mk}{E'k'}. \tag{K.38}$$

Thus we obtain finally the result

$$\boxed{\mathrm{dLips}(s;k',p') = \frac{1}{(4\pi)^2}\frac{k'^2}{Mk}\,\mathrm{d}\Omega} \tag{K.39}$$

for two-body elastic scattering in terms of 'laboratory' variables, neglecting lepton masses.

Putting all these elements together yields the advertised result

$$\left(\frac{\mathrm{d}\sigma}{\mathrm{d}\Omega}\right)_{\mathrm{ns}} \equiv \frac{\mathrm{d}\bar{\sigma}}{\mathrm{d}\Omega} = \frac{\alpha^2}{4k^2\sin^4(\theta/2)}\frac{k'}{k}\cos^2(\theta/2). \tag{K.40}$$

As a final twist to this calculation let us consider the change of variables from $\mathrm{d}\Omega$ to $\mathrm{d}q^2$ in this elastic scattering example. In the unpolarized case

$$\mathrm{d}\Omega = 2\pi\mathrm{d}(\cos\theta) \tag{K.41}$$

and

$$q^2 = -2kk'(1 - \cos\theta) \tag{K.42}$$

where

$$k' = \frac{k}{1 + (2k/M)\sin^2(\theta/2)}. \tag{K.43}$$

Thus, since k' and $\cos\theta$ are not independent variables, we have

$$\mathrm{d}q^2 = 2kk'\,\mathrm{d}(\cos\theta) + (1 - \cos\theta)(-2k)\frac{\mathrm{d}k'}{\mathrm{d}(\cos\theta)}\,\mathrm{d}(\cos\theta). \tag{K.44}$$

From (K.20) we find

$$\frac{\mathrm{d}k'}{\mathrm{d}(\cos\theta)} = \frac{k'^2}{M} \tag{K.45}$$

and, after some routine juggling, arrive at the result

$$\boxed{\mathrm{d}q^2 = 2k'^2\,\mathrm{d}(\cos\theta).} \tag{K.46}$$

If we introduce the variable ν defined, for elastic scattering, by

$$2p\cdot q \equiv 2M\nu = -q^2 \tag{K.47}$$

we have immediately

$$\mathrm{d}\nu = \frac{k'^2}{M}\,\mathrm{d}(\cos\theta). \tag{K.48}$$

Similarly, if we introduce the variable y defined by

$$y = \nu/k \tag{K.49}$$

we find

$$\mathrm{d}y = \frac{k'^2}{2\pi kM}\,\mathrm{d}\Omega \tag{K.50}$$

for elastic scattering.

L

Feynman Rules for Tree Graphs in QED

$2 \to 2$ cross section formula

$$\mathrm{d}\sigma = \frac{1}{4[(p_1 \cdot p_2)^2 - m_1^2 m_2^2]^{1/2}} |\mathcal{M}|^2 \mathrm{dLips}(s; p_3, p_4).$$

$1 \to 2$ decay formula

$$\mathrm{d}\Gamma = \frac{1}{2m_1} |\mathcal{M}|^2 \mathrm{dLips}(m_1^2; p_2, p_3).$$

Note that for two identical particles in the final state an extra factor of $\frac{1}{2}$ must be included in these formulae.

The amplitude $\mathrm{i}\mathcal{M}$ is the invariant matrix element for the process under consideration, and is given by the Feynman rules of the relevant theory. For particles with non-zero spin, unpolarized cross sections are formed by averaging over initial spin components and summing over final.

L.1 External particles

Spin-$\frac{1}{2}$

For each fermion or antifermion line entering the graph, include the spinor

$$u(p, s) \qquad \text{or} \qquad v(p, s) \tag{L.1}$$

and for spin-$\frac{1}{2}$ particles leaving the graph the spinor

$$\bar{u}(p', s') \qquad \text{or} \qquad \bar{v}(p', s'). \tag{L.2}$$

Photons

For each photon line entering the graph include a polarization vector

$$\epsilon_\mu(k, \lambda) \tag{L.3}$$

and for photons leaving the graph the vector

$$\epsilon_\mu^*(k', \lambda'). \tag{L.4}$$

L.2 Propagators

Spin-0

$$\text{----}\!\!\rightarrow\!\!\text{----} = \frac{\mathrm{i}}{p^2 - m^2 + \mathrm{i}\epsilon}. \tag{L.5}$$

Spin-$\frac{1}{2}$

$$\longrightarrow = \frac{\mathrm{i}}{\not{p} - m} = \mathrm{i}\frac{\not{p} + m}{p^2 - m^2 + \mathrm{i}\epsilon}. \tag{L.6}$$

Photon

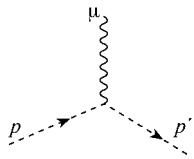
$$\sim\!\!\!\rightarrow\!\!\!\sim = \frac{\mathrm{i}}{k^2}\left(-g^{\mu\nu} + (1-\xi)\frac{k^\mu k^\nu}{k^2 + \mathrm{i}\epsilon}\right) \tag{L.7}$$

for a general ξ. Calculations are usually performed in the Lorentz or Feynman gauge with $\xi = 1$ and photon propagator equal to

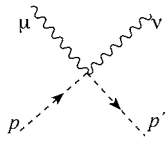
$$\mathrm{i}\frac{(-g^{\mu\nu})}{k^2 + \mathrm{i}\epsilon}. \tag{L.8}$$

L.3 Vertices

Spin-0

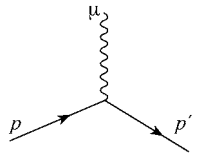


$-\mathrm{i}e(p + p')_\mu$ (for charge $+e$)



$2\mathrm{i}e^2 g_{\mu\nu}$

Spin-$\frac{1}{2}$



$-\mathrm{i}e\gamma_\mu$ (for charge $+e$)


Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

References

Aad G *et al.* 2012 (ATLAS Collaboration) *Phys. Lett.* B **716** 1
Abbiendi G *et al.* 1999 *Eur. Phys. J.* C **6** 1
——2000 (OPAL Collaboration) *Eur. Phys. J.* C **13** 553
Abe F *et al.* 1995 *Phys. Rev. Lett.* **74** 2626
Abachi S *et al.* 1995 *Phys. Rev. Lett.* **74** 2632
Abouzaid E *et al.* 2011 (KTeV Collaboration) *Phys. Rev.* D **83** 092001
Acciari M *et al.* 2000 (L3 Collaboration) *Phys. Lett.* B **476** 40
Achard P *et al.* 2005 (L3 Collaboration) *Phys. Lett.* B **623** 26
Ackerstaff K *et al.* 1998 *Eur. Phys. J.* C **2** 441
Aharonov Y and Bohm D 1959 *Phys. Rev.* **115** 485
Aitchison I J R 1985 *Contemp. Phys.* **26** 333
Altarelli G *et al.* (ed) 1989 *Z. Phys. at LEP 1* vol 1 (CERN 89-08, Geneva)
Althoff M *et al.* 1984 *Z. Phys.* C **22** 307
Alvarez A and Synkman A 2008 *Mod.Phys. Lett.* A **23** 2085
Anderson C D 1932 *Science* **76** 238
Anderson C D and Neddermeyer S 1937 *Phys. Rev.* **51** 884
Angelopoulos A *et al* 1998 [CPLEAR Collaboration] *Phys. Lett.* B **444** 43
Antoniadis I *et al.* 1998 *Phys. Lett.* B **436** 257
Aoyama T *et al.* 2007 *Phys. Rev. Lett.* **99** 110406
——2008 *Phys. Rev.* D **77** 053012
Arkani-Hamad N *et al.* 1998 *Phys. Lett.* B **429** 263
——1999 *Phys. Rev.* D **59** 086004
Arnison G *et al.* (UA1 Collaboration) 1985 *Phys. Lett.* B **158** 494
Attwood W B 1980 *Proc. 1979 SLAC Summer Institute on Particle Physics* (SLAC-R-224) ed A Mosher, vol 3
Aubert J J *et al.* 1974 *Phys. Rev. Lett.* **33** 1404
Augustin J E *et al.* 1974 *Phys. Rev. Lett.* **33** 1406
Banuls M C and Bernabeu 1999 *Phys. Lett.* B **464** 117
——2000 *Nucl. Phys.* B **590** 19
Bardeen J, Cooper L N and Schrieffer J R 1957 *Phys. Rev.* **108** 1175
Bennett G W *et al. Phys. Rev.* D **73** 072003
Bernstein J 1968 *Elementary Particles and their Currents* (San Francisco, CA: Freeman)
Berry M V 1984 *Proc. R. Soc.* A **392** 45
Bettini A 2008 *Introduction to Elementary Particle Physics* (Cambridge: Cambridge University Press)
Bjorken J D 1969 *Phys. Rev.* **179** 1547
Bjorken J D and Drell S D 1964 *Relativistic Quantum Mechanics* (New York: McGraw-Hill)
Bjorken J D and Glashow S L 1964 *Phys. Lett.* **11** 255

Bleuler K 1950 *Helv. Phys. Acta* **23** 567
Boas M L 1983 *Mathematical Methods in the Physical Sciences* 2nd edn (New York: Wiley)
Born M *et al.* 1926 *Z. f. Phys.* **35** 557
Bouchendira R *et al.* 2011 *Phys. Rev. Lett.* **106** 080801
Burkhardt H and Pietrzyk B 2001 *Phys. Lett.* B **513** 46
Callan C G and Gross D 1969 *Phys. Rev. Lett.* **22** 156
Carter A B and Sanda A I 1980 *Phys. Rev. Lett.* **45** 952
Casimir H B G 1948 *Koninkl. Ned. Akad. Wetenschap. Proc.* **51** 793
Chambers R G 1960 *Phys. Rev. Lett.* **5** 3
CHARM Collaboration 1981 *Phys. Lett.* B **99** 265
Chatrchyan S *et al.* 2012 (CMS Collaboration) *Phys. Lett.* B **716** 30
Christenson J H *et al.* 1964 *Phys. Rev. Lett.* **13** 138
Close F E 1979 *An Introduction to Quarks and Partons* (London: Academic)
Conversi M E *et al.* 1947 *Phys. Rev.* **71** 209
Cowan C L *et al.* 1956 *Science* **124** 103
Danby *et al.* 1962 *Phys. Rev. Lett.* **9** 36
Dirac P A M 1928 *Proc. R. Soc.* A **117** 610
——1933 *Phys. Z. der Sow.* **3** 64
——1981 *The Principles of Quantum Mechanics* 4th edn (Oxford: Oxford University Press) (reprinted)
Drell S D and Yan T M 1970 *Phys. Rev. Lett.* **25** 316
Dyson F J 1949a *Phys. Rev.* **75** 486
——1949b *Phys. Rev.* **75** 1736
Ehrenburg W and Siday R E 1949 *Proc. Phys. Soc.* B **62** 8
Englert F and Brout R 1964 *Phys. Rev. Lett.* **13** 321
Fermi E 1934a *Nuovo Cimento* **11** 1
——1934b *Z. Phys.* **88** 161
Feynman R P 1949a *Phys. Rev.* **76** 749
——1949b *Phys. Rev.* **76** 769
——1964 *The Feynman Lectures on Physics* vol 2 (Reading, MA: Addison-Wesley)
——1965a *The Feynman Lectures on Physics* vol 3 (Reading MA: Addison-Wesley)
——1965b *Symmetries in Elementary Particle Physics* 1964 Int. School of Physics 'Ettore Majorana' ed. A Zichichi (New York: Academic Press)
——1966 *Science* **153** 699
——1969 *Phys. Rev. Lett.* **23** 1415
Feynman R P and Hibbs A R 1965 *Quantum Mechanics and Path Integrals* (New York: McGraw-Hill) *Phys. Rev.* **105** 1681
Friedman J I and Kendall H W 1972 *Ann. Rev. Nucl. Sci.* **22** 203
Friedman J I and Telegdi V L 1957 *Phys. Rev.* **105** 1681
Gaillard M K and Lee B 1974 *Phys. Rev.* D **10** 897
Garwin R L *et al.* 1957 *Phys. Rev.* **105** 1415
Gell-Mann M 1962 *Phys. Rev.* **125** 1067
——1964 *Phys. Lett.* **8** 214
Ginzburg V I and Landau L D 1950 *Zh. Eksp. Teor. Fiz.* **20** 1064
Glashow S L, Iliopoulos J and Maiani L 1970 *Phys. Rev.* D **2** 1285
Goldhaber G *et al* 1976 *Phys. Rev. Lett.* **37** 255
Goldstein H 1980 *Classical Mechanics* 2nd edn (Reading, MA: Addison-Wesley) sections 1–3

Gottfried K 1966 *Quantum Mechanics* vol I (New York: Benjamin)
Green M B, Schwarz J H and Witten E 1987 *Superstring Theory* vols I and II (Cambridge: Cambridge University Press)
Griffiths D 2008 *Introduction to Elementary Particles* 2nd edn (New York: Wiley)
Gross D J and Wilczek F 1973 *Phys. Rev. Lett.* **30** 1343
Gupta S N 1950 *Proc. R. Soc.* A **63** 681
Guralnik G S *et al.* 1964 *Phys. Rev. Lett.* **13** 585
Hanneke D *et al.* 2008 *Phys. Rev. Lett.* **100** 120801
Heisenberg W 1939 *Z. Phys.* **113** 61
Herb S W *et al.* 1977 *Phys. Rev. Lett.* **39** 252
Higgs P W 1964 *Phys. Rev. Lett.* **13** 508
Hofstadter R 1963 *Electron Scattering and Nuclear and Nucleon Structure* (New York: Benjamin)
't Hooft G 1971a *Nucl. Phys.* B **33** 173
——1971b *Nucl. Phys.* B **35** 167
——1971c *Phys. Lett.* B **37** 195
——1980 *Sci. Am.* **242** (6) 90
't Hooft G and Veltman M 1972 *Nucl. Phys.* B **44** 189
Huet P and Sather E 1995 *Phys. Rev.* D **51** 379
Innes W R *et al.* 1977 *Phys. Rev. Lett.* **39** 1240
Itzykson C and Zuber J-B 1980 *Quantum Field Theory* (New York: McGraw-Hill)
Jegerlehner F and Nyffeler A 2009 *Phys. Rep.* **477** 1
Kabir P K 1970 *Phys. Rev.* D 2 540
Kanesawa S and Tomonaga S-i 1948a *Prog. Theor. Phys.* **3** 1
——1948b *Prog. Theor. Phys.* **3** 101
Kinoshita T and Nio M 2006 *Phys. Rev.* D **73** 053007
Kitchener J A and Prosser A P 1957 *Proc. R. Soc.* A **242** 403
Koba Z, Tati T and Tomonaga S-i 1947a *Prog. Theor. Phys.* **2** 101
——1947b *Prog. Theor. Phys.* **2** 198
Koba Z and Takeda G 1948 *Prog. Theor. Phys.* **3** 407
——1949 *Prog. Theor. Phys.* **4** 60, 130
Koba Z and Tomonaga S-i 1948 *Prog. Theor. Phys.* **3** 290
Kobayashi M and Maskawa K 1973 *Prog. Theor. Phys.* **49** 652
Kusch P and Foley H M 1947 *Phys. Rev.* **72** 1256
Landau L D 1948 *Dokl. Akad. Nauk. USSR* **60** 207
——1955 in *Niels Bohr and the Development of Physics* (New York: Pergamon) p 52
Landau L D and Lifshitz E M 1977 *Quantum Mechanics* 3rd edn, translated by J B Sykes and J S Bell (Oxford: Pergamon)
Lamoreaux S 1997 *Phys. Rev. Lett.* **78** 5
——1998 *Phys. Rev. Lett.* **81** 5475
La Rue G, Fairbank W M and Hebard A F 1977 *Phys. Rev. Lett.* **38** 1011
La Rue G, Phillips J D and Fairbank W M 1981 *Phys. Rev. Lett.* **46** 967
Leader E and Predazzi E 1996 *An Introduction to Gauge Theories and Modern Particle Physics* vol 1 (Cambridge: Cambridge University Press) ch 15–17
Lee T D 1981 *Particle Physics and Introduction to Field Theory* (Chur, London and New York: Harwood Academic Publishers)
Lee T D and Yang C N 1956 *Phys. Rev.* **104** 254

LEP 2003 (The LEP Working Group for Higgs Searches, ALEPH, DELPHI, L3 and OPAL collaborations) *Phys. Lett.* B **565** 61
Lüders G 1954 *Kong. Dansk. Vid. Selskab, Mat.-Fys. Medd.* **28** 5
——1957 *Ann. Phys.* **2** 1
Maiani L 1991 Z^0 physics *Proc. 1990 NATO ASI, Cargèse, France* ed M Lèvy *et al* (New York: Plenum) p 237
Majorana E 1937 *Nuovo Cimento* **14** 171
Mandelstam S 1958 *Phys. Rev.* **112** 1344
——1959 *Phys. Rev.* **115** 1741
Mandl F 1992 *Quantum Mechanics* (Chichester: Wiley)
Martin A D *et al.* 2009 *Eur. Phys. J.* C **63** 189
Maxwell J C 1864 *Phil. Trans. R. Soc.* **155** 459
Mott N F 1929 *Proc. R. Soc.* A **124** 425
Miyabayashi K *et al.* 1995 *Phys. Lett.* **347B** 171
Nakamura K *et al.* (Particle Data Group) 2010 *J. Phys. G: Nucl. Phys.* **37** 075021
Nambu Y 1960 *Phys. Rev. Lett.* **4** 380
Noether E 1918 *Nachr. Ges. Wiss. Göttingen* 171
Panofsky W K H and Phillips M 1962 *Classical Electricity and Magnetism* 2nd edn (Reading, MA: Addison-Wesley)
Pauli W 1933 *Handb. Phys.* **24** 246
——1940 *Phys. Rev.* **58** 716
——1957 *Nuovo Cimento* **6** 204
Perkins D H 1987 *Introduction to High Energy Physics* 3rd edn (Reading, MA: Addison-Wesley)
——2000 *Introduction to High Energy Physics* 4th edn (Cambridge: Cambridge University Press)
Perl M *et al.* 1975 *Phys. Rev. Lett.* **35** 1489
Peruzzi I *et al.* 1976 *Phys. Rev. Lett.* **37** 569
Peskin M E and Schroeder D V 1995 *An Introduction to Quantum Field Theory* (Reading, MA: Addison-Wesley)
Polchinski J 1998 *String Theory* vols I and II (Cambridge: Cambridge University Press)
Politzer H D 1973 *Phys. Rev. Lett.* **30** 1346
Sachrajda C T C 1983 *Gauge Theories in High Energy Physics, Les Houches Lectures, Session XXXVII* ed. M K Gaillard and R Stora (Amsterdam: North-Holland)
Sakharov A D 1967 *JETP Letters* **5** 24
Salam A 1968 *Elementary Particle Physics* ed. N Svartholm (Stockholm: Almqvist and Wiksells)
Schweber S S 1994 *QED and the Men Who Made It: Dyson, Feynman, Schwinger and Tomonaga* (Princeton, NJ: Princeton University Press)
Schwinger J 1948a *Phys. Rev.* **73** 416L
——1948b *Phys. Rev.* **74** 1439
——1949a *Phys. Rev.* **75** 651
——1949b *Phys. Rev.* **76** 790
——1962 *Phys. Rev.* **125** 397
Scott D M 1985 *Proc. School for Young High Energy Physicists, Rutherford Appleton Laboratory, 1984* (RAL-85-010) ed. J B Dainton

Shaw R 1955 The problem of particle types and other contributions to the theory of elementary particles *PhD Thesis* University of Cambridge
Sparnaay M J 1958 *Physica* **24** 751
Streater R F *et al.* 1964 *PCT, Spin and Statistics, and All That* (New York: Benjamin)
Street J C and Stevenson E C 1937 *Phys. Rev.* **52** 1003
Swanson M 1992 *Path Integrals and Quantum Processes* (Boston, MA: Academic)
Tomonaga S-i 1946 *Prog. Theor. Phys.* **1** 27
Uehling E A 1935 *Phys. Rev.* **48** 55
Uhlenbeck G F and Goudsmit S 1925 *Nature* **13** 953
Utiyama R 1956 *Phys. Rev.* **101** 1597
Ward J C 1950 *Phys. Rev.* **78** 182
Weinberg S 1967 *Phys. Rev. Lett.* **19** 1264
——1979 *Physica* A **96** 327
——1995 *The Quantum Theory of Fields* vol I (Cambridge: Cambridge University Press)
——1996 *The Quantum Theory of Fields* vol II (Cambridge: Cambridge University Press)
Wick G C 1938 *Nature* **142** 993
——1950 *Phys. Rev.* **80** 268
Wigner E P 1949 *Proc. Am. Phil. Soc.* **93** 521; also reprinted in *Symmetries and Reflections* (Bloomington, IN: Indiana University Press) pp 3–13
——1964 *Group Theory* 1959 edn., 4th printing (New York and London: Academic Press)
Wilson K G 1969 *Phys. Rev.* **179** 1499
Wu C S *et al.* 1957 *Phys. Rev.* **105** 1413
Yang C N 1950 *Phys. Rev.* **77** 242
Yang C N and Mills R L 1954 *Phys. Rev.* **96** 191
Yukawa H 1935 *Proc. Phys. Math. Soc. Japan* **17** 48
Zweig G 1964 *CERN Preprint* TH-401
Zwiebach B 2004 *A First Course in String Theory* (Cambridge: Cambridge University Press)


Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

Index

FOURTH EDITION

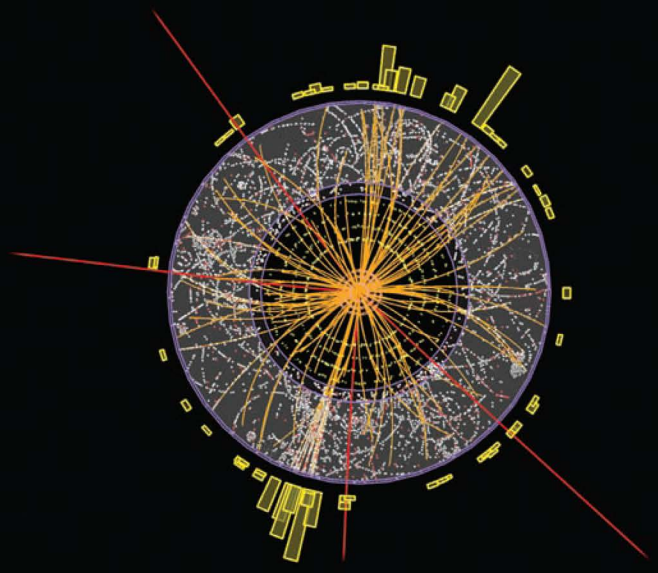
GAUGE THEORIES IN PARTICLE PHYSICS

A PRACTICAL INTRODUCTION

VOLUME 2

Non-Abelian Gauge Theories

QCD and The Electroweak Theory



Ian J.R. Aitchison • Anthony J.G. Hey

CRC Press
Taylor & Francis Group

FOURTH EDITION

GAUGE THEORIES IN PARTICLE PHYSICS

A PRACTICAL INTRODUCTION

VOLUME 2

Non-Abelian Gauge Theories

QCD and The Electroweak Theory


Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

FOURTH EDITION

GAUGE THEORIES IN PARTICLE PHYSICS

A PRACTICAL INTRODUCTION

VOLUME 2

Non-Abelian Gauge Theories

QCD and The Electroweak Theory

Ian J.R. Aitchison • Anthony J.G. Hey

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business

CRC Press
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

Printed in the United States of America on acid-free paper
Version Date: 2012912

International Standard Book Number: 978-1-4665-1307-5 (Hardback)

Library of Congress Cataloging-in-Publication Data

Aitchison, Ian Johnston Rhind, 1936-
Gauge theories in particle physics : a practical introduction / Ian J R Aitchison, Anthony J.G. Hey. -- 4th ed.
v. cm.
Includes bibliographical references and index.
Contents: v. 1. From relativistic quantum mechanics to QED -- v. 2. Non-Abelian gauge theories : QCD and the electroweak theory.
ISBN 978-1-4665-1299-3 (v. 1 : hardback) -- ISBN 978-1-4665-1307-5 (v. 2 : hardback)
1. Gauge fields (Physics) 2. Particles (Nuclear physics) 3. Weak interactions (Nuclear physics) 4. Quantum electrodynamics. 5. Feynman diagrams. I. Hey, Anthony J. G. II. Title.

QC793.3.F5A34 2012
539.7'21--dc23 2012031181

Visit the Taylor & Francis Web site at
http://www.taylorandfrancis.com

and the CRC Press Web site at
http://www.crcpress.com

To Jessie
and to
Jean, Katherine and Elizabeth


Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

Contents

Preface to Volume 2 of the Fourth Edition

The main focus of the second volume of this fourth edition, as in the third, is on the two non-Abelian quantum gauge field theories of the Standard Model – that is, QCD and the electroweak theory of Glashow, Salam and Weinberg. We preserve the same division into four parts: non-Abelian symmetries, both global and local; QCD and the renormalization group; spontaneously broken symmetry; and weak interaction phenomenology and the electroweak theory.

However, the book has always combined theoretical development with discussion of relevant experimental results. And it is on the experimental side that most progress has been made in the ten years since the third edition appeared – first of all, in the study of CP violation in B-meson physics, and in neutrino oscillations. The inclusion of these results, and the increasing importance of the topics, have required some reorganization, and a new chapter (21) devoted wholly to them. We concentrate mainly on CP-violation in B-meson decays, particularly on the determination of the angles of the unitarity triangle from B-meson oscillations. CP-violation in K-meson systems is also discussed. In the neutrino sector, we describe some of the principal experiments which have led to our current knowledge of the mass-squared differences and the mixing angles. In discussing weak interaction phenomenology, we keep in view the possibility that neutrinos may turn out to be Majorana particles, an outcome for which we have prepared the reader in (new) chapters 4 and 7 of volume 1.

More recently, on July 4, 2012, the ATLAS and CMS collaborations at the CERN LHC announced the discovery of a boson of mass between 125 and 126 GeV, with production and decay characteristics which are consistent (at the 1σ level) with those of the Standard Model Higgs boson. We can now conclude our treatment of the electroweak theory, and this volume, with a discussion of this historic discovery, which opens a new era in particle physics – one in which the electroweak symmetry-breaking (Higgs) sector of the SM will be rigorously tested.

Our treatment of a number of topics has been updated and, we hope, improved. In QCD, the definition of 2-jet cross sections in e^+e^- annihilation is explained, and used in a short discussion of jet algorithms (sections 14.5 and 14.6). Progress in lattice QCD is recognized with the inclusion of some of the recent impressive results using dynamical fermions (section 16.5). In the chapter on chiral symmetry breaking, a new section (18.3) introduces the

important technique of effective Lagrangians, including the extension to the three-flavour case and the associated mass relations. A much fuller account is given of three-generation quark mixing and the CKM matrix (section 20.7.3), as preparation for chapter 21. The essential points in chapter 21 of the previous edition, relating to problems with the current–current and IVB models, now provide the introductory motivation for the GSW theory in chapter 22.

One item has been banished to an appendix: geometrical aspects of gauge theories, which did after all seem to interrupt the flow of chapter 13 too much (but we hope readers will not ignore it). And another has been brought in from the cold: as already mentioned, Majorana fermions now find themselves appearing for the first time in volume 1.

Acknowledgements

We are very grateful to Paolo Strolin for providing a list of misprints and a very thorough catalogue of excellent comments for volume 2 of the third edition, which has resulted in a large number of improvements in the present text. The CP-violation sections in chapters 20 and 21 were much improved following detailed comments by Abi Soffer, and the neutrino sections in chapter 21 likewise benefited greatly from careful readings by Francesco Tramontano and Tim Cohen; we thank all three for their generous help. The eps files for figures 16.11 and 16.12 were kindly supplied by Christine Davies and Stephan Dürr, respectively. IJRA thanks Michael Peskin and Stan Brodsky for welcoming him as a visitor to the SLAC National Accelerator Laboratory Particle Theory group (supported by the Department of Energy under contract DE-AC02-76SF00515), and Bill Dunwoodie and BaBar colleagues for very kindly arranging for him to be a BaBar Associate; these connections have been invaluable. On a more technical note, IJRA thanks Xing-Gang Wu for some crucial help with JaxoDraw.

Ian J R Aitchison and Anthony J G Hey
October 2012

Part V

Non-Abelian Symmetries


Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

12

Global Non-Abelian Symmetries

12.1 The Standard Model

In the preceding volume, a very successful dynamical theory – QED – has been introduced, based on the remarkably simple *gauge principle*: namely, that the theory should be invariant under local phase transformations on the wavefunctions (chapter 2) or field operators (chapter 7) of charged particles. Such transformations were characterized as *Abelian* in section 2.6, since the phase factors commuted. The second volume of this book will be largely concerned with the formulation and elementary application of the remaining two dynamical theories within the Standard Model – that is, QCD and the electroweak theory. They are built on a generalization of the gauge principle, in which the transformations involve more than one state, or field, at a time. In that case, the 'phase factors' become matrices, which generally do not commute with each other, and the associated symmetry is called a '*non-Abelian*' one. When the phase factors are independent of the space-time coordinate x, the symmetry is a 'global non-Abelian' one; when they are allowed to depend on x, one is led to a non-Abelian gauge theory. Both QCD and the electroweak theory are of the latter type, providing generalizations of the Abelian U(1) gauge theory which is QED. It is a striking fact that all three dynamical theories in the Standard Model are based on a gauge principle of local phase invariance.

In this chapter we shall be mainly concerned with two global non-Abelian symmetries, which lead to useful conservation laws but not to any specific dynamical theory. We begin in section 12.1 with the first non-Abelian symmetry to be used in particle physics, the *hadronic isospin* 'SU(2) symmetry' proposed by Heisenberg (1932) in the context of nuclear physics, and now understood as following from QCD and the smallness of the u and d quark masses as compared with the QCD scale parameter $\Lambda_{\overline{\rm MS}}$ (see section 18.3.3). In section 12.2 we extend this to $\mathrm{SU(3)_f}$ flavour symmetry, as was first done by Gell-Mann (1961) and Ne'eman (1961) – an extension seen, in its turn, as reflecting the smallness of the u, d and s quark masses as compared with $\Lambda_{\overline{\rm MS}}$. The 'wavefunction' approach of sections 12.1 and 12.2 is then reformulated in field-theoretic language in section 12.3.

In the last section of this chapter, we shall introduce the idea of a global *chiral* symmetry, which is a symmetry of theories with massless fermions. This may be expected to be a good approximate symmetry for the u and d quarks.

But the anticipated observable consequences of this symmetry (for example, nucleon parity doublets) appear to be absent. This puzzle will be resolved in Part VII, via the profoundly important concept of 'spontaneous symmetry breaking'.

The formalism introduced in this chapter for SU(2) and SU(3) will be required again in the following one, when we consider the local versions of these non-Abelian symmetries and the associated dynamical gauge theories. The whole modern development of non-Abelian gauge theories began with the attempt by Yang and Mills (1954) (see also Shaw 1955) to make hadronic isospin into a local symmetry. However, the beautiful formalism developed by these authors turned out *not* to describe interactions between hadrons. Instead, it describes the interactions between the *constituents* of the hadrons, namely quarks – and this in two respects. First, a local SU(3) symmetry (called $SU(3)_c$) governs the strong interactions of quarks, binding them into hadrons (see Part VI). Secondly, a local SU(2) symmetry (called *weak isospin*) governs the weak interactions of quarks (and leptons); together with QED, this constitutes the electroweak theory (see Part VIII). It is important to realize that, despite the fact that each of these two local symmetries is based on the same group as one of the earlier global (flavour) symmetries, the physics involved is completely different. In the case of the strong quark interactions, the $SU(3)_c$ group refers to a new degree of freedom ('colour') which is quite distinct from flavour u, d, s (see chapter 14). In the weak interaction case, since the group is an SU(2), it is natural to use 'isospin language' in talking about it, particularly since flavour degrees of freedom are involved. But we must always remember that it is *weak* isospin, which (as we shall see in chapter 20) is an attribute of leptons as well as of quarks, and hence physically quite distinct from hadronic isospin. Furthermore, it is a parity-violating chiral gauge theory.

Despite the attractive conceptual unity associated with the gauge principle, the way in which each of QCD and the electroweak theory 'works' is actually quite different from QED, and from each other. Indeed it is worth emphasizing very strongly that it is, *a priori*, far from obvious why either the strong interactions between quarks, or the weak interactions, should have anything to do with gauge theories at all. Just as in the U(1) (electromagnetic) case, gauge invariance forbids a mass term in the Lagrangian for non-Abelian gauge fields, as we shall see in chapter 13. Thus it would seem that gauge field quanta are necessarily massless. But this, in turn, would imply that the associated forces must have a long-range (Coulombic) part, due to exchange of these massless quanta – and of course in neither the strong nor the weak interaction case is that what is observed.[1] As regards the former, the gluon quanta are indeed massless, but the contradiction is resolved by *non-perturbative* effects which lead to *confinement*, as we indicated in chapter 1. We shall discuss

[1] Pauli had independently developed the theory of non-Abelian gauge fields during 1953, but did not publish any of this work because of the seeming physical irrelevancy associated with the masslessness problem (Enz 2002, pages 474-82; Pais 2000, pages 242-5).

this further in chapter 16. In weak interactions, a third realization appears: the gauge quanta acquire mass via (it is believed) a second instance of *spontaneous symmetry breaking*, as will be explained in Part VII. In fact a further application of this idea is required in the electroweak theory, because of the chiral nature of the gauge symmetry in this case: the quark and lepton masses also must be 'spontaneously generated'.

12.2 The flavour symmetry $\mathbf{SU(2)_f}$

12.2.1 The nucleon isospin doublet and the group SU(2)

The transformations initially considered in connection with the gauge principle in section 2.5 were just global phase transformations on a single wavefunction

$$\psi' = \mathrm{e}^{\mathrm{i}\alpha}\psi. \tag{12.1}$$

The generalization to non-Abelian invariances comes when we take the simple step – but one with many ramifications – of considering more than one wavefunction, or state, at a time. Quite generally in quantum mechanics, we know that whenever we have a set of states which are *degenerate* in energy (or mass) there is no unique way of specifying the states: any linear combination of some initially chosen set of states will do just as well, provided the normalization conditions on the states are still satisfied. Consider, for example, the simplest case of just two such states – to be specific, the neutron and proton (figure 12.1). This single near coincidence of the masses was enough to suggest to Heisenberg (1932) that, as far as the strong nuclear forces were concerned (electromagnetism being negligible by comparison), the two states could be regarded as truly degenerate, so that any arbitrary linear combination of neutron and proton wavefunctions would be entirely equivalent, as far as this force was concerned, for a single 'neutron' or single 'proton' wavefunction. This hypothesis became known as 'charge independence of nuclear forces'. Thus redefinitions of neutron and proton wavefunctions could be allowed, of the form

$$\psi_\mathrm{p} \to \psi'_\mathrm{p} = \alpha\psi_\mathrm{p} + \beta\psi_\mathrm{n} \tag{12.2}$$

$$\psi_\mathrm{n} \to \psi'_\mathrm{n} = \gamma\psi_\mathrm{p} + \delta\psi_\mathrm{n} \tag{12.3}$$

for complex coefficients α, β, γ, and δ. In particular, since ψ_p and ψ_n are degenerate, we have

$$H\psi_\mathrm{p} = E\psi_\mathrm{p}, \qquad H\psi_\mathrm{n} = E\psi_\mathrm{n} \tag{12.4}$$

from which it follows that

$$H\psi'_\mathrm{p} = H(\alpha\psi_\mathrm{p} + \beta\psi_\mathrm{n}) = \alpha H\psi_\mathrm{p} + \beta H\psi_\mathrm{n} \tag{12.5}$$

$$= E(\alpha\psi_\mathrm{p} + \beta\psi_\mathrm{n}) = E\psi'_\mathrm{p} \tag{12.6}$$

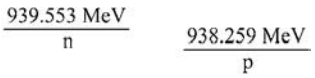


FIGURE 12.1
Early evidence for isospin symmetry.

and similarly

$$H\psi'_{\rm n} = E\psi'_{\rm n} \tag{12.7}$$

showing that the redefined wavefunctions still describe two states with the same energy degeneracy.

The two-fold degeneracy seen in figure 12.1 is suggestive of that found in spin-$\frac{1}{2}$ systems in the absence of any magnetic field; the $s_z = \pm\frac{1}{2}$ components are degenerate. The analogy can be brought out by introducing the *two-component nucleon isospinor*

$$\psi^{(1/2)} \equiv \begin{pmatrix} \psi_{\rm p} \\ \psi_{\rm n} \end{pmatrix} \equiv \psi_{\rm p}\chi_{\rm p} + \psi_{\rm n}\chi_{\rm n} \tag{12.8}$$

where

$$\chi_{\rm p} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \qquad \chi_{\rm n} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}. \tag{12.9}$$

In $\psi^{(1/2)}$, $\psi_{\rm p}$ is the amplitude for the nucleon to have 'isospin up', and $\psi_{\rm n}$ is that for it to have 'isospin down'.

As far as the states are concerned, this terminology arises, of course, from the formal identity between the 'isospinors' of (12.9) and the two-component eigenvectors (3.60) corresponding to eigenvalues $\pm\frac{1}{2}\hbar$ of (true) spin: compare also (3.61) and (12.8). It is important to be clear, however, that the degrees of freedom involved in the two cases are quite distinct; in particular, even though both the proton and the neutron have (true) spin$-\frac{1}{2}$, the transformations (12.2) and (12.3) leave the (true) spin part of their wavefunctions completely untouched. Indeed, we are suppressing the spinor part of both wavefunctions altogether (they are of course 4-component Dirac spinors). As we proceed, the precise mathematical nature of this 'spin-1/2' analogy will become clear.

Equations (12.2) and (12.3) can be compactly written in terms of $\psi^{(1/2)}$ as

$$\psi^{(1/2)} \to \psi^{(1/2)'} = \mathbf{V}\psi^{(1/2)}, \qquad \mathbf{V} = \begin{pmatrix} \alpha & \beta \\ \gamma & \delta \end{pmatrix} \tag{12.10}$$

where $\mathbf{V}$ is the indicated complex 2×2 matrix. Heisenberg's proposal, then, was that the physics of strong interactions between nucleons remained the same under the transformation (12.10): in other words, a symmetry was involved. We must emphasise that such a symmetry can *only* be exact in the *absence* of electromagnetic interactions: it is therefore an intrinsically approximate symmetry, though presumably quite a useful one in view of the relative weakness of electromagnetic interactions as compared to hadronic ones.

We now consider the general form of the matrix $\mathbf{V}$, as constrained by various relevant restrictions: quite remarkably, we shall discover that (after extracting an overall phase) $\mathbf{V}$ has essentially the same mathematical form as the matrix $\mathbf{U}$ of (4.33), which we encountered in the discussion of the transformation of (real) spin wavefunctions under rotations of the (real) space axes. It will be instructive to see how the present discussion leads to the same form (4.33).

We first note that $\mathbf{V}$ of (12.10) depends on four arbitrary complex numbers, or alternatively on eight real parameters. By contrast, the matrix $\mathbf{U}$ of (4.33) depends on only three real parameters, which we may think of in terms of two to describe the direction of the axis of rotation, and a third for the angle of rotation. However, $\mathbf{V}$ is subject to certain restrictions, and these reduce the number of free parameters in $\mathbf{V}$ to three, as we now discuss. First, in order to preserve the normalization of $\psi^{(1/2)}$ we require

$$\psi^{(1/2)'\dagger}\psi^{(1/2)'} = \psi^{(1/2)\dagger}\mathbf{V}^\dagger\mathbf{V}\psi^{(1/2)} = \psi^{(1/2)\dagger}\psi^{(1/2)} \tag{12.11}$$

which implies that $\mathbf{V}$ has to be *unitary*:

$$\mathbf{V}^\dagger\mathbf{V} = \mathbf{1}_2, \tag{12.12}$$

where $\mathbf{1}_2$ is the unit 2×2 matrix. Clearly this unitarity property is in no way restricted to the case of two states: the transformation coefficients for n degenerate states will form the entries of an $n \times n$ unitary matrix. A trivialization is the case $n = 1$, for which, as we noted in section 2.6, $\mathbf{V}$ reduces to a single phase factor as in (12.1), indicating how all the previous work is going to be contained as a special case of these more general transformations. Indeed, from elementary properties of determinants we have

$$\det\mathbf{V}^\dagger\mathbf{V} = \det\mathbf{V}^\dagger \cdot \det\mathbf{V} = \det\mathbf{V}^* \cdot \det\mathbf{V} = \mid \det\mathbf{V} \mid^2 = 1 \tag{12.13}$$

so that

$$\det\mathbf{V} = \exp(\mathrm{i}\theta) \tag{12.14}$$

where θ is a real number. We can separate off such an overall phase factor from the transformations mixing 'p' and 'n', because it corresponds to a rotation of the phase of both p and n wavefunctions by the *same* amount:

$$\psi'_\mathrm{p} = \mathrm{e}^{\mathrm{i}\alpha}\psi_\mathrm{p}, \quad \psi'_\mathrm{n} = \mathrm{e}^{\mathrm{i}\alpha}\psi_\mathrm{n}. \tag{12.15}$$

The $\mathbf{V}$ corresponding to (12.15) is $\mathbf{V} = \mathrm{e}^{\mathrm{i}\alpha}\mathbf{1}_2$, which has determinant $\exp(2\mathrm{i}\alpha)$ and is therefore of the form (12.1) with $\theta = 2\alpha$. In the field-theoretic formalism of section 7.2, such a symmetry can be shown to lead to the conservation of baryon number $N_\mathrm{u} + N_\mathrm{d} - N_{\bar{\mathrm{u}}} - N_{\bar{\mathrm{d}}}$, where bar denotes the antiparticle.

The new physics will lie in the remaining transformations which satisfy

$$\det\mathbf{V} = +1. \tag{12.16}$$

Such a matrix is said to be a *special* unitary matrix, which simply means it has unit determinant. Thus, finally, the $\mathbf{V}$'s we are dealing with are *special, unitary, 2×2 matrices*. The set of all such matrices form a *group*. The general defining properties of a group are given in appendix M. In the present case, the elements of the group are all such 2×2 matrices, and the 'law of combination' is just ordinary matrix multiplication. It is straightforward to verify (problem 12.1) that all the defining properties are satisfied here; the group is called 'SU(2)', the 'S' standing for 'special', the 'U' for 'unitary', and the '2' for '2×2'.

SU(2) is actually an example of a *Lie group* (see appendix M). Such groups have the important property that their physical consequences may be found by considering 'infinitesimal' transformations, that is – in this case – matrices $\mathbf{V}$ which differ only slightly from the 'no-change' situation corresponding to $\mathbf{V} = \mathbf{1}_2$. For such an infinitesimal SU(2) matrix $\mathbf{V}_{\text{infl}}$ we may therefore write

$$\mathbf{V}_{\text{infl}} = \mathbf{1}_2 + i\boldsymbol{\xi} \tag{12.17}$$

where $\boldsymbol{\xi}$ is a 2×2 matrix whose entries are all first-order small quantities. The condition $\det \mathbf{V}_{\text{infl}} = 1$ now reduces, on neglect of second-order terms $0(\boldsymbol{\xi}^2)$, to the condition (see problem 12.2)

$$\text{Tr}\boldsymbol{\xi} = 0. \tag{12.18}$$

The condition that $\mathbf{V}_{\text{infl}}$ be unitary, i.e.

$$(\mathbf{1}_2 + i\boldsymbol{\xi})(\mathbf{1}_2 - i\boldsymbol{\xi}^\dagger) = \mathbf{1}_2 \tag{12.19}$$

similarly reduces (in first order) to the condition

$$\boldsymbol{\xi} = \boldsymbol{\xi}^\dagger. \tag{12.20}$$

Thus $\boldsymbol{\xi}$ is a 2×2 traceless Hermitian matrix, which means it must have the form

$$\boldsymbol{\xi} = \begin{pmatrix} a & b - \mathrm{i}c \\ b + \mathrm{i}c & -a \end{pmatrix}, \tag{12.21}$$

where a, b, c are infinitesimal real parameters. Writing

$$a = \epsilon_3/2, \quad b = \epsilon_1/2, \quad c = \epsilon_2/2, \tag{12.22}$$

(12.21) can be put in the more suggestive form

$$\boldsymbol{\xi} = \boldsymbol{\epsilon} \cdot \boldsymbol{\tau}/2 \tag{12.23}$$

where $\boldsymbol{\epsilon}$ stands for the three real quantities

$$\boldsymbol{\epsilon} = (\epsilon_1, \epsilon_2, \epsilon_3) \tag{12.24}$$

which are all first-order small. The three matrices $\boldsymbol{\tau}$ are just the familiar Hermitian Pauli matrices

$$\tau_1 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \tau_2 = \begin{pmatrix} 0 & -\mathrm{i} \\ \mathrm{i} & 0 \end{pmatrix}, \tau_3 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, \tag{12.25}$$

here called 'tau' precisely in order to distinguish them from the mathematically identical 'sigma' matrices which are associated with the real spin degree of freedom. Hence a general infinitesimal SU(2) matrix takes the form

$$\mathbf{V}_\mathrm{infl} = (\mathbf{1}_2 + \mathrm{i}\boldsymbol{\epsilon} \cdot \boldsymbol{\tau}/2), \tag{12.26}$$

and an infinitesimal SU(2) transformation of the p-n doublet is specified by

$$\begin{pmatrix} \psi'_\mathrm{p} \\ \psi'_\mathrm{n} \end{pmatrix} = (\mathbf{1}_2 + \mathrm{i}\boldsymbol{\epsilon} \cdot \boldsymbol{\tau}/2) \begin{pmatrix} \psi_\mathrm{p} \\ \psi_\mathrm{n} \end{pmatrix}. \tag{12.27}$$

The $\boldsymbol{\tau}$-matrices clearly play an important role, since they determine the forms of the three independent infinitesimal SU(2) transformations. They are called the *generators* of infinitesimal SU(2) transformations; more precisely, the matrices $\boldsymbol{\tau}/2$ provide a particular *matrix representation* of the generators, namely the two-dimensional, or 'fundamental' one (see appendix M). We note that they do not commute amongst themselves: rather, introducing $\mathbf{T}^{(\frac{1}{2})} \equiv \boldsymbol{\tau}/2$, we find (see problem 12.3)

$$[T_i^{(\frac{1}{2})}, T_j^{(\frac{1}{2})}] = \mathrm{i}\epsilon_{ijk} T_k^{(\frac{1}{2})}, \tag{12.28}$$

where i, j and k run from 1 to 3, and a sum on the repeated index k is understood as usual. The reader will recognize the commutation relations (12.28) as being precisely the same as those of angular momentum operators in quantum mechanics:

$$[J_i, J_j] = \mathrm{i}\epsilon_{ijk} J_k. \tag{12.29}$$

In that case, the choice $J_i = \sigma_i/2 \equiv J_i^{(\frac{1}{2})}$ would correspond to a (real) spin-1/2 system. Here the identity between the tau's and the sigma's gives us a good reason to regard our 'p-n' system as formally analogous to a 'spin-1/2' one. Of course, the 'analogy' was made into a mathematical identity by the judicious way in which $\boldsymbol{\xi}$ was parametrised in (12.23).

The form for a *finite* SU(2) transformation $\mathbf{V}$ may then be obtained from the infinitesimal form using the result

$$\mathrm{e}^A = \lim_{n\to\infty} (1 + A/n)^n \tag{12.30}$$

generalized to matrices. Let $\boldsymbol{\epsilon} = \boldsymbol{\alpha}/n$, where $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \alpha_3)$ are three real finite (not infinitesimal) parameters, apply the infinitesimal transformation n times, and let n tend to infinity. We obtain

$$\mathbf{V} = \exp(\mathrm{i}\boldsymbol{\alpha} \cdot \boldsymbol{\tau}/2) \tag{12.31}$$

so that

$$\psi^{(1/2)'} \equiv \begin{pmatrix} \psi'_{\rm p} \\ \psi'_{\rm n} \end{pmatrix} = \exp(\mathrm{i}\boldsymbol{\alpha}\cdot\boldsymbol{\tau}/2) \begin{pmatrix} \psi_{\rm p} \\ \psi_{\rm n} \end{pmatrix} = \exp(\mathrm{i}\boldsymbol{\alpha}\cdot\boldsymbol{\tau}/2)\psi^{(1/2)}. \quad (12.32)$$

Note that in the finite transformation, the generators appear in the exponent. Indeed, (12.31) has the form

$$\mathbf{V} = \exp(\mathrm{i}G) \quad (12.33)$$

where $G = \boldsymbol{\alpha}\cdot\boldsymbol{\tau}/2$, from which the unitary property of $\mathbf{V}$ easily follows:

$$\mathbf{V}^\dagger = \exp(-\mathrm{i}G^\dagger) = \exp(-\mathrm{i}G) = \mathbf{V}^{-1} \quad (12.34)$$

where we used the Hermiticity of the tau's. Equation (12.33) has the general form

$$\text{unitary matrix} = \exp(\text{i Hermitian matrix}) \quad (12.35)$$

where the 'Hermitian matrix' is composed of the generators and the transformation parameters. We shall meet generalizations of this structure in the following sub-section for SU(2), again in section 12.2 for SU(3), and a field theoretic version of it in section 12.3.

As promised, (12.32) has essentially the same mathematical form as (4.33). In each case, three real parameters appear. In (4.33) they describe the axis and angle of a physical rotation in real three-dimensional space: we can always write $\boldsymbol{\alpha} = |\boldsymbol{\alpha}|\hat{\boldsymbol{\alpha}}$ and identify $|\boldsymbol{\alpha}|$ with the angle θ and $\hat{\boldsymbol{\alpha}}$ with the axis $\hat{\boldsymbol{n}}$ of the rotation. In (12.32) there are just the three parameters in $\boldsymbol{\alpha}$.[2]

In the form (12.32), it is clear that our 2×2 isospin transformation is a generalization of the global phase transformation of (12.1), except that:

(i) there are now *three* 'phase angles' $\boldsymbol{\alpha}$;

(ii) there are non-commuting matrix operators (the $\boldsymbol{\tau}$'s) appearing in the exponent.

The last fact is the reason for the description 'non-Abelian' phase invariance. As the commutation relations for the $\boldsymbol{\tau}$ matrices show, SU(2) is a non-Abelian group in that two SU(2) transformations do not in general commute. By contrast, in the case of electric charge or particle number, successive transformations clearly commute: this corresponds to an Abelian phase invariance and, as noted in section 2.6, to an Abelian U(1) group.

We may now put our initial 'spin-1/2' analogy on a more precise mathematical footing. In quantum mechanics, states within a degenerate multiplet may conveniently be characterized by the eigenvalues of a complete set of Hermitian operators which commute with the Hamiltonian and with each other.

[2]It is not obvious that the general SU(2) matrix can be parametrized by an angle θ with $0 \le \theta \le 2\pi$, and $\hat{\boldsymbol{n}}$: for further discussion of the relation between SU(2) and the three-dimensional rotation group, see appendix M, section M.7.

In the case of the p-n doublet, it is easy to see what these operators are. We may write (12.4), (12.6) and (12.7) as

$$H_2\psi^{(1/2)} = E\psi^{(1/2)} \tag{12.36}$$

and

$$H_2\psi^{(1/2)'} = E\psi^{(1/2)'}, \tag{12.37}$$

where H_2 is the 2×2 matrix

$$H_2 = \begin{pmatrix} H & 0 \\ 0 & H \end{pmatrix}. \tag{12.38}$$

Hence H_2 is proportional to the unit matrix in this two-dimensional space, and it therefore commutes with the tau's:

$$[H_2, \boldsymbol{\tau}] = 0. \tag{12.39}$$

It then also follows that H_2 commutes with $\mathbf{V}$, or equivalently

$$\mathbf{V}H_2\mathbf{V}^{-1} = H_2 \tag{12.40}$$

which is the statement that H_2 is invariant under the transformation (12.32). Now the tau's are Hermitian, and hence correspond to possible observables. Equation (12.39) implies that their eigenvalues are constants of the motion (i.e. conserved quantities), associated with the invariance (12.40). But the tau's do not commute amongst themselves and so according to the general principles of quantum mechanics we cannot give definite values to more than one of them at a time. The problem of finding a classification of the states which makes the maximum use of (12.39), given the commutation relations (12.28), is easily solved by making use of the formal identity between the operators $\tau_i/2$ and angular momentum operators J_i (cf (12.29)). The answer is[3] that the total squared 'spin'

$$(\mathbf{T}^{(1/2)})^2 = \left(\frac{1}{2}\boldsymbol{\tau}\right)^2 = \frac{1}{4}(\tau_1^2 + \tau_2^2 + \tau_3^2) = \frac{3}{4}\mathbf{1}_2 \tag{12.41}$$

and one component of spin, say $T_3^{(1/2)} = \frac{1}{2}\tau_3$, can be given definite values simultaneously. The corresponding eigenfunctions are just the χ_p's and χ_n's of (12.9), which satisfy

$$\frac{1}{4}\boldsymbol{\tau}^2\chi_\mathrm{p} = \frac{3}{4}\chi_\mathrm{p}, \qquad \frac{1}{2}\boldsymbol{\tau}_3\chi_\mathrm{p} = \frac{1}{2}\chi_\mathrm{p} \tag{12.42}$$

$$\frac{1}{4}\boldsymbol{\tau}^2\chi_\mathrm{n} = \frac{3}{4}\chi_\mathrm{n}, \qquad \frac{1}{2}\boldsymbol{\tau}_3\chi_\mathrm{n} = -\frac{1}{2}\chi_\mathrm{n}. \tag{12.43}$$

The reason for the 'spin' part of the name 'isospin' should by now be clear; the term is actually a shortened version of the historical one 'isotopic spin'.

[3]See for example Mandl (1992).

In concluding this section we remark that, in this two-dimensional n-p space, the electromagnetric charge operator is represented by the matrix

$$\mathbf{Q}_{\rm em} = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} = \frac{1}{2}(\mathbf{1}_2 + \tau_3). \tag{12.44}$$

It is clear that although $\mathbf{Q}_{\rm em}$ commutes with τ_3, it does not commute with either τ_1 or τ_2. Thus, as we would expect, electromagnetic corrections to the strong interaction Hamiltonian will violate SU(2) symmetry.

12.2.2 Larger (higher-dimensional) multiplets of SU(2) in nuclear physics

For the single nucleon states considered so far, the foregoing is really nothing more than the general quantum mechanics of a two-state system, phrased in 'spin-1/2' language. The real power of the isospin (SU(2)) symmetry concept becomes more apparent when we consider states of *several* nucleons. For A nucleons in the nucleus, we introduce three 'total isospin operators' $\mathbf{T} = (T_1, T_2, T_3)$ via

$$\mathbf{T} = \frac{1}{2}\boldsymbol{\tau}_{(1)} + \frac{1}{2}\boldsymbol{\tau}_{(2)} + \ldots + \frac{1}{2}\boldsymbol{\tau}_{(A)}, \tag{12.45}$$

which are Hermitian. Here $\boldsymbol{\tau}_{(n)}$ is the $\boldsymbol{\tau}$-matrix for the nth nucleon. The Hamiltonian H describing the strong interactions of this system is presumed to be invariant under the transformation (12.40) for all the nucleons independently. It then follows that

$$[H, \mathbf{T}] = 0. \tag{12.46}$$

Thus the eigenvalues of the $\mathbf{T}$ operators are constants of the motion. Further, since the isospin operators for different nucleons commute with each other (they are quite independent), the commutation relations (12.28) for each of the individual $\boldsymbol{\tau}$'s imply (see problem 12.4) that the components of $\mathbf{T}$ defined by (12.45) satisfy the commutation relations

$$[T_i, T_j] = \mathrm{i}\epsilon_{ijk}T_k \tag{12.47}$$

for $i, j, k = 1, 2, 3$, which are simply the standard angular momentum commutation relations, once more. Thus the energy levels of nuclei ought to be characterized – after allowance for electromagnetic effects, and correcting for the slight neutron-proton mass difference – by the eigenvalues of $\mathbf{T}^2$ and T_3, say, which can be simultaneously diagonalized along with H. These eigenvalues should then be, to a good approximation, 'good quantum numbers' for nuclei, if the assumed isospin invariance is true.

What are the possible eigenvalues? We know that the $\mathbf{T}$'s are Hermitian and satisfy exactly the same commutation relations (12.47) as the angular momentum operators. These conditions are all that are needed to show that the eigenvalues of $\mathbf{T}^2$ are of the form $T(T+1)$, where $T = 0, \frac{1}{2}, 1, \ldots$, and that

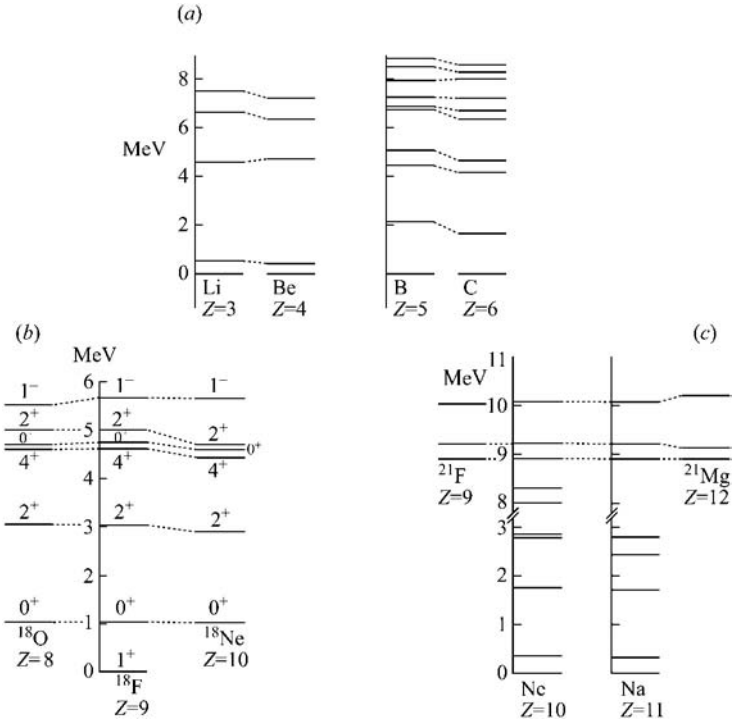


FIGURE 12.2
Energy levels (adjusted for Coulomb energy and neutron-proton mass differences) of nuclei of the same mass number but different charge, showing (a) 'mirror' doublets, (b) triplets and (c) doublets and quartets.

for a given T the eigenvalues of T_3 are $-T, -T+1, \ldots, T-1, T$; that is, there are $2T+1$ *degenerate states* for a given T. These states all have the same A value, and since T_3 counts $+\frac{1}{2}$ for every proton and $-\frac{1}{2}$ for every neutron, it is clear that successive values of T_3 correspond physically to changing one neutron into a proton or vice versa. Thus we expect to see 'charge multiplets' of levels in neighbouring nuclear isobars. These are indeed observed; figure 12.2 shows some examples. These level schemes (which have been adjusted for Coulomb energy differences, and for the neutron-proton mass difference), provide clear evidence of $T = \frac{1}{2}$ (doublet), $T = 1$ (triplet) and $T = \frac{3}{2}$ (quartet) multiplets. It is important to note that states in the same T-multiplet must have the same J^P quantum numbers (these are indicated on the levels for ^{18}F); obviously the nuclear forces will depend on the space and spin degrees of freedom of the nucleons, and will only be the same between different nucleons if the space-spin part of the wavefunction is the same.

Thus the assumed invariance of the nucleon-nucleon force produces a richer nuclear multiplet structure, going beyond the original n-p doublet. These higher-dimensional multiplets ($T = 1, \frac{3}{2}, \ldots$) are called 'irreducible representations' of SU(2). The commutation relations (12.47) are called the *Lie algebra* of SU(2)[4] (see appendix M), and the general group theoretical problem of understanding all *possible* multiplets for SU(2) is equivalent to the problem of finding matrices which satisfy these commutation relations. These are, in fact, precisely the angular momentum matrices of dimension $(2T+1) \times (2T+1)$ which are generalizations of the $\boldsymbol{\tau}/2$'s, which themselves correspond to $T = \frac{1}{2}$, as indicated in the notation $\mathbf{T}^{(\frac{1}{2})}$. For example, the $T = 1$ matrices are 3×3 and can be compactly summarised by (problem 12.5)

$$(T_i^{(1)})_{jk} = -\mathrm{i}\epsilon_{ijk} \tag{12.48}$$

where the numbers $-\mathrm{i}\epsilon_{ijk}$ are deliberately chosen to be the *same* numbers (with a minus sign) that specify the algebra in (12.47); the latter are called the *structure constants* of the SU(2) group (see appendix M, sections M.3–M.5). In general there will be matrices $\mathbf{T}^{(T)}$ of dimensionality $(2T+1)\times(2T+1)$ which satisfy (12.47), and correspondingly $(2T+1)$-dimensional wavefunctions $\psi^{(T)}$ analogous to the two-dimensional ($T = \frac{1}{2}$) case of (12.8). The generalization of (12.32) to these higher-dimensional multiplets is then

$$\psi^{(T)\prime} = \exp(\mathrm{i}\boldsymbol{\alpha} \cdot \mathbf{T}^{(T)})\psi^{(T)}, \tag{12.49}$$

which has the general form of (12.35). In this case, the matrices $\mathbf{T}^{(T)}$ provide a $(2T+1)$-dimensional matrix representation of the generators of SU(2). We shall meet field-theoretic representations of the generators in section 12.3.

We now proceed to consider isospin in our primary area of interest, which is particle physics.

12.2.3 Isospin in particle physics: flavour $\mathbf{SU(2)_f}$

The neutron and proton states themselves are actually only the ground states of a whole series of corresponding $B = 1$ levels with isospin $\frac{1}{2}$ (i.e. doublets). Another series of baryonic levels comes in *four* charge states, corresponding to $T = \frac{3}{2}$; and in the meson sector, the π's appear as the lowest states of a sequence of mesonic triplets ($T = 1$). Many other examples also exist, but with one remarkable difference as compared to the nuclear physics case: no baryon states are known with $T > \frac{3}{2}$, nor any meson states with $T > 1$.

The most natural interpretation of these facts is that the observed states are composites of more basic entities which carry different charges but are nearly degenerate in mass, while the forces between these entities are charge-independent, just as in the nuclear (p,n) case. These entities are, of course,

[4] Likewise, the angular momentum commutation relations (12.29) are the Lie algebra of the rotation group SO(3). The Lie algebras of the two groups are therefore the same. For an indication of how, nevertheless, the groups do differ, see appendix M, section M.7.

the quarks: the n contains (udd), the p is (uud), and the Δ-quartet is (uuu, uud, udd, ddd). The u-d isospin doublet plays the role of the p-n doublet in the nuclear case, and this degree of freedom is what we now call SU(2) isospin flavour symmetry at the quark level, denoted by $\mathrm{SU}(2)_\mathrm{f}$. We shall denote the u-d quark doublet wavefunction by

$$q = \begin{pmatrix} u \\ d \end{pmatrix} \tag{12.50}$$

omitting now the explicit representation label '$(\frac{1}{2})$', and shortening 'ψ_u' to just 'u', and similarly for 'd'. Then, under an $\mathrm{SU}(2)_\mathrm{f}$ transformation,

$$q \to q' = \mathbf{V}q = \exp(\mathrm{i}\boldsymbol{\alpha}\cdot\boldsymbol{\tau}/2)\, q. \tag{12.51}$$

The limitation $T \leq \frac{3}{2}$ for baryonic states can be understood in terms of their being composed of three $T = \frac{1}{2}$ constituents (two of them pair to $T = 1$ or $T = 0$, and the third adds to $T = 1$ to make $T = \frac{3}{2}$ or $T = \frac{1}{2}$, and to $T = 0$ to make $T = \frac{1}{2}$, by the usual angular momentum addition rules). It is, however, a challenge for QCD to explain why, for example, states with four or five quarks should not exist (nor states of one or two quarks!), and why a state of six quarks, for example, appears as the deuteron, which is a loosely bound state of n and p, rather than as a compact $B = 2$ analogue of the n and p themselves.

Meson states such as the pion are formed from a quark and an antiquark, and it is therefore appropriate at this point to explain how *antiparticles* are described in isospin terms. An antiparticle is characterized by having the signs of all its additively conserved quantum numbers reversed, relative to those of the corresponding particle. Thus if a u-quark has $B = \frac{1}{3}, T = \frac{1}{2}, T_3 = \frac{1}{2}$, a $\bar{\mathrm{u}}$-quark has $B = -\frac{1}{3}, T = \frac{1}{2}, T_3 = -\frac{1}{2}$. Similarly, the $\bar{\mathrm{d}}$ has $B = -\frac{1}{3}, T = \frac{1}{2}$ and $T_3 = \frac{1}{2}$. Note that, while T_3 is an additively conserved quantum number, the magnitude of the isospin is not additively conserved: rather, it is 'vectorially' conserved according to the rules of combining angular-momentum-like quantum numbers, as we have seen. Thus the antiquarks $\bar{\mathrm{d}}$ and $\bar{\mathrm{u}}$ form the $T_3 = +\frac{1}{2}$ and $T_3 = -\frac{1}{2}$ members of an $\mathrm{SU}(2)_\mathrm{f}$ doublet, just as u and d themselves do, and the question arises: given that the (u, d) doublet transforms as in (12.51), how does the $(\bar{u}, \bar{d})$ doublet transform?

The answer is that antiparticles are assigned to the *complex conjugate* of the representation to which the corresponding particles belong. Thus identifying $\bar{u} \equiv u^*$ and $\bar{d} \equiv d^*$ we have[5]

$$q^{*\prime} = \mathbf{V}^* q^*, \quad \text{or} \begin{pmatrix} \bar{u} \\ \bar{d} \end{pmatrix}' = \exp(-\mathrm{i}\boldsymbol{\alpha}\cdot\boldsymbol{\tau}^*/2) \begin{pmatrix} \bar{u} \\ \bar{d} \end{pmatrix} \tag{12.52}$$

for the $\mathrm{SU}(2)_\mathrm{f}$ transformation law of the antiquark doublet. In mathematical terms, this means (compare (12.32)) that the three matrices $-\frac{1}{2}\boldsymbol{\tau}^*$ must

[5]The overbar ($\bar{u}$ etc.) here stands only for 'antiparticle', and has nothing to do with the Dirac conjugate $\bar{\psi}$ introduced in section 4.4.

represent the generators of $\mathrm{SU}(2)_\mathrm{f}$ in the $\mathbf{2}^*$ representation (i.e. the complex conjugate of the original two-dimensional representation, which we will now call $\mathbf{2}$). Referring to (12.25), we see that $\tau_1^* = \tau_1, \tau_2^* = -\tau_2$ and $\tau_3^* = \tau_3$. It is then easy to check that the three matrices $-\tau_1/2, +\tau_2/2$ and $-\tau_3/2$ do indeed satisfy the required commutation relations (12.28), and thus provide a valid matrix representation of the SU(2) generators. Also, since the third component of isospin is here represented by $-\tau_3^*/2 = -\tau_3/2$, the desired reversal in sign of the additively conserved eigenvalue does occur.

Although the quark doublet (u, d) and antiquark doublet $(\bar{u}, \bar{d})$ do transform differently under $\mathrm{SU}(2)_\mathrm{f}$ transformations, there is nevertheless a sense in which the $\mathbf{2}^*$ and $\mathbf{2}$ representations are somehow the 'same': after all, the quantum numbers $T = \frac{1}{2}, T_3 = \pm\frac{1}{2}$ describe them both. In fact, the two representations are 'unitarily equivalent', in that we can find a unitary matrix U_C such that

$$U_C \exp(-\mathrm{i}\boldsymbol{\alpha}\cdot\boldsymbol{\tau}^*/2)U_C^{-1} = \exp(\mathrm{i}\boldsymbol{\alpha}\cdot\boldsymbol{\tau}/2). \tag{12.53}$$

This requirement is easier to disentangle if we consider infinitesimal transformations, for which (12.53) becomes

$$U_C(-\boldsymbol{\tau}^*)U_C^{-1} = \boldsymbol{\tau}, \tag{12.54}$$

or

$$U_C\tau_1U_C^{-1} = -\tau_1, \quad U_C\tau_2U_C^{-1} = \tau_2, \quad U_C\tau_3U_C^{-1} = -\tau_3. \tag{12.55}$$

Bearing the commutation relations (12.28) in mind, and the fact that $\tau_i^{-1} = \tau_i$, it is clear that we can choose U_C proportional to τ_2, and set

$$U_C = \mathrm{i}\tau_2 = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \tag{12.56}$$

to obtain a convenient unitary form. From (12.52) and (12.53) we obtain $(U_C q^{*\prime}) = \mathbf{V}(U_C q^*)$, which implies that the doublet

$$U_C \begin{pmatrix} \bar{u} \\ \bar{d} \end{pmatrix} = \begin{pmatrix} \bar{d} \\ -\bar{u} \end{pmatrix} \tag{12.57}$$

transforms in exactly the same way as (u, d). This result is useful, because it means that we can use the familiar tables of (Clebsch-Gordan) angular momentum coupling coefficients for combining quark and antiquark states together, *provided* we include the relative minus sign between the $\bar{d}$ and $\bar{u}$ components which has appeared in (12.57). Note that, as expected, the $\bar{d}$ is in the $T_3 = +\frac{1}{2}$ position, and the $\bar{u}$ is in the $T_3 = -\frac{1}{2}$ position.

As an application of these results, let us compare the $T = 0$ combination of the p and n states to form the (isoscalar) deuteron, and the combination of (u, d) and $(\bar{u}, \bar{d})$ states to form the isoscalar ω-meson. In the first, the isospin part of the wavefunction is $\frac{1}{\sqrt{2}}(\psi_\mathrm{p}\psi_\mathrm{n} - \psi_\mathrm{n}\psi_\mathrm{p})$, corresponding to the $S = 0$ combination of two spin-$\frac{1}{2}$ particles in quantum mechanics given by

$\frac{1}{\sqrt{2}}(|\uparrow\rangle\,|\downarrow\rangle - |\downarrow\rangle\,|\uparrow\rangle)$. But in the second case the corresponding wavefunction is $\frac{1}{\sqrt{2}}(\bar{d}d - (-\bar{u})u) = \frac{1}{\sqrt{2}}(\bar{d}d + \bar{u}u)$. Similarly, the $T = 1\ T_3 = 0$ state describing the π^0 is $\frac{1}{\sqrt{2}}(\bar{d}d + (-\bar{u})u) = \frac{1}{\sqrt{2}}(\bar{d}d - \bar{u}u)$.

There is a very convenient alternative way of obtaining these wavefunctions, which we include here because it generalizes straightforwardly to SU(3); its advantage is that it avoids the use of the explicit C-G coupling coefficients, and of their (more complicated) analogues in SU(3).

Bearing in mind the identifications $\bar{u} \equiv u^*, \bar{d} \equiv d^*$, we see that the $T = 0$ $\bar{q}q$ combination $\bar{u}u + \bar{d}d$ can be written as $u^*u + d^*d$ which is just $q^\dagger q$, (recall that † means transpose and complex conjugate). Under an $\mathrm{SU}(2)_\mathrm{f}$ transformation, $q \to q' = \mathbf{V}q$, so $q^\dagger \to q'^\dagger = q^\dagger \mathbf{V}^\dagger$ and

$$q^\dagger q \to q'^\dagger q' = q^\dagger \mathbf{V}^\dagger \mathbf{V} q = q^\dagger q \tag{12.58}$$

using $\mathbf{V}^\dagger \mathbf{V} = \mathbf{1}_2$; thus $q^\dagger q$ is indeed an $\mathrm{SU}(2)_\mathrm{f}$ invariant, which means it has $T = 0$ (no multiplet partners).

We may also construct the $T = 1$ $q - \bar{q}$ states in a similar way. Consider the three quantities v_i defined by

$$v_i = q^\dagger \tau_i q \quad i = 1, 2, 3. \tag{12.59}$$

Under an infinitesimal $\mathrm{SU}(2)_\mathrm{f}$ transformation

$$q' = (\mathbf{1}_2 + \mathrm{i}\boldsymbol{\epsilon} \cdot \boldsymbol{\tau}/2)q, \tag{12.60}$$

the three quantities v_i transform to

$$v_i' = q^\dagger(\mathbf{1}_2 - \mathrm{i}\boldsymbol{\epsilon} \cdot \boldsymbol{\tau}/2)\tau_i(\mathbf{1}_2 + \mathrm{i}\boldsymbol{\epsilon} \cdot \boldsymbol{\tau}/2)q, \tag{12.61}$$

where we have used $q'^\dagger = q^\dagger(\mathbf{1}_2 + \mathrm{i}\boldsymbol{\epsilon} \cdot \boldsymbol{\tau}/2)^\dagger$ and then $\boldsymbol{\tau}^\dagger = \boldsymbol{\tau}$. Retaining only the first-order terms in $\boldsymbol{\epsilon}$ gives (problem 12.6)

$$v_i' = v_i + \mathrm{i}\frac{\epsilon_j}{2} q^\dagger(\tau_i\tau_j - \tau_j\tau_i)q \tag{12.62}$$

where the sum on $j = 1, 2, 3$ is understood. But from (12.28) we know the commutator of two τ's, so that (12.62) becomes

$$\begin{aligned} v_i' &= v_i + \mathrm{i}\frac{\epsilon_j}{2} q^\dagger . 2\mathrm{i}\epsilon_{ijk}\tau_k q \quad (\text{sum on } k = 1, 2, 3) \\ &= v_i - \epsilon_{ijk}\epsilon_j q^\dagger \tau_k q \\ &= v_i - \epsilon_{ijk}\epsilon_j v_k, \end{aligned} \tag{12.63}$$

which may also be written in 'vector' notation as

$$\boldsymbol{v}' = \boldsymbol{v} - \boldsymbol{\epsilon} \times \boldsymbol{v}. \tag{12.64}$$

Equation (12.63) states that, under an (infinitesimal) $\mathrm{SU}(2)_\mathrm{f}$ transformation, the three quantities v_i $(i = 1, 2, 3)$ transform into *specific linear combinations* of themselves, as determined by the coefficients ϵ_{ijk} (the ϵ's are just

the parameters of the infinitesimal transformation). This is precisely what is needed for a set of quantities to *form the basis for a representation*. In this case, it is the $T = 1$ representation as we can guess from the multiplicity of three, but we can also directly verify it, as follows. Equation (12.49) with $T = 1$, together with (12.48), tell us how a $T = 1$ triplet should transform: namely, under an infinitesimal transformation (with $\mathbf{1}_3$ the unit 3×3 matrix),

$$\begin{aligned}
\psi_i^{(1)\prime} &= (\mathbf{1}_3 + \mathrm{i}\boldsymbol{\epsilon} \cdot \mathbf{T}^{(1)})_{ik}\psi_k^{(1)} \quad \text{(sum on } k = 1,2,3) \\
&= (\mathbf{1}_3 + \mathrm{i}\epsilon_j T_j^{(1)})_{ik}\psi_k^{(1)} \quad \text{(sum on } j = 1,2,3) \\
&= (\delta_{ik} + \mathrm{i}\epsilon_j (T_j^{(1)})_{ik})\psi_k^{(1)} \\
&= (\delta_{ik} + \mathrm{i}\epsilon_j . - \mathrm{i}\epsilon_{jik})\psi_k^{(1)} \quad \text{using (12.48)} \\
&= \psi_i^{(1)} - \epsilon_{ijk}\epsilon_j \psi_k^{(1)} \quad \text{using the antisymmetry of } \epsilon_{ijk} \qquad (12.65)
\end{aligned}$$

which is exactly the same as (12.63).

The reader who has worked through problem 4.2(a) will recognize the exact analogy between the $T = 1$ transformation law (12.64) for the isospin bilinear $q^\dagger \boldsymbol{\tau} q$, and the 3-vector transformation law (cf (4.9)) for the Pauli spinor bilinear $\phi^\dagger \boldsymbol{\sigma} \phi$.

Returning to the physics of v_i, inserting (12.50) into (12.59) we find explicitly

$$v_1 = \bar{u}d + \bar{d}u, \; v_2 = -\mathrm{i}\,\bar{u}d + \mathrm{i}\,\bar{d}u, \; v_3 = \bar{u}u - \bar{d}d. \qquad (12.66)$$

Apart from the normalization factor of $\frac{1}{\sqrt{2}}$, v_3 may therefore be identified with the $T_3 = 0$ member of the $T = 1$ triplet, having the quantum numbers of the π^0. Neither v_1 nor v_2 has a definite value of T_3, however: rather, we need to consider the linear combinations

$$\frac{1}{2}(v_1 + \mathrm{i}v_2) = \bar{u}d \quad T_3 = -1 \qquad (12.67)$$

and

$$\frac{1}{2}(v_1 - \mathrm{i}v_2) = \bar{d}u \quad T_3 = +1 \qquad (12.68)$$

which have the quantum numbers of the π^- and π^+. The use of $v_1 \pm \mathrm{i}v_2$ here is precisely analogous to the use of the 'spherical basis' wavefunctions $x \pm \mathrm{i}y = r \sin\theta \mathrm{e}^{\pm \mathrm{i}\phi}$ for $\ell = 1$ states in quantum mechanics, rather than the 'Cartesian' ones x and y.

We are now ready to proceed to SU(3).

12.3 Flavour SU(3)$_\mathrm{f}$

Larger hadronic multiplets also exist, in which strange particles are grouped with non-strange ones. Gell-Mann (1961) and Ne'eman (1961) (see also Gell-Mann and Ne'eman 1964) were the first to propose SU(3)$_\mathrm{f}$ as the correct

generalization of isospin $\mathrm{SU(2)_f}$ to include strangeness. Like SU(2), SU(3) is a group whose elements are matrices – in this case, unitary 3×3 ones, of unit determinant. The general group-theoretic analysis of SU(3) is quite complicated, but is fortunately not necessary for the physical applications we require. We can, in fact, develop all the results needed by mimicking the steps followed for SU(2).

We start by finding the general form of an SU(3) matrix. Such matrices obviously act on 3-component column vectors, the generalization of the 2-component isospinors of SU(2). In more physical terms, we regard the three quark wavefunctions u, d and s as being approximately degenerate, and we consider unitary 3×3 transformations among them via

$$q' = \mathbf{W} q \tag{12.69}$$

where q now stands for the 3-component column vector

$$q = \begin{pmatrix} u \\ d \\ s \end{pmatrix} \tag{12.70}$$

and $\mathbf{W}$ is a 3×3 unitary matrix of determinant 1 (again, an overall phase has been extracted). The representation provided by this triplet of states is called the 'fundamental' representation of $\mathrm{SU(3)_f}$ (just as the isospinor representation is the fundamental one of $\mathrm{SU(2)_f}$).

To determine the general form of an SU(3) matrix $\mathbf{W}$, we follow exactly the same steps as in the SU(2) case. An infinitesimal SU(3) matrix has the form

$$\mathbf{W}_{\mathrm{infl}} = \mathbf{1}_3 + \mathrm{i}\boldsymbol{\chi} \tag{12.71}$$

where $\boldsymbol{\chi}$ is a 3×3 traceless Hermitian matrix. Such a matrix involves *eight* independent parameters (problem (12.7)) and can be written as

$$\boldsymbol{\chi} = \boldsymbol{\eta} \cdot \boldsymbol{\lambda}/2 \tag{12.72}$$

where $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_8)$ and the $\boldsymbol{\lambda}$'s are eight matrices generalizing the $\boldsymbol{\tau}$ matrices of (12.25). They are the generators of SU(3) in the three-dimensional fundamental representation, and their commutation relations define the *algebra of SU(3)* (compare (12.28) for SU(2)):

$$[\lambda_a/2, \lambda_b/2] = \mathrm{i} f_{abc} \lambda_c/2, \tag{12.73}$$

where a, b and c run from 1 to 8.

The λ-matrices (often called the *Gell-Mann matrices*), are given in appendix M, along with the *SU(3) structure constants* $\mathrm{i}f_{abc}$; the constants f_{abc} are all real.

A finite SU(3) transformation on the quark triplet is then (cf (12.32))

$$q' = \exp(\mathrm{i}\boldsymbol{\alpha} \cdot \boldsymbol{\lambda}/2) q, \tag{12.74}$$

which also has the 'generalized phase transformation' character of (12.35), now with *eight* 'phase angles'. Thus $\mathbf{W}$ is parametrized as $\mathbf{W} = \exp(\mathrm{i}\boldsymbol{\alpha} \cdot \boldsymbol{\lambda}/2)$.

As in the case of $\mathrm{SU}(2)_\mathrm{f}$, exact symmetry under $\mathrm{SU}(3)_\mathrm{f}$ would imply that the three states u, d and s were degenerate in mass. Actually, of course, this is not the case: in particular, while the u and d quark masses are of order 1-5 MeV, the s quark mass is greater, of order 100 MeV. Nevertheless it is still possible to regard this as relatively small on a typical hadronic mass scale, so we may proceed to explore the physical consequences of this (approximate) $\mathrm{SU}(3)_\mathrm{f}$ flavour symmetry.

Such a symmetry implies that the eigenvalues of the $\boldsymbol{\lambda}$'s are constants of the motion, but because of the commutation relations (12.73) not all of these operators have simultaneous eigenstates. This happened for SU(2) too, but there the very close analogy with SO(3) told us how the states were to be correctly classified, by the eigenvalues of the relevant complete set of mutually commuting operators. Here it is more involved – for a start, there are 8 matrices λ_a. A glance at appendix M, section M.4.5, shows that *two* of the λ's are diagonal (in the chosen representation), namely λ_3 and λ_8. This means physically that for SU(3) there are *two* additively conserved quantum numbers, which in this case are of course the third component of hadronic isospin (since λ_3 is simply τ_3 bordered by zeros), and a quantity related to strangeness. Defining the hadronic hypercharge Y by $Y = B + S$, where B is the baryon number ($\frac{1}{3}$ for each quark) and the strangeness values are $S(\mathrm{u}) = S(\mathrm{d}) = 0$, $S(\mathrm{s}) = -1$, we find that the physically required eigenvalues imply that the matrix representing the hypercharge operator is $Y^{(\mathbf{3})} = \frac{1}{\sqrt{3}}\lambda_8$, in this fundamental (three-dimensional) representation, denoted by the symbol $\mathbf{3}$. Identifying $T_3^{(\mathbf{3})} = \frac{1}{2}\lambda_3$ then gives the Gell-Mann–Nishijima relation $Q = T_3 + Y/2$ for the quark charges in units of $\mid e \mid$.

So λ_3 and λ_8 are analogous to τ_3; what about the analogue of $\boldsymbol{\tau}^2$, which is diagonalizable simultaneously with τ_3 in the case of SU(2)? Indeed, (cf (12.41)) $\boldsymbol{\tau}^2$ is a multiple of the 2×2 unit matrix. In just the same way one finds that $\boldsymbol{\lambda}^2$ is also proportional to the unit matrix:

$$(\boldsymbol{\lambda}/2)^2 = \sum_{a=1}^{8} (\lambda_a/2)^2 = \frac{4}{3}\mathbf{1}_3, \tag{12.75}$$

as can be verified from the explicit forms of the λ-matrices given in appendix M, section M.4.5. Thus we may characterize the 'fundamental triplet' (12.70) by the eigenvalues of $(\boldsymbol{\lambda}/2)^2$, λ_3 and λ_8. The conventional way of representing this pictorially is to plot the states in a $Y - T_3$ diagram, as shown in figure 12.3.

We may now consider other representations of $\mathrm{SU}(3)_\mathrm{f}$. The first important one is that to which the *antiquarks* belong. If we denote the fundamental three-dimensional representation accommodating the quarks by $\mathbf{3}$, then the antiquarks have quantum numbers appropriate to the 'complex conjugate' of

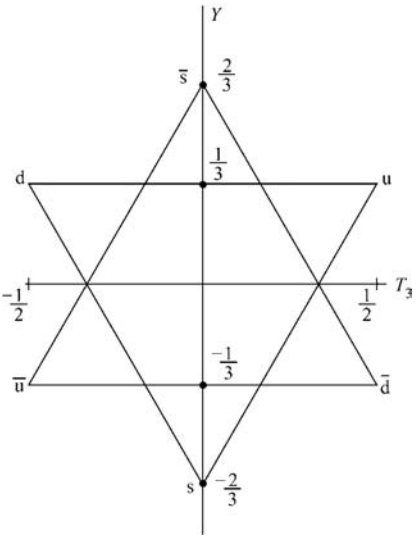


FIGURE 12.3
The $Y - T_3$ quantum numbers of the fundamental triplet **3** of quarks, and of the antitriplet $\mathbf{3}^*$ of antiquarks.

this representation, denoted by $\mathbf{3}^*$ just as in the SU(2) case. The $\bar{q}$ wavefunctions identified as $\bar{u} \equiv u^*, \bar{d} \equiv d^*$ and $\bar{s} \equiv s^*$, then transform by

$$\bar{q}' = \begin{pmatrix} \bar{u} \\ \bar{d} \\ \bar{s} \end{pmatrix}' = \mathbf{W}^* \bar{q} = \exp(-\mathrm{i}\boldsymbol{\alpha} \cdot \boldsymbol{\lambda}^*/2)\bar{q} \tag{12.76}$$

instead of by (12.74). As for the $\mathbf{2}^*$ representation of SU(2), (12.76) means that the eight quantities $-\boldsymbol{\lambda}^*/2$ represent the SU(3) generators in this $\mathbf{3}^*$ representation. Referring to appendix M, section M.4.5, one quickly sees that λ_3 and λ_8 are real, so that the eigenvalues of the physical observables $T_3^{(\mathbf{3}^*)} = -\lambda_3/2$ and $Y^{(\mathbf{3}^*)} = -\frac{1}{\sqrt{3}}\lambda_8/2$ (in this representation) are reversed relative to those in the **3**, as expected for antiparticles. The $\bar{\mathrm{u}}, \bar{\mathrm{d}}$ and $\bar{\mathrm{s}}$ states may also be plotted on the $Y - T_3$ diagram, figure 12.3, as shown.

Here is already one important difference between SU(3) and SU(2): the fundamental SU(3) representation **3** and its complex conjugate $\mathbf{3}^*$ are *not* equivalent. This follows immediately from figure 12.3, where it is clear that the extra quantum number Y distinguishes the two representations.

Larger $\mathrm{SU(3)_f}$ representations can be created by combining quarks and antiquarks, as in $\mathrm{SU(2)_f}$. For our present purposes, an important one is the eight-dimensional ('octet') representation which appears when one combines the $\mathbf{3}^*$ and **3** representations, in a way which is very analogous to the three-

dimensional ('triplet') representation obtained by combining the $\mathbf{2}^*$ and $\mathbf{2}$ representations of SU(2).

Consider first the quantity $\bar{u}u + \bar{d}d + \bar{s}s$. As in the SU(2) case, this can be written equivalently as $q^\dagger q$, which is invariant under $q \to q' = \mathbf{W}q$ since $\mathbf{W}^\dagger\mathbf{W} = \mathbf{1}_3$. So this combination is an SU(3) *singlet*. The *octet* coupling is formed by a straightforward generalization of the SU(2) triplet coupling $q^\dagger\boldsymbol{\tau}q$ of (12.59),

$$w_a = q^\dagger\lambda_a q \qquad a = 1, 2, \ldots 8. \tag{12.77}$$

Under an infinitesimal $\mathrm{SU(3)_f}$ transformation (compare (12.61) and (12.62)),

$$\begin{aligned} w_a \to w'_a &= q^\dagger(\mathbf{1}_3 - \mathrm{i}\boldsymbol{\eta}\cdot\boldsymbol{\lambda}/2)\lambda_a(\mathbf{1}_3 + \mathrm{i}\boldsymbol{\eta}\cdot\boldsymbol{\lambda}/2)q \\ &\approx q^\dagger\lambda_a q + \mathrm{i}\frac{\eta_b}{2}q^\dagger(\lambda_a\lambda_b - \lambda_b\lambda_a)q \end{aligned} \tag{12.78}$$

where the sum on $b = 1$ to 8 is understood. Using (12.73) for the commutator of two λ's we find

$$w'_a = w_a + \mathrm{i}\frac{\eta_b}{2}q^\dagger.2\mathrm{i}f_{abc}\lambda_c q \tag{12.79}$$

or

$$w'_a = w_a - f_{abc}\eta_b w_c \tag{12.80}$$

which may usefully be compared with (12.63). Just as in the $\mathrm{SU(2)_f}$ triplet case, equation (12.80) shows that, under an $\mathrm{SU(3)_f}$ transformation, the eight quantities $w_a (a = 1, 2, \ldots 8)$ transform into specific linear combinations of themselves, as determined by the coefficients f_{abc} (the η's are just the parameters of the infinitesimal transformation).

This is, again, precisely what is needed for a set of quantities to form the basis for a representation – in this case, an eight-dimensional representation of $\mathrm{SU(3)_f}$. For a finite $\mathrm{SU(3)_f}$ transformation, we can 'exponentiate' (12.80) to obtain

$$\boldsymbol{w}' = \exp(\mathrm{i}\boldsymbol{\alpha}\cdot\mathbf{G}^{(8)})\boldsymbol{w} \tag{12.81}$$

where $\boldsymbol{w}$ is an 8-component column vector

$$\boldsymbol{w} = \begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_8 \end{pmatrix} \tag{12.82}$$

such that $w_a = q^\dagger\lambda_a q$, and where (cf (12.49) for $\mathrm{SU(2)_f}$) the quantities $\mathbf{G}^{(8)} = (G_1^{(8)}, G_2^{(8)}, \ldots G_8^{(8)})$ are 8×8 matrices, acting on the 8-component vector $\boldsymbol{w}$, and forming an 8-dimensional representation of the algebra of SU(3): that is to say, the $\mathbf{G}^{(8)}$'s satisfy (cf (12.73))

$$\left[G_a^{(8)}, G_b^{(8)}\right] = \mathrm{i}f_{abc}G_c^{(8)}. \tag{12.83}$$

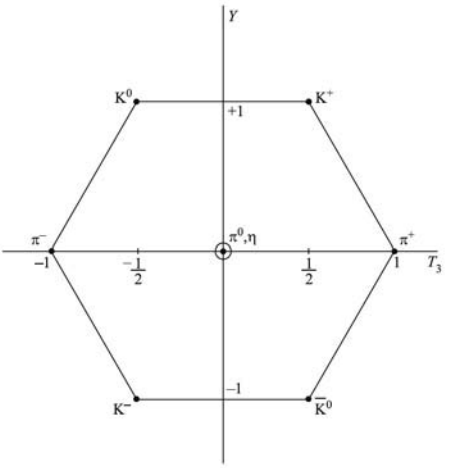


FIGURE 12.4
The $Y - T_3$ quantum numbers of the pseudoscalar meson octet.

The actual form of the $G_a^{(8)}$ matrices is given by comparing the infinitesimal version of (12.81) with (12.80)

$$\left(G_a^{(8)}\right)_{bc} = -\mathrm{i} f_{abc}, \tag{12.84}$$

as may be checked in problem 12.8, where it is also verified that the matrices specified by (12.84) do obey the commutation relations (12.83).

As in the $\mathrm{SU(2)_f}$ case, the 8 states generated by the combinations $q^\dagger \lambda_a q$ are not necessarily the ones with the physically desired quantum numbers. To get the $\pi^\pm$, for example, we again need to form $(w_1 \pm \mathrm{i}w_2)/2$. Similarly, w_4 produces $\bar{u}s + \bar{s}u$ and w_5 the combination $-\mathrm{i}\,\bar{u}s + \mathrm{i}\,\bar{s}u$, so the $\mathrm{K}^\pm$ states are $w_4 \mp \mathrm{i}w_5$. Similarly the K^0, $\bar{\mathrm{K}}^0$ states are $w_6 - \mathrm{i}w_7$, and $w_6 + \mathrm{i}w_7$, while the η (in this simple model) would be $w_8 \sim (\bar{u}u + \bar{d}d - 2\bar{s}s)$, which is orthogonal to both the π^0 state and the $\mathrm{SU(3)_f}$ singlet. In this way all the pseudoscalar octet of π-partners has been identified, as shown on the $Y - T$ diagram of figure 12.4. We say 'octet of π-partners', but a reader knowing the masses of these particles might well query why we should feel justified in regarding them as (even approximately) degenerate. By contrast, a similar octet of vector ($J^P 1^-$) mesons (the $\omega, \rho, \mathrm{K}^*$ and $\bar{\mathrm{K}}^*$) are all much closer in mass, averaging around 800 MeV; in these states the $\bar{\mathrm{q}}\mathrm{q}$ spins add to $S = 1$, while the orbital angular momentum is still zero. The pion, and to a much lesser extent the kaons, seem to be 'anomalously light' for some reason: we shall learn the likely explanation for this in chapter 15.

There is a deep similarity between (12.84) and (12.48). In both cases, a representation has been found in which the matrix element of a generator is

minus the corresponding structure constant. Such a representation is always possible for a Lie group, and is called the *adjoint*, or *regular*, representation (see appendix M, section M.5). These representations are of particular importance in gauge theories, as we will see, since gauge quanta always belong to the adjoint representation of the gauged group (for example, the 8 gluons in $\mathrm{SU(3)_c}$).

Further flavours c, b and t of course exist, but the mass differences are now so large that it is generally not useful to think about higher flavour groups such as $\mathrm{SU(4)_f}$ etc. Instead, we now move on to consider the field-theoretic formulation of global $\mathrm{SU(2)_f}$ and $\mathrm{SU(3)_f}$.

12.4 Non-Abelian global symmetries in Lagrangian quantum field theory

12.4.1 $\mathrm{SU(2)_f}$ and $\mathrm{SU(3)_f}$

As may already have begun to be apparent in chapter 7, Lagrangian quantum field theory is a formalism which is especially well adapted for the description of symmetries. Without going into any elaborate general theory, we shall now give a few examples showing how global flavour symmetry is very easily built into a Lagrangian, generalizing in a simple way the global U(1) symmetries considered in section 7.1 and section 7.2. This will also prepare the way for the (local) gauge case, to be considered in the following chapter.

Consider, for example, the Lagrangian

$$\hat{\mathcal{L}} = \bar{\hat{u}}(\mathrm{i}\,\partial\!\!\!/ - m)\hat{u} + \bar{\hat{d}}(\mathrm{i}\,\partial\!\!\!/ - m)\hat{d} \tag{12.85}$$

describing two free fermions 'u' and 'd' of equal mass m, with the overbar now meaning the Dirac conjugate for the four-component spinor fields. Note carefully that we are suppressing the space-time arguments of the quantum fields $\hat{u}(x), \hat{d}(x)$. As in (12.50), we are using the convenient shorthand $\hat{\psi}_{\mathrm{u}} = \hat{u}$ and $\hat{\psi}_{\mathrm{d}} = \hat{d}$. Let us introduce

$$\hat{q} = \begin{pmatrix} \hat{u} \\ \hat{d} \end{pmatrix} \tag{12.86}$$

so that $\hat{\mathcal{L}}$ can be compactly written as

$$\hat{\mathcal{L}} = \bar{\hat{q}}(\mathrm{i}\,\partial\!\!\!/ - m)\hat{q}. \tag{12.87}$$

In this form it is obvious that $\hat{\mathcal{L}}$ – and hence the associated Hamiltonian $\hat{\mathcal{H}}$ – is invariant under the global U(1) transformation

$$\hat{q}' = \mathrm{e}^{\mathrm{i}\alpha}\hat{q} \tag{12.88}$$

(cf (12.1)) which is associated with baryon number conservation. It is also invariant under global $\mathrm{SU}(2)_\mathrm{f}$ transformations acting in the flavour u-d space (cf (12.32)):

$$\hat{q}' = \exp(-\mathrm{i}\boldsymbol{\alpha}\cdot\boldsymbol{\tau}/2)\hat{q} \tag{12.89}$$

(for the change in sign with respect to (12.51), compare section 7.1 and section 7.2 in the U(1) case). In (12.89), the three parameters $\boldsymbol{\alpha}$ are independent of x.

What are the conserved quantities associated with the invariance of $\hat{\mathcal{L}}$ under (12.89) ? Let us recall the discussion of the simpler U(1) cases studied in sections 7.1 and 7.2. Considering the complex scalar field of section 7.1, the analogue of (12.89) was just $\hat{\phi} \to \hat{\phi}' = \mathrm{e}^{-\mathrm{i}\alpha}\hat{\phi}$, and the conserved quantity was the Hermitian operator $\hat{N}_\phi$ which appeared in the exponent of the unitary operator $\hat{U}$ that effected the transformation $\hat{\phi} \to \hat{\phi}'$ via

$$\hat{\phi}' = \hat{U}\hat{\phi}\hat{U}^\dagger, \tag{12.90}$$

with

$$\hat{U} = \exp(\mathrm{i}\alpha\hat{N}_\phi). \tag{12.91}$$

For an infinitesimal α, we have

$$\hat{\phi}' \approx (1-\mathrm{i}\epsilon)\hat{\phi}, \quad \hat{U} \approx 1 + \mathrm{i}\epsilon\hat{N}_\phi, \tag{12.92}$$

so that (12.90) becomes

$$(1-\mathrm{i}\epsilon)\hat{\phi} = (1+\mathrm{i}\epsilon\hat{N}_\phi)\hat{\phi}(1-\mathrm{i}\epsilon\hat{N}_\phi) \approx \hat{\phi} + \mathrm{i}\epsilon[\hat{N}_\phi, \hat{\phi}]; \tag{12.93}$$

hence we require

$$[\hat{N}_\phi, \hat{\phi}] = -\hat{\phi} \tag{12.94}$$

for consistency. Insofar as $\hat{N}_\phi$ determines the form of an infinitesimal version of the unitary transformation operator $\hat{U}$, it seems reasonable to call it the *generator* of these global U(1) transformations (compare the discussion after (12.27) and (12.35), but note that here $\hat{N}_\phi$ is a quantum field operator, not a matrix).

Consider now the $\mathrm{SU}(2)_\mathrm{f}$ transformation (12.89), in the infinitesimal case:

$$\hat{q}' = (1 - \mathrm{i}\boldsymbol{\epsilon}\cdot\boldsymbol{\tau}/2)\hat{q}. \tag{12.95}$$

Since the single U(1) parameter ϵ is now replaced by the three parameters $\boldsymbol{\epsilon} = (\epsilon_1, \epsilon_2, \epsilon_3)$, we shall need three analogues of $\hat{N}_\phi$, which we call

$$\hat{\boldsymbol{T}}^{(\frac{1}{2})} = (\hat{T}_1^{(\frac{1}{2})}, \hat{T}_2^{(\frac{1}{2})}, \hat{T}_3^{(\frac{1}{2})}), \tag{12.96}$$

corresponding to the three independent infinitesimal SU(2) transformations. The generalizations of (12.90) and (12.91) are then

$$\hat{q}' = \hat{U}^{(\frac{1}{2})}\hat{q}\hat{U}^{(\frac{1}{2})\dagger} \tag{12.97}$$

and

$$\hat{U}^{(\frac{1}{2})} = \exp(\mathrm{i}\boldsymbol{\alpha} \cdot \hat{\boldsymbol{T}}^{(\frac{1}{2})}) \tag{12.98}$$

where the $\hat{\boldsymbol{T}}^{(\frac{1}{2})}$'s are Hermitian, so that $\hat{U}^{(\frac{1}{2})}$ is unitary (cf (12.35)). It would seem reasonable in this case too to regard the $\hat{\boldsymbol{T}}^{(\frac{1}{2})}$'s as providing a *field theoretic representation* of the generators of $\mathrm{SU}(2)_\mathrm{f}$, an interpretation we shall shortly confirm. In the infinitesimal case, (12.97) and (12.98) become

$$(1 - \mathrm{i}\boldsymbol{\epsilon} \cdot \boldsymbol{\tau}/2)\hat{q} = (1 + \mathrm{i}\boldsymbol{\epsilon} \cdot \hat{\boldsymbol{T}}^{(\frac{1}{2})})\hat{q}(1 - \mathrm{i}\boldsymbol{\epsilon} \cdot \hat{\boldsymbol{T}}^{(\frac{1}{2})}), \tag{12.99}$$

using the Hermiticity of the $\hat{\boldsymbol{T}}^{(\frac{1}{2})}$'s. Expanding the right-hand side of (12.99) to first order in $\boldsymbol{\epsilon}$, and equating coefficients of $\boldsymbol{\epsilon}$ on both sides, (12.99) reduces to (problem 12.9)

$$[\hat{\boldsymbol{T}}^{(\frac{1}{2})}, \hat{q}] = -(\boldsymbol{\tau}/2)\hat{q}, \tag{12.100}$$

which is the analogue of (12.94). Equation (12.100) expresses a very specific *commutation* property of the operators $\hat{\boldsymbol{T}}^{(\frac{1}{2})}$, which turns out to be satisfied by the expression

$$\hat{\boldsymbol{T}}^{(\frac{1}{2})} = \int \hat{q}^\dagger(\boldsymbol{\tau}/2)\hat{q}\,\mathrm{d}^3\boldsymbol{x} \tag{12.101}$$

as can be checked (problem 12.10) from the anticommutation relations of the fermionic fields in $\hat{q}$. We shall derive (12.101) from Noether's theorem (Noether 1918) in a little while. Note that if '$\boldsymbol{\tau}/2$' is replaced by 1, (12.101) reduces to the sum of the u and d number operators, as required for the one-parameter U(1) case. The '$\hat{q}^\dagger\boldsymbol{\tau}\hat{q}$' combination is precisely the field-theoretic version of the $q^\dagger\boldsymbol{\tau}q$ coupling we discussed in section 12.1.3. It means that the three operators $\hat{\boldsymbol{T}}^{(\frac{1}{2})}$ themselves belong to a $T = 1$ triplet of $\mathrm{SU}(2)_\mathrm{f}$.

It is possible to verify that these $\hat{\boldsymbol{T}}^{(\frac{1}{2})}$'s do indeed commute with the Hamiltonian $\hat{H}$:

$$\mathrm{d}\hat{\boldsymbol{T}}^{(\frac{1}{2})}/\mathrm{d}t = -\mathrm{i}[\hat{\boldsymbol{T}}^{(\frac{1}{2})}, \hat{H}] = 0 \tag{12.102}$$

so that their eigenvalues are conserved. That the $\hat{\boldsymbol{T}}^{(\frac{1}{2})}$ are, as already suggested, a field theoretic representation of the generators of SU(2), appropriate to the case $T = \frac{1}{2}$, follows from the fact that they obey the SU(2) algebra (problem 12.11):

$$[\hat{T}_i^{(\frac{1}{2})}, \hat{T}_j^{(\frac{1}{2})}] = \mathrm{i}\epsilon_{ijk}\hat{T}_k^{(\frac{1}{2})}. \tag{12.103}$$

For many purposes it is more useful to consider the raising and lowering operators

$$\hat{T}_\pm^{(\frac{1}{2})} = (\hat{T}_1^{(\frac{1}{2})} \pm \mathrm{i}\hat{T}_2^{(\frac{1}{2})}). \tag{12.104}$$

For example, we easily find

$$\hat{T}_+^{(\frac{1}{2})} = \int \hat{u}^\dagger \hat{d}\ \mathrm{d}^3\boldsymbol{x}, \tag{12.105}$$

which destroys a d quark and creates a u, or destroys a $\bar{\mathrm{u}}$ and creates a $\bar{\mathrm{d}}$, in either case raising the $\hat{T}_3^{(\frac{1}{2})}$ eigenvalue by +1, since

$$\hat{T}_3^{(\frac{1}{2})} = \frac{1}{2}\int (\hat{u}^\dagger \hat{u} - \hat{d}^\dagger \hat{d}) \mathrm{d}^3\boldsymbol{x} \tag{12.106}$$

which counts $+\frac{1}{2}$ for each u (or $\bar{\mathrm{d}}$) and $-\frac{1}{2}$ for each d (or $\bar{\mathrm{u}}$). Thus these operators certainly 'do the job' expected of field theoretic isospin operators, in this isospin-1/2 case.

In the U(1) case, considering now the fermionic example of section 7.2 for variety, we could go further and associate the conserved operator $\hat{N}_\psi$ with a *conserved current* $\hat{N}_\psi^\mu$:

$$\hat{N}_\psi = \int \hat{N}_\psi^0 \mathrm{d}^3\boldsymbol{x}, \qquad \hat{N}_\psi^\mu = \bar{\hat{\psi}}\gamma^\mu\hat{\psi} \tag{12.107}$$

where

$$\partial_\mu \hat{N}_\psi^\mu = 0. \tag{12.108}$$

The obvious generalization appropriate to (12.101) is

$$\hat{\boldsymbol{T}}^{(\frac{1}{2})} = \int \hat{\boldsymbol{T}}^{(\frac{1}{2})0} \mathrm{d}^3\boldsymbol{x}, \quad \hat{\boldsymbol{T}}^{(\frac{1}{2})\mu} = \bar{\hat{q}}\gamma^\mu \frac{\boldsymbol{\tau}}{2}\hat{q}. \tag{12.109}$$

Note that both $\hat{N}_\psi^\mu$ and $\hat{\boldsymbol{T}}^{(\frac{1}{2})\mu}$ are of course functions of the space-time coordinate x, via the (suppressed) dependence of the $\hat{q}$-fields on x. Indeed one can verify from the equations of motion that

$$\partial_\mu \hat{\boldsymbol{T}}^{(\frac{1}{2})\mu} = 0. \tag{12.110}$$

Thus $\hat{\boldsymbol{T}}^{(\frac{1}{2})\mu}$ is a *conserved isospin current operator* appropriate to the $T = \frac{1}{2}$ (u, d) system; it transforms as a 4-vector under Lorentz transformations, and as a $T = 1$ triplet under $\mathrm{SU}(2)_\mathrm{f}$ transformations.

Clearly there should be some general formalism for dealing with all this more efficiently, and it is provided by a generalization of the steps followed, in the U(1) case, in equations (7.6)–(7.8). Suppose the Lagrangian involves a set of fields $\hat{\psi}_r$ (they could be bosons or fermions) and suppose that it is *invariant* under the infinitesimal transformation

$$\delta\hat{\psi}_r = -\mathrm{i}\epsilon T_{rs}\hat{\psi}_s \tag{12.111}$$

for some set of numerical coefficients T_{rs}. Equation (12.111) generalizes (7.5). Then since $\hat{\mathcal{L}}$ is invariant under this change,

$$0 = \delta\hat{\mathcal{L}} = \frac{\partial\hat{\mathcal{L}}}{\partial\hat{\psi}_r}\delta\hat{\psi}_r + \frac{\partial\hat{\mathcal{L}}}{\partial(\partial^\mu\hat{\psi}_r)}\partial^\mu(\delta\hat{\psi}_r). \tag{12.112}$$

But

$$\frac{\partial\hat{\mathcal{L}}}{\partial\hat{\psi}_r} = \partial^\mu\left(\frac{\partial\hat{\mathcal{L}}}{\partial(\partial^\mu\hat{\psi}_r)}\right) \tag{12.113}$$

from the equations of motion. Hence

$$\partial^\mu\left(\frac{\partial\hat{\mathcal{L}}}{\partial(\partial^\mu\hat{\psi}_r)}\delta\hat{\psi}_r\right) = 0 \tag{12.114}$$

which is precisely a current conservation law of the form

$$\partial^\mu\hat{j}_\mu = 0. \tag{12.115}$$

Indeed, disregarding the irrelevant constant small parameter ϵ, the conserved current is

$$\hat{j}_\mu = -\mathrm{i}\frac{\partial\hat{\mathcal{L}}}{\partial(\partial^\mu\hat{\psi}_r)}T_{rs}\hat{\psi}_s. \tag{12.116}$$

Let us try this out on (12.87) with

$$\delta\hat{q} = (-\mathrm{i}\boldsymbol{\epsilon}\cdot\boldsymbol{\tau}/2)\hat{q}. \tag{12.117}$$

As we know already, there are now three ϵ's, and so three T_{rs}'s, namely $\frac{1}{2}(\tau_1)_{rs}, \frac{1}{2}(\tau_2)_{rs}, \frac{1}{2}(\tau_3)_{rs}$. For each one we have a current, for example

$$\hat{T}_{1\mu}^{(\frac{1}{2})} = -\mathrm{i}\frac{\partial\hat{\mathcal{L}}}{\partial(\partial^\mu\hat{q})}\ \frac{\tau_1}{2}\ \hat{q} = \bar{\hat{q}}\gamma_\mu\frac{\tau_1}{2}\hat{q} \tag{12.118}$$

and similarly for the other τ's, and so we recover (12.109). From the invariance of the Lagrangian under the transformation (12.117) there follows the conservation of an associated symmetry current. This is the quantum field theory version of Noether's theorem.

This theorem is of fundamental significance as it tells us how to relate symmetries (under transformations of the general form (12.111)) to 'current' conservation laws (of the form (12.115), and it constructs the actual currents for us. In gauge theories, the *dynamics* is generated from a symmetry, in the sense that (as we have seen in the local U(1) of electromagnetism) the symmetry currents are the dynamical currents that drive the equations for the force field. Thus the symmetries of the Lagrangian are basic to gauge field theories.

Let us look at another example, this time involving spin-0 fields. Suppose we have three spin-0 fields all with the same mass, and take

$$\hat{\mathcal{L}} = \frac{1}{2}\partial_\mu\hat{\phi}_1\partial^\mu\hat{\phi}_1 + \frac{1}{2}\partial_\mu\hat{\phi}_2\partial^\mu\hat{\phi}_2 + \frac{1}{2}\partial_\mu\hat{\phi}_3\partial^\mu\hat{\phi}_3 - \frac{1}{2}m^2(\hat{\phi}_1^2 + \hat{\phi}_2^2 + \hat{\phi}_3^2). \tag{12.119}$$

It is obvious that $\hat{\mathcal{L}}$ is invariant under an arbitrary rotation of the three $\hat{\phi}$'s among themselves, generalizing the 'rotation about the 3-axis' considered for

the $\hat{\phi}_1 - \hat{\phi}_2$ system of section 7.1. An infinitesimal such rotation is (cf (12.64), and noting the sign change in the field theory case)

$$\hat{\boldsymbol{\phi}}' = \hat{\boldsymbol{\phi}} + \boldsymbol{\epsilon} \times \hat{\boldsymbol{\phi}} \tag{12.120}$$

which implies

$$\delta\hat{\phi}_r = -\mathrm{i}\epsilon_a T^{(1)}_{ars}\hat{\phi}_s, \tag{12.121}$$

with

$$T^{(1)}_{ars} = -\mathrm{i}\epsilon_{ars} \tag{12.122}$$

as in (12.48). There are of course three conserved $\hat{\boldsymbol{T}}$ operators again, and three $\hat{\boldsymbol{T}}^{\mu}$'s, which we call $\hat{\boldsymbol{T}}^{(1)}$ and $\hat{\boldsymbol{T}}^{(1)\mu}$ respectively, since we are now dealing with a $T = 1$ isospin case. The $a = 1$ component of the conserved current in this case is, from (12.116),

$$\hat{T}^{(1)\mu}_1 = \hat{\phi}_2\partial^{\mu}\hat{\phi}_3 - \hat{\phi}_3\partial^{\mu}\hat{\phi}_2. \tag{12.123}$$

Cyclic permutations give us the other components which can be summarised as

$$\hat{\boldsymbol{T}}^{(1)\mu} = \mathrm{i}(\hat{\phi}^{(1)\mathrm{tr}}\,\mathbf{T}^{(1)}\partial^{\mu}\hat{\phi}^{(1)} - (\partial^{\mu}\hat{\phi}^{(1)})^{\mathrm{tr}}\,\mathbf{T}^{(1)}\hat{\phi}^{(1)}) \tag{12.124}$$

where we have written

$$\hat{\phi}^{(1)} = \begin{pmatrix} \hat{\phi}_1 \\ \hat{\phi}_2 \\ \hat{\phi}_3 \end{pmatrix} \tag{12.125}$$

and $^{\mathrm{tr}}$ denotes transpose. Equation (12.124) has the form expected of a bosonic spin-0 current, but with the matrices $\mathbf{T}^{(1)}$ appearing, appropriate to the $T = 1$ (triplet) representation of $\mathrm{SU}(2)_{\mathrm{f}}$.

The general form of such SU(2) currents should now be clear. For an isospin T-multiplet of bosons we shall have the form

$$\mathrm{i}(\hat{\phi}^{(T)\dagger}\mathbf{T}^{(T)}\partial^{\mu}\hat{\phi}^{(T)}) - (\partial^{\mu}\hat{\phi}^{(T)})^{\dagger}\mathbf{T}^{(T)}\hat{\phi}^{(T)}) \tag{12.126}$$

where we have put the $\dagger$ to allow for possibly complex fields; and for an isospin T-multiplet of fermions we shall have

$$\bar{\hat{\psi}}^{\,(T)}\gamma^{\mu}\mathbf{T}^{(T)}\hat{\psi}^{(T)} \tag{12.127}$$

where in each case the $(2T + 1)$ components of $\hat{\phi}$ or $\hat{\psi}$ transforms as a T-multiplet under SU(2), i.e.

$$\hat{\psi}^{(T)\prime} = \exp(-\mathrm{i}\boldsymbol{\alpha}\cdot\mathbf{T}^{(T)})\hat{\psi}^{(T)} \tag{12.128}$$

and similarly for $\hat{\phi}^{(T)}$, where $\mathbf{T}^{(T)}$ are the $2T+1 \times 2T+1$ matrices representing the generators of $\mathrm{SU}(2)_{\mathrm{f}}$ in this representation. In all cases, the integral over

all space of the $\mu = 0$ component of these currents results in a triplet of isospin operators obeying the SU(2) algebra (12.47), as in (12.103).

The cases considered so far have all been *free* field theories, but SU(2)-invariant interactions can be easily formed. For example, the interaction $g_1\bar{\hat{\psi}}\boldsymbol{\tau}\hat{\psi}\cdot\hat{\boldsymbol{\phi}}$ describes SU(2)-invariant interactions between a $T = \frac{1}{2}$ isospinor (spin- $\frac{1}{2}$) field $\hat{\psi}$, and a $T = 1$ isotriplet (Lorentz scalar) $\hat{\boldsymbol{\phi}}$. An effective interaction between pions and nucleons could take the form $g_\pi\bar{\hat{\psi}}\boldsymbol{\tau}\gamma_5\hat{\psi}\cdot\hat{\boldsymbol{\phi}}$, allowing for the pseudoscalar nature of the pions (we shall see in the following section that $\bar{\hat{\psi}}\gamma_5\hat{\psi}$ is a pseudoscalar, so the product is a true scalar as is required for a parity-conserving strong interaction). In these examples the 'vector' analogy for the $T = 1$ states allows us to see that the 'dot product' will be invariant. A similar dot product occurs in the interaction between the isospinor $\hat{\psi}^{(\frac{1}{2})}$ and the weak SU(2) gauge field $\hat{\boldsymbol{W}}_\mu$, which has the form

$$g\bar{\hat{q}}\gamma^\mu\frac{\boldsymbol{\tau}}{2}\hat{q}\cdot\hat{\boldsymbol{W}}_\mu \tag{12.129}$$

as will be discussed in the following chapter. This is just the SU(2) dot product of the symmetry current (12.109) and the gauge field triplet, both of which are in the adjoint ($T = 1$) representation of SU(2).

All of the foregoing can be generalized straightforwardly to $\mathrm{SU(3)_f}$. For example, the Lagrangian

$$\hat{\mathcal{L}} = \bar{\hat{q}}(\mathrm{i}\,\partial\!\!\!/ - m)\hat{q} \tag{12.130}$$

with $\hat{q}$ now extended to

$$\hat{q} = \begin{pmatrix} \hat{u} \\ \hat{d} \\ \hat{s} \end{pmatrix} \tag{12.131}$$

describes free u, d and s quarks of equal mass m. $\hat{\mathcal{L}}$ is clearly invariant under global $\mathrm{SU(3)_f}$ transformations

$$\hat{q}' = \exp(-\mathrm{i}\boldsymbol{\alpha}\cdot\boldsymbol{\lambda}/2)\hat{q}, \tag{12.132}$$

as well as the usual global U(1) transformation associated with quark number conservation. The associated Noether currents are (in somewhat informal notation)

$$\hat{G}_a^{(\mathrm{q})\mu} = \bar{\hat{q}}\gamma^\mu\frac{\lambda_a}{2}\hat{q} \qquad a = 1, 2, \ldots 8 \tag{12.133}$$

(note that there are eight of them), and the associated conserved 'charge operators' are

$$\hat{G}_a^{(\mathrm{q})} = \int \hat{G}_a^{(\mathrm{q})0}\mathrm{d}^3\boldsymbol{x} = \int \hat{q}^\dagger\frac{\lambda_a}{2}\hat{q} \;\; a = 1, 2, \ldots 8, \tag{12.134}$$

which obey the SU(3) commutation relations

$$[\hat{G}_a^{(\mathrm{q})}, \hat{G}_b^{(\mathrm{q})}] = \mathrm{i}f_{abc}\hat{G}_c^{(\mathrm{q})}. \tag{12.135}$$

SU(3)-invariant interactions can also be formed. A particularly important one is the 'SU(3) dot-product' of two octets (the analogues of the SU(2) triplets), which arises in the quark-gluon vertex of QCD (see chapters 13 and 14):

$$-\mathrm{i}g_{\mathrm{s}} \sum_{\mathrm{f}} \bar{\hat{q}}_{\mathrm{f}} \gamma^{\mu} \frac{\lambda_a}{2} \hat{q}_{\mathrm{f}} \hat{A}^a_{\mu}. \tag{12.136}$$

In (12.136), $\hat{q}_{\mathrm{f}}$ stands for the $\mathrm{SU}(3)_{\mathrm{c}}$ *colour* triplet

$$\hat{q}_{\mathrm{f}} = \begin{pmatrix} \hat{f}_{\mathrm{r}} \\ \hat{f}_{\mathrm{b}} \\ \hat{f}_{\mathrm{g}} \end{pmatrix} \tag{12.137}$$

where '$\hat{f}$' is any of the six quark flavour fields $\hat{u}, \hat{d}, \hat{c}, \hat{s}, \hat{t}, \hat{b}$, and $\hat{A}^a_{\mu}$ are the 8 ($a = 1, 2, \ldots 8$) gluon fields. Once again, (12.136) has the form 'symmetry current · gauge field' characteristic of all gauge interactions.

12.4.2 Chiral symmetry

As our final example of a global non-Abelian symmetry, we shall introduce the idea of *chiral symmetry*, which is an exact symmetry for fermions in the limit in which their masses may be neglected. We have seen that the u and d quarks have indeed very small masses (≤ 5 MeV) on hadronic scales, and even the s quark mass (~ 100 MeV) is relatively small. Thus we may certainly expect some physical signs of the symmetry associated with $m_{\mathrm{u}} \approx m_{\mathrm{d}} \approx 0$, and possibly also of the larger symmetry holding when $m_{\mathrm{u}} \approx m_{\mathrm{d}} \approx m_{\mathrm{s}} \approx 0$. As we shall see, however, this expectation leads to a puzzle, the resolution of which will have to be postponed until the concept of 'spontaneous symmetry breaking' has been developed in Part VII.

We begin with the simplest case of just one fermion. Since we are interested in the 'small mass' regime, it is sensible to use the representation (3.40) of the Dirac matrices, in which the momentum part of the Dirac Hamiltonian is 'diagonal' and the mass appears as an 'off-diagonal' coupling:

$$\boldsymbol{\alpha} = \begin{pmatrix} \boldsymbol{\sigma} & 0 \\ 0 & -\boldsymbol{\sigma} \end{pmatrix}, \quad \beta = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}. \tag{12.138}$$

Writing the general Dirac spinor ω as

$$\omega = \begin{pmatrix} \phi \\ \chi \end{pmatrix}, \tag{12.139}$$

we have (as in (4.14), (4.15))

$$E\phi = \boldsymbol{\sigma} \cdot \boldsymbol{p}\phi + m\chi \tag{12.140}$$

$$E\chi = -\boldsymbol{\sigma} \cdot \boldsymbol{p}\chi + m\phi. \tag{12.141}$$

We now recall the matrix γ_5 introduced in section 4.2.1

$$\gamma_5 = \mathrm{i}\gamma^0\gamma^1\gamma^2\gamma^3, \tag{12.142}$$

which takes the form

$$\gamma_5 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \tag{12.143}$$

in this representation. The matrix γ_5 plays a prominent role in chiral symmetry, as we shall see. Its defining property is that it anticommutes with the γ^μ matrices:

$$\{\gamma_5, \gamma^\mu\} = 0. \tag{12.144}$$

'Chirality' means 'handedness', from the Greek word for hand, $\chi\epsilon\iota\rho$. Its use here stems from the fact that, in the limit $m \to 0$ the 2-component spinors ϕ, χ become helicity eigenstates (cf problem 9.4), having definite 'handedness'. As $m \to 0$ we have $E \to |\boldsymbol{p}|$, and (12.140) and (12.141) reduce to

$$(\boldsymbol{\sigma}\cdot\boldsymbol{p}/|\boldsymbol{p}|)\tilde{\phi} = \tilde{\phi} \tag{12.145}$$

$$(\boldsymbol{\sigma}\cdot\boldsymbol{p}/|\boldsymbol{p}|)\tilde{\chi} = -\tilde{\chi}, \tag{12.146}$$

so that the limiting spinor $\tilde{\phi}$ has positive helicity, and $\tilde{\chi}$ negative helicity (cf (3.68) and (3.69)). In this $m \to 0$ limit, the two helicity spinors are *decoupled*, reflecting the fact that no Lorentz transformation can reverse the helicity of a massless particle. Also in this limit, the Dirac energy operator is

$$\boldsymbol{\alpha}\cdot\boldsymbol{p} = \begin{pmatrix} \boldsymbol{\sigma}\cdot\boldsymbol{p} & 0 \\ 0 & -\boldsymbol{\sigma}\cdot\boldsymbol{p} \end{pmatrix} \tag{12.147}$$

which is easily seen to commute with γ_5. Thus the massless states may equivalently be classified by the eigenvalues of γ_5, which are clearly ± 1 since $\gamma_5^2 = I$.

Consider then a massless fermion with positive helicity. It is described by the 'u'-spinor $\begin{pmatrix} \tilde{\phi} \\ 0 \end{pmatrix}$ which is an eigenstate of γ_5 with eigenvalue $+1$. Similarly, a fermion with negative helicity is described by $\begin{pmatrix} 0 \\ \tilde{\chi} \end{pmatrix}$ which has $\gamma_5 = -1$. Thus for these states chirality equals helicity. We have to be more careful for antifermions, however. A physical antifermion of energy E and momentum $\boldsymbol{p}$ is described by a 'v'- spinor corresponding to $-E$ and $-\boldsymbol{p}$; but with $m = 0$ in (12.140) and (12.141) the equations for ϕ and χ remain the same for $-E, -\boldsymbol{p}$ as for $E, \boldsymbol{p}$. Consider the spin, however. If the physical antiparticle has positive helicity, with $\boldsymbol{p}$ along the z-axis say, then $s_z = +\frac{1}{2}$. The corresponding v-spinor must then have $s_z = -\frac{1}{2}$ (see section 3.4.3) and must therefore be of $\tilde{\chi}$ type (12.146). So the v-spinor for this antifermion of positive helicity is $\begin{pmatrix} 0 \\ \tilde{\chi} \end{pmatrix}$ which has $\gamma_5 = -1$. In summary, for fermions the γ_5 eigenvalue is equal to the helicity, and for antifermions it is equal to minus the helicity. It is the γ_5 eigenvalue that is called the 'chirality'.

In the massless limit, the chirality of $\tilde{\phi}$ and $\tilde{\chi}$ is a good quantum number (γ_5 commuting with the energy operator), and we may say that 'chirality is conserved' in this massless limit. On the other hand, the massive spinor ω is clearly *not* an eigenstate of chirality:

$$\gamma_5\omega = \begin{pmatrix} \phi \\ -\chi \end{pmatrix} \neq \lambda \begin{pmatrix} \phi \\ \chi \end{pmatrix}. \tag{12.148}$$

Referring to (12.140) and (12.141), we may therefore regard the mass terms as 'coupling the states of different chirality'.

It is usual to introduce operators $P_{\mathrm{R,L}} = \left(\frac{1\pm\gamma_5}{2}\right)$ which 'project' out states of definite chirality from ω:

$$\omega = \left(\frac{1+\gamma_5}{2}\right)\omega + \left(\frac{1-\gamma_5}{2}\right)\omega \equiv P_{\mathrm{R}}\omega + P_{\mathrm{L}}\omega \equiv \omega_{\mathrm{R}} + \omega_{\mathrm{L}}, \tag{12.149}$$

so that

$$\omega_{\mathrm{R}} = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}\begin{pmatrix} \phi \\ \chi \end{pmatrix} = \begin{pmatrix} \phi \\ 0 \end{pmatrix}, \quad \omega_{\mathrm{L}} = \begin{pmatrix} 0 \\ \chi \end{pmatrix}. \tag{12.150}$$

Then clearly $\gamma_5\omega_{\mathrm{R}} = \omega_{\mathrm{R}}$ and $\gamma_5\omega_{\mathrm{L}} = -\omega_{\mathrm{L}}$; slightly confusingly, the notation 'R', 'L' is used for the *chirality* eigenvalue.

We now reformulate the above in field-theoretic terms. The Dirac Lagrangian for a single massless fermion is

$$\hat{\mathcal{L}}_0 = \bar{\hat{\psi}}\mathrm{i}\,\not{\partial}\hat{\psi}. \tag{12.151}$$

This is invariant not only under the now familiar global U(1) transformation $\hat{\psi} \to \hat{\psi}' = \mathrm{e}^{-\mathrm{i}\alpha}\hat{\psi}$, but also under the 'global *chiral* U(1)' transformation

$$\hat{\psi} \to \hat{\psi}' = \mathrm{e}^{-\mathrm{i}\theta\gamma_5}\hat{\psi} \tag{12.152}$$

where θ is an arbitrary (x-independent) real parameter. The invariance is easily verified: using $\{\gamma^0, \gamma_5\} = 0$ we have

$$\bar{\hat{\psi}}' = \hat{\psi}'^{\dagger}\gamma^0 = \hat{\psi}^{\dagger}\mathrm{e}^{\mathrm{i}\theta\gamma_5}\gamma^0 = \hat{\psi}^{\dagger}\gamma^0\mathrm{e}^{-\mathrm{i}\theta\gamma_5} = \bar{\hat{\psi}}\mathrm{e}^{-\mathrm{i}\theta\gamma_5}, \tag{12.153}$$

and then using $\{\gamma^\mu, \gamma_5\} = 0$,

$$\begin{aligned} \bar{\hat{\psi}}'\gamma^\mu\partial_\mu\hat{\psi}' &= \bar{\hat{\psi}}\mathrm{e}^{-\mathrm{i}\theta\gamma_5}\gamma^\mu\partial_\mu\mathrm{e}^{-\mathrm{i}\theta\gamma_5}\hat{\psi} \\ &= \bar{\hat{\psi}}\gamma^\mu\mathrm{e}^{\mathrm{i}\theta\gamma_5}\partial_\mu\mathrm{e}^{-\mathrm{i}\theta\gamma_5}\hat{\psi} \\ &= \bar{\hat{\psi}}\gamma^\mu\partial_\mu\hat{\psi} \end{aligned} \tag{12.154}$$

as required. The corresponding Noether current is

$$\hat{j}^\mu_5 = \bar{\hat{\psi}}\gamma^\mu\gamma_5\hat{\psi}, \tag{12.155}$$

and the spatial integral of its $\mu = 0$ component is the (conserved) chirality operator

$$\hat{Q}_5 = \int \hat{\psi}^\dagger \gamma_5 \hat{\psi} \mathrm{d}^3\boldsymbol{x} = \int \left(\hat{\phi}^\dagger \hat{\phi} - \hat{\chi}^\dagger \hat{\chi} \right) \mathrm{d}^3\boldsymbol{x}. \tag{12.156}$$

We denote this chiral U(1) by $\mathrm{U}(1)_5$.

It is interesting to compare the form of $\hat{Q}_5$ with that of the corresponding operator $\int \hat{\psi}^\dagger \hat{\psi} \mathrm{d}^3\boldsymbol{x}$ in the non-chiral case (cf (7.51)). The difference has to do with their behaviour under a transformation already discussed in section 4.2.1, namely *parity*. Under the parity transformation $\boldsymbol{p} \to -\boldsymbol{p}$ and thus, for (12.140) and (12.141) to be covariant under parity, we require $\phi \to \chi,\ \chi \to \phi$; this will ensure (as we saw in section 4.2.1) that the Dirac equation in the parity-transformed frame will be consistent with the one in the original frame. In the representation (12.138), this is equivalent to saying that the spinor $\omega_{\mathbf{P}}$ in the parity-transformed frame is given by

$$\omega_{\mathbf{P}} = \gamma^0 \omega. \tag{12.157}$$

which implies $\phi_{\mathbf{P}} = \chi,\ \chi_{\mathbf{P}} = \phi$.

All this carries over to the field theory case, with $\hat{\psi}_{\mathbf{P}}(\boldsymbol{x}, t) = \gamma^0 \hat{\psi}(-\boldsymbol{x}, t)$, as we saw in section 7.5.1. Consider then the operator $\hat{Q}_5$ in the parity-transformed frame:

$$\begin{aligned} (\hat{Q}_5)_{\mathbf{P}} &= \int \hat{\psi}^\dagger_{\mathbf{P}}(\boldsymbol{x}, t) \gamma_5 \psi_{\mathbf{P}}(\boldsymbol{x}, t) \mathrm{d}^3\boldsymbol{x} = \int \hat{\psi}^\dagger(-\boldsymbol{x}, t) \gamma^0 \gamma_5 \gamma^0 \hat{\psi}(-\boldsymbol{x}, t) \mathrm{d}^3\boldsymbol{x} \\ &= -\int \hat{\psi}^\dagger(\boldsymbol{y}, t) \gamma_5 \hat{\psi}(\boldsymbol{y}, t) \mathrm{d}^3\boldsymbol{y} = -\hat{Q}_5 \end{aligned} \tag{12.158}$$

where we used $\{\gamma^0, \gamma_5\} = 0$ and $\left(\gamma^0\right)^2 = 1$, and changed the integration variable to $\boldsymbol{y} = -\boldsymbol{x}$. Hence $\hat{Q}_5$ is a 'pseudoscalar' operator, meaning that it changes sign in the parity-transformed frame. We can also see this directly from (12.156), making the interchange $\hat{\phi} \leftrightarrow \hat{\chi}$. In contrast, the non-chiral operator $\int \hat{\psi}^\dagger \hat{\psi} \mathrm{d}^3\boldsymbol{x}$ is a (true) scalar, remaining the same in the parity-transformed frame.

In a similar way, the appearance of the γ_5 in the current operator $\hat{j}^\mu_5 = \bar{\hat{\psi}} \gamma^\mu \gamma_5 \hat{\psi}$ affects its parity properties: for example, the $\mu = 0$ component $\hat{\psi}^\dagger \gamma_5 \hat{\psi}$ is a pseudoscalar, as we have seen. Problem 4.4(b) showed that the spatial parts $\bar{\hat{\psi}} \boldsymbol{\gamma} \gamma_5 \hat{\psi}$ behave as an *axial vector* rather than a normal (*polar*) vector under parity: that is, they behave like $\boldsymbol{r} \times \boldsymbol{p}$ for example, rather than like $\boldsymbol{r}$, in that they do *not* reverse sign under parity. Such a current is referred to generally as an 'axial vector current', as opposed to the ordinary vector currents with no γ_5.

As a consequence of (12.158), the operator $\hat{Q}_5$ changes the parity of any state on which it acts. We can see this formally by introducing the (unitary) parity operator $\hat{\mathbf{P}}$ in field theory, such that states of definite parity $|+\rangle, |-\rangle$ satisfy

$$\hat{\mathbf{P}}|+\rangle = |+\rangle, \quad \hat{\mathbf{P}}|-\rangle = -|-\rangle. \tag{12.159}$$

Equation (12.158) then implies that $\hat{\mathbf{P}}\hat{Q}_5\hat{\mathbf{P}}^{-1} = -\hat{Q}_5$, following the normal rule for operator transformations in quantum mechanics. Consider now the state $\hat{Q}_5|+\rangle$. We have

$$\begin{aligned}\hat{\mathbf{P}}\hat{Q}_5|+\rangle &= \left(\hat{\mathbf{P}}\hat{Q}_5\hat{\mathbf{P}}^{-1}\right)\hat{\mathbf{P}}|+\rangle \\ &= -\hat{Q}_5|+\rangle \end{aligned} \tag{12.160}$$

showing that $\hat{Q}_5|+\rangle$ is an eigenstate of $\hat{\mathbf{P}}$ with the opposite eigenvalue, -1.

A very important physical consequence now follows from the fact that (in this simple $m = 0$ model) $\hat{Q}_5$ is a symmetry operator commuting with the Hamiltonian $\hat{H}$. We have

$$\hat{H}\hat{Q}_5|\psi\rangle = \hat{Q}_5\hat{H}|\psi\rangle = E\hat{Q}_5|\psi\rangle. \tag{12.161}$$

Hence for every state $|\psi\rangle$ with energy eigenvalue E, there should exist a state $\hat{Q}_5|\psi\rangle$ with the same eigenvalue E and the opposite parity: that is, chiral symmetry apparently implies the existence of 'parity doublets'.

Of course, it may reasonably be objected that all of the above refers not only to the massless, but also the *non-interacting* case. However, this is just where the analysis begins to get interesting. Suppose we allow the fermion field $\hat{\psi}$ to interact with a U(1)-gauge field $\hat{A}^\mu$ via the standard electromagnetic coupling

$$\hat{\mathcal{L}}_{\rm int} = q\bar{\hat{\psi}}\gamma^\mu\hat{\psi}\hat{A}_\mu. \tag{12.162}$$

Remarkably enough, $\hat{\mathcal{L}}_{\rm int}$ is *also* invariant under the chiral transformation (12.152), for the simple reason that the 'Dirac' structure of (12.162) is exactly the same as that of the free kinetic term $\bar{\hat{\psi}}\ \not{\partial}\hat{\psi}$: the 'covariant derivative' prescription $\partial^\mu \to D^\mu = \partial^\mu + \mathrm{i}q\hat{A}^\mu$ automatically means that any 'Dirac' (e.g. γ_5) symmetry of the kinetic part will be preserved when the gauge interaction is included. Thus chirality remains a 'good symmetry' in the presence of a U(1) gauge interaction.

The generalization of this to the more physical $m_{\rm u} \approx m_{\rm d} \approx 0$ case is quite straightforward. The Lagrangian (12.87) becomes

$$\hat{\mathcal{L}} = \bar{\hat{q}}\,\mathrm{i}\,\not{\partial}\hat{q} \tag{12.163}$$

as $m \to 0$, which is invariant under the γ_5-version of (12.89),[6] namely

$$\hat{q}' = \exp(-\mathrm{i}\boldsymbol{\beta}\cdot\boldsymbol{\tau}/2\gamma_5)\hat{q}. \tag{12.164}$$

There are three associated Noether currents (compare (12.109))

$$\hat{\boldsymbol{T}}_5^{(\frac{1}{2})\,\mu} = \bar{\hat{q}}\gamma^\mu\gamma_5\frac{\boldsymbol{\tau}}{2}\hat{q} \tag{12.165}$$

[6] $\hat{\mathcal{L}}_0$ is also invariant under $\hat{q}' = \mathrm{e}^{-\mathrm{i}\theta\gamma_5}\hat{q}$ which is an 'axial' version of the global U(1) associated with quark number conservation. We shall discuss this additional U(1)-symmetry in section 18.1.1.

which are axial vectors, and three associated 'charge' operators

$$\hat{\boldsymbol{T}}_5^{(\frac{1}{2})} = \int \hat{q}^\dagger \gamma_5 \frac{\boldsymbol{\tau}}{2} \hat{q} \mathrm{d}^3\boldsymbol{x} \tag{12.166}$$

which are pseudoscalars, belonging to the T=1 representation of SU(2). We have a new non-Abelian global symmetry, called chiral $\mathrm{SU}(2)_{\mathrm{f}}$, which we shall denote by $\mathrm{SU}(2)_{\mathrm{f}\,5}$. As far as their action in the isospinor u-d space is concerned, these chiral charges have exactly the same effect as the ordinary flavour isospin operators of (12.109). But they are pseudoscalars rather than scalars, and hence they flip the parity of a state on which they act. Thus, whereas the isospin raising operator $\hat{T}_+^{(\frac{1}{2})}$ is such that

$$\hat{T}_+^{(\frac{1}{2})}|d\rangle = |u\rangle, \tag{12.167}$$

$\hat{T}_{+\,5}^{(\frac{1}{2})}$ will also produce a u-type state from a d-type one via

$$\hat{T}_{+\,5}^{(\frac{1}{2})}|d\rangle = |\tilde{u}\rangle, \tag{12.168}$$

but the $|\tilde{u}\rangle$ state will have opposite parity from $|u\rangle$. Further, since $[\hat{T}_{+\,5}^{(\frac{1}{2})}, \hat{H}] = 0$, this state $|\tilde{u}\rangle$ will be degenerate with $|d\rangle$. Similarly, the state $|\tilde{d}\rangle$ produced via $\hat{T}_{-\,5}^{(\frac{1}{2})}|u\rangle$ will have opposite parity from $|d\rangle$, and will be degenerate with $|u\rangle$. The upshot is that we have two massless states $|u\rangle,\ |d\rangle$ of (say) positive parity, and a further two massless states $|\tilde{u}\rangle,\ |\tilde{d}\rangle$ of negative parity, in this simple model.

Suppose we now let the quarks interact, for example by an interaction of the QCD type, already indicated in (12.136). In that case, the interaction terms have the form

$$\bar{\hat{u}}\gamma^\mu \frac{\lambda_a}{2} \hat{u} \hat{A}_\mu^a + \bar{\hat{d}}\gamma^\mu \frac{\lambda_a}{2} \hat{d} \hat{A}_\mu^a \tag{12.169}$$

where

$$\hat{u} = \begin{pmatrix} \hat{u}_{\mathrm{r}} \\ \hat{u}_{\mathrm{b}} \\ \hat{u}_{\mathrm{g}} \end{pmatrix}, \hat{d} = \begin{pmatrix} \hat{d}_{\mathrm{r}} \\ \hat{d}_{\mathrm{b}} \\ \hat{d}_{\mathrm{g}} \end{pmatrix} \tag{12.170}$$

and the 3×3 λ's act in the r-b-g space. Just as in the previous U(1) case, the interaction (12.169) is invariant under the global $\mathrm{SU}(2)_{\mathrm{f}\,5}$ chiral symmetry (12.164), acting in the u-d space. Note that, somewhat confusingly, (12.169) is *not* a simple 'gauging' of (12.163): a covariant derivative is being introduced, but in the space of a new (colour) degree of freedom, not in flavour space. In fact, the flavour degrees of freedom are 'inert' in (12.169), so that it is invariant under $\mathrm{SU}(2)_{\mathrm{f}}$ transformations, while the Dirac structure implies that it is also invariant under chiral $\mathrm{SU}(2)_{\mathrm{f}\,5}$ transformations (12.164). All the foregoing can be extended unchanged to chiral $\mathrm{SU}(3)_{\mathrm{f}\,5}$, given that QCD is 'flavour blind', and supposing that $m_{\mathrm{s}} \approx 0$.

The effect of the QCD interactions must be to bind the quark into nucleons, such as the proton (uud) and neutron (udd). But what about the equally possible states ($\tilde{u}\tilde{u}\tilde{d}$) and ($\tilde{u}\tilde{d}\tilde{d}$), for example? These would have to be degenerate in mass with (uud) and (udd), and of opposite parity. Yet such 'parity doublet' partners of the physical p and n are not observed, and so we have a puzzle.

One might feel that this whole discussion is unrealistic, based as it is on massless quarks. Are the baryons then supposed to be massless too? If so, perhaps the discussion is idle, as they are evidently by no means massless. But it is not necessary to suppose that the mass of a relativistic bound state has any very simple relation to the masses of its constituents: its mass may derive, in part at least, from the interaction energy in the fields. Alternatively, one might suppose that somehow the finite mass of the u and d quarks, which of course breaks the chiral symmetry, splits the degeneracy of the nucleon parity doublets, promoting the negative parity 'nucleon' state to an acceptably high mass. But this seems very implausible, in view of the actual magnitudes of m_{u} and m_{d}, compared to the nucleon masses.

In short, we have here a situation in which a *symmetry of the Lagrangian* (to an apparently good approximation) does *not* seem to result in the expected *multiplet structure of the states*. The resolution of this puzzle will have to await our discussion of 'spontaneous symmetry breaking', in Part VII.

In conclusion, we note an important feature of the flavour symmetry currents $\hat{\boldsymbol{T}}^{(\frac{1}{2})\mu}$ and $\hat{\boldsymbol{T}}_5^{(\frac{1}{2})\mu}$ discussed in this and the preceding section. Although these currents have been introduced entirely within the context of *strong* interaction symmetries, it is a remarkable fact that exactly these currents also appear in strangeness-conserving semileptonic *weak* interactions such as β-decay, as we shall see in chapter 20. (The fact that *both* appear is precisely a manifestation of *parity violation* in weak interactions, as we noted in section 4.2.1). Thus some of the physical consequences of 'spontaneously broken chiral symmetry' will involve weak interaction quantities.

Problems

12.1 Verify that the set of all unitary 2×2 matrices with determinant equal to $+1$ form a group, the law of combination being matrix multiplication.

12.2 Derive (12.18).

12.3 Check the commutation relations (12.28).

12.4 Show that the T_i's defined by (12.45) satisfy (12.47).

12.5 Write out each of the 3×3 matrices $T_i^{(1)} (i = 1, 2, 3)$ whose matrix

elements are given by (12.48), and verify that they satisfy the SU(2) commutation relations (12.47).

12.6 Verify (12.62).

12.7 Show that a general Hermitian traceless 3×3 matrix is parametrized by 8 real numbers.

12.8 Check that (12.84) is consistent with (12.80) and the infinitesimal form of (12.81), and verify that the matrices $G_a^{(8)}$ defined by (12.84) satisfy the commutation relations (12.83).

12.9 Verify, by comparing the coefficients of ϵ_1, ϵ_2 and ϵ_3 on both sides of (12.99), that (12.100) follows from (12.99).

12.10 Verify that the operators $\hat{\boldsymbol{T}}^{(\frac{1}{2})}$ defined by (12.101) satisfy (12.100). (Note: use the anticommutation relations of the fermionic operators.)

12.11 Verify that the operators $\hat{\boldsymbol{T}}^{(\frac{1}{2})}$ given by (12.101) satisfy the commutation relations (12.103).

13

Local Non-Abelian (Gauge) Symmetries

> . . . The difference between a neutron and a proton is then a purely arbitrary process. As usually conceived, however, this arbitrariness is subject to the following limitations: once one chooses what to call a proton, what a neutron, at one space time point, one is then not free to make any choices at other space time points.
>
> It seems that this is not consistent with the localized field concept that underlies the usual physical theories. In the present paper we wish to explore the possibility of requiring all interactions to be invariant under *independent* rotations of the isotopic spin at all space time points . . . .
>
> —Yang and Mills (1954)

Consider the global SU(2) isospinor transformation (12.32), written here again,

$$\psi^{(\frac{1}{2})\prime}(x) = \exp(\mathrm{i}\boldsymbol{\alpha}\cdot\boldsymbol{\tau}/2)\psi^{(\frac{1}{2})}(x) \tag{13.1}$$

for an isospin doublet wavefunction $\psi^{(\frac{1}{2})}(x)$. The dependence of $\psi^{(\frac{1}{2})}(x)$ on the space-time coordinate x has now been included explicitly, but the parameters $\boldsymbol{\alpha}$ are independent of x, which is why the transformation is called a 'global' one. As we have seen in the previous chapter, invariance under this transformation amounts to the assertion that the choice of *which* two base states – $(n,p),(u,d),\ldots$ – to use is a matter of convention; any such non-Abelian phase transformation on a chosen pair produces another equally good pair. However, the choice cannot be made independently at all space-time points, only *globally*. To Yang and Mills (1954) (cf the quotation above) this seemed somehow an unaesthetic limitation of symmetry: 'Once one chooses what to call a proton, what a neutron, at one space-time point, one is then not free to make any choices at other space-time points.' They even suggested that this could be viewed as 'inconsistent with the localised field concept', and they therefore 'explored the possibility' of replacing this global (space-time independent) phase transformation by the local (space-time dependent) one

$$\psi^{(\frac{1}{2})\prime}(x) = \exp[\mathrm{i}g\boldsymbol{\tau}\cdot\boldsymbol{\alpha}(x)/2]\psi^{(\frac{1}{2})}(x) \tag{13.2}$$

in which the phase parameters $\boldsymbol{\alpha}(x)$ are also now functions of $x=(t,\boldsymbol{x})$ as

indicated. Notice that we have inserted a parameter g in the exponent to make the analogy with the electromagnetic U(1) case

$$\psi'(x) = \exp[\mathrm{i}q\chi(x)]\psi(x) \tag{13.3}$$

even stronger: g will be a coupling strength, analogous to the electromagnetic charge q. The consideration of theories based on (13.2) was the fundamental step taken by Yang and Mills (1954); see also Shaw (1955).

Global symmetries and their associated (possibly approximate) conservation laws are certainly important, but they do not have the *dynamical* significance of local symmetries. We saw in section 7.4 how the 'requirement' of local U(1) phase invariance led almost automatically to the local gauge theory of QED, in which the conserved current $\bar{\hat{\psi}}\gamma^\mu\hat{\psi}$ of the global U(1) symmetry is 'promoted' to the role of dynamical current which, when dotted into the gauge field $\hat{A}^\mu$, gave the interaction term in $\hat{\mathcal{L}}_{\mathrm{QED}}$. A similar link between symmetry and dynamics appears if, following Yang and Mills, we generalize the non-Abelian global symmetries of the preceding chapter to local non-Abelian symmetries, which are the subject of the present one.

However, as mentioned in the introduction to chapter 12, the original Yang-Mills attempt to get a theory of hadronic interactions by 'localizing' the flavour symmetry group SU(2) turned out not to be phenomenologically viable (although a remarkable attempt was made to push the idea further by Sakurai (1960)). In the event, the successful application of a local SU(2) symmetry was to the *weak* interactions. But this is complicated by the fact that the symmetry is 'spontaneously broken', and consequently we shall delay the discussion of this application until after QCD – which *is* the theory of strong interactions, but at the quark, rather than the composite (hadronic) level. QCD is based on the local form of an SU(3) symmetry; once again, however, it is *not* the flavour SU(3) of section 12.2, but a symmetry with respect to a totally new degree of freedom, colour. This will be introduced in the following chapter.

Although the application of local SU(2) symmetry to the weak interactions will follow that of local SU(3) to the strong, we shall begin our discussion of local non-Abelian symmetries with the local SU(2) case, since the group theory is more familiar. We shall also start with the 'wavefunction' formalism, deferring the field theory treatment until section 13.3.

13.1 Local SU(2) symmetry

13.1.1 The covariant derivative and interactions with matter

In this section we shall introduce the main ideas of the non-Abelian SU(2) gauge theory which results from the demand of invariance, or covariance,

under transformations such as (13.2). We shall generally use the language of isospin when referring to the physical states and operators, bearing in mind that this will eventually mean *weak* isospin.

We shall mimic as literally as possible the discussion of electromagnetic gauge covariance in sections 2.4 and 2.5 of volume 1. As in that case, no free particle wave equation can be covariant under the transformation (13.2) (taking the isospinor example for definiteness), since the gradient terms in the equation will act on the phase factor $\boldsymbol{\alpha}(x)$. However, wave equations with a suitably defined *covariant derivative* can be covariant under (13.2); physically this means that, just as for electromagnetism, covariance under local non-Abelian phase transformations requires the introduction of a definite force field.

In the electromagnetic case the covariant derivative is

$$D^\mu = \partial^\mu + \mathrm{i}qA^\mu(x). \tag{13.4}$$

For convenience we recall here the crucial property of D^μ. Under a local U(1) phase transformation, a wavefunction transforms as (cf (13.3))

$$\psi(x) \to \psi'(x) = \exp(\mathrm{i}q\chi(x))\psi(x), \tag{13.5}$$

from which it easily follows that the derivative (gradient) of ψ transforms as

$$\partial^\mu\psi(x) \to \partial^\mu\psi'(x) = \exp(\mathrm{i}q\chi(x))\partial^\mu\psi(x) + \mathrm{i}q\partial^\mu\chi(x)\exp(\mathrm{i}q\chi(x))\psi(x). \tag{13.6}$$

Comparing (13.6) with (13.5), we see that, in addition to the expected first term on the right-hand side of (13.6), which has the same form as the right-hand side of (13.5), there is an *extra* term in (13.6). By contrast, the covariant derivative of ψ transforms as (see section 2.4 of volume 1)

$$D^\mu\psi(x) \to D'^\mu\psi'(x) = \exp(\mathrm{i}q\chi(x))D^\mu\psi(x) \tag{13.7}$$

exactly as in (13.5), with no additional term on the right-hand side. Note that D^μ has to carry a prime also, since it contains A^μ which transforms to $A'^\mu = A^\mu - \partial^\mu\chi(x)$ when ψ transforms by (13.5). The property (13.7) ensured the gauge covariance of wave equations in the U(1) case; the similar property in the quantum field case meant that a globally U(1)-invariant Lagrangian could be converted immediately to a locally U(1)-invariant one by replacing ∂^μ by $\hat{D}^\mu$ (section 7.4).

In appendix D of volume 1 we introduced the idea of 'covariance' in the context of coordinate transformations of 3- and 4-vectors. The essential notion was of something 'maintaining the same form', or 'transforming the same way'. The transformations being considered here are gauge transformations rather than coordinate ones; nevertheless it is true that, under them, $D^\mu\psi$ transforms in the same way as ψ, while $\partial^\mu\psi$ does not. Thus the term covariant derivative seems appropriate. In fact, there is a much closer analogy between the 'coordinate' and the 'gauge' cases, which we did not present in volume 1, but give now in appendix N, for the interested reader.

We need the local SU(2) generalization of (13.4), appropriate to the local SU(2) transformation (13.2). Just as in the U(1) case (13.6), the ordinary gradient acting on $\psi^{(\frac{1}{2})}(x)$ does not transform in the same way as $\psi^{(\frac{1}{2})}(x)$: taking ∂^μ of (13.2) leads to

$$\begin{aligned}\partial^\mu\psi^{(\frac{1}{2})\prime}(x) &= \exp[\mathrm{i}g\boldsymbol{\tau}\cdot\boldsymbol{\alpha}(x)/2]\partial^\mu\psi^{(\frac{1}{2})}(x) \\ &\quad + \mathrm{i}g\boldsymbol{\tau}\cdot\partial^\mu\boldsymbol{\alpha}(x)/2\exp[\mathrm{i}g\boldsymbol{\tau}\cdot\boldsymbol{\alpha}(x)/2]\psi^{(\frac{1}{2})}(x) \qquad (13.8)\end{aligned}$$

as can be checked by writing the matrix exponential exp[A] as the series

$$\exp[\mathrm{A}] = \sum_{n=0}^{\infty}\mathrm{A}^n/n!$$

and differentiating term by term. By analogy with (13.7), the key property we demand for our *SU(2) covariant derivative* $D^\mu\psi^{(\frac{1}{2})}$ is that this quantity should transform like $\psi^{(\frac{1}{2})}$ – i.e. without the second term in (13.8). So we require

$$(D'^\mu\psi^{(\frac{1}{2})\prime}(x)) = \exp[\mathrm{i}g\boldsymbol{\tau}\cdot\boldsymbol{\alpha}(x)/2](D^\mu\psi^{(\frac{1}{2})}(x)). \qquad (13.9)$$

The definition of D^μ which generalizes (13.4) so as to fulfil this requirement is

$$D^\mu(\text{acting on an isospinor}) = \partial^\mu + \mathrm{i}g\boldsymbol{\tau}\cdot\boldsymbol{W}^\mu(x)/2. \qquad (13.10)$$

The definition (13.10), as indicated on the left-hand side, is only appropriate for isospinors $\psi^{(\frac{1}{2})}$; it has to be suitably generalized for other $\psi^{(t)}$'s (see (13.44)).

We now discuss (13.9) and (13.10) in detail. The ∂^μ is multiplied implicitly by the unit 2 matrix, and the $\boldsymbol{\tau}$'s act on the two-component space of $\psi^{(\frac{1}{2})}$. The $\boldsymbol{W}^\mu(x)$ are *three* independent gauge fields

$$\boldsymbol{W}^\mu = (W_1^\mu, W_2^\mu, W_3^\mu), \qquad (13.11)$$

generalizing the single electromagnetic gauge field A^μ. They are called SU(2) gauge fields, or more generally *Yang-Mills fields*. The term $\boldsymbol{\tau}\cdot\boldsymbol{W}^\mu$ is then the 2×2 matrix

$$\boldsymbol{\tau}\cdot\boldsymbol{W}^\mu = \begin{pmatrix} W_3^\mu & W_1^\mu - \mathrm{i}W_2^\mu \\ W_1^\mu + \mathrm{i}W_2^\mu & -W_3^\mu \end{pmatrix} \qquad (13.12)$$

using the τ's of (12.25); the x-dependence of the W^μ's is understood. Let us 'decode' the desired property (13.9), for the algebraically simpler case of an infinitesimal local SU(2) transformation with parameters $\boldsymbol{\epsilon}(x)$, which are of course functions of x since the transformation is local. In this case, $\psi^{(\frac{1}{2})}$ transforms by

$$\psi^{(\frac{1}{2})\prime} = (1 + \mathrm{i}g\boldsymbol{\tau}\cdot\boldsymbol{\epsilon}(x)/2)\psi^{(\frac{1}{2})} \qquad (13.13)$$

and the 'uncovariant' derivative $\partial^\mu\psi^{(\frac{1}{2})}$ transforms by

$$\partial^\mu\psi^{(\frac{1}{2})\prime} = (1 + \mathrm{i}g\boldsymbol{\tau}\cdot\boldsymbol{\epsilon}(x)/2)\partial^\mu\psi^{(\frac{1}{2})} + \mathrm{i}g\boldsymbol{\tau}\cdot\partial^\mu\boldsymbol{\epsilon}(x)/2\,\psi^{(\frac{1}{2})}, \qquad (13.14)$$

where we have retained only the terms linear in $\boldsymbol{\epsilon}$ from an expansion of (13.8) with $\boldsymbol{\alpha} \to \boldsymbol{\epsilon}$. We have now dropped the x-dependence of the $\psi^{(\frac{1}{2})}$'s, but kept that of $\boldsymbol{\epsilon}(x)$, and we have used the simple '1' for the unit matrix in the two-dimensional isospace. Equation (13.14) exhibits again an 'extra piece' on the right-hand side, as compared to (13.13). On the other hand, inserting (13.10) and (13.13) into our covariant derivative requirement (13.9) yields, for the left-hand side in the infinitesimal case,

$$D'^{\mu}\psi^{(\frac{1}{2})\prime} = (\partial^{\mu} + \mathrm{i}g\boldsymbol{\tau}\cdot\boldsymbol{W}'^{\mu}/2)[1 + \mathrm{i}g\boldsymbol{\tau}\cdot\boldsymbol{\epsilon}(x)/2]\psi^{(\frac{1}{2})} \tag{13.15}$$

while the right-hand side is

$$[1 + \mathrm{i}g\boldsymbol{\tau}\cdot\boldsymbol{\epsilon}(x)/2](\partial^{\mu} + \mathrm{i}g\boldsymbol{\tau}\cdot\boldsymbol{W}^{\mu}/2)\psi^{(\frac{1}{2})}. \tag{13.16}$$

In order to verify that these are the same, however, we would need to know $\boldsymbol{W}'^{\mu}$ – that is, the transformation law for the three $\boldsymbol{W}^{\mu}$ fields. Instead, we shall proceed 'in reverse', and use the *imposed* equality between (13.15) and (13.16) to determine the transformation law of $\boldsymbol{W}^{\mu}$.

Suppose that, under this infinitesimal transformation,

$$\boldsymbol{W}^{\mu} \to \boldsymbol{W}'^{\mu} = \boldsymbol{W}^{\mu} + \delta\boldsymbol{W}^{\mu}. \tag{13.17}$$

Then the condition of equality is

$$\begin{aligned}&[\partial^{\mu} + \mathrm{i}g\boldsymbol{\tau}/2\cdot(\boldsymbol{W}^{\mu} + \delta\boldsymbol{W}^{\mu})][1 + \mathrm{i}g\boldsymbol{\tau}\cdot\boldsymbol{\epsilon}(x)/2]\psi^{(\frac{1}{2})}\\ &\quad = [1 + \mathrm{i}g\boldsymbol{\tau}\cdot\boldsymbol{\epsilon}(x)/2](\partial^{\mu} + \mathrm{i}g\boldsymbol{\tau}\cdot\boldsymbol{W}^{\mu}/2)\psi^{(\frac{1}{2})}.\end{aligned} \tag{13.18}$$

Multiplying out the terms, neglecting the term of second order involving the product of $\delta\boldsymbol{W}^{\mu}$ and $\boldsymbol{\epsilon}$ and noting that

$$\partial^{\mu}(\boldsymbol{\epsilon}\psi) = (\partial^{\mu}\boldsymbol{\epsilon})\psi + \boldsymbol{\epsilon}(\partial^{\mu}\psi) \tag{13.19}$$

we see that many terms cancel and we are left with

$$\begin{aligned}\mathrm{i}g\frac{\boldsymbol{\tau}\cdot\delta\boldsymbol{W}^{\mu}}{2} \quad &= \quad -\mathrm{i}g\frac{\boldsymbol{\tau}\cdot\partial^{\mu}\boldsymbol{\epsilon}(x)}{2}\\ &\quad + (\mathrm{i}g)^2\left[\left(\frac{\boldsymbol{\tau}\cdot\boldsymbol{\epsilon}(x)}{2}\right)\left(\frac{\boldsymbol{\tau}\cdot\boldsymbol{W}^{\mu}}{2}\right) - \left(\frac{\boldsymbol{\tau}\cdot\boldsymbol{W}^{\mu}}{2}\right)\left(\frac{\boldsymbol{\tau}\cdot\boldsymbol{\epsilon}(x)}{2}\right)\right].\end{aligned} \tag{13.20}$$

Using the identity for Pauli matrices (see problem 3.4(b))

$$\boldsymbol{\sigma}\cdot\boldsymbol{a}\boldsymbol{\sigma}\cdot\boldsymbol{b} = \boldsymbol{a}\cdot\boldsymbol{b} + \mathrm{i}\boldsymbol{\sigma}\cdot\boldsymbol{a}\times\boldsymbol{b} \tag{13.21}$$

this yields

$$\boldsymbol{\tau}\cdot\delta\boldsymbol{W}^{\mu} = -\boldsymbol{\tau}\cdot\partial^{\mu}\boldsymbol{\epsilon}(x) - g\boldsymbol{\tau}\cdot(\boldsymbol{\epsilon}(x)\times\boldsymbol{W}^{\mu}). \tag{13.22}$$

Equating components of $\boldsymbol{\tau}$ on both sides, we deduce

$$\boxed{\delta \boldsymbol{W}^{\mu} = -\partial^{\mu}\boldsymbol{\epsilon}(x) - g[\boldsymbol{\epsilon}(x) \times \boldsymbol{W}^{\mu}].} \tag{13.23}$$

The reader may note the close similarity between these manipulations and those encountered in section 12.1.3.

Equation (13.23) defines the way in which the SU(2) gauge fields $\boldsymbol{W}^{\mu}$ transform under an infinitesimal SU(2) gauge transformation. If it were not for the presence of the first term $\partial^{\mu}\boldsymbol{\epsilon}(x)$ on the right-hand side, (13.23) would be simply the (infinitesimal) transformation law for the $T = 1$ triplet representation of SU(2) – see (12.64) and (12.65) in section 12.1.3. As mentioned at the end of section 12.2, the $T = 1$ representation is the 'adjoint', or 'regular', representation of SU(2), and this is the one to which gauge fields belong, in general. But there is the extra term $-\partial^{\mu}\boldsymbol{\epsilon}(x)$. Clearly this is directly analogous to the $-\partial^{\mu}\chi(x)$ term in the transformation of the U(1) gauge field A^{μ}; here, an independent infinitesimal function $\epsilon_i(x)$ is required for each component $W_i^{\mu}(x)$. If the ϵ's were independent of x, then $\partial^{\mu}\boldsymbol{\epsilon}(x)$ would of course vanish and the transformation law (13.23) would indeed be just that of an SU(2) triplet. Thus we can say that under global SU(2) transformations, the $\boldsymbol{W}^{\mu}$ behave as a normal triplet. But under *local* SU(2) transformations they acquire the additional $-\partial^{\mu}\boldsymbol{\epsilon}(x)$ piece, and thus no longer transform 'properly', as an SU(2) triplet. In exactly the same way, $\partial^{\mu}\psi^{(\frac{1}{2})}$ did not transform 'properly' as an SU(2) doublet, under a local SU(2) transformation, because of the second term in (13.14), which also involves $\partial^{\mu}\boldsymbol{\epsilon}(x)$. The remarkable result behind the fact that $D^{\mu}\psi^{(\frac{1}{2})}$ *does* transform 'properly' under local SU(2) transformations, is that the extra term in (13.23) precisely cancels that in (13.14)!

To summarize progress so far: we have shown that, for infinitesimal transformations, the relation

$$(D'^{\mu}\psi^{(\frac{1}{2})\prime}) = [1 + \mathrm{i}g\boldsymbol{\tau} \cdot \boldsymbol{\epsilon}(x)/2](D^{\mu}\psi^{(\frac{1}{2})}) \tag{13.24}$$

(where D^{μ} is given by (13.10)) holds true if in addition to the infinitesimal local SU(2) phase transformation on $\psi^{(\frac{1}{2})}$

$$\psi^{(\frac{1}{2})\prime} = [1 + \mathrm{i}g\boldsymbol{\tau} \cdot \boldsymbol{\epsilon}(x)/2]\psi^{(\frac{1}{2})} \tag{13.25}$$

the gauge fields transform according to

$$\boldsymbol{W}'^{\mu} = \boldsymbol{W}^{\mu} - \partial^{\mu}\boldsymbol{\epsilon}(x) - g[\boldsymbol{\epsilon}(x) \times \boldsymbol{W}^{\mu}]. \tag{13.26}$$

In obtaining these results, the form (13.10) for the covariant derivative has been assumed, and only the infinitesimal version of (13.2) has been treated explicitly. It turns out that (13.10) is still appropriate for the finite (non-infinitesimal) transformation (13.2), but the associated transformation law for the gauge fields is then slightly more complicated than (13.26). Let us write

$$\mathbf{U}(\boldsymbol{\alpha}(x)) \equiv \exp[\mathrm{i}g\boldsymbol{\tau} \cdot \boldsymbol{\alpha}(x)/2] \tag{13.27}$$

so that $\psi^{(\frac{1}{2})}$ transforms by

$$\psi^{(\frac{1}{2})\prime} = \mathbf{U}(\boldsymbol{\alpha}(x))\psi^{(\frac{1}{2})}. \tag{13.28}$$

Then we require

$$D'^{\mu}\psi^{(\frac{1}{2})\prime} = \mathbf{U}(\boldsymbol{\alpha}(x))D^{\mu}\psi^{(\frac{1}{2})}. \tag{13.29}$$

The left-hand side is

$$\begin{aligned} & (\partial^{\mu} + \mathrm{i}g\boldsymbol{\tau}\cdot\boldsymbol{W}'^{\mu}/2)\mathbf{U}(\boldsymbol{\alpha}(x))\psi^{(\frac{1}{2})} \\ = \; & (\partial^{\mu}\mathbf{U})\psi^{(\frac{1}{2})} + \mathbf{U}\partial^{\mu}\psi^{(\frac{1}{2})} + \mathrm{i}g\boldsymbol{\tau}\cdot\boldsymbol{W}'^{\mu}/2\,\mathbf{U}\psi^{(\frac{1}{2})}, \end{aligned} \tag{13.30}$$

while the right-hand side is

$$\mathbf{U}(\partial^{\mu} + \mathrm{i}g\boldsymbol{\tau}\cdot\boldsymbol{W}^{\mu}/2)\psi^{(\frac{1}{2})}. \tag{13.31}$$

The $\mathbf{U}\partial^{\mu}\psi^{(\frac{1}{2})}$ terms cancel leaving

$$(\partial^{\mu}\mathbf{U})\psi^{(\frac{1}{2})} + \mathrm{i}g\boldsymbol{\tau}\cdot\boldsymbol{W}'^{\mu}/2\,\mathbf{U}\psi^{(\frac{1}{2})} = \mathbf{U}\mathrm{i}g\boldsymbol{\tau}\cdot\boldsymbol{W}^{\mu}/2\,\psi^{(\frac{1}{2})}. \tag{13.32}$$

Since this has to be true for all (two-component) $\psi^{(\frac{1}{2})}$'s, we can treat it as an operator equation acting in the space of $\psi^{(\frac{1}{2})}$'s to give

$$\partial^{\mu}\mathbf{U} + \mathrm{i}g\boldsymbol{\tau}\cdot\boldsymbol{W}'^{\mu}/2\,\mathbf{U} = \mathbf{U}\mathrm{i}g\boldsymbol{\tau}\cdot\boldsymbol{W}^{\mu}/2, \tag{13.33}$$

or equivalently

$$\frac{1}{2}\boldsymbol{\tau}\cdot\boldsymbol{W}'^{\mu} = \frac{\mathrm{i}}{g}(\partial^{\mu}\mathbf{U})\mathbf{U}^{-1} + \mathbf{U}\frac{1}{2}\boldsymbol{\tau}\cdot\boldsymbol{W}^{\mu}\mathbf{U}^{-1}, \tag{13.34}$$

which defines the (finite) transformation law for SU(2) gauge fields. Problem 13.1 verifies that (13.34) reduces to (13.26) in the infinitesimal case $\boldsymbol{\alpha}(x) \to \boldsymbol{\epsilon}(x)$.

Suppose now that we consider a Dirac equation for $\psi^{(\frac{1}{2})}$:

$$(\mathrm{i}\gamma_{\mu}\partial^{\mu} - m)\psi^{(\frac{1}{2})} = 0 \tag{13.35}$$

where both the 'isospinor' components of $\psi^{(\frac{1}{2})}$ are four-component Dirac spinors. We assert that we can ensure *local SU(2) gauge covariance by replacing* ∂^{μ} *in this equation by the covariant derivative of* (13.10). Indeed, we have

$$\begin{aligned} \mathbf{U}(\boldsymbol{\alpha}(x))[\mathrm{i}\gamma_{\mu}D^{\mu} - m]\psi^{(\frac{1}{2})} &= \mathrm{i}\gamma_{\mu}\mathbf{U}(\boldsymbol{\alpha}(x))[D^{\mu}\psi^{(\frac{1}{2})}) - m\boldsymbol{U}(\boldsymbol{\alpha}(x)]\psi^{(\frac{1}{2})} \\ &= \mathrm{i}\gamma_{\mu}D'^{\mu}\psi^{(\frac{1}{2})\prime} - m\psi^{(\frac{1}{2})\prime} \end{aligned} \tag{13.36}$$

using equations (13.9) and (13.28). Thus if

$$(\mathrm{i}\gamma_{\mu}D^{\mu} - m)\psi^{(\frac{1}{2})} = 0 \tag{13.37}$$

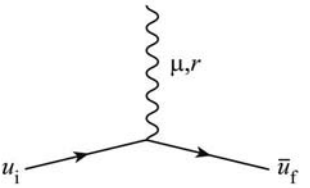


FIGURE 13.1
Vertex for isospinor-W interaction.

then

$$(\mathrm{i}\gamma_\mu D'^\mu - m)\psi^{(\frac{1}{2})\prime} = 0, \tag{13.38}$$

proving the asserted covariance. In the same way, any free particle wave equation satisfied by an 'isospinor' $\psi^{(\frac{1}{2})}$ – the relevant equation is determined by the Lorentz spin of the particles involved – can be made locally covariant by the use of the covariant derivative D^μ, just as in the U(1) case.

The essential point here, of course, is that the locally covariant form includes *interactions* between the $\psi^{(\frac{1}{2})}$'s and the gauge fields $\boldsymbol{W}^\mu$, which are determined by the local phase invariance requirement (the 'gauge principle'). Indeed, we can already begin to find some of the Feynman rules appropriate to tree graphs for SU(2) gauge theories. Consider again the case of an SU(2) isospinor fermion, $\psi^{(\frac{1}{2})}$, obeying equation (13.38). This can be written as

$$(\mathrm{i}\ \partial\!\!\!/ - m)\psi^{(\frac{1}{2})} = g(\boldsymbol{\tau}/2)\cdot \boldsymbol{W}\psi^{(\frac{1}{2})}. \tag{13.39}$$

In lowest-order perturbation theory the one-W emission/absorption process is given by the amplitude (cf (8.39)) for the electromagnetic case)

$$-\mathrm{i}g\int \bar{\psi}_\mathrm{f}^{(\frac{1}{2})}(\boldsymbol{\tau}/2)\gamma_\mu\psi_\mathrm{i}^{(\frac{1}{2})}\cdot\boldsymbol{W}^\mu \mathrm{d}^4x \tag{13.40}$$

exactly as advertized (for the field-theoretic vertex) in (12.129). The matrix degree of freedom in the τ's is sandwiched between the two-component isospinors $\psi^{(\frac{1}{2})}$; the γ matrix acts on the four-component (Dirac) parts of $\psi^{(\frac{1}{2})}$. The external $\boldsymbol{W}^\mu$ field is now specified by a spin-1 polarization vector ϵ^μ, like a photon, and by an 'SU(2) polarization vector' $a^r (r = 1, 2, 3)$ which tells us which of the three SU(2) W-states is participating. The Feynman rule for figure 13.1 is therefore

$$-\mathrm{i}g(\tau^r/2)\gamma_\mu \tag{13.41}$$

which is to be sandwiched between spinors/isospinors $u_\mathrm{i}, \bar{u}_\mathrm{f}$ and dotted into ϵ^μ and a^r. (13.41) is a very economical generalization of rule (ii) in Comment (3) of section 8.3.1.

The foregoing is easily generalized to SU(2) multiplets other than doublets. We shall change the notation slightly to use t instead of T for the 'isospin'

quantum number, so as to emphasize that it is *not* the hadronic isospin, for which we retain T; t will be the symbol used for the *weak isospin* to be introduced in chapter 20. The general local SU(2) transformation for a t-multiplet is then

$$\psi^{(t)} \to \psi^{(t)\prime} = \exp[\mathrm{i}g\boldsymbol{\alpha}(x) \cdot \mathbf{T}^{(t)}]\psi^{(t)} \tag{13.42}$$

where the $(2t+1) \times (2t+1)$ matrices $T_i^{(t)}$ $(i = 1, 2, 3)$ satisfy (cf (12.47))

$$[T_i^{(t)}, T_j^{(t)}] = \mathrm{i}\epsilon_{ijk} T_k^{(t)}. \tag{13.43}$$

The appropriate covariant derivative is

$$D^\mu = \partial^\mu + \mathrm{i}g\mathbf{T}^{(t)} \cdot \boldsymbol{W}^\mu \tag{13.44}$$

which is a $(2t+1) \times (2t+1)$ matrix acting on the $(2t+1)$ components of $\psi^{(t)}$. The gauge fields interact with such 'isomultiplets' in a *universal* way – only one g, the same for all the particles – which is prescribed by the local covariance requirement to be simply that interaction which is generated by the covariant derivatives. The fermion vertex corresponding to (13.44) is obtained by replacing $\boldsymbol{\tau}/2$ in (13.40) by $\boldsymbol{T}^{(t)}$.

We end this section with some comments:

(i) It is a remarkable fact that only one constant g is needed. This is *not* the same as in electromagnetism. There, each charged field interacts with the gauge field A^μ via a coupling whose strength is its charge $(e, -e, 2e, -5e \ldots)$. The crucial point is the appearance of the quadratic g^2 multiplying the *commutator* of the $\boldsymbol{\tau}$'s, $[\boldsymbol{\tau} \cdot \boldsymbol{\epsilon}, \boldsymbol{\tau} \cdot \boldsymbol{W}]$, in the $\boldsymbol{W}^\mu$ transformation (equation (13.20)). In the electromagnetic case, there is no such commutator – the associated U(1) phase group is Abelian. As signalled by the presence of g^2, a commutator is a non-linear quantity, and the scale of quantities appearing in such commutation relations is not arbitrary. It is an instructive exercise to check that, once $\delta\boldsymbol{W}^\mu$ is given by equation (13.23) – in the SU(2) case – then the g's appearing in $\psi^{(\frac{1}{2})\prime}$ (equation (13.13)) and $\psi^{(t)\prime}$ (via the infinitesimal version of equation (13.42)) must be the *same* as the one appearing in $\delta\boldsymbol{W}^\mu$.

(ii) According to the foregoing argument, it is actually a mystery why electric charge should be quantized. Since it is the coupling constant of an Abelian group, each charged field could have an arbitrary charge from this point of view: there are no commutators to fix the scale. This is one of the motivations of attempts to 'embed' the electromagnetic gauge transformations inside a larger non-Abelian group structure. Such is the case, for example, in 'grand unified theories' of strong, weak and electromagnetic interactions.

(iii) Finally we draw attention to the extremely important physical significance of the second term $\delta \boldsymbol{W}^\mu$ (equation (13.23)). The gauge fields themselves are not 'inert' as far as the gauge group is concerned: in the SU(2) case they have 'isospin' 1, while for a general group they belong to the regular representation of the group. This is profoundly different from the electromagnetic case, where the gauge field A^μ for the photon is of course uncharged: quite simply, $e = 0$ for a photon, and the second term in (13.23) is absent for A^μ. The fact that non-Abelian (Yang-Mills) gauge fields carry non-Abelian 'charge' degrees of freedom means that, since they are also the quanta of the force field, *they will necessarily interact with themselves*. Thus a non-Abelian gauge theory of gauge fields alone, with no 'matter' fields, has non-trivial interactions and is not a free theory.

We shall examine the form of these 'self-interactions' in section 13.3.2. First, we need to find the equivalent, for the Yang-Mills field, of the Maxwell field strength tensor $F^{\mu\nu}$, which gave us the gauge-invariant formulation of Maxwell's equations, and in terms of which the Maxwell Lagrangian can be immediately written down.

13.1.2 The non-Abelian field strength tensor

A simple way of arriving at the desired quantity is to consider the commutator of two covariant derivatives, as we can see by calculating it for the U(1) case. We find

$$[D^\mu, D^\nu]\,\psi \equiv (D^\mu D^\nu - D^\nu D^\mu)\psi = \mathrm{i}eF^{\mu\nu}\psi \tag{13.45}$$

as is verified in problem 13.2. Equation (13.45) suggests that we will find the SU(2) analogue of $F^{\mu\nu}$ by evaluating

$$[D^\mu, D^\nu]\,\psi^{(\frac{1}{2})} \tag{13.46}$$

where as usual

$$D^\mu(\text{on } \psi^{(\frac{1}{2})}) = \partial^\mu + \mathrm{i}g\boldsymbol{\tau}\cdot\boldsymbol{W}^\mu/2. \tag{13.47}$$

Problem 13.3 confirms that the result is

$$[D^\mu, D^\nu]\,\psi^{(\frac{1}{2})} = \mathrm{i}g\boldsymbol{\tau}/2\cdot(\partial^\mu\boldsymbol{W}^\nu - \partial^\nu\boldsymbol{W}^\mu - g\boldsymbol{W}^\mu\times\boldsymbol{W}^\nu)\,\psi^{(\frac{1}{2})}; \tag{13.48}$$

the manipulations are very similar to those in (13.20)–(13.23). Noting the analogy between the right-hand side of (13.48) and (13.45), we accordingly expect the SU(2) 'curvature' or field strength tensor, to be given by

$$\boldsymbol{F}^{\mu\nu} = \partial^\mu\boldsymbol{W}^\nu - \partial^\nu\boldsymbol{W}^\mu - g\boldsymbol{W}^\mu\times\boldsymbol{W}^\nu \tag{13.49}$$

or, in component notation,

$$F_i^{\mu\nu} = \partial^\mu W_i^\nu - \partial^\nu W_i^\mu - g\epsilon_{ijk}W_j^\mu W_k^\nu. \tag{13.50}$$

This tensor is of fundamental importance in a (non-Abelian) gauge theory. Since it arises from the commutator of two gauge-covariant derivatives, we are guaranteed that it itself is gauge covariant – that is to say, 'it transforms under local SU(2) transformations in the way its SU(2) structure would indicate'. Now $\boldsymbol{F}^{\mu\nu}$ has clearly three SU(2) components and must be an SU(2) triplet: indeed, it is true that under an infinitesimal local SU(2) transformation

$$\boldsymbol{F}'^{\mu\nu} = \boldsymbol{F}^{\mu\nu} - g\boldsymbol{\epsilon}(x) \times \boldsymbol{F}^{\mu\nu} \tag{13.51}$$

which is the expected law (cf (12.64)) for an SU(2) triplet. Problem 13.4 verifies that (13.51) follows from (13.49) and the transformation law (13.23) for the $\boldsymbol{W}^{\mu}$ fields. Note particularly that $\boldsymbol{F}^{\mu\nu}$ transforms 'properly', as an SU(2) triplet should, *without* the ∂^{μ} part which appears in $\delta\boldsymbol{W}^{\mu}$.

This non-Abelian $\boldsymbol{F}^{\mu\nu}$ is a much more interesting object than the Abelian $F^{\mu\nu}$ (which is actually U(1)-gauge *invariant*, of course: $F'^{\mu\nu} = F^{\mu\nu}$). $\boldsymbol{F}^{\mu\nu}$ contains the gauge coupling constant g, confirming (cf comment(c) in section 13.1.1) that the gauge fields themselves carry SU(2) 'charge', and act as sources for the field strength. Appendix N shows how these field strength tensors may be regarded as analogous to geometrical curvatures.

It is now straightforward to move to the quantum field case and construct the SU(2) Yang-Mills analogue of the Maxwell Lagrangian $-\frac{1}{4}\hat{F}_{\mu\nu}\hat{F}^{\mu\nu}$. It is simply $-\frac{1}{4}\hat{\boldsymbol{F}}_{\mu\nu} \cdot \hat{\boldsymbol{F}}^{\mu\nu}$, the SU(2) 'dot product' ensuring SU(2) invariance (see problem 13.5), even under *local* transformation, in view of the transformation law (13.51). But before proceeding in this way we first need to introduce local SU(3) symmetry.

13.2 Local SU(3) Symmetry

Using what has been done for global SU(3) symmetry in section 12.2, and the preceding discussion of how to make a global SU(2) into a local one, it is straightforward to develop the corresponding theory of local SU(3). This is the gauge group of QCD, the three degrees of freedom of the fundamental quark triplet now referring to 'colour', as will be further discussed in chapter 14. We denote the basic triplet by ψ, which transforms under a local SU(3) transformation according to

$$\psi' = \exp[\mathrm{i}g_{\mathrm{s}}\boldsymbol{\lambda} \cdot \boldsymbol{\alpha}(x)/2]\psi, \tag{13.52}$$

which is the same as the global transformation (12.74) but with the 8 constant parameters $\boldsymbol{\alpha}$ replaced by x-dependent ones, and with a coupling strength g_{s} inserted. The SU(3)-covariant derivative, when acting on an SU(3) triplet ψ, is given by the indicated generalization of (13.10), namely

$$D^{\mu}(\text{acting on SU(3) triplet}) = \partial^{\mu} + \mathrm{i}g_{\mathrm{s}}\boldsymbol{\lambda}/2 \cdot \boldsymbol{A}^{\mu} \tag{13.53}$$

where $A_1^\mu, A_2^\mu, \ldots A_8^\mu$ are eight gauge fields which are called *gluons*. The coupling is denoted by 'g_s' in anticipation of the application to strong interactions via QCD.

The infinitesimal version of (13.52) is (cf (13.13))

$$\psi' = (1 + \mathrm{i}g_\mathrm{s}\boldsymbol{\lambda} \cdot \boldsymbol{\eta}(x)/2)\psi \tag{13.54}$$

where '1' stands for the unit matrix in the three-dimensional space of components of the triplet ψ. As in (13.14), it is clear that $\partial^\mu\psi'$ will involve an 'unwanted' term $\partial^\mu\boldsymbol{\eta}(x)$. By contrast, the desired covariant derivative $D^\mu\psi$ should transform according to

$$D'^\mu\psi' = (1 + \mathrm{i}g_\mathrm{s}\boldsymbol{\lambda} \cdot \boldsymbol{\eta}(x)/2)D^\mu\psi \tag{13.55}$$

without the $\partial^\mu\boldsymbol{\eta}(x)$ term. Problem 13.6 verifies that this is fulfilled by having the gauge fields transform by

$$A_a'^\mu = A_a^\mu - \partial^\mu\eta_a(x) - g_\mathrm{s}f_{abc}\eta_b(x)A_c^\mu. \tag{13.56}$$

Comparing (13.56) with (12.80) we can identify the term in f_{abc} as telling us that the 8 fields A_a^μ transform as an SU(3) octet, the η's now depending on x, of course. This is the adjoint, or regular representation of SU(3), as we have now come to expect for gauge fields. However, the $\partial^\mu\eta_a(x)$ piece spoils this simple transformation property under local transformations. But it *is* just what is needed to cancel the corresponding $\partial^\mu\boldsymbol{\eta}(x)$ term in $\partial^\mu\psi'$, leaving $D^\mu\psi$ transforming as a proper triplet via (13.55). The finite version of (13.56) can be derived as in section 13.1 for SU(2), but we shall not need the result here.

As in the SU(2) case, the free Dirac equation for an SU(3)-triplet ψ,

$$(\mathrm{i}\gamma_\mu\partial^\mu - m)\psi = 0, \tag{13.57}$$

can be 'promoted' into one which is covariant under local SU(3) transformations by replacing ∂^μ by D^μ of (13.53), leading to

$$(\mathrm{i}\not{\partial} - m)\psi = g_\mathrm{s}\boldsymbol{\lambda}/2 \cdot \not{\boldsymbol{A}}\psi \tag{13.58}$$

(compare (13.39)). This leads immediately to the one gluon emission amplitude (see figure 13.2)

$$-\mathrm{i}g_\mathrm{s}\int \bar{\psi}_\mathrm{f}\boldsymbol{\lambda}/2\gamma^\mu\psi_\mathrm{i} \cdot \boldsymbol{A}_\mu \mathrm{d}^4x \tag{13.59}$$

as already suggested in section 12.3.1: the SU(3) current of (12.133) – but this time in *colour* space – is 'dotted' with the gauge field. The Feynman rule for figure 13.2 is therefore

$$-\mathrm{i}g_\mathrm{s}\lambda_a/2\ \gamma^\mu. \tag{13.60}$$

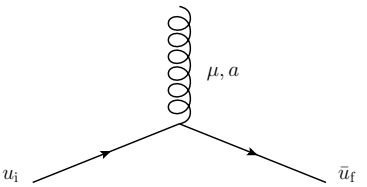


FIGURE 13.2
Quark-gluon vertex.

The SU(3) field strength tensor can be calculated by evaluating the commutator of two D's of the form (13.53); the result (problem 13.7) is

$$F_a^{\mu\nu} = \partial^\mu A_a^\nu - \partial^\nu A_a^\mu - g_{\rm s} f_{abc} A_b^\mu A_c^\nu \tag{13.61}$$

which is closely analogous to the SU(2) case (13.50) (the structure constants of SU(2) are given by $\mathrm{i}\epsilon_{ijk}$, and of SU(3) by $\mathrm{i}f_{abc}$). Once again, the crucial property of $F_a^{\mu\nu}$ is that, under *local* SU(3) transformations it develops no '$\partial^\mu \eta_a$' part, but transforms as a 'proper' octet:

$$F_a^{\mu\nu} = F_a^{\mu\nu} - g_{\rm s} f_{abc} \eta_b(x) F_c^{\mu\nu}. \tag{13.62}$$

This allows us to write down a locally SU(3)-invariant analogue of the Maxwell Lagrangian

$$-\frac{1}{4} F_a^{\mu\nu} F_{a\mu\nu} \tag{13.63}$$

by dotting the two octets together.

It is now time to consider locally SU(2)- and SU(3)-invariant quantum field Lagrangians and, in particular, the resulting self-interactions among the gauge quanta.

13.3 Local non-Abelian symmetries in Lagrangian quantum field theory

13.3.1 Local SU(2) and SU(3) Lagrangians

We consider here only the particular examples relevant to the strong and electroweak interactions of quarks: namely, a (weak) SU(2) doublet of fermions interacting with SU(2) gauge fields W_i^μ, and a (strong) SU(3) triplet of fermions interacting with the gauge fields A_a^μ. We follow the same steps as in the U(1) case of chapter 7, noting again that for quantum fields the sign of the exponents in (13.2) and (13.52) is reversed, by convention; thus (12.89) is replaced

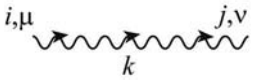


FIGURE 13.3
SU(2) gauge-boson propagator.

by its local version

$$\hat{q}' = \exp(-\mathrm{i}g\hat{\boldsymbol{\alpha}}(x)\cdot\boldsymbol{\tau}/2)\hat{q} \tag{13.64}$$

and (12.132) by

$$\hat{q}' = \exp(-\mathrm{i}g_{\mathrm{s}}\hat{\boldsymbol{\alpha}}(x)\cdot\boldsymbol{\lambda}/2)\hat{q}. \tag{13.65}$$

Correspondingly, the $\boldsymbol{\epsilon}$ in (13.23) and the η's in (13.56) become field operators, with a reversal of sign.

The globally SU(2)-invariant Lagrangian (12.87) becomes locally SU(2)-invariant if we replaced ∂^μ by D^μ of (13.10), with W^μ now a quantum field:

$$\begin{aligned}\hat{\mathcal{L}}_{\mathrm{D,local\ SU(2)}} &= \bar{\hat{q}}(\mathrm{i}\,\hat{\not{D}} - m)\hat{q} \\ &= \bar{\hat{q}}(\mathrm{i}\,\not{\partial} - m)\hat{q} - g\bar{\hat{q}}\gamma^\mu\boldsymbol{\tau}/2\hat{q}\cdot\hat{\boldsymbol{W}}_\mu\end{aligned} \tag{13.66}$$

with an interaction of the form 'symmetry current (12.109) dotted into the gauge field'. To this we must add the SU(2) Yang-Mills term

$$\mathcal{L}_{\mathrm{Y-M,SU(2)}} = -\frac{1}{4}\hat{\boldsymbol{F}}_{\mu\nu}\cdot\hat{\boldsymbol{F}}^{\mu\nu} \tag{13.67}$$

to get the local SU(2) analogue of $\mathcal{L}_{\mathrm{QED}}$. It is *not* possible to add a mass term for the gauge fields of the form $\frac{1}{2}\hat{\boldsymbol{W}}^\mu\cdot\hat{\boldsymbol{W}}_\mu$, since such a term would not be invariant under the gauge transformations (13.26) or (13.34) of the W-fields. Thus, just as in the U(1) (electromagnetic) case, the W-quanta of this theory are *massless*. We presumably also need a gauge-fixing term for the gauge fields, as in section 7.3.2, which we can take to be[1]

$$\mathcal{L}_{\mathrm{gf}} = -\frac{1}{2\xi}\left(\partial_\mu\hat{\boldsymbol{W}}^\mu\cdot\partial_\nu\hat{\boldsymbol{W}}^\nu\right). \tag{13.68}$$

The Feynman rule for the fermion-W vertex is then the same as already given in (13.41), while the W-propagator is (figure 13.3)

$$\frac{\mathrm{i}\left[-g^{\mu\nu}+(1-\xi)k^\mu k^\nu/k^2\right]}{k^2+\mathrm{i}\epsilon}\,\delta^{ij}. \tag{13.69}$$

Before proceeding to the SU(3) case, we must now emphasize three respects

[1] We shall see in section 13.5.3 that in the non-Abelian case this gauge-fixing term does *not* completely solve the problem of quantizing such gauge fields; however, it is adequate for tree graphs.

in which our local SU(2) Lagrangian is not suitable (yet) for describing weak interactions. First, weak interactions violate parity, in fact 'maximally', by which is meant that only the 'left-handed' part $\hat{\psi}_{\mathrm{L}}$ of the fermion field enters the interactions with the $\boldsymbol{W}^{\mu}$ fields, where $\hat{\psi}_{\mathrm{L}} \equiv \left(\frac{1-\gamma_5}{2}\right)\hat{\psi}$; for this reason the weak isospin group is called $\mathrm{SU(2)_L}$. Secondly, the physical $\mathrm{W}^{\pm}$ are of course not massless, and therefore cannot be described by propagators of the form (13.69). And thirdly, the *fermion* mass term violates the 'left-handed' SU(2) gauge symmetry, as the discussion in section 12.3.2 shows. In this case, however, the chiral symmetry which is broken by fermion masses in the Lagrangian is a local, or gauge, symmetry (in section 12.3.2 the chiral flavour symmetry was a global symmetry). If we want to preserve the chiral gauge symmetry $\mathrm{SU(2)_L}$ – and it is necessary for renormalizability – then we shall have to replace the simple fermion mass term in (13.66) by something else, as will be explained in chapter 22.

The locally $\mathrm{SU(3)_c}$-invariant Lagrangian for one quark triplet (cf (12.137))

$$\hat{q}_{\mathrm{f}} = \begin{pmatrix} \hat{f}_{\mathrm{r}} \\ \hat{f}_{\mathrm{b}} \\ \hat{f}_{\mathrm{g}} \end{pmatrix}, \tag{13.70}$$

where 'f' stands for 'flavour', and 'r, b, and g' for 'red, blue, and green', is

$$\bar{\hat{q}}_{\mathrm{f}}(\mathrm{i}\hat{\not{D}} - m_{\mathrm{f}})\hat{q}_{\mathrm{f}} - \frac{1}{4}\hat{F}_{a\mu\nu}\hat{F}_a^{\mu\nu} - \frac{1}{2\xi}(\partial_\mu \hat{A}_a^{\mu})(\partial_\nu \hat{A}_a^{\nu}) \tag{13.71}$$

where $\hat{D}^{\mu}$ is given by (13.53) with $\boldsymbol{A}^{\mu}$ replaced by $\hat{\boldsymbol{A}}^{\mu}$, and the footnote before equation (13.68) also applies here. This leads to the interaction term (cf (13.59))

$$-g_{\mathrm{s}}\bar{\hat{q}}_{\mathrm{f}}\gamma^{\mu}\boldsymbol{\lambda}/2\hat{q}_{\mathrm{f}} \cdot \hat{\boldsymbol{A}}_{\mu} \tag{13.72}$$

and the Feynman rule (13.60) for figure 13.2. Once again, the gluon quanta must be *massless*, and their propagator is the same as (13.69), with $\delta_{ij} \to \delta_{ab}$ $(a, b = 1, 2, \ldots 8)$. The different quark flavours are included by simply repeating the first term of (13.71) for all flavours:

$$\sum_{\mathrm{f}} \bar{\hat{q}}_{\mathrm{f}}(\mathrm{i}\hat{\not{D}} - m_{\mathrm{f}})\hat{q}_{\mathrm{f}}, \tag{13.73}$$

which incorporates the hypothesis that the $\mathrm{SU(3)_c}$-gauge interaction is 'flavour-blind', i.e. exactly the same for each flavour. Note that although the flavour masses are different, the masses of different 'coloured' quarks of the same flavour are the same ($m_{\mathrm{u}} \neq m_{\mathrm{d}}, m_{\mathrm{u,r}} = m_{\mathrm{u,b}} = m_{\mathrm{u,g}}$).

The Lagrangians (13.66)–(13.68), and (13.71), though easily written down after all this preparation, are unfortunately not adequate for anything but tree graphs. We shall indicate why this is so in section 13.3.3. Before that, we want to discuss in more detail the nature of the gauge-field self-interactions contained in the Yang-Mills pieces.

13.3.2 Gauge field self-interactions

We start by pointing out an interesting ambiguity in the prescription for 'covariantizing' wave equations which we have followed, namely 'replace ∂^μ by D^μ'. Suppose we wished to consider the electromagnetic interactions of charged massless spin-1 particles, call them X's, carrying charge e. The standard wave equation for such free massless vector particles would be the same as for A^μ, namely

$$\Box X^\mu - \partial^\mu \partial^\nu X_\nu = 0. \tag{13.74}$$

To 'covariantize' this (i.e. introduce the electromagnetic coupling) we would replace ∂^μ by $D^\mu = \partial^\mu + \mathrm{i}eA^\mu$ so as to obtain

$$D^2 X^\mu - D^\mu D^\nu X_\nu = 0. \tag{13.75}$$

But this procedure is not unique: if we had started from the perfectly equivalent wave equation

$$\Box X^\mu - \partial^\nu \partial^\mu X_\nu = 0 \tag{13.76}$$

we would have arrived at

$$D^2 X^\mu - D^\nu D^\mu X_\nu = 0 \tag{13.77}$$

which is not the same as (13.75), since (cf (13.45))

$$[D^\mu, D^\nu] = \mathrm{i}eF^{\mu\nu}. \tag{13.78}$$

The simple prescription $\partial^\mu \to D^\mu$ has, in this case, failed to produce a unique wave equation. We can allow for this ambiguity by introducing an arbitrary parameter δ in the wave equation, which we write as

$$D^2 X^\mu - D^\nu D^\mu X_\nu + \mathrm{i}e\delta F^{\mu\nu} X_\nu = 0. \tag{13.79}$$

The δ term in (13.79) contributes to the magnetic moment coupling of the X-particle to the electromagnetic field, and is called the 'ambiguous magnetic moment'. Just such an ambiguity would seem to arise in the case of the charged weak interaction quanta $\mathrm{W}^\pm$ (their masses do not affect this argument). For the photon itself, of course, $e = 0$ and there is no such ambiguity.

It is important to be clear that (13.79) is fully U(1) gauge-covariant, so that δ cannot be fixed by further appeal to the local U(1) symmetry. Moreover, it turns out that the theory for arbitrary δ is *not renormalizable* (though we shall not show this here): thus the quantum electrodynamics of charged massless vector bosons is in general non-renormalizable.

However, the theory *is* renormalizable if – to continue with the present terminology – the photon, the X-particle, and its antiparticle the $\bar{\mathrm{X}}$ are the members of an SU(2) gauge triplet (like the W's), with gauge coupling constant e. This is, indeed, very much how the photon and the $\mathrm{W}^\pm$ are 'unified', but there is a complication (as always!) in that case, having to do with the

necessity for finding room in the scheme for the neutral weak boson Z^0 as well. We shall see how this works in chapter 19; meanwhile we continue with this $\mathrm{X}-\gamma$ model. We shall show that when the $\mathrm{X}-\gamma$ interaction contained in (13.79) is regarded as a $3-\mathrm{X}$ vertex in a local SU(2) gauge theory, the value of δ has to equal 1; for this value the theory is renormalizable. In this interpretation, the X^μ wave function is identified with '$\frac{1}{\sqrt{2}}(X_1^\mu + \mathrm{i}X_2^\mu)$' and $\bar{X}^\mu$ with '$\frac{1}{\sqrt{2}}(X_1^\mu - \mathrm{i}X_2^\mu)$' in terms of components of the SU(2) triplet X_i^μ, while A^μ is identified with X_3^μ.

Consider then equation (13.79) written in the form[2]

$$\Box X^\mu - \partial^\nu \partial^\mu X_\nu = \hat{V} X^\mu \tag{13.80}$$

where

$$\begin{aligned} \hat{V}X^\mu \;=\; & -\mathrm{i}e\,\{[\partial^\nu(A_\nu X^\mu) + A^\nu \partial_\nu X^\mu] \\ & -(1+\delta)\,[\partial^\nu(A^\mu X_\nu) + A^\nu \partial^\mu X_\nu] \\ & +\delta\,[\partial^\mu(A^\nu X_\nu) + A^\mu \partial^\nu X_\nu]\,\}, \end{aligned} \tag{13.81}$$

and we have dropped terms of $O(e^2)$ which appear in the 'D^2' term; we shall come back to them later. The terms inside the { } brackets have been written in such a way that each [ ] bracket has the structure

$$\partial(AX) + A(\partial X) \tag{13.82}$$

which will be convenient for the following evaluation.

The lowest-order ($O(e)$) perturbation theory amplitude for '$\mathrm{X} \to \mathrm{X}$' under the potential $\hat{V}$ is then

$$-\mathrm{i}\int X_\mu^*(\mathrm{f})\hat{V}X^\mu(\mathrm{i})\mathrm{d}^4x. \tag{13.83}$$

Inserting (13.81) into (13.83) clearly gives something involving two 'X'-wave-functions and one 'A' one, i.e. a triple-X vertex (with $A^\mu \equiv X_3^\mu$), shown in figure 13.4. To obtain the rule for this vertex from (13.83), consider the first [ ] bracket in (13.81). It contributes

$$-\mathrm{i}(-\mathrm{i}e)\int X_\mu^*(2)\{\partial^\nu(X_{3\nu}(3)X^\mu(1)) + X_3^\nu(3)\partial_\nu X^\mu(1)\}\mathrm{d}^4x \tag{13.84}$$

where the (1), (2), (3) refer to the momenta as shown in figure 13.4, and for reasons of symmetry are all taken to be ingoing; thus

$$X_3^\mu(3) = \epsilon_3^\mu \exp(-\mathrm{i}k_3 \cdot x) \tag{13.85}$$

[2]The sign chosen for $\hat{V}$ here apparently *differs* from that in the KG case (3.101), but it does agree when allowance is made, in the amplitude (13.83), for the fact that the dot product of the polarization vectors is negative (cf (7.87)).

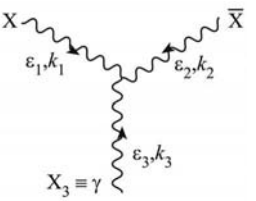


FIGURE 13.4
Triple-X vertex.

for example. The first term in (13.84) can be easily evaluated by a partial integration to turn the ∂^ν onto the $X^*_\mu(2)$, while in the second term ∂_ν acts straightforwardly on $X^\mu(1)$. Omitting the usual $(2\pi)^4\ \delta^4$ energy-momentum conserving factor, we find (problem 13.8) that (13.84) leads to the amplitude

$$\mathrm{i}e\epsilon_1 \cdot \epsilon_2\ (k_1 - k_2) \cdot \epsilon_3. \tag{13.86}$$

In a similar way, the other terms in (13.83) give

$$-\mathrm{i}e\delta(\epsilon_1 \cdot \epsilon_3\ \epsilon_2 \cdot k_2 - \epsilon_2 \cdot \epsilon_3\ \epsilon_1 \cdot k_1) \tag{13.87}$$

and

$$+\mathrm{i}e(1+\delta)(\epsilon_2 \cdot \epsilon_3\ \epsilon_1 \cdot k_2 - \epsilon_1 \cdot \epsilon_3\ \epsilon_2 \cdot k_1). \tag{13.88}$$

Adding all the terms up and using the 4-momentum conservation condition

$$k_1 + k_2 + k_3 = 0 \tag{13.89}$$

we obtain the vertex

$$+\mathrm{i}e\{\epsilon_1 \cdot \epsilon_2\ (k_1 - k_2) \cdot \epsilon_3 + \epsilon_2 \cdot \epsilon_3\ (\delta k_2 - k_3) \cdot \epsilon_1 + \epsilon_3 \cdot \epsilon_1\ (k_3 - \delta k_1) \cdot \epsilon_2\}. \tag{13.90}$$

It is quite evident from (13.90) that the value $\delta = 1$ has a privileged role, and we strongly suspect that this will be the value selected by the proposed SU(2) gauge symmetry of this model. We shall check this in two ways: in the first, we consider a 'physical' process involving the vertex (13.90), and show how requiring it to be SU(2)-gauge invariant fixes δ to be 1; in the second, we 'unpack' the relevant vertex from the compact Yang-Mills Lagrangian $-\frac{1}{4}\hat{\boldsymbol{X}}_{\mu\nu} \cdot \hat{\boldsymbol{X}}^{\mu\nu}$.

The process we shall choose is $\mathrm{X} + \mathrm{d} \to \mathrm{X} + \mathrm{d}$ where d is a fermion (which we call a quark) transforming as the $T_3 = -\frac{1}{2}$ component of a doublet under the SU(2) gauge group, its $T_3 = +\frac{1}{2}$ partner being the u. There are two contributing Feynman graphs, shown in figure 13.5(a) and (b). Consider first the amplitude for figure 13.5(a). We use the rule of figure 13.1, with the τ-matrix combination $\tau_+ = (\tau_1 + \mathrm{i}\tau_2)/\sqrt{2}$ corresponding to the absorption of

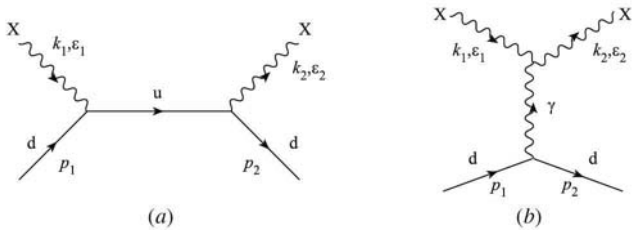


FIGURE 13.5
Tree graphs contributing to $\mathrm{X} + \mathrm{d} \to \mathrm{X} + \mathrm{d}$.

the positively charged X, and $\tau_- = (\tau_1 - \mathrm{i}\tau_2)/\sqrt{2}$ for the emission of the X. Then figure 13.5(a) is

$$(-\mathrm{i}e)^2 \bar{\psi}^{(\frac{1}{2})}(p_2) \frac{\tau_-}{2} \not{\epsilon}_2 \frac{\mathrm{i}}{\not{p}_1 + \not{k}_1 - m} \frac{\tau_+}{2} \not{\epsilon}_1 \psi^{(\frac{1}{2})}(p_1) \tag{13.91}$$

where

$$\psi^{(\frac{1}{2})} = \begin{pmatrix} u \\ d \end{pmatrix}, \tag{13.92}$$

and we have chosen real polarization vectors. Using the explicit forms (12.25) for the τ-matrices, (13.91) becomes

$$(-\mathrm{i}e)^2 \bar{d}(p_2) \frac{1}{\sqrt{2}} \not{\epsilon}_2 \frac{\mathrm{i}}{\not{p}_1 + \not{k}_1 - m} \frac{1}{\sqrt{2}} \not{\epsilon}_1 d(p_1). \tag{13.93}$$

We must now discuss how to implement gauge invariance. In the QED case of electron Compton scattering (section 8.6.2), the test of gauge invariance was that the amplitude should vanish if any photon polarization vector $\epsilon^\mu(k)$ was replaced by k^μ – see (8.165). This requirement was derived from the fact that a gauge transformation on the photon A^μ took the form $A^\mu \to A'^\mu = A^\mu - \partial^\mu \chi$, so that, consistently with the Lorentz condition, ϵ^μ could be replaced by $\epsilon'^\mu = \epsilon^\mu + \beta k^\mu$ (cf 8.163) without changing the physics. But the SU(2) analogue of the U(1) gauge transformation is given by (13.26), for infinitesimal ϵ's, and although there is indeed an analogous '$-\partial^\mu \boldsymbol{\epsilon}$' part, there is also an additional part (with $g \to e$ in our case) expressing the fact that the X's carry SU(2) charge. However this extra part does involve the coupling e. Hence, if we were to make the *full* change corresponding to (13.26) in a tree graph of order e^2, the extra part would produce a term of order e^3. We shall take the view that gauge invariance should hold at each order of perturbation theory separately; thus we shall demand that the tree graphs for X-d scattering, for example, should be invariant under $\epsilon^\mu \to k^\mu$ for any ϵ.

The replacement $\epsilon_1 \to k_1$ in (13.93) produces the result (problem 13.9)

$$(-\mathrm{i}e)^2 \frac{\mathrm{i}}{2} \bar{d}(p_2) \not{\epsilon}_2 d(p_1) \tag{13.94}$$

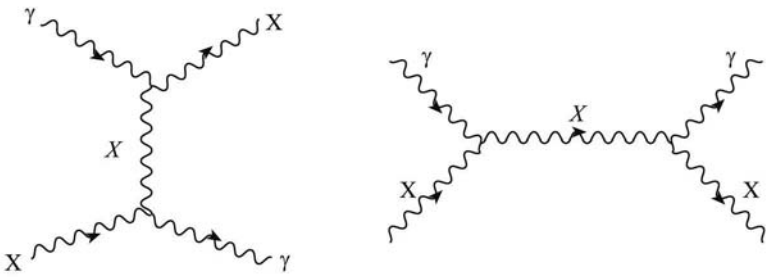


FIGURE 13.6
Tree graphs contributing to $\gamma + \mathrm{X} \to \gamma + \mathrm{X}$.

where we have used the Dirac equation for the quark spinors of mass m. The term (13.94) is certainly not zero, but we must of course also include the amplitude for figure 13.5(b). Using the vertex of (13.90) with suitable sign changes of momenta, and the photon propagator of (7.119), and remembering that d has $\tau_3 = -1$, the amplitude for figure 13.5(b) is

$$\begin{aligned} \mathrm{i}e[\epsilon_1 \cdot \epsilon_2 \ (k_1 + k_2)_\mu \quad &+ \quad \epsilon_{2\mu}\epsilon_1 \cdot (-\delta k_2 - k_2 + k_1) + \epsilon_{1\mu}\epsilon_2 \cdot (k_2 - k_1 - \delta k_1)] \\ &\times \quad \frac{-\mathrm{i}g^{\mu\nu}}{q^2} \times [-\mathrm{i}e\bar{d}(p_2)\left(-\frac{1}{2}\right)\gamma_\nu d(p_1)], \end{aligned} \tag{13.95}$$

where $q^2 = (k_1 - k_2)^2 = -2k_1 \cdot k_2$ using $k_1^2 = k_2^2 = 0$, and where the ξ-dependent part of the γ-propagator vanishes since $\bar{d}(p_2) \not{q} d(p_1) = 0$. We now leave it as an exercise (problem 13.10) to verify that, when $\epsilon_1 \to k_1$ in (13.95), the resulting amplitude does exactly cancel the contribution (13.94), *provided that* $\delta = 1$. Thus the $\mathrm{X} - \bar{\mathrm{X}} - \gamma$ vertex is, assuming the SU(2) gauge symmetry,

$$\mathrm{i}e[\epsilon_1 \cdot \epsilon_2 \ (k_1 - k_2) \cdot \epsilon_3 + \epsilon_2 \cdot \epsilon_3 \ (k_2 - k_3) \cdot \epsilon_1 + \epsilon_3 \cdot \epsilon_1 \ (k_3 - k_1) \cdot \epsilon_2]. \tag{13.96}$$

The verification of this non-Abelian gauge invariance to order e^2 is, of course, not a proof that the entire theory of massless X quanta, γ's and quark isospinors will be gauge invariant if $\delta = 1$. Indeed, having obtained the $\mathrm{X} - \mathrm{X} - \gamma$ vertex, we immediately have something new to check: we can see if the lowest-order $\gamma - \mathrm{X}$ scattering amplitude is gauge invariant. The $\mathrm{X} - \mathrm{X} - \gamma$ vertex will generate the $O(e^2)$ graphs shown in figure 13.6, and the dedicated reader may check that the sum of these amplitudes is *not* gauge invariant, again in the (tree-graph) sense of not vanishing when any ϵ is replaced by the corresponding k. But this is actually correct. In obtaining the $\mathrm{X} - \mathrm{X} - \gamma$ vertex we dropped an $O(e^2)$ term involving the three fields A, A and X, in going from (13.81) to (13.90): this will generate an $O(e^2)$ $\gamma - \gamma - \mathrm{X} - \mathrm{X}$ interaction, figure 13.7, when used in lowest-order perturbation theory. One can find the amplitude for figure 13.7 by the gauge invariance requirement

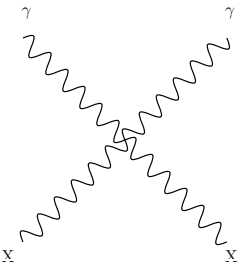


FIGURE 13.7
$\gamma - \gamma - \mathrm{X} - \mathrm{X}$ vertex.

applied to figures 13.6 and 13.7, but it has to be admitted that this approach is becoming laborious. It is, of course, far more efficient to deduce the vertices from the compact Yang-Mills Lagrangian $-\frac{1}{4}\hat{\boldsymbol{X}}_{\mu\nu}\cdot\hat{\boldsymbol{X}}^{\mu\nu}$, which we shall now do; nevertheless, some of the physical implications of those couplings, such as we have discussed above, are worth exposing.

The SU(2) Yang-Mills Lagrangian for the SU(2) triplet of gauge fields $\hat{\boldsymbol{X}}^{\mu}$ is

$$\hat{\mathcal{L}}_{2,\mathrm{YM}} = -\frac{1}{4}\hat{\boldsymbol{X}}_{\mu\nu}\cdot\hat{\boldsymbol{X}}^{\mu\nu}, \tag{13.97}$$

where

$$\hat{\boldsymbol{X}}^{\mu\nu} = \partial^{\mu}\hat{\boldsymbol{X}}^{\nu} - \partial^{\nu}\hat{\boldsymbol{X}}^{\mu} - e\hat{\boldsymbol{X}}^{\mu}\times\hat{\boldsymbol{X}}^{\nu}. \tag{13.98}$$

$\hat{\mathcal{L}}_{2,\mathrm{YM}}$ can be unpacked a bit into

$$\begin{aligned} &- \frac{1}{2}(\partial_{\mu}\hat{\boldsymbol{X}}_{\nu} - \partial_{\nu}\hat{X}_{\mu})\cdot(\partial^{\mu}\hat{\boldsymbol{X}}^{\nu}) \\ &+ e(\hat{\boldsymbol{X}}_{\mu}\times\hat{\boldsymbol{X}}_{\nu})\cdot\partial^{\mu}\hat{\boldsymbol{X}}^{\nu} \\ &- \frac{1}{4}e^{2}\left[(\hat{\boldsymbol{X}}^{\mu}\cdot\hat{\boldsymbol{X}}_{\mu})^{2} - (\hat{\boldsymbol{X}}^{\mu}\cdot\hat{\boldsymbol{X}}^{\nu})(\hat{\boldsymbol{X}}_{\mu}\cdot\hat{\boldsymbol{X}}_{\nu})\right]. \end{aligned} \tag{13.99}$$

The $\mathrm{X}-\mathrm{X}-\gamma$ vertex is in the 'e' term, the $\mathrm{X}-\mathrm{X}-\gamma-\gamma$ one in the 'e^2' term. We give the form of the latter using SU(2) 'i, j, k' labels, as shown in figure 13.8:

$$\begin{aligned} &-\mathrm{i}e^{2}[\epsilon_{ij\ell}\epsilon_{mn\ell}(\epsilon_{1}\cdot\epsilon_{3}\,\epsilon_{2}\cdot\epsilon_{4} - \epsilon_{1}\cdot\epsilon_{4}\,\epsilon_{2}\cdot\epsilon_{3}) \\ &+ \epsilon_{in\ell}\epsilon_{jm\ell}(\epsilon_{1}\cdot\epsilon_{2}\,\epsilon_{3}\cdot\epsilon_{4} - \epsilon_{1}\cdot\epsilon_{3}\,\epsilon_{2}\cdot\epsilon_{4}) \\ &+ \epsilon_{im\ell}\epsilon_{nj\ell}(\epsilon_{1}\cdot\epsilon_{4}\,\epsilon_{2}\cdot\epsilon_{3} - \epsilon_{1}\cdot\epsilon_{2}\,\epsilon_{3}\cdot\epsilon_{4})] \end{aligned} \tag{13.100}$$

The reason for the collection of terms seen in (13.96) and (13.100) can be understood as follows. Consider the $3-\mathrm{X}$ vertex

$$\langle k_{2},\epsilon_{2},j;k_{3},\epsilon_{3},k \mid e(\hat{\boldsymbol{X}}_{\mu}\times\hat{\boldsymbol{X}}_{\nu})\cdot\partial^{\mu}\hat{\boldsymbol{X}}^{\nu} \mid k_{1},\epsilon_{1},i\rangle \tag{13.101}$$

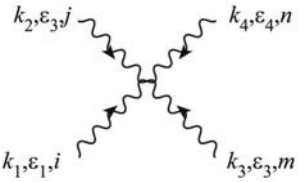


FIGURE 13.8
$4-\mathrm{X}$ vertex.

for example. When each $\hat{\boldsymbol{X}}$ is expressed as a mode expansion, and the initial and final states are also written in terms of appropriate $\hat{a}$'s and $\hat{a}^\dagger$'s, the amplitude will be a vacuum expectation value (vev) of six $\hat{a}$'s and $\hat{a}^\dagger$'s; the different terms in (13.96) arise from the different ways of getting a non-zero value for this vev, by manipulations similar to those in section 6.3.

We end this chapter by presenting an introduction to the problem of quantizing non-Abelian gauge field theories. Our aim will be, first, to indicate where the approach followed for the Abelian gauge field $\hat{A}^\mu$ in section 7.3.2 fails; and then to show how the assumption (nevertheless) that the Feynman rules we have established for tree graphs work for loops as well, leads to violations of unitarity. This calculation will indicate a very curious way of remedying the situation 'by hand', through the introduction of *ghost particles*, only present in loops.

13.3.3 Quantizing non-Abelian gauge fields

We consider for definiteness the SU(2) gauge theory with massless gauge fields $\hat{\boldsymbol{W}}^\mu(x)$, which we shall call gluons, by a slight abuse of language. We try to carry through for the Yang-Mills Lagrangian

$$\hat{\mathcal{L}}_2 = -\frac{1}{4}\hat{\boldsymbol{F}}_{\mu\nu}\cdot\hat{\boldsymbol{F}}^{\mu\nu}, \tag{13.102}$$

where

$$\hat{\boldsymbol{F}}_{\mu\nu} = \partial_\mu\hat{\boldsymbol{W}}_\nu - \partial_\nu\hat{\boldsymbol{W}}_\mu - g\hat{\boldsymbol{W}}_\mu\times\hat{\boldsymbol{W}}_\nu, \tag{13.103}$$

the same steps we followed for the Maxwell one in section 7.3.2.

We begin by re-formulating the prescription arrived at in (7.119), which we reproduce again here for convenience:

$$\hat{\mathcal{L}}_\xi = -\frac{1}{4}\hat{F}_{\mu\nu}\hat{F}^{\mu\nu} - \frac{1}{2\xi}(\partial_\mu\hat{A}^\mu)^2. \tag{13.104}$$

$\hat{\mathcal{L}}_\xi$ leads to the equation of motion

$$\Box\hat{A}^\mu - \partial^\mu\partial_\nu\hat{A}^\nu + \frac{1}{\xi}\partial^\mu\partial_\nu\hat{A}^\nu = 0. \tag{13.105}$$

This has the drawback that the limit $\xi \to 0$ appears to be singular (though the propagator (7.122) is well-behaved as $\xi \to 0$). To avoid this unpleasantness, consider the Lagrangian (Lautrup 1967)

$$\hat{\mathcal{L}}_{\xi B} = -\frac{1}{4}\hat{F}_{\mu\nu}\hat{F}^{\mu\nu} + \hat{B}\partial_\mu \hat{A}^\mu + \frac{1}{2}\xi \hat{B}^2 \tag{13.106}$$

where $\hat{B}$ is a scalar field. We may think of the '$\hat{B}\partial \cdot \hat{A}$' term as a field theory analogue of the procedure followed in classical Lagrangian mechanics, whereby a constraint (in this case the gauge-fixing one $\partial \cdot \hat{A} = 0$) is brought into the Lagrangian with a 'Lagrange multiplier' (here the *auxiliary* field $\hat{B}$). The momentum conjugate to $\hat{A}^0$ is now

$$\hat{\pi}^0 = \hat{B} \tag{13.107}$$

while the Euler-Lagrange equations for $\hat{A}^{\mu\nu}$ read

$$\Box \hat{A}^\mu - \partial^\mu \partial_\nu \hat{A}^\nu = \partial^\mu \hat{B}, \tag{13.108}$$

and for $\hat{B}$ yield

$$\partial_\mu \hat{A}^\mu + \xi \hat{B} = 0. \tag{13.109}$$

Eliminating $\hat{B}$ from (13.106) by means of (13.109) we recover (13.104). Taking ∂_μ of (13.108) we learn that $\Box \hat{B} = 0$, so that $\hat{B}$ is a free massless field. Applying $\Box$ to (13.109) then shows that $\Box \partial_\mu \hat{A}^\mu = 0$, so that $\partial_\mu \hat{A}^\mu$ is also a free massless field.

In this formulation, the appropriate subsidiary condition for getting rid of the unphysical (non-transverse) degrees of freedom is (cf (7.111))

$$\hat{B}^{(+)}(x) \mid \Psi\rangle = 0. \tag{13.110}$$

Kugo and Ojima (1979) have shown that (13.110) provides a satisfactory definition of the Hilbert space of states. In addition to this it is also essential to prove that all physical results are independent of the gauge parameter ξ.

We now try to generalize the foregoing in a straightforward way to (13.102). The obvious analogue of (13.106) would be to consider

$$\hat{\mathcal{L}}_{2,\xi\ B} = -\frac{1}{4}\hat{\boldsymbol{F}}_{\mu\nu} \cdot \hat{\boldsymbol{F}}^{\mu\nu} + \hat{\boldsymbol{B}} \cdot (\partial_\mu \hat{\boldsymbol{W}}^\mu) + \frac{1}{2}\xi \hat{\boldsymbol{B}} \cdot \hat{\boldsymbol{B}} \tag{13.111}$$

where $\hat{\boldsymbol{B}}$ is an SU(2) triplet of scalar fields. Equation (13.111) gives (cf (13.108))

$$(\hat{D}^\nu)_{ij}\hat{F}_{j\mu\nu} + \partial_\mu \hat{B}_i = 0 \tag{13.112}$$

where the covariant derivative is now the one appropriate to the SU(2) triplet $\boldsymbol{F}_{\mu\nu}$ (see (13.44) with $t = 1$, and (12.48)), and i, j are the SU(2) labels. Similarly, (13.109) becomes

$$\partial_\mu \hat{\boldsymbol{W}}^\mu + \xi \hat{\boldsymbol{B}} = \boldsymbol{0}. \tag{13.113}$$

It is possible to verify that

$$(\hat{D}^\mu)_{ki}(\hat{D}^\nu)_{ij}\hat{F}_{j\mu\nu} = 0 \tag{13.114}$$

where i, j, k are the SU(2) matrix indices, which implies that

$$(\hat{D}^\mu)_{ki}\partial_\mu \hat{B}_i = 0. \tag{13.115}$$

This is the crucial result: it implies that the auxiliary field $\hat{\boldsymbol{B}}$ is *not* a free field in this non-Abelian case, and so neither (from (13.113)) is $\partial_\mu \hat{\boldsymbol{W}}^\mu$. In consequence, the obvious generalizations of (7.108) or (13.110) cannot be used to define the physical (transverse) states. The reason is that a condition like (13.110) must hold for all times, and only if the field is free is its time variation known (and essentially trivial).

Let us press ahead nevertheless, and assume that the rules we have derived so far are the correct Feynman rules for this gauge theory. We will see that this leads to physically unacceptable consequences, namely to the *violation of unitarity.*

In fact, this is a problem which threatens all gauge theories if the gauge field is treated covariantly, i.e. as a 4-vector. As we saw in section 7.3.2, this introduces *unphysical degrees of freedom* which must somehow be eliminated from the theory, or at least prevented from affecting physical processes. In QED we do this by imposing the condition (7.111), or (13.110), but as we have seen the analogous conditions will not work in the non-Abelian case, and so unphysical states may make their presence felt, for example in the 'sum over intermediate states' which arises in the unitarity relation. This relation determines the imaginary part of an amplitude via an equation of the form (cf (11.65))

$$2\ \mathrm{Im}\ \langle \mathrm{f} \mid \mathcal{M} \mid \mathrm{i}\rangle = \int \sum_{\mathrm{n}} \langle \mathrm{f} \mid \mathcal{M} \mid \mathrm{n}\rangle\langle \mathrm{n} \mid \mathcal{M}^\dagger \mid \mathrm{i}\rangle \mathrm{d}\rho_{\mathrm{n}} \tag{13.116}$$

where $\langle \mathrm{f} \mid \mathcal{M} \mid \mathrm{i}\rangle$ is the (Feynman) amplitude for the process $\mathrm{i} \to \mathrm{f}$, and the sum is over a complete set of physical intermediate states $\mid \mathrm{n}\rangle$, which can enter at the given energy; $\mathrm{d}\rho_{\mathrm{n}}$ represents the phase space element for the general intermediate state $\mid \mathrm{n}\rangle$. Consider now the possibility of gauge quanta appearing in the states $\mid \mathrm{n}\rangle$. Since unitarity deals only with physical states, such quanta can have only the two degrees of freedom (polarizations) allowed for a physical massless gauge field (cf section 7.3.1). Now part of the power of the 'Feynman rules' approach to perturbation theory is that it is manifestly covariant. But there is no completely covariant way of selecting out just the two physical components of a massless polarization vector ϵ_μ, from the four originally introduced precisely for reasons of covariance. In fact, when gauge quanta appear as virtual particles in *intermediate* states in Feynman graphs, they will not be restricted to having only two polarization states (as we shall see explicitly in a moment). Hence there is a real chance

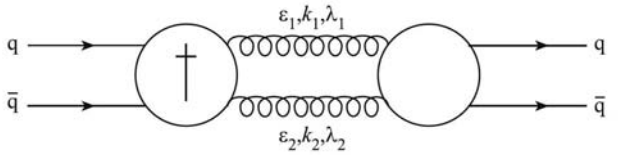


FIGURE 13.9
Two-gluon intermediate state in the unitarity relation for the amplitude for $q\bar{q} \to q\bar{q}$.

that when the imaginary part of such graphs is calculated, a contribution from the unphysical polarization states will be found, which has no counterpart at all in the physical unitarity relation, so that unitarity will not be satisfied. Since unitarity is an expression of conservation of probability, its violation is a serious disease indeed.

Consider, for example, the process $q\bar{q} \to q\bar{q}$ (where the 'quarks' are an SU(2) doublet), whose imaginary part has a contribution from a state containing two gluons (figure 13.9):

$$2\ \mathrm{Im}\ \langle q\bar{q} \mid \mathcal{M} \mid q\bar{q}\rangle = \int \sum \langle q\bar{q} \mid \mathcal{M} \mid gg\rangle\langle gg \mid \mathcal{M}^\dagger \mid q\bar{q}\rangle \mathrm{d}\rho_2 \qquad (13.117)$$

where $\mathrm{d}\rho_2$ is the 2-body phase space for the g-g state. The 2-gluon amplitudes in (13.117) must have the form

$$\mathcal{M}_{\mu_1\nu_1}\epsilon_1^{\mu_1}(k_1,\lambda_1)\epsilon_2^{\nu_1}(k_2,\lambda_2) \qquad (13.118)$$

where $\epsilon^\mu(k,\lambda)$ is the polarization vector for the gluon with polarization λ and 4-momentum k. The sum in (13.117) is then to be performed over $\lambda_1 = 1,2$ and $\lambda_2 = 1,2$ which are the physical polarization states (cf section 7.3.1). Thus (13.117) becomes

$$\begin{aligned} 2\ \mathrm{Im}\ \mathcal{M}_{q\bar{q}\to q\bar{q}} \quad = \quad & \int \sum_{\lambda_1=1,2;\lambda_2=1,2} \mathcal{M}_{\mu_1\nu_1}\epsilon_1^{\mu_1}(k_1,\lambda_1)\epsilon_2^{\nu_1}(k_2,\lambda_2) \\ & \times\ \mathcal{M}^*_{\mu_2\nu_2}\epsilon_1^{\mu_2}(k_1,\lambda_1)\epsilon_2^{\nu_2}(k_2,\lambda_2)\mathrm{d}\rho_2. \end{aligned} \qquad (13.119)$$

For later convenience we are using real polarization vectors as in (7.81) and (7.82): $\epsilon(k_i,\lambda_i=+1) = (0,1,0,0), \epsilon(k_i,\lambda_i=-1) = (0,0,1,0)$; and of course $k_1^2 = k_2^2 = 0$.

We now wish to find out whether or not a result of the form (13.119) will hold when the $\mathcal{M}$'s represent some suitable Feynman graphs. We first note that we want the unitarity relation (13.119) to be satisfied order by order in perturbation theory: that is to say, when the $\mathcal{M}$'s on both sides are expanded in powers of the coupling strengths (as in the usual Feynman graph expansion), the coefficients of corresponding powers on each side should be

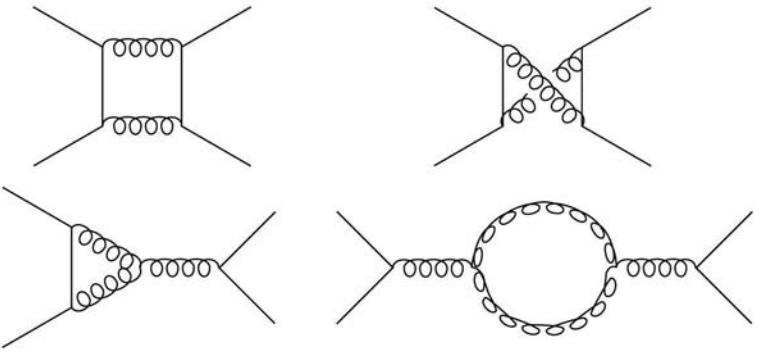


FIGURE 13.10
Some $O(g^4)$ contributions to $\mathrm{q\bar{q}} \to \mathrm{q\bar{q}}$.

equal. Since each emission or absorption of a gluon produces one power of the SU(2) coupling g, the right-hand side of (13.119) involves at least the power g^4. Thus the lowest-order process in which (13.119) may be tested is for the fourth-order amplitude $\mathcal{M}^{(4)}_{\mathrm{q\bar{q}}\to\mathrm{q\bar{q}}}$. There are quite a number of contributions to $\mathcal{M}^{(4)}_{\mathrm{q\bar{q}}\to\mathrm{q\bar{q}}}$, some of which are shown in Figure 13.10; all contain a loop. On the right-hand side of (13.119), each $\mathcal{M}$ involves two polarization vectors, and so each must represent the $0(g^2)$ contribution to $\mathrm{q\bar{q}} \to \mathrm{gg}$, which we call $\mathcal{M}^{(2)}_{\mu\nu}$; thus both sides are consistently of order g^4. There are three contributions to $\mathcal{M}^{(2)}_{\mu\nu}$ shown in figure 13.11; when these are placed in (13.119), contributions to the imaginary part of $\mathcal{M}^{(4)}_{\mathrm{q\bar{q}}\to\mathrm{q\bar{q}}}$ are generated, which should agree with the imaginary part of the total $0(g^4)$ loop-graph contribution. Let us see if this works out. We choose to work in the gauge $\xi = 1$, so that the gluon propagator takes the familiar form $-\mathrm{i}g^{\mu\nu}\delta_{ij}/k^2$. According to the rules for propagators and vertices already given, each of the loop amplitudes $\mathcal{M}^{(4)}_{\mathrm{q\bar{q}}\to\mathrm{q\bar{q}}}$ (e.g. those of figure 13.10) will be proportional to the product of the propagators for the quarks and the gluons, together with appropriate 'γ' and 'τ' vertex factors, the whole being integrated over the loop momentum. The extraction of the imaginary part of a Feynman diagram is a technical matter, having to do with careful consideration of the '$\mathrm{i}\epsilon$' in the propagators. Rules for doing this exist (Eden *et al.* 1966, section 2.9), and in the present case the result is that, to compute the imaginary part of the amplitudes of figure 13.10, one replaces each gluon propagator of momentum k by

$$\pi(-g^{\mu\nu})\delta(k^2)\theta(k_0)\delta_{ij}. \tag{13.120}$$

That is, the propagator is replaced by a condition stating that, in evaluating the imaginary part of the diagram, the gluon's mass is constrained to have

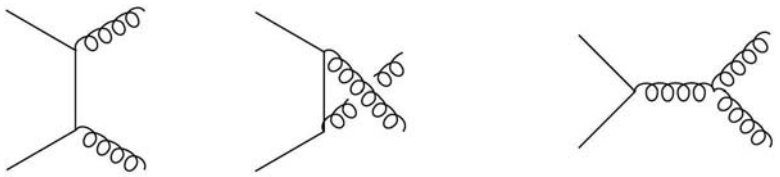


FIGURE 13.11
$O(g^2)$ contributions to $q\bar{q} \to gg$.

the physical (free-field) value of zero, instead of varying freely as the loop momentum varies, and its energy is positive. These conditions (one for each gluon) have the effect of converting the loop integral with a standard two-body phase space integral for the gg intermediate state, so that eventually

$$2\text{Im}\ \mathcal{M}^{(4)}_{q\bar{q}\to q\bar{q}} = \int \mathcal{M}^{(2)}_{\mu_1\nu_1}(-g^{\mu_1\mu_2})\mathcal{M}^{(2)}_{\mu_2\nu_2}(-g^{\nu_1\nu_2})\mathrm{d}\rho_2 \tag{13.121}$$

where $\mathcal{M}^{(2)}_{\mu_1\nu_1}$ is the sum of the three $O(g^2)$ tree graphs shown in figure 13.11, with all external legs satisfying the 'mass-shell' conditions.

So, the imaginary part of the loop contribution to $\mathcal{M}^{(4)}_{q\bar{q}\to q\bar{q}}$ does seem to have the form (13.116) as required by unitarity, with $|n\rangle$ the gg intermediate state as in (13.119). But there is one essential difference between (13.121) and (13.119): the place of the factor $-g^{\mu\nu}$ in (13.121) is taken in (13.119) by the gluon polarization sum

$$P^{\mu\nu}(k) \equiv \sum_{\lambda=1,2} \epsilon^{\mu}(k,\lambda)\epsilon^{\nu}(k,\lambda) \tag{13.122}$$

for $k = k_1, k_2$ and $\lambda = \lambda_1, \lambda_2$ respectively. Thus we have to investigate whether this difference matters.

To proceed further, it is helpful to have an explicit expression for $P^{\mu\nu}$. We might think of calculating the necessary sum over λ by brute force, using two ϵ's specified by the conditions (cf (7.87))

$$\epsilon^{\mu}(k,\lambda)\epsilon_{\mu}(k,\lambda') = -\delta_{\lambda\lambda'}, \qquad \epsilon \cdot k = 0. \tag{13.123}$$

The trouble is that conditions (13.123) *do not fix the* ϵ's *uniquely if* $k^2 = 0$. (Note the $\delta(k^2)$ in (13.120)). Indeed, it is precisely the fact that any given ϵ_μ satisfying (13.123) can be replaced by $\epsilon_\mu + \lambda k_\mu$ that both reduces the degrees of freedom to two (as we saw in section 7.3.1), and evinces the essential arbitrariness in the ϵ_μ specified only by (13.123). In order to calculate (13.122), we need to put another condition on ϵ_μ, so as to fix it uniquely. A standard choice (see e.g. Taylor 1976, pp 14–15) is to supplement (13.123) with the further condition

$$t \cdot \epsilon = 0 \tag{13.124}$$

where t is some 4-vector. This certainly fixes ϵ_μ, and enables us to calculate (13.122), but of course now two further difficulties have appeared: namely, the physical results seem to depend on t_μ; and have we not lost Lorentz covariance, because the theory involves a special 4-vector t_μ?

Setting these questions aside for the moment, we can calculate (13.122) using the conditions (13.123) and (13.124), finding (problem 13.11)

$$P_{\mu\nu} = -g_{\mu\nu} - [t^2 k_\mu k_\nu - k\cdot t(k_\mu t_\nu + k_\nu t_\mu)]/(k\cdot t)^2. \tag{13.125}$$

But only the *first* term on the right-hand side of (13.125) is to be seen in (13.121). A crucial quantity is clearly

$$\begin{aligned} U_{\mu\nu}(k,t) &\equiv -g_{\mu\nu} - P_{\mu\nu} \\ &= [t^2 k_\mu k_\nu - k\cdot t(k_\mu t_\nu + k_\nu t_\mu)]/(k\cdot t)^2. \end{aligned} \tag{13.126}$$

We note that whereas

$$k^\mu P_{\mu\nu} = k^\nu P_{\mu\nu} = 0 \tag{13.127}$$

(from the condition $k\cdot\epsilon = 0$), the same is *not* true of $k^\mu U_{\mu\nu}$ – in fact,

$$k^\mu U_{\mu\nu} = -k_\nu \tag{13.128}$$

where we have used $k^2 = 0$. It follows that $U_{\mu\nu}$ may be regarded as including polarization states for which $\epsilon\cdot k \neq 0$. In physical terms, therefore, a gluon appearing internally in a Feynman graph has to be regarded as existing in more than just the two polarization states available to an external gluon (cf section 7.3.1). $U_{\mu\nu}$ characterizes the contribution of these unphysical polarization states.

The discrepancy between (13.121) and (13.119) is then

$$2\mathrm{Im}\ \mathcal{M}^{(4)}_{\mathrm{q\bar{q}}\to\mathrm{q\bar{q}}} = \int \mathcal{M}^{(2)}_{\mu_1\nu_1}[U^{\mu_1\mu_2}(k_1,t_1)]\mathcal{M}^{(2)}_{\mu_2\nu_2}[U^{\nu_1\nu_2}(k_2,t_2)]\mathrm{d}\rho_2, \tag{13.129}$$

together with similar terms involving one P and one U. It follows that these unwanted contributions will, in fact, vanish if

$$k_1^{\mu_1}\mathcal{M}^{(2)}_{\mu_1\nu_1} = 0, \tag{13.130}$$

and similarly for k_2. This will also ensure that amplitudes are independent of t_μ.

Condition (13.130) is apparently the same as the U(1) gauge invariance requirement of (8.165), already recalled in the previous section. As discussed there, it can be interpreted here also as expressing gauge invariance in the non-Abelian case, working to this given order in perturbation theory. Indeed, the diagrams of figure 13.11 are essentially 'crossed' versions of those in figure 13.5. However, there is one crucial difference here. In figure 13.5, both the X's were physical, their polarizations satisfying the condition $\epsilon\cdot k = 0$. In figure 13.11, by contrast, neither of the gluons, in the discrepant contribution

(13.129), satisfies $\epsilon \cdot k = 0$ – see the sentence following (13.128). Thus the crucial point is that (13.130) must be true for each gluon, *even when the other gluon has* $\epsilon \cdot k \neq 0$. And, in fact, we shall now see that whereas the (crossed) version of (13.130) did hold for our dX $\to$ dX amplitudes of section 13.3.2, (13.130) *fails* for states with $\epsilon \cdot k \neq 0$.

The three graphs of figure 13.11 together yield

$$\begin{aligned}
&\mathcal{M}^{(2)}_{\mu_1\nu_1}\epsilon_1^{\mu}(k_1,\lambda_1)\epsilon_2^{\nu_1}(k_2,\lambda_2) = g^2\bar{v}(p_2)\frac{\tau_j}{2}\not{\epsilon}_2 a_{2j}\frac{1}{\not{p}_1 - \not{k}_1 - m}\frac{\tau_i}{2}a_{1i}\not{\epsilon}_1 u(p_1)\\
+\quad & g^2\bar{v}(p_2)\frac{\tau_i}{2}a_{1i}\not{\epsilon}_1\frac{1}{\not{p}_1 - \not{k}_2 - m}\frac{\tau_j}{2}a_{2j}\not{\epsilon}_2 u(p_1)\\
+\quad & (-\mathrm{i})g^2\epsilon_{kij}[(p_1+p_2+k_1)^{\nu_1}g^{\mu_1\rho} + (-k_2-p_1-p_2)^{\mu_1}g^{\rho\nu_1}\\
+\quad & (-k_1+k_2)^{\rho}g^{\mu_1\nu_1}]\epsilon_{1\mu_1}a_{1i}a_{2j}\epsilon_{2\nu_1}\frac{-1}{(p_1+p_2)^2}\bar{v}(p_2)\frac{\tau_k}{2}\gamma_\rho u(p_1) \qquad (13.131)
\end{aligned}$$

where we have written the gluon polarization vectors as a product of a Lorentz 4-vector ϵ_μ and an 'SU(2) polarization vector' a_i to specify the triplet state label. Now replace ϵ_1, say, by k_1. Using the Dirac equation for $u(p_1)$ and $\bar{v}(p_2)$ the first two terms reduce to (cf (13.94))

$$\begin{aligned}
& g^2\bar{v}(p_2)\not{\epsilon}_2[\tau_i/2,\tau_j/2]u(p_1)a_{1i}a_{2j}\\
= \quad & \mathrm{i}g^2\bar{v}(p_2)\not{\epsilon}_2\epsilon_{ijk}(\tau_k/2)u(p_1)a_{1i}a_{2j} \qquad (13.132)
\end{aligned}$$

using the SU(2) algebra of the τ's. The third term in (13.131) gives

$$-\mathrm{i}g^2\epsilon_{ijk}\bar{v}(p_2)\not{\epsilon}_2(\tau_k/2)u(p_1)a_{1i}a_{2j} \qquad (13.133)$$

$$+\mathrm{i}g^2\frac{\epsilon_{ijk}}{2k_1\cdot k_2}\bar{v}(p_2)\not{k}_1(\tau_k/2)u(p_1)k_2\cdot\epsilon_2 a_{1i}a_{2j}. \qquad (13.134)$$

We see that the first part (13.133) certainly does cancel (13.132), but there remains the second piece (13.134), *which only vanishes if* $k_2 \cdot \epsilon_2 = 0$. This is not sufficient to guarantee the absence of all unphysical contributions to the imaginary part of the 2-gluon graphs, as the preceding discussion shows. *We conclude that loop diagrams involving two (or, in fact, more) gluons, if constructed according to the simple rules for tree diagrams, will violate unitarity.*

The correct rule for such loops must be as to satisfy unitarity. Since there seems no other way in which the offending piece in (13.134) can be removed, we must infer that the rule for loops will have to involve some extra term, or terms, over and above the simple tree-type constructions, which will cancel the contributions of unphysical polarization states. To get an intuitive idea of what such extra terms might be, we return to expression (13.126) for the sum over unphysical polarization states $U_{\mu\nu}$, and make a specific choice for t. We take $t_\mu = \bar{k}_\mu$, where the 4-vector $\bar{k}$ is defined by $\bar{k} = (-\mid \boldsymbol{k} \mid, \boldsymbol{k})$, and $\boldsymbol{k} = (0, 0, \mid \boldsymbol{k} \mid)$. This choice obviously satisfies (13.124). Then

$$U_{\mu\nu}(k,\bar{k}) = (k_\mu\bar{k}_\nu + k_\nu\bar{k}_\mu)/(2 \mid \boldsymbol{k} \mid^2) \qquad (13.135)$$

and unitarity (cf (13.129)) requires

$$\int \mathcal{M}^{(2)}_{\mu_1\nu_1}\mathcal{M}^{(2)}_{\mu_2\nu_2}\frac{(k_1^{\mu_1}\bar{k}_1^{\mu_2}+k_1^{\mu_2}\bar{k}_1^{\mu_1})}{2\mid \boldsymbol{k}_1\mid^2}\ \frac{(k_2^{\nu_1}\bar{k}_2^{\nu_2}+k_2^{\nu_2}\bar{k}_2^{\nu_1})}{2\mid \boldsymbol{k}_2\mid^2}\mathrm{d}\rho_2 \tag{13.136}$$

to vanish, but it does not. Let us work in the centre of momentum (CM) frame of the two gluons, with $k_1=(\mid \boldsymbol{k}\mid,0,0,\mid \boldsymbol{k}\mid), k_2=(\mid \boldsymbol{k}\mid,0,0,-\mid \boldsymbol{k}\mid), \bar{k}_1=(-\mid \boldsymbol{k}\mid,0,0,\mid \boldsymbol{k}\mid), \bar{k}_2=(-\mid \boldsymbol{k}\mid,0,0,-\mid \boldsymbol{k}\mid)$, and consider for definiteness the contractions with the $\mathcal{M}^{(2)}_{\mu_1\nu_1}$ term. These are $\mathcal{M}^{(2)}_{\mu_1\nu_1}k_1^{\mu_1}k_2^{\nu_1}, \mathcal{M}^{(2)}_{\mu_1\nu_1}k_1^{\mu_1}\bar{k}_2^{\nu_1}$ etc. Such quantities can be calculated from expression (13.131) by setting $\epsilon_1=k_1, \epsilon_2=k_2$ for the first, $\epsilon_1=k_1, \epsilon_2=\bar{k}_2$ for the second, and so on. We have already obtained the result of putting $\epsilon_1=k_1$. From (13.134) it is clear that a term in which ϵ_2 is replaced by k_2 as well as ϵ_1 by k_1 will vanish, since $k_2^2=0$. A typical non-vanishing term is of the form $\mathcal{M}^{(2)}_{\mu_1\nu_1}k_1^{\mu_1}\bar{k}_2^{\nu_1}/2\mid \boldsymbol{k}\mid^2$. From (13.134) this reduces to

$$-\mathrm{i}g^2\frac{\epsilon_{ijk}}{2k_1\cdot k_2}\bar{v}(p_2)\ \not{k}_1(\tau_k/2)u(p_1)a_{1i}a_{2j} \tag{13.137}$$

using $k_2\cdot\bar{k}_2/2\mid \boldsymbol{k}\mid^2=-1$. We may rewrite (13.137) as

$$j_{\mu k}\ \frac{-g^{\mu\nu}\delta_{k\ell}}{(k_1+k_2)^2}\mathrm{i}g\epsilon_{ij\ell}a_{1i}a_{2j}k_{1\nu} \tag{13.138}$$

where

$$j_{\mu k}=g\bar{v}(p_2)\gamma_\mu(\tau_k/2)u(p_1) \tag{13.139}$$

is the SU(2) current associated with the $\mathrm{q\bar{q}}$ pair.

The unwanted terms of the form (13.138) can be eliminated if we adopt the following rule (on the grounds of 'forcing the theory to make sense'). In addition to the fourth-order diagrams of the type shown in figure 13.10, constructed according to the simple 'tree' prescriptions, there must exist a previously unknown fourth-order contribution, *only present in loops*, such that it has an imaginary part which is non-zero in the same physical region as the two-gluon intermediate state, and moreover is of just the right magnitude to cancel all the contributions to (13.136) from terms like (13.138). Now (13.138) has the appearance of a one-gluon intermediate state amplitude. The $\mathrm{q\bar{q}}\to\mathrm{g}$ vertex is represented by the current (13.139), the gluon propagator appears in Feynman gauge $\xi=1$, and the rest of the expression would have the interpretation of a coupling between the intermediate gluon and two scalar particles with SU(2) polarizations a_{1i}, a_{2j}. Thus (13.138) can be interpreted as the amplitude for the tree graph shown in figure 13.12, where the dotted lines represent the scalar particles. It seems plausible, therefore, that the fourth-order graph we are looking for has the form shown in figure 13.13. The new scalar particles must be massless, so that this new amplitude has an imaginary part in the same physical region as the gg state. When the imaginary part of figure 13.13 is calculated in the usual way, it will involve

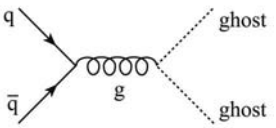


FIGURE 13.12
Tree graph interpretation of the expression (13.138).

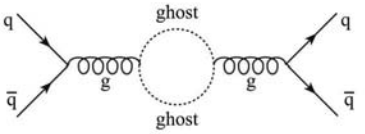


FIGURE 13.13
Ghost loop diagram contributing in fourth order to $q\bar{q} \to q\bar{q}$.

contributions from the tree graph of figure 13.12, and these can be arranged to cancel the unphysical polarization pieces like (13.138).

For this cancellation to work, the scalar particle loop graph of figure 13.13 must enter with the opposite sign from the three-gluon loop graph of figure 13.10, which in retrospect was the cause of all the trouble. Such a relative minus sign between single closed loop graphs would be expected if the scalar particles in figure 13.13 were in fact fermions! (Recall the rule given in section 11.3 and problem 11.2). Thus we appear to need *scalar* particles obeying *Fermi* statistics. Such particles are called 'ghosts'. We must emphasize that although we have introduced the tree graph of figure 13.12, which apparently involves ghosts as external lines, in reality the ghosts are always confined to loops, their function being to cancel unphysical contributions from intermediate gluons.

The preceding discussion has, of course, been entirely heuristic. It can be followed through so as to yield the correct prescription for eliminating unphysical contributions from a single closed gluon loop. But, as Feynman recognized (1963, 1977), unitarity alone is not a sufficient constraint to provide the prescription for more than one closed gluon loop. Clearly what is required is some additional term in the Lagrangian, which will do the job in general. Such a term indeed exists, and was first derived using the path integral form of quantum field theory (see chapter 16) by Faddeev and Popov (1967). The result is that the covariant gauge-fixing term (13.68) must be supplemented by the 'ghost Lagrangian'

$$\hat{\mathcal{L}}_g = \partial_\mu \hat{\eta}_i^\dagger \hat{D}_{ij}^\mu \hat{\eta}_j \tag{13.140}$$

where the η field is an SU(2) triplet, and spinless, but obeying *anti*commutation relations; the covariant derivative is the one appropriate for an SU(2) triplet,

namely (from (13.44) and (12.48))

$$\hat{D}^{\mu}_{ij} = \partial^{\mu}\delta_{ij} + g\epsilon_{kij}\hat{W}^{\mu}_{k}, \tag{13.141}$$

in this case. The result (13.140) is derived in standard books of quantum field theory, for example Cheng and Li (1984), Peskin and Schroeder (1995) or Ryder (1996). We should add the caution that the form of the ghost Lagrangian depends on the choice of the gauge-fixing term; there are gauges in which the ghosts are absent. Feynman rules for non-Abelian gauge field theories are given in Cheng and Li (1984), for example. We give the rules for tree diagrams, for which there are no problems with ghosts, in appendix Q.

Problems

13.1 Verify that (13.34) reduces to (13.26) in the infinitesimal case.

13.2 Verify equation (13.45).

13.3 Using the expression for D^{μ} in (13.47), verify (13.48).

13.4 Verify the transformation law (13.51) of $\boldsymbol{F}^{\mu\nu}$ under local SU(2) transformations.

13.5 Verify that $\boldsymbol{F}_{\mu\nu} \cdot \boldsymbol{F}^{\mu\nu}$ is invariant under local SU(2) transformations.

13.6 Verify that the (infinitesimal) transformation law (13.56) for the SU(3) gauge field A^{μ}_{a} is consistent with (13.55).

13.7 By considering the commutator of two D^{μ}'s of the form (13.53), verify (13.61).

13.8 Verify that (13.84) reduces to (13.86) (omitting the $(2\pi)^4\delta^4$ factors).

13.9 Verify that the replacement of ϵ_1 by k_1 in (13.93) leads to (13.94).

13.10 Verify that when ϵ_1 is replaced by k_1 in (13.95), the resulting amplitude cancels the contribution (13.94), provided that $\delta = 1$.

13.11 Show that $P^{\mu\nu}$ of (13.122), with the ϵ's specified by the conditions (13.123) and (13.124), is given by (13.125).

Part VI

QCD and the Renormalization Group


Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

14

QCD I: Introduction, Tree Graph Predictions, and Jets

In the previous chapter we have introduced the elementary concepts and formalism associated with non-Abelian quantum gauge field theories. It is now well established that the strong interactions between quarks are described by a theory of this type, in which the gauge group is an $SU(3)_c$, acting on a degree of freedom called 'colour' (indicated by the subscript c). This theory is called Quantum Chromodynamics, or QCD for short. QCD will be our first application of the theory developed in chapter 13, and we shall devote the next two chapters, and much of chapter 16, to it.

In the present chapter we introduce QCD and discuss some of its simpler experimental consequences. We briefly recall the evidence for the 'colour' degree of freedom in section 14.1, and then proceed to the dynamics of colour, and the QCD Lagrangian, in section 14.2. Perhaps the most remarkable thing about the dynamics of QCD is that, despite its being a theory of the *strong* interactions, there are certain kinematic regimes – roughly speaking, short distances or high energies – in which it is effectively a quite *weakly* interacting theory. This is a consequence of a fundamental property, possessed only by non-Abelian gauge theories, whereby the effective interaction strength becomes progressively smaller in such regimes. This property is called 'asymptotic freedom', and was already mentioned in section 11.5.3 of volume 1. In appropriate cases, therefore, the lowest-order perturbation theory amplitudes (tree graphs) provide a very convincing qualitative, or even 'semi-quantitative', orientation to the data. In sections 14.3 and 14.4 we shall see how the tree graph techniques acquired for QED in volume 1 produce more useful physics when applied to QCD.

However, most of the quantitative experimental support for QCD has come from comparison with predictions which include higher-order QCD corrections; indeed, the asymptotic freedom property itself emerges from summing a whole class of higher-order contributions, as we shall indicate at the beginning of chapter 15. This immediately involves all the apparatus of *renormalization*. The necessary calculations quite rapidly become too technical for the intended scope of this book, but in chapter 15 we shall try to provide an elementary introduction to the issues involved, and to the necessary techniques, by building on the discussion of renormalization given in chapters 10 and 11 of volume 1. The main new concept will be the *renormalization group* (and related ideas),

which is an essential tool in the modern confrontation of perturbative QCD with data. Some of the simpler predictions of the renormalization group technique will be compared with experimental data in the last part of chapter 15.

In chapter 16 we work towards understanding some non-perturbative aspects of QCD. As a natural concomitant of asymptotic freedom, it is to be expected that the effective coupling strength becomes progressively larger at longer distances or lower energies, ultimately being strong enough to lead (presumably) to the confinement of quarks and gluons; this is sometimes referred to as 'infrared slavery'. In this regime perturbation theory clearly fails. An alternative, purely numerical, approach is available however, namely the method of 'lattice' QCD, which involves replacing the space-time continuum by a *discrete lattice* of points. At first sight, this may seem a topic rather disconnected from everything that has preceded it. But we shall see that in fact it provides some powerful new insights into several aspects of quantum field theory in general, and in particular of renormalization, by revisiting it in coordinate (rather than momentum) space. Quite apart from this, however, results from lattice QCD now provide independent confirmation of the theory, in the non-perturbative regime.

14.1 The colour degree of freedom

The first intimation of a new, unrevealed degree of freedom of matter came from baryon spectroscopy (Greenberg 1964; see also Han and Nambu 1965, and Tavkhelidze 1965). For a baryon made of three spin-$\frac{1}{2}$ quarks, the original non-relativistic quark model wave-function took the form

$$\psi_{3q} = \psi_{3q,\text{space}}\psi_{3q,\text{spin}}\psi_{3q,\text{flavour}}. \tag{14.1}$$

It was soon realized (e.g. Dalitz 1965) that the product of these space, spin and flavour wavefunctions for the ground state baryons was *symmetric* under interchange of any two quarks. For example, the Δ^{++} state mentioned in section 12.2.3 is made of three u quarks (flavour symmetric) in the $J^P = \frac{3}{2}^+$ state, which has zero orbital angular momentum and is hence spatially symmetric, and a symmetric $S = \frac{3}{2}$ spin wavefunction. But we saw in section 7.2 that quantum field theory requires fermions to obey the exclusion principle – i.e. the wavefunction ψ_{3q} should be *anti*symmetric with respect to quark interchange. A simple way of implementing this requirement is to suppose that the quarks carry a further degree of freedom, called colour, with respect to which the 3q wavefunction can be antisymmetrized, as follows (Fritzsch and Gell-Mann 1972, Bardeen, Fritzsch and Gell-Mann 1973). We introduce a *colour wavefunction* with colour index α:

$$\psi_\alpha \quad (\alpha = 1, 2, 3).$$

We are here writing the three labels as '1, 2, 3', but they are often referred to by colour names such as 'red, blue, green'; it should be understood that this is merely a picturesque way of referring to the three basic states of this degree of freedom, and has nothing to do with real colour! With the addition of this degree of freedom we can certainly form a three-quark wavefunction which is antisymmetric in colour by using the antisymmetric symbol $\epsilon_{\alpha\beta\gamma}$, namely[1]

$$\psi_{3\mathrm{q},\ \mathrm{colour}} = \epsilon_{\alpha\beta\gamma}\psi_\alpha\psi_\beta\psi_\gamma \tag{14.2}$$

and this must then be multiplied into (14.1) to give the full 3q wavefunction. To date, *all* known baryon states can be described this way, i.e. the symmetry of the 'traditional' space-spin-flavour wavefunction (14.1) is symmetric overall, while the required antisymmetry is restored by the additional factor (14.2). As far as meson ($\bar{\mathrm{q}}\mathrm{q}$) states are concerned, what was previously a π^+ wavefunction d^*u is now

$$\frac{1}{\sqrt{3}}(d_1^*u_1 + d_2^*u_2 + d_3^*u_3) \tag{14.3}$$

which we write in general as $(1/\sqrt{3})d_\alpha^\dagger u_\alpha$. We shall shortly see the group theoretical significance of this 'neutral superposition', and of (14.2). Meanwhile, we note that (14.2) is actually the *only* way of making an antisymmetric combination of the three ψ's; it is therefore called a (colour) *singlet*. It is reassuring that there is only one way of doing this – otherwise, we would have obtained more baryon states than are physically observed. As we shall see in section 14.2.1, (14.3) is also a colour singlet combination.

The above would seem a somewhat artificial device unless there were some physical consequences of this increase in the number of quark types – and there are. In any process which we can describe in terms of creation or annihilation of quarks, the *multiplicity* of quark types will enter into the relevant observable cross section or decay rate. For example, at high energies the ratio

$$R = \frac{\sigma(\mathrm{e^+e^-} \to \mathrm{hadrons})}{\sigma(\mathrm{e^+e^-} \to \mu^+\mu^-)} \tag{14.4}$$

will, in the quark parton model (see section 9.5), reflect the magnitudes of the individual quark couplings to the photon:

$$R = \sum_a e_a^2 \tag{14.5}$$

where a runs over all quark types. For five quarks u, d, s, c, b with respective charges $\frac{2}{3}, -\frac{1}{3}, -\frac{1}{3}, \frac{2}{3}, -\frac{1}{3}$, this yields

$$R_{\mathrm{no\ colour}} = \frac{11}{9} \tag{14.6}$$

[1] In (14.2) each ψ refers to a different quark, but we have not indicated the quark labels explicitly.

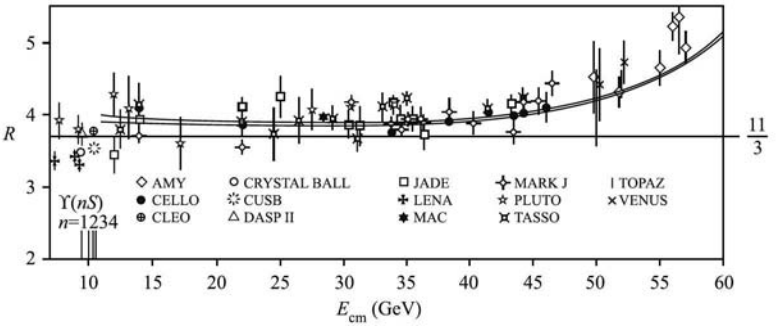


FIGURE 14.1
The ratio R (see (14.4)). Figure reprinted with permission from L. Montanet *et al. Physical Review* D **50** 1173 (1994). Copyright 1994 by the American Physical Society.

and

$$R_{\text{colour}} = \frac{11}{3} \tag{14.7}$$

for the two cases, as we saw in section 9.5. (The values $R = 2$ below the charm threshold, and $R = 10/3$ below the b threshold, were predicted by Bardeen *et al.* 1973). The data (figure 14.1) rule out (14.6), and are in good agreement with (14.7) at energies well above the b threshold, and well below the Z^0 resonance peak. There is an indication that the data tend to lie *above* the parton model prediction; this is actually predicted by QCD via higher-order corrections, as will be discussed in section 15.1.

A number of branching fractions also provide simple ways of measuring the number of colours N_{c}. For example, consider the branching fraction for $\tau^- \to \mathrm{e}^-\bar{\nu}_{\mathrm{e}}\nu_\tau$ (i.e. the ratio of the rate for $\tau^- \to \mathrm{e}^-\bar{\nu}_{\mathrm{e}}\nu_\tau$ to that for all other decays). τ^- decays proceed via the weak process shown in figure 14.2, where the final fermions can be $\mathrm{e}^-\bar{\nu}_{\mathrm{e}}, \mu^-\bar{\nu}_\mu$, or $\bar{\mathrm{u}}\mathrm{d}$, the last with multiplicity N_{c}. Thus

$$B(\tau^- \to \mathrm{e}^-\bar{\nu}_{\mathrm{e}}\nu_\tau) \approx \frac{1}{2 + N_{\mathrm{c}}}. \tag{14.8}$$

Experiments give $B \approx 18$ % and hence $N_{\mathrm{c}} \approx 3$.

Similarly, the branching fraction $B(\mathrm{W}^- \to \mathrm{e}^-\bar{\nu}_{\mathrm{e}})$ is $\sim \frac{1}{3+2N_{\mathrm{c}}}$ (from $f = \mathrm{e}, \mu, \tau, \mathrm{u}$ and c). Experiment gives a value of 10.7 %, so again $\mathrm{N_c} \approx 3$.

In chapter 9 we also discussed the Drell–Yan process in the quark parton model; it involves the subprocess $\mathrm{q}\bar{\mathrm{q}} \to \mathrm{l}\bar{\mathrm{l}}$ which is the inverse of the one in (14.4). We mentioned that a factor of $\frac{1}{3}$ appears in this case: it arises because we must average over the nine possible initial $\mathrm{q}\bar{\mathrm{q}}$ combinations (factor $\frac{1}{9}$) and then sum over the number of such states that lead to the colour neutral photon, which is 3 ($\bar{\mathrm{q}}_1\mathrm{q}_1, \bar{\mathrm{q}}_2\mathrm{q}_2$ and $\bar{\mathrm{q}}_3\mathrm{q}_3$). With this factor, and using quark

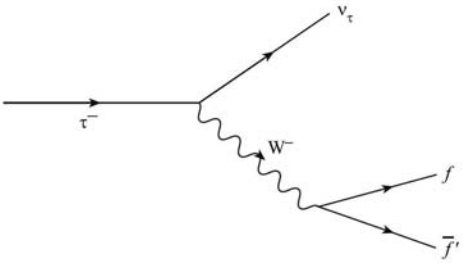


FIGURE 14.2
τ decay.

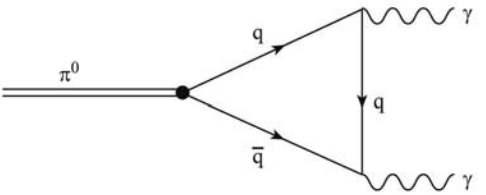


FIGURE 14.3
Triangle graph for π^0 decay.

distribution functions consistent with deep inelastic scattering, the parton model gives a good first approximation to the data.

Finally, we mention the rate for $\pi^0 \to \gamma\gamma$. As will be discussed in section 18.4, this process is entirely calculable from the graph shown in figure 14.3 (and the one with the γ's 'crossed'), where 'q' is u or d. The amplitude is proportional to the square of the quark charges, but because the π^0 is an isovector, the contributions from the $u\bar{u}$ and $d\bar{d}$ states have opposite signs (see section 12.1.3). Thus the rate contains a factor

$$((2/3)^2 - (1/3)^2)^2 = \frac{1}{9}. \tag{14.9}$$

However, the original calculation of this rate by Steinberger (1949) used a model in which the proton and neutron replaced the u and d in the loop, in which case the factor corresponding to (14.9) is just 1 (since the n has zero charge). Experimentally the rate agrees well with Steinberger's calculation, indicating that (14.9) needs to be multiplied by 9, which corresponds to $N_c = 3$ identical amplitudes of the form shown in figure 14.3, as was noted by Bardeen, Fritzsch and Gell-Mann (1973).

14.2 The dynamics of colour

14.2.1 Colour as an SU(3) group

We now want to consider the possible dynamical role of colour – in other words, the way in which the forces between quarks depend on their colours. We have seen that we seem to need three different quark types for each given flavour. They must all have the same mass, or else we would observe some 'fine structure' in the hadronic levels. Furthermore, and for the same reason, 'colour' must be an exact symmetry of the Hamiltonian governing the quark dynamics. What symmetry group is involved? We shall consider how some empirical facts suggest that the answer is $\mathrm{SU}(3)_\mathrm{c}$.

To begin with, it is certainly clear that the interquark force must depend on colour, since we do *not* observe 'colour multiplicity' of hadronic states: for example we do not see eight other coloured π^+'s $(d_1^* u_2, d_3^* u_1, \dots)$ degenerate with the one 'colourless' physical π^+ whose wavefunction was given previously. The observed hadronic states are all *colour singlets*, and the force must somehow be responsible for this. More particularly, the force has to produce only those very restricted *types* of quark configuration which are observed in the hadron spectrum. Consider again the isospin multiplets in nuclear physics discussed in section 12.1.2. There is one very striking difference in the particle physics case: for mesons *only* $T = 0, \frac{1}{2}$ and 1 occur, and for baryons *only* $T = 0, \frac{1}{2}$, 1 and $\frac{3}{2}$, while in nuclei there is nothing in principle to stop us finding $T = \frac{5}{2}, 3, \dots$ states. (In fact such nuclear states are hard to identify experimentally, because they occur at high excitation energy for some of the isobars – cf figure 1.8(c) – where the levels are very dense). The same restriction holds for $\mathrm{SU}(3)_\mathrm{f}$ also – only **1**'s and **8**'s occur for mesons; and only **1**'s, **8**'s and **10**'s for baryons. In quark terms, this of course is what is translated into the recipe: 'mesons are $\bar{\mathrm{q}}\mathrm{q}$, baryons are qqq'. It is as if we said, in nuclear physics, that only $A = 2$ and $A = 3$ nuclei exist! Thus the quark forces must have a dramatic saturation property: apparently no $\bar{\mathrm{q}}\mathrm{qq}$, no qqqq, qqqqq, ... states exist. Furthermore, no qq or $\bar{\mathrm{q}}\bar{\mathrm{q}}$ states exist either – nor, for that matter, do single q's or $\bar{\mathrm{q}}$'s. All this can be summarized by saying that the quark colour degree of freedom must be *confined*, a property we shall now assume and return to in chapter 16.

If we assume that only colour singlet states exist (Fritzsch and Gell-Mann 1972, Bardeen, Fritzsch and Gell-Mann 1973), and that the strong interquark force depends only on colour, the fact that $\bar{\mathrm{q}}\mathrm{q}$ states are seen but qq and $\bar{\mathrm{q}}\bar{\mathrm{q}}$ are not gives us an important clue as to what group to associate with colour. One simple possibility might be that the three colours correspond to the components of an $\mathrm{SU}(2)_\mathrm{c}$ triplet '$\boldsymbol{\psi}$'. The antisymmetric, colour singlet, three-quark baryon wavefunction of (14.2) is then just the triple scalar product $\boldsymbol{\psi}_1 \cdot \boldsymbol{\psi}_2 \times \boldsymbol{\psi}_3$, which seems satisfactory. But what about the meson wavefunction? Mesons are formed of quarks and antiquarks, and we recall from sections 12.1.3 and

12.2 that antiquarks belong to the complex conjugate of the representation (or multiplet) to which quarks belong. Thus if a quark colour triplet wavefunction ψ_α transforms under a colour transformation as

$$\psi_\alpha \to \psi'_\alpha = V^{(1)}_{\alpha\beta}\psi_\beta \tag{14.10}$$

where $\mathbf{V}^{(1)}$ is a 3×3 unitary matrix appropriate to the $T = 1$ representation of SU(2) (cf (12.48) and (12.49)), then the wavefunction for the 'anti'-triplet is ψ^*_α, which transforms as

$$\psi^*_\alpha \to \psi^{*\prime}_\alpha = V^{(1)*}_{\alpha\beta}\psi^*_\beta. \tag{14.11}$$

Given this information, we can now construct colour singlet wavefunctions for mesons, built from $\bar{\text{q}}\text{q}$. Consider the quantity (cf (14.3)) $\sum_\alpha \psi^*_\alpha\psi_\alpha$ where ψ^* represents the antiquark and ψ the quark. This may be written in matrix notation as $\psi^\dagger\psi$ where the $\psi^\dagger$ as usual denotes the transpose of the complex conjugate of the column vector ψ. Then, taking the transpose of (14.11), we find that $\psi^\dagger$ transforms by

$$\psi^\dagger \to \psi^{\dagger\prime} = \psi^\dagger \mathbf{V}^{(1)\dagger} \tag{14.12}$$

so that the combination $\psi^\dagger\psi$ transforms as

$$\psi^\dagger\psi \to \psi^{\dagger\prime}\psi' = \psi^\dagger \mathbf{V}^{(1)\dagger}\mathbf{V}^{(1)}\psi = \psi^\dagger\psi \tag{14.13}$$

where the last step follows since $\mathbf{V}^{(1)}$ is unitary (compare (12.58)). Thus the product is *invariant* under (14.10) and (14.11) – that is, it is a colour singlet, as required. This is the meaning of the superposition (14.3).

All this may seem fine, but there is a problem. The three-dimensional representation of SU(2)_c which we are using here has a very special nature: the matrix $\mathbf{V}^{(1)}$ can be chosen to be *real*. This can be understood 'physically' if we make use of the great similarity between SU(2) and the group of rotations in three dimensions (which is the reason for the geometrical language of isospin 'rotations', and so on). We know very well how real three-dimensional vectors transform, namely by an orthogonal 3×3 matrix. It is the same in SU(2). It is always possible to choose the wavefunctions $\boldsymbol{\psi}$ to be real, and the transformation matrix $\mathbf{V}^{(1)}$ to be real also. Since $\mathbf{V}^{(1)}$ is, in general, unitary, this means that it must be orthogonal. But now the basic difficulty appears: there is no distinction between $\boldsymbol{\psi}$ and $\boldsymbol{\psi}^*$! They both transform by the real matrix $\mathbf{V}^{(1)}$. This means that we can make SU(2) invariant (colour singlet) combinations for $\bar{\text{q}}\bar{\text{q}}$ states, and for qq states, just as well as for $\bar{\text{q}}\text{q}$ states – indeed they are formally identical. But such 'diquark' (or 'antidiquark') states are not found, and hence – by assumption – should *not* be colour singlets.

The next simplest possibility seems to be that the three colours correspond to the components of an SU(3)_c triplet. In this case the quark colour wavefunction ψ_α transforms as (cf (12.74))

$$\psi \to \psi' = \mathbf{W}\psi \tag{14.14}$$

where $\mathbf{W}$ is a special unitary 3×3 matrix parametrized as

$$\mathbf{W} = \exp(\mathrm{i}\boldsymbol{\alpha} \cdot \boldsymbol{\lambda}/2), \tag{14.15}$$

and $\psi^\dagger$ transforms as

$$\psi^\dagger \to \psi^{\dagger\prime} = \psi^\dagger \mathbf{W}^\dagger. \tag{14.16}$$

The proof of the invariance of $\psi^\dagger\psi$ goes through as in (14.13), and it can be shown (problem 14.1(a)) that the antisymmetric 3q combination (14.2) is also an $\mathrm{SU(3)_c}$ invariant. Thus both the proposed meson and baryon states are colour singlets. It is *not* possible to choose the $\boldsymbol{\lambda}$'s to be pure imaginary in (14.15), and thus the 3×3 $\mathbf{W}$ matrices of $\mathrm{SU(3)_c}$ cannot be real, so that there is a distinction between ψ and ψ^*, as we learned in section 12.2. Indeed, it can be shown (see Carruthers 1966, chapter 3, Jones 1990, chapter 8, and also problem 14.1(b)) that, unlike the case of $\mathrm{SU(2)_c}$ triplets, it is not possible to form an $\mathrm{SU(3)_c}$ colour singlet combination out of two colour triplets qq or anti-triplets $\bar{\mathrm{q}}\bar{\mathrm{q}}$. Thus $\mathrm{SU(3)_c}$ seems to be a possible and economical choice for the colour group.

14.2.2 Global $\mathrm{SU(3)_c}$ invariance, and 'scalar gluons'

As stated above, we are assuming, on empirical grounds, that the only physically observed hadronic states are colour singlets – and this now means singlets under $\mathrm{SU(3)_c}$. What sort of interquark force could produce this dramatic result? Consider an SU(2) analogy again, the interaction of two nucleons belonging to the lowest (doublet) representation of SU(2). Labelling the states by an isospin T, the possible T values for two nucleons are $T = 1$ (triplet) and $T = 0$ (singlet). We know of an isospin-dependent force which can produce a splitting between these states, namely $V\boldsymbol{\tau}_1 \cdot \boldsymbol{\tau}_2$, where the '1' and '2' refer to the two nucleons. The total isospin is $\boldsymbol{T} = \frac{1}{2}(\boldsymbol{\tau}_1 + \boldsymbol{\tau}_2)$, and we have

$$\boldsymbol{T}^2 = \frac{1}{4}(\boldsymbol{\tau}_1^2 + 2\boldsymbol{\tau}_1 \cdot \boldsymbol{\tau}_2 + \boldsymbol{\tau}_2^2) = \frac{1}{4}(3 + 2\boldsymbol{\tau}_1 \cdot \boldsymbol{\tau}_2 + 3) \tag{14.17}$$

whence

$$\boldsymbol{\tau}_1 \cdot \boldsymbol{\tau}_2 = 2\boldsymbol{T}^2 - 3. \tag{14.18}$$

In the triplet state $\boldsymbol{T}^2 = 2$, and in the singlet state $\boldsymbol{T}^2 = 0$. Thus

$$(\boldsymbol{\tau}_1 \cdot \boldsymbol{\tau}_2)_{T=1} = 1 \tag{14.19}$$

$$(\boldsymbol{\tau}_1 \cdot \boldsymbol{\tau}_2)_{T=0} = -3 \tag{14.20}$$

and if V is positive the $T = 0$ state is pulled down. A similar thing happens in $\mathrm{SU(3)_c}$. Suppose this interquark force depended on the quark colours via a term proportional to

$$\boldsymbol{\lambda}_1 \cdot \boldsymbol{\lambda}_2. \tag{14.21}$$

Then, in just the same way, we can introduce the total colour operator

$$\boldsymbol{F} = \frac{1}{2}(\boldsymbol{\lambda}_1 + \boldsymbol{\lambda}_2), \tag{14.22}$$

so that

$$\boldsymbol{F}^2 = \frac{1}{4}(\boldsymbol{\lambda}_1^2 + 2\boldsymbol{\lambda}_1 \cdot \boldsymbol{\lambda}_2 + \boldsymbol{\lambda}_2^2) \tag{14.23}$$

and

$$\boldsymbol{\lambda}_1 \cdot \boldsymbol{\lambda}_2 = 2\boldsymbol{F}^2 - \boldsymbol{\lambda}^2, \tag{14.24}$$

where $\boldsymbol{\lambda}_1^2 = \boldsymbol{\lambda}_2^2 = \boldsymbol{\lambda}^2$, say. Here $\boldsymbol{\lambda}^2 \equiv \sum_{a=1}^{8}(\lambda_a)^2$ is found (see (12.75)) to have the value 16/3 (the unit matrix being understood). The operator $\boldsymbol{F}^2$ commutes with all components of $\boldsymbol{\lambda}_1$ and $\boldsymbol{\lambda}_2$ (as $\boldsymbol{T}^2$ does with $\boldsymbol{\tau}_1$ and $\boldsymbol{\tau}_2$) and represents the quadratic Casimir operator $\hat{C}_2$ of $\mathrm{SU(3)_c}$ (see section M.5 of appendix M), in the colour space of the two quarks considered here. The eigenvalues of $\hat{C}_2$ play a very important role in $\mathrm{SU(3)_c}$, analogous to that of the total spin/angular momentum in SU(2). They depend on the $\mathrm{SU(3)_c}$ representation: indeed, they are one of the defining labels of SU(3) representations in general (see section M.5). Two quarks, each in the representation $\mathbf{3}_\mathrm{c}$, combine to give a $\mathbf{6}_\mathrm{c}$-dimensional representation and a $\mathbf{3}_\mathrm{c}^*$ (see problem 14.1(b), and Jones (1990) chapter 8). The value of $\hat{C}_2$ for the singlet $\mathbf{6}_\mathrm{c}$ representation is 10/3, and for the $\mathbf{3}_\mathrm{c}^*$ representation is 4/3. Thus the '$\boldsymbol{\lambda}_1 \cdot \boldsymbol{\lambda}_2$' interaction will produce a negative (attractive) eigenvalue -8/3 in the $\mathbf{3}_\mathrm{c}^*$ states, but a repulsive eigenvalue +4/3 in the $\mathbf{6}_\mathrm{c}$ states, for two quarks.

The maximum attraction will clearly be for states in which $\boldsymbol{F}^2$ is zero. This is the singlet representation $\mathbf{1}_\mathrm{c}$. Two quarks cannot combine to give a colour singlet state, but we have seen in section 12.2 that a quark and an antiquark can: they combine to give $\mathbf{1}_\mathrm{c}$ and $\mathbf{8}_\mathrm{c}$. In this case (14.24) is replaced by

$$\boldsymbol{\lambda}_1 \cdot \boldsymbol{\lambda}_2 = 2\boldsymbol{F}^2 - \frac{1}{2}(\boldsymbol{\lambda}_1^2 + \boldsymbol{\lambda}_2^2), \tag{14.25}$$

where '1' refers to the quark and '2' to the antiquark. Thus the '$\boldsymbol{\lambda}_1 \cdot \boldsymbol{\lambda}_2$' interaction will give a repulsive eigenvalue +2/3 in the $\mathbf{8}_\mathrm{c}$ channel, for which $\hat{C}_2 = 3$, and a 'maximally attractive' eigenvalue -16/3 in the $\mathbf{1}_\mathrm{c}$ channel, for a quark and an antiquark.

In the case of baryons, built from three quarks, we have seen that when two of them are coupled to the $\mathbf{3}_\mathrm{c}^*$ state, the eigenvalue of $\boldsymbol{\lambda}_1 \cdot \boldsymbol{\lambda}_2$ is -8/3, one half of the attraction in the $\bar{q}q$ colour singlet state, but still strongly attractive. The (qq) pair in the $\mathbf{3}_\mathrm{c}^*$ state can then couple to the remaining third quark to make the overall colour singlet state (14.2), with maximum binding.

Of course, such a simple potential model does not imply that the energy difference between the $\mathbf{1}_\mathrm{c}$ states and all coloured states is *infinite*, as our strict 'colour singlets only' hypothesis would demand, and which would be one (rather crude) way of interpreting confinement. Nevertheless, we can ask: what single particle exchange process between quark (or antiquark) colour triplets produces a $\boldsymbol{\lambda}_1 \cdot \boldsymbol{\lambda}_2$ type of term? The answer is the exchange of

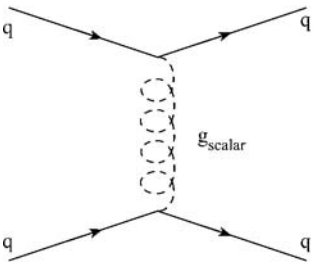


FIGURE 14.4
Scalar gluon exchange between two quarks.

an $\mathrm{SU(3)_c}$ octet ($\mathbf{8}_\mathrm{c}$) of particles, which (anticipating somewhat) we shall call gluons. Since colour is an exact symmetry, the quark wave equation describing the colour interactions must be $\mathrm{SU(3)_c}$ covariant. A simple such equation is

$$(\mathrm{i}\,\partial\!\!\!/ - m)\psi = g_\mathrm{s}\frac{\lambda_a}{2}A_a\psi \tag{14.26}$$

where g_s is a 'strong charge' and A_a ($a = 1, 2, \ldots, 8$) is an octet of *scalar* 'gluon potentials'. Equation (14.26) may be compared with (13.58): in the latter, $A\!\!\!/_a$ appears on the right-hand side, because the gauge field quanta are vectors rather than scalars. In (14.26), we are dealing at this stage only with a *global* SU(3) symmetry, not a local SU(3) gauge symmetry, and so the potentials may be taken to be scalars, for simplicity. As in (13.60), the vertex corresponding to (14.26) is

$$-\mathrm{i}g_\mathrm{s}\lambda_a/2. \tag{14.27}$$

(14.27) differs from (13.60) simply in the absence of the γ^μ factor, due to the assumed scalar, rather than vector, nature of the 'gluon' here. When we put two such vertices together and join them with a gluon propagator (figure 14.4), the $\mathrm{SU(3)_c}$ structure of the amplitude will be

$$\frac{\lambda_{1a}}{2}\delta_{ab}\frac{\lambda_{2b}}{2} = \frac{\boldsymbol{\lambda}_1}{2}\cdot\frac{\boldsymbol{\lambda}_2}{2} \tag{14.28}$$

the δ_{ab} arising from the fact that the freely propagating gluon does not change its colour. This interaction has exactly the required '$\boldsymbol{\lambda}_1 \cdot \boldsymbol{\lambda}_2$' character in the colour space.

14.2.3 Local $\mathrm{SU(3)_c}$ invariance: the QCD Lagrangian

It is tempting to suppose (Fritzsch and Gell-Mann 1972, Fritzsch, Gell-Mann and Leutwyler 1973) that the 'scalar gluons' introduced in (14.26) are, in fact, vector particles, like the photons of QED. Equation (14.26) then becomes

$$(\mathrm{i}\,\partial\!\!\!/ - m)\psi = g_\mathrm{s}\frac{\lambda_a}{2}A\!\!\!/_a\psi \tag{14.29}$$

as in (13.58), and the vertex (14.27) becomes

$$-\mathrm{i}g_\mathrm{s}\frac{\lambda_a}{2}\gamma^\mu \tag{14.30}$$

as in (13.60). One motivation for this is the desire to make the colour dynamics as much as possible like the highly successful theory of QED, and to derive the dynamics from a gauge principle. As we have seen in the last chapter, this involves the simple but deep step of supposing that the quark wave equation is covariant under *local* $\mathrm{SU(3)_c}$ transformations of the form

$$\psi \to \psi' = \exp(\mathrm{i}g_\mathrm{s}\boldsymbol{\alpha}(x)\cdot\boldsymbol{\lambda}/2)\psi. \tag{14.31}$$

This is implemented by the replacement

$$\partial_\mu \to \partial_\mu + \mathrm{i}g_\mathrm{s}\frac{\lambda_a}{2}A_{a\mu}(x) \tag{14.32}$$

in the Dirac equation for the quarks, which leads immediately to (14.29) and the vertex (14.30).

Of course, the assumption of local $\mathrm{SU(3)_c}$ covariance leads to a great deal more: for example, it implies that the gluons are *massless vector* (spin 1) particles, and that they interact with themselves via *three-gluon* and *four-gluon* vertices, which are the $\mathrm{SU(3)_c}$ analogues of the SU(2) vertices discussed in section 13.3.2. The most compact way of summarizing all this structure is via the Lagrangian, most of which we have already introduced in chapter 13. Gathering together (13.71) and (13.140) (adapted to $\mathrm{SU(3)_c}$), we write it out here for convenience:

$$\begin{aligned}\mathcal{L}_\mathrm{QCD} \;=\; & \sum_{\text{flavours f}} \bar{\hat{q}}_{\mathrm{f},\alpha}(\mathrm{i}\hat{\not{D}} - m_\mathrm{f})_{\alpha\beta}\hat{q}_{\mathrm{f},\beta} - \frac{1}{4}\hat{F}_{a\mu\nu}\hat{F}_a^{\mu\nu} \\ & - \frac{1}{2\xi}(\partial_\mu\hat{A}^\mu_a)(\partial_\nu\hat{A}^\nu_a) + \partial_\mu\hat{\eta}^\dagger_a\hat{D}^\mu_{ab}\hat{\eta}_b. \end{aligned}\tag{14.33}$$

In (14.33), repeated indices are as usual summed over: α and β are $\mathrm{SU(3)_c}$-triplet indices running from 1 to 3, and a, b are $\mathrm{SU(3)_c}$-octet indices running from 1 to 8. The covariant derivatives are defined by

$$(\hat{D}_\mu)_{\alpha\beta} = \partial_\mu\delta_{\alpha\beta} + \mathrm{i}g_\mathrm{s}\frac{1}{2}(\lambda_a)_{\alpha\beta}\hat{A}_{a\mu} \tag{14.34}$$

when acting on the quark $\mathrm{SU(3)}_c$ triplet, as in (13.53), and by

$$(\hat{D}_\mu)_{ab} = \partial_\mu\delta_{ab} + g_s f_{cab}\hat{A}_{c\mu} \tag{14.35}$$

when acting on the octet of ghost fields. For the second of these, note that the matrices representing the SU(3) generators in the octet representation are as given in (12.84), and these take the place of the '$\lambda/2$' in (14.34) (compare (13.141) in the SU(2) case). We remind the reader that the last two terms

in (14.33) are the gauge-fixing and ghost terms, respectively, appropriate to a gauge field propagator of the form (13.69) (with δ_{ij} replaced by δ_{ab} here). The Feynman rules following from (14.33) are given in appendix Q.

As remarked in section 12.3.2, the fact that the QCD interactions (14.33) are 'flavour-blind' implies that the global flavour symmetries discussed in chapter 12 are all preserved by QCD. These include the conservation of each quark flavour (for example, the number of strange quarks minus the number of strange antiquarks is conserved); and the symmetries $\mathrm{SU}(2)_\mathrm{f}$ and $\mathrm{SU}(3)_\mathrm{f}$, and the chiral symmetries $\mathrm{SU}(2)_{5\mathrm{f}}$ and $\mathrm{SU}(3)_{5\mathrm{f}}$, to the extent that these latter are good symmetries. Further, (14.33) conserves the discrete symmetries **P**, **C** and **T**, in a manner quite analogous to QED, already covered in section 7.5. In the case of **P** and **T**, the gluon fields $\hat{A}_{a\mu}$ have the same transformation properties as the photon field $\hat{A}_\mu$, and the (normally ordered) $\mathrm{SU}(3)_\mathrm{c}$ currents $\hat{j}^\mu_{\mathrm{f}a} = \bar{\hat{q}}_\mathrm{f}\gamma^\mu \frac{1}{2}\lambda_a \hat{q}_\mathrm{f}$ transform in the same way as the electromagnetic current $\hat{q}\gamma^\mu\hat{q}$, ensuring **P** and **T** invariance. Under **C**, the quark fields transform as usual according to (7.151). Charge conjugation for the gluon field needs a little more care. The required rule is

$$\hat{\mathbf{C}}\lambda_a \hat{A}_{a\mu}\hat{\mathbf{C}}^{-1} = -\lambda_a^* \hat{A}_{a\mu}. \tag{14.36}$$

The overall minus sign in (14.36) is analogous to that for the photon field (cf (7.152)). To understand the complex conjugate on the right-hand side of (14.36), recall from (7.153) that the complex scalar field $\hat{\phi} = \frac{1}{\sqrt{2}}(\hat{\phi}_1 - \mathrm{i}\hat{\phi}_2)$ transforms according to

$$\hat{\mathbf{C}}(\hat{\phi}_1 - \mathrm{i}\hat{\phi}_2)\hat{\mathbf{C}}^{-1} = \hat{\phi}_1 + \mathrm{i}\hat{\phi}_2. \tag{14.37}$$

Problem 14.2(a) verifies that the (normally ordered) interaction $\hat{j}^\mu_{\mathrm{f}a}\hat{A}_{a\mu}$ is then **C**-invariant. As regards the term $\hat{F}_{a\mu\nu}\hat{F}_a^{\mu\nu}$, we can write it as

$$\frac{1}{2}\mathrm{Tr}(\lambda_a \hat{F}_{a\mu\nu}\lambda_b \hat{F}_b^{\mu\nu}) \tag{14.38}$$

using the relation

$$\mathrm{Tr}(\lambda_a\lambda_b) = 2\delta_{ab}. \tag{14.39}$$

A short calculation (problem 14.2(b)) shows that $\lambda_a\hat{F}_{a\mu\nu}$ transforms under **C** the same way as $\lambda_a\hat{A}_{a\mu}$ (i.e. according to (14.36)). Using the complex conjugate of (14.39), it then follows that (14.38) is invariant under **C**.

14.2.4 The θ-term

In arriving at (14.33) we have relied essentially on the 'gauge principle' (invariance under a local symmetry) and the requirement of renormalizability (to forbid the presence of terms with mass dimension higher than 4). The renormalizability of such a theory was proved by 't Hooft (1971a, b). However,

there is in fact one more gauge invariant term of mass dimension 4 which can be written down, namely

$$\hat{\mathcal{L}}_\theta = \frac{\theta g_{\rm s}^2}{64\pi^2}\epsilon_{\mu\nu\rho\sigma}\hat{F}_a^{\mu\nu}\hat{F}_a^{\rho\sigma}; \tag{14.40}$$

this is the 'θ-term' of QCD. A full discussion of this term (see for example Weinberg 1996, section 23.6) is beyond our scope, but we shall give a brief introduction to the main ideas.

The reader may wonder, first of all, whether the θ-term should give rise to a new Feynman rule. The answer to this begins by noting that (14.40) can actually be written as a total divergence:

$$\epsilon_{\mu\nu\rho\sigma}\hat{F}_a^{\mu\nu}\hat{F}_a^{\rho\sigma} = \partial_\mu \hat{K}^\mu. \tag{14.41}$$

This is more easily seen in the analogous term for QED, namely $\epsilon_{\mu\nu\rho\sigma}\hat{F}^{\mu\nu}\hat{F}^{\rho\sigma}$. We have

$$\begin{aligned}
\epsilon_{\mu\nu\rho\sigma}\hat{F}^{\mu\nu}\hat{F}^{\rho\sigma} &= \epsilon_{\mu\nu\rho\sigma}(\partial^\mu\hat{A}^\nu - \partial^\nu\hat{A}^\mu)(\partial^\rho\hat{A}^\sigma - \partial^\sigma\hat{A}^\rho) && (14.42)\\
&= 4\epsilon_{\mu\nu\rho\sigma}\partial^\mu\hat{A}^\nu\partial^\rho\hat{A}^\sigma && (14.43)\\
&= \partial^\mu(4\epsilon_{\mu\nu\rho\sigma}\hat{A}^\nu\partial^\rho\hat{A}^\sigma), && (14.44)
\end{aligned}$$

where we have used the antisymmetry of the ϵ symbol in (14.43), and also in (14.44) since the contraction of ϵ with the symmetric tensor $\partial^\mu\partial^\rho$ vanishes. We shall not need the explicit form of $\hat{K}^\mu$.

Any total divergence in a Lagrangian can be integrated to give only a 'surface' term in the action, which can usually be discarded, making conventional assumptions about the vanishing of the fields at spatial infinity. There are, however, field configurations ('instantons') which do contribute to the θ-term. Such configurations are not reachable in perturbation theory, and so no perturbative Feynman rules are associated with (14.40). They approach a pure gauge form at spatial infinity, and are therefore associated with the QCD vacuum state; their effect is equivalent to including the term (14.40) in the QCD Lagrangian (see for example Rajaraman 1982).

The term (14.40) has potentially important phenomenological implications, since it conserves **C** but violates both **P** and **T** (and hence also **CP**). Again, this is easy to see in the QED analogue term (14.42), which equals $8\hat{\boldsymbol{E}}\cdot\hat{\boldsymbol{B}}$ (problem 14.3): we recall that under **P**, $\hat{\boldsymbol{E}} \to -\hat{\boldsymbol{E}}$ and $\hat{\boldsymbol{B}} \to \hat{\boldsymbol{B}}$, while under **T**, $\hat{\boldsymbol{E}} \to \hat{\boldsymbol{E}}$ and $\hat{\boldsymbol{B}} \to -\hat{\boldsymbol{B}}$. But we know (section 4.2) that strong interactions conserve both **P** and **T** to a high degree of accuracy. In particular, the neutron electric dipole moment $d_{\rm n}$, which would violate both **P** and **T**, is extremely small (see (4.133)). A very crude estimate of the size of $d_{\rm n}$, induced by the θ-term, is given by dimensional analysis as

$$d_{\rm n} \sim \frac{e}{M_{\rm n}}\theta, \tag{14.45}$$

where $M_{\rm n}$ is the neutron mass. This would imply $\theta < 10^{-12}$. In fact, this estimate is too restrictive, since it turns out (Weinberg 1996, section 23.6) that if any quark has zero mass, θ can be reduced to zero by a global chiral U(1) transformation on that quark field. Although neither of the u and d quark masses are zero, they are small on a hadronic scale, and a suppression of (14.45) is expected, increasing the bound on theta. Estimates suggest $\theta < 10^{-9} - 10^{-10}$.

This may seem an unsatisfactorily special value to force on a dimensionless Lagrangian parameter, when there is nothing in the theory, *a priori*, to prevent something of order unity. This perceived difficulty is referred to as the 'strong **CP** problem'. A possible solution to the problem, in which a very small value of θ could arise naturally was suggested by Peccei and Quinn (1977a, 1977b). Their idea goes beyond the Standard Model, and involves the existence of a new very light pseudoscalar particle, the *axion* (Wilczek 1978, Winberg 1978).

We proceed now with the main topic of this chapter, which is the application of perturbative QCD.

14.3 Hard scattering processes, QCD tree graphs, and jets

14.3.1 Introduction

The fundamental distinctive feature of non-Abelian gauge theories is that they are 'asymptotically free', meaning that the effective coupling strength becomes progressively smaller at short distances, or high energies (Gross and Wilczek 1973, Politzer 1973). This property is the most compelling theoretical motivation for choosing a non-Abelian gauge theory for the strong interactions, and it enables a quantitative perturbative approach to be followed (in appropriate circumstances) even in strong interaction physics. This programme has indeed been phenomenally successful, firmly establishing QCD as the theory of strong interactions, and now – in the era of the LHC – serving as a precision tool to guide searches for new physics.

A proper understanding of how this works necessitates a considerable detour, however, into the physics of renormalization. In particular, we need to understand the important cluster of ideas going under the general heading of the 'renormalization group', and this will be the topic of chapter 15. For the moment we proceed with a discussion of some simple tree-level applications of QCD, which provided early confrontation of QCD with experiment.

Let us begin by recapitulating, from a QCD-informed viewpoint, how the parton model successfully interpreted deep inelastic and large-Q^2 data in terms of almost free point-like partons – now to be identified with the QCD quanta: quarks, antiquarks, and gluons.

In section 9.5 we briefly introduced the idea of *jets* in e^+e^- physics: two well collimated sprays of hadrons, apparently created as a quark–antiquark pair separate from each other at high speed. The angular distribution of the two jets followed closely the distribution expected from the parton-level process $e^+e^- \to \bar{q}q$. The dynamics at the parton level was governed by QED, but QCD is responsible for the way the emerging q and $\bar{q}$ turn themselves into hadrons, a process called parton fragmentation (it occurs for gluons too). We may think of it as proceeding in two stages. First, as the rapidly moving q and $\bar{q}$ begin to separate, they develop perturbative showers of narrowly collimated gluons and quark–antiquark pairs. Then, as the partons separate further, the strength of the forces between them increases, becoming strongly non-perturbative at a separation of about 1 fm, and ensuring that the coloured quanta are all confined into hadrons. As yet we do not have a completely quantitative dynamical understanding of the second, hadronization, stage: it is implemented by means of a model. Nevertheless, we can argue that for the forces to be strong enough to produce the observed hadrons, the dominant processes in hadronization must involve small momentum transfers – that is, the exchange of 'soft' quanta. Thus the emerging hadrons are also well collimated into two jets, whose energy and angular distributions reflect the short-distance physics at the parton level. This simple 2-jet picture will be extended in section 14.4, where we consider $e^+e^- \to 3$ jets.

A somewhat different aspect of parton physics arose in sections 9.2–9.3, where we considered deep inelastic electron scattering from nucleons. There the initial state contained one hadron. Correspondingly, one parton appeared in the *initial* state of the parton-level interaction, and the analysis required new functions measuring the probabilities of finding a particular parton in the parent hadron – the parton distribution functions. These too are beyond the reach of perturbation theory.

We may also consider, finally, hadron-hadron collisions. In this case, we need all three of the features we have been discussing: the parton distribution functions, to provide the intial parton-parton state from the two-hadron state; the perturbative short-distance parton-parton interaction; and the parton fragmentation process in the final state. These three parts to the process are pictured in figure 14.5. The identification and analysis of short distance parton-parton interactions provide direct tests of the tree-graph structure of QCD, and perturbative corrections to it.

This three-part schematization of certain features of hadronic interactions is useful, because although we cannot yet calculate from first principles either the parton distribution functions or the fragmentation process, both are *universal*. The quark and gluon composition of hadrons is the same for all processes, and so measurements in one experiment can be used to predict the results of others. We saw an example of this in the Drell–Yan process of section 9.4. As regards the fragmentation stage, this too will be universal, provided one is interested in sufficiently inclusive aspects of the final state. The

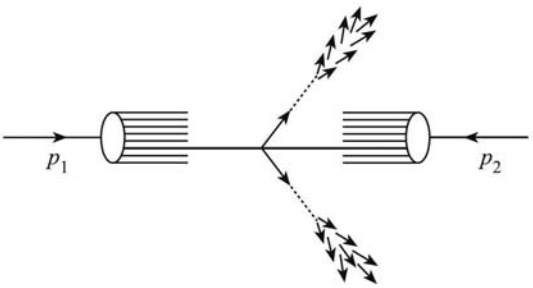


FIGURE 14.5
Hadron-hadron collision involving parton-parton interaction followed by parton fragmentation.

three-part scheme is called *factorization*, and it has been rigorously proved for some cases. We shall return to factorization in section 15.7.

Let us turn now to some of the early data on parton-parton interactions in hadron-hadron collisions.

14.3.2 Two-jet events in $\bar{p}p$ collisions

How are short-distance parton-parton interactions to be identified experimentally? The answer is: in just the same way as Rutherford distinguished the presence of a small heavy scattering centre (the nucleus) in the atom: by looking at secondary particles emerging at large angles with respect to the beam direction. For each secondary particle we can define a transverse momentum $p_T = p \sin\theta$ where p is the particle momentum and θ is the emission angle with respect to the beam axis. If hadronic matter were smooth and uniform (cf the Thomson atom), the distribution of events in p_T would be expected to fall off very rapidly at large p_T values – perhaps exponentially. This is just what is observed in the vast majority of events: the average value of p_T measured for charged particles is very low ($\langle p_T \rangle \sim 0.4$ GeV), but in a small fraction of collisions the emission of high-p_T secondaries is observed. They were first seen (Büsser *et al.* 1972, 1973, Alper *et al.* 1973, Banner *et al.* 1982) at the CERN ISR (CMS energies 30-62 GeV), and were interpreted in parton terms as previously indicated. Referring to figure 14.5, a parton from one hadron undergoes a short-distance 'hard scattering' interaction with a parton from the other, leading in lowest-order perturbation theory to two wide-angle partons, which then fragment into two jets.

We now face the experimental problem of picking out, from the enormous multiplicity of total events, just these hard scattering ones, in order to analyse them further. Early experiments used a trigger based on the detection of a single high-p_T particle. But it turns out that such triggering really reduces

the probability of observing jets, since the probability that a single hadron in a jet will actually carry most of the jet's total transverse momentum is quite small (Jacob and Landshoff 1978; Collins and Martin 1984, Chapter 5). It is much better to surround the collision volume with an array of calorimeters which measure the total energy deposited. *Wide-angle jets* can then be identified by the occurrence of a large amount of total transverse energy deposited in a number of adjacent calorimeter cells: this is then a 'jet trigger'. The importance of calorimetric triggers was first emphasized by Bjorken (1973), following earlier work by Berman, Bjorken and Kogut (1971). The application of this method to the detection and analysis of wide-angle jets was first reported by the UA2 collaboration at the CERN $\bar{\text{p}}\text{p}$ collider (Banner *et al.* 1982). An impressive body of quite remarkably clean jet data was subsequently accumulated by both the UA1 and UA2 collaborations (at $\sqrt{s} = 546$ GeV and 630 GeV), and by the CDF and D0 collaborations at the FNAL Tevatron collider ($\sqrt{s} = 1.8$ TeV).

For each event the total transverse energy $\sum E_{\text{T}}$ is measured where

$$\sum E_{\text{T}} = \sum_i E_i \sin\theta_i. \tag{14.46}$$

E_i is the energy deposited in the ith calorimeter cell and θ_i is the polar angle of the cell centre; the sum extends over all cells. Figure 14.6 shows the $\sum E_{\text{T}}$ distribution observed by UA2: it follows the 'soft' exponential form for $\sum E_{\text{T}} \leq 60$ GeV, but thereafter departs from it, showing clear evidence of the wide-angle collisions characteristic of hard processes.

As we shall see shortly, the majority of 'hard' events are of two-jet type, with the jets sharing the $\sum E_{\text{T}}$ approximately equally. Thus a 'local' trigger set to select events with localized transverse energy ≥ 30 GeV and/or a 'global' trigger set at ≥ 60 GeV can be used. At $\sqrt{s} \geq 500$–600 GeV there is plenty of energy available to produce such events.

The total $\sqrt{s}$ value is important for another reason. Consider the kinematics of the two-parton collision (figure 14.5) in the $\bar{\text{p}}\text{p}$ CMS. As in the Drell–Yan process of section 9.4, the right-moving parton has 4-momentum

$$x_1 p_1 = x_1(P, 0, 0, P) \tag{14.47}$$

and the left-moving one

$$x_2 p_2 = x_2(P, 0, 0, -P) \tag{14.48}$$

where $P = \sqrt{s}/2$ and we are neglecting parton transverse momenta, which are approximately limited by the observed $\langle p_{\text{T}} \rangle$ value (~ 0.4 GeV, and thus negligible on this energy scale). Consider the simple case of 90^0 scattering, which requires (for massless partons) $x_1 = x_2$, equal to x say. The total outgoing transverse energy is then $2xP = x\sqrt{s}$. If this is to be greater than 50 GeV, then partons with $x \geq 0.1$ will contribute to the process. The parton distribution functions are large at these relatively small x values, due to sea

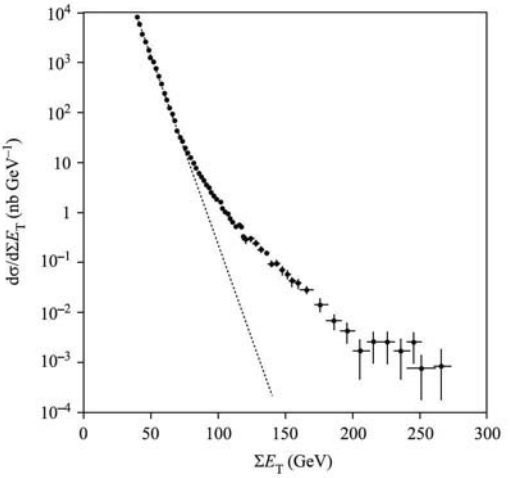


FIGURE 14.6
Distribution of the total transverse energy $\sum E_{\mathrm{T}}$ observed in the UA2 central calorimeter (DiLella 1985).

quarks (section 9.3) and gluons (figure 9.9), and thus we expect to obtain a reasonable cross section.

What are the characteristics of jet events? When $\sum E_{\mathrm{T}}$ is large enough (≥ 150 GeV), it is found that essentially all of the transverse energy is indeed split roughly equally between two approximately back-to-back jets. A typical such event is shown in figure 14.7. Returning to the kinematics of (14.47) and (14.48), x_1 will not in general be equal to x_2, so that – as is apparent in figure 14.7 – the jets will not be collinear. However, to the extent that the transverse parton momenta can be neglected, the jets will be coplanar with the beam direction, i.e. their relative azimuthal angle will be 180^0. Figure 14.8 shows a number of examples in which the distribution of the transverse energy over the calorimeter cells is analyzed as a function of the jet opening angle θ and the azimuthal angle ϕ. It is strikingly evident that we are seeing precisely a kind of 'Rutherford' process, or – to vary the analogy – we might say that hadronic jets are acting as the modern counterpart of Faraday's iron filings, in rendering visible the underlying field dynamics!

We may now consider more detailed features of these two-jet events – in particular, the expectations based on QCD tree graphs. The initial hadrons provide wide-band beams of quarks, antiquarks and gluons[2]; thus we shall have many parton subprocesses, such as $\mathrm{qq} \to \mathrm{qq}$, $\mathrm{q\bar{q}} \to \mathrm{q\bar{q}}$, $\mathrm{q\bar{q}} \to \mathrm{gg}$, $\mathrm{gg} \to \mathrm{gg}$, etc. The most important, numerically, for a $\mathrm{p\bar{p}}$ collider are $\mathrm{q\bar{q}} \to \mathrm{q\bar{q}}$, $\mathrm{gq} \to \mathrm{gq}$

[2]In the sense that the partons in hadrons have momentum or energy distributions, which are characteristic of their localization to hadronic dimensions.

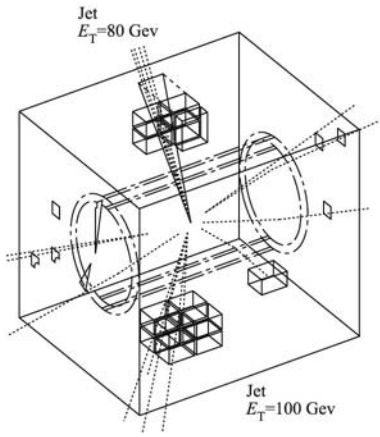


FIGURE 14.7
Two-jet event. Two tightly collimated groups of reconstructed charged tracks can be seen in the cylindrical central detector of UA1, associated with two large clusters of calorimeter energy depositions. Figure reprinted with permission from S Geer in *High Energy Physics 1985, Proc. Yale Advanced Study Institute* eds M J Bowick and F Gursey; copyright 1986 World Scientific Publishing Company.

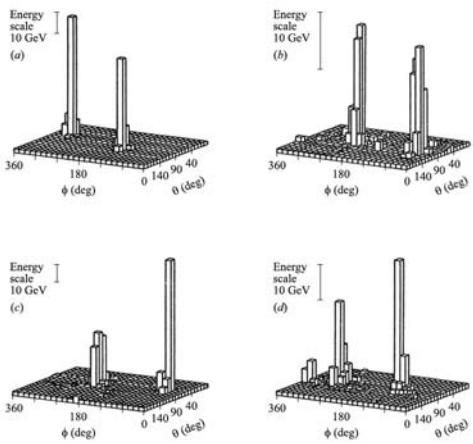


FIGURE 14.8
Four transverse energy distributions for events with $\sum E_T > 100$ GeV, in the θ, ϕ plane (UA2, DiLella 1985). Each bin represents a cell of the UA2 calorimeter. Note that the sum of the ϕ's equals 180^0 (mod 360^0).

TABLE 14.1
Spin-averaged squared matrix elements for one-gluon exchange ($\hat{t}$-channel) processes.

Subprocess	$\lvert\mathcal{M}\rvert^2$
$\left.\begin{matrix} qq \to qq \\ q\bar{q} \to q\bar{q} \end{matrix}\right\}$	$\frac{4}{9}\left(\frac{\hat{s}^2+\hat{u}^2}{\hat{t}^2}\right)$
$qg \to qg$	$\frac{\hat{s}^2+\hat{u}^2}{\hat{t}^2} + \ldots$
$gg \to gg$	$\frac{9}{4}\left(\frac{\hat{s}^2+\hat{u}^2}{\hat{t}^2}\right) + \ldots$

and $gg \to gg$. The cross section will be given, in the parton model, by a formula of the Drell–Yan type, except that the electromagnetic annihilation cross section

$$\sigma(q\bar{q} \to \mu^+\mu^-) = 4\pi\alpha^2/3q^2 \tag{14.49}$$

is replaced by the various QCD subprocess cross sections, each one being weighted by the appropriate distribution functions. At first sight this seems to be a very complicated story, with so many contributing parton processes. But a significant simplification comes from the fact that in the CMS of the parton collision, all processes involving one gluon exchange will lead to essentially the same dominant angular distribution of Rutherford-type, $\sim \sin^{-4}\theta/2$, where θ is the parton CMS scattering angle (recall section 1.3.6). This is illustrated in table 14.1 (taken from Combridge *et al.* 1977), which lists the different relevant spin averaged, squared, one-gluon-exchange matrix elements $|\,\mathcal{M}\,|^2$, where the parton differential cross section is given by (cf (6.129))

$$\frac{\mathrm{d}\sigma}{\mathrm{d}\cos\theta} = \frac{\pi\alpha_s^2}{2\hat{s}}\,|\,\mathcal{M}\,|^2\,. \tag{14.50}$$

Here $\alpha_s = g_s^2/4\pi$, and $\hat{s}$, $\hat{t}$ and $\hat{u}$ are the subprocess invariants, so that

$$\hat{s} = (x_1p_1 + x_2p_2)^2 = x_1x_2s \quad \text{(cf (9.84))}. \tag{14.51}$$

Continuing to neglect the parton transverse momenta, the initial parton configuration shown in figure 14.5 can be brought to the parton CMS by a Lorentz transformation along the beam direction, the outgoing partons then emerging back-to-back at an angle θ to the beam axis, so $\hat{t} \propto (1-\cos\theta) \propto \sin^2\theta/2$. Only the terms in $(\hat{t})^{-2} \sim \sin^{-4}\theta/2$ are given in table 14.1. We note that the $\hat{s}, \hat{t}, \hat{u}$ dependence of these terms is the same for the three types of process (and is in fact the same as that found for the 1γ exchange process $e^-\mu^- \to e^-\mu^-$: see problem 8.17, converting $\mathrm{d}\sigma/\mathrm{d}t$ into $\mathrm{d}\sigma/\mathrm{d}\cos\theta$). Figure 14.9 shows the two jet angular distribution measured by UA1 (Arnison *et al.* 1985). The broken

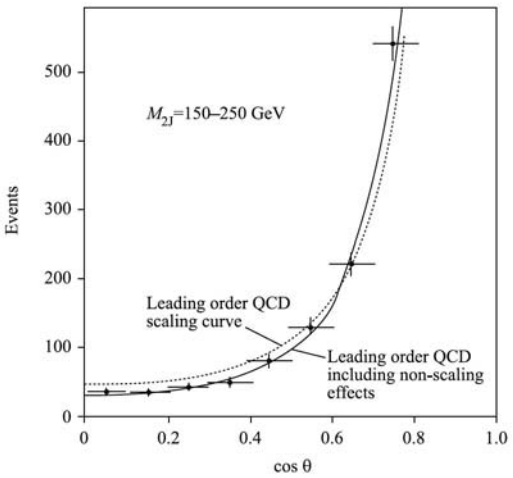


FIGURE 14.9
Two-jet angular distribution plotted against $\cos\theta$ (Arnison *et al.* 1985).

curve is the exact angular distribution predicted by all the QCD tree graphs – it actually follows the $\sin^{-4}\theta/2$ shape quite closely.

It is interesting to compare this angular distribution with the one predicted on the assumption that the exchanged gluon is a spinless particle, so that the vertices have the form '$\bar{u}u$' rather than '$\bar{u}\gamma_\mu u$'. Problem 14.4 shows that in this case the $1/\hat{t}^2$ factor in the cross section is completely cancelled, thus ruling out such a model.

This analysis provides compelling evidence for elementary hard scattering events proceeding via the exchange of a massless vector quantum. It is possible to go much further. Anticipating our later discussion, the small discrepancy between 'tree graph' theory (which is labelled 'leading order QCD scaling curve' in figure 14.9) and experiment can be accounted for by including corrections which are of higher order in α_s. The solid curve in figure 14.9 includes QCD corrections beyond the tree level, involving the 'running' of the coupling constant α_s and 'scaling violation' in the effective parton distribution functions, both of which effects will be discussed in the following chapter. The corrections lead to good agreement with experiment.

The fact that the angular distributions of all the subprocesses are so similar allows further information to be extracted from these two-jet data. In general, the parton model cross section will have the form (cf (9.91))

$$\frac{\mathrm{d}^3\sigma}{\mathrm{d}x_1\mathrm{d}x_2\mathrm{d}\cos\theta} = \sum_{\mathrm{a,b}} \frac{F_\mathrm{a}(x_1)}{x_1}\,\frac{F_\mathrm{b}(x_2)}{x_2} \sum_{\mathrm{c,d}} \frac{\mathrm{d}\sigma_{\mathrm{ab}\to\mathrm{cd}}}{\mathrm{d}\cos\theta} \tag{14.52}$$

where$F_\mathrm{a}(x_1)/x_1$ is the distribution function for partons of type 'a' (q, $\bar{\mathrm{q}}$ or g), and similarly for $F_\mathrm{b}(x_2)/x_2$. Using the near identity of all $\mathrm{d}\sigma/\mathrm{d}\cos\theta$'s, and

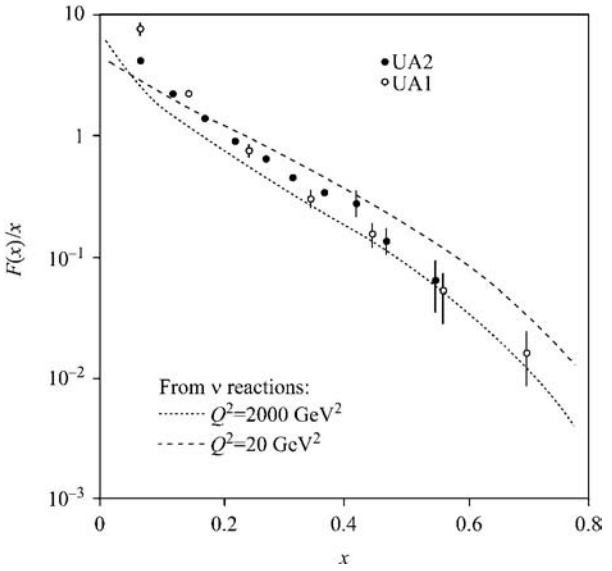


FIGURE 14.10
Effective distribution function measured from two-jet events (Arnison *et al.* 1984 and Bagnaia *et al.* 1984). The broken and chain curves are obtained from deep inelastic neutrino scattering. Taken from DiLella (1985).

noting the numerical factors in table 14.1, the sums over parton types reduce to

$$\frac{9}{4}\{g(x_1)+\frac{4}{9}[q(x_1)+\bar{q}(x_1)]\}\{g(x_2)+\frac{4}{9}[q(x_2)+\bar{q}(x_2)]\} \tag{14.53}$$

where $g(x)$, $q(x)$ and $\bar{q}(x)$ are the gluon, quark and antiquark distribution functions. Thus effectively the weighted distribution function[3]

$$\frac{F(x)}{x} = g(x)+\frac{4}{9}[q(x)+\bar{q}(x)] \tag{14.54}$$

is measured (Combridge and Maxwell, 1984); in fact, with the weights as in (14.53),

$$\frac{\mathrm{d}^3\sigma}{\mathrm{d}x_1\mathrm{d}x_2\mathrm{d}\cos\theta} = \frac{F(x_1)}{x_1}\cdot\frac{F(x_2)}{x_2}\cdot\frac{\mathrm{d}\sigma_{\mathrm{gg}\to\mathrm{gg}}}{\mathrm{d}\cos\theta}. \tag{14.55}$$

x_1 and x_2 are kinematically determined from the measured jet variables: from (14.51),

$$x_1x_2 = \hat{s}/s \tag{14.56}$$

where $\hat{s}$ is the invariant [mass]^2 of the two-jet system and

$$x_1 - x_2 = 2P_{\mathrm{L}}/\sqrt{s} \qquad (\text{cf (9.82)}) \tag{14.57}$$

[3]The $\frac{4}{9}$ reflects the relative strengths of the quark-gluon and gluon-gluon couplings in QCD; see problem 14.5.

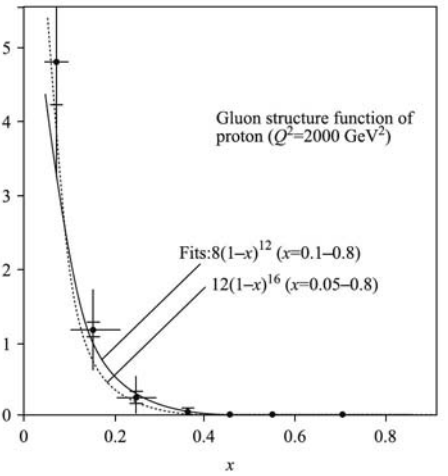


FIGURE 14.11
The gluon distribution function $g(x)$ extracted from the effective distribution function $F(x)$ by subtracting the expected contribution from the quarks and antiquarks. Figure reprinted with permission from S Geer in *High Energy Physics 1985, Proc. Yale Theoretical Advanced Study Institute*, eds M J Bowick and F Gursey; copyright 1986 World Scientific Publishing Company.

with P_{L} the total two-jet longitudinal momentum. Figure 14.10 shows $F(x)/x$ obtained in the UA1 (Arnison *et al.* 1984) and UA(2) (Bagnaia *et al.* 1984) experiments. Also shown in this figure is the expected $F(x)/x$ based on contemporary fits to the deep inelastic neutrino scattering data at $Q^2 = 20$ GeV^2 and 2000 GeV^2 (Abramovicz *et al.* 1982a,b, 1983); the reason for the change with Q^2 will be discussed in section 15.6. The agreement is qualitatively very satisfactory. Subtracting the distributions for quarks and antiquarks as found in deep inelastic lepton scattering, UA1 were able to deduce the gluon structure function $g(x)$ shown in figure 14.11. It is clear that gluon processes will dominate at small x – and even at larger x will be important because of the colour factors in table 14.1.

14.3.3 Three-jet events in $\bar{\mathrm{p}}\mathrm{p}$ collisions

Although most of the high- E_{T} events at hadron colliders are two-jet events, in some 10–30% of the cases the energy is shared between three jets. An example is included as (*d*) in the collection of figure 14.8; a clearer one is shown in figure 14.12. In QCD such events are interpreted as arising from a 2 parton $\rightarrow$ 2 parton + 1 gluon process of the type gg $\rightarrow$ ggg, gq $\rightarrow$ ggq, etc. Once again, one can calculate (Kunszt and Piétarinen 1980, Gottschalk

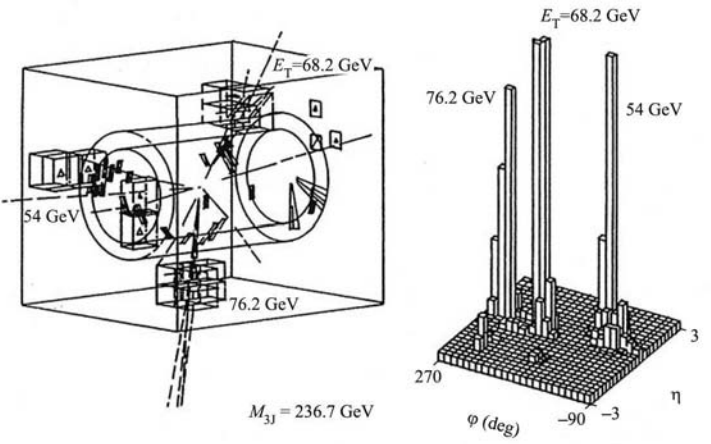


FIGURE 14.12
Three-jet event in the UA1 detector, and the associated transverse energy flow plot. Figure reprinted with permission from S Geer in *High Energy Physics 1985, Proc. Yale Theoretical Advanced Study Institute*, eds M J Bowick and F Gursey; copyright 1986 World Scientific Publishing Company.

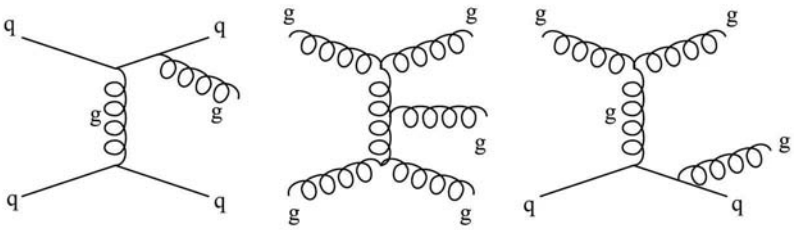


FIGURE 14.13
Some tree graphs associated with three-jet events.

and Sivers 1980, Berends *et al.* 1981) all possible contributing tree graphs, of the kind shown in figure 14.13, which should dominate at small α_s. They are collectively known as QCD single-bremsstrahlung diagrams. Analysis of triple jets which are well separated both from each other and from the beam directions shows that the data are in good agreement with these lowest-order QCD predictions. For example, figure 14.14 shows the production angular distribution of UA2 (Appel *et al.* 1986) as a function of $\cos\theta^*$, where θ^* is the angle between the leading (most energetic) jet momentum and the beam axis, in the three-jet CMS. It follows just the same $\sin^{-4}\theta^*/2$ curve as in the two-jet case (the data for which are also shown in the figure), as expected

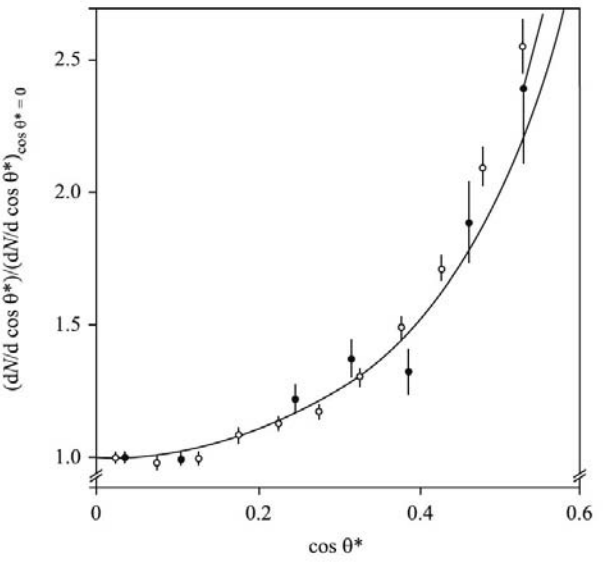


FIGURE 14.14
The distribution of $\cos\theta^*$(•), the angle of the leading jet with respect to the beam line (normalized to unity at $\cos\theta^* = 0$), for three-jet events in $\bar{p}p$ collisions (Appel *et al.* 1986). The distribution for two-jet events is also shown (○). The full curve is a parton model calculation using the tree graph amplitudes for gg → ggg, and cut-offs in transverse momentum and angular separation to eliminate divergences (see remarks following equation (14.73)).

for massless quantum exchange; the particular curve is for the representative process gg → ggg.

Another qualitative feature is that the ratio of three-jet to two-jet events is controlled, roughly, by α_s (compare figure 14.13 with the graphs in table 14.1). Thus an estimate of α_s can be obtained by comparing the rates of 3-jet to 2-jet events in $\bar{p}p$ collisions. Other interesting predictions concern the characteristics of the 3-jet final state (for example, the distributions in the jet energy variables). At this point, however, it is convenient to leave $\bar{p}p$ collisions and consider instead 3-jet events in e^+e^- collisions, for which the complications associated with the initial state hadrons are absent.

14.4 3-jet events in e^+e^- annihilation

Three-jet events in e^+e^- collisions originate, according to QCD, from gluon bremsstrahlung corrections to the two-jet parton level process $e^+e^- \to \gamma^* \to$

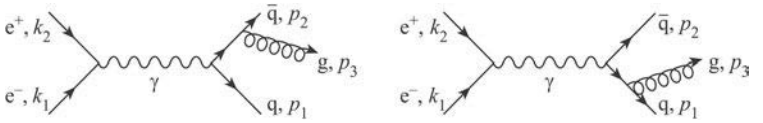


FIGURE 14.15
Gluon brehmsstrahlung corrections to two-jet parton level process.

$q\bar{q}$, as shown in figure 14.15.[4] This phenomenon was predicted by Ellis *et al.* (1976) and subsequently observed by Brandelik *et al.* (1979) with the TASSO detector at PETRA, and Barber *et al.* (1979) with MARK-J at PETRA, thus providing early encouragement for QCD. The situation here is in many ways simpler and cleaner than in the $\bar{p}p$ case; the initial state 'partons' are perfectly physical QED quanta, and their total 4-momentum is zero, so that the three jets have to be coplanar; further, there is only one type of diagram compared to the large number in the $\bar{p}p$ case, and much of that diagram involves the easier vertices of QED. Since the calculation of the cross section predicted from figure 14.15 is relevant not only to three-jet production in e^+e^- collisions, but also to a satisfactory definition of the two-jet production cross section, to QCD corrections to the *total* e^+e^- annihilation cross section, and to scaling violations in deep inelastic scattering as well, we shall now consider it in some detail. It is important to emphasize at the outset that *quark masses will be neglected* in this calculation.

14.4.1 Calculation of the parton-level cross section

The quark, antiquark and gluon 4-momenta are p_1, p_2 and p_3 respectively, as shown in figure 14.15; the e^- and e^+ 4-momenta are k_1 and k_2. The cross section is then (cf (6.110) and (6.112))

$$\mathrm{d}\sigma = \frac{1}{(2\pi)^5}\delta^4(k_1+k_2-p_1-p_2-p_3)\frac{|\,\mathcal{M}_{\mathrm{q\bar{q}g}}\,|^2}{2Q^2}\,\frac{\mathrm{d}^3p_1}{2E_1}\,\frac{\mathrm{d}^3p_2}{2E_2}\,\frac{\mathrm{d}^3p_3}{2E_3} \qquad (14.58)$$

where (neglecting all masses)

$$\begin{aligned}\mathcal{M}_{\mathrm{q\bar{q}g}} \;=\; & \frac{e_{\mathrm{a}}e^2g_{\mathrm{s}}}{Q^2}\bar{v}(k_2)\gamma^\mu u(k_1)\left(\bar{u}(p_1)\gamma_\nu\frac{\lambda_c}{2}\cdot\frac{(\not{p}_1+\not{p}_3)}{2p_1\cdot p_3}\cdot\gamma_\mu v(p_2)\right.\\ & \left.-\bar{u}(p_1)\gamma_\mu\frac{\lambda_c}{2}\cdot\frac{(\not{p}_2+\not{p}_3)}{2p_2\cdot p_3}\cdot\gamma_\nu v(p_2)\right)\epsilon^{*\nu}(\lambda)a_c \qquad (14.59)\end{aligned}$$

and $Q^2 = 4E^2$ is the square of the total e^+e^- energy, and also the square of the virtual photon's 4-momentum Q, and e_{a} (in units of e) is the charge of a

[4] This is assuming that the total e^+e^- energy is far from the Z^0 mass; if not, the contribution from the intermediate Z^0 must be added to that from the photon.

quark of type 'a'. Note the minus sign in (14.59): the antiquark coupling is $-g_s$. In (14.59), $\epsilon^{*\nu}(\lambda)$ is the polarization vector of the outgoing gluon with polarization λ; a_c is the colour wavefunction of the gluon ($c = 1 \ldots\ldots 8$), and λ_c is the corresponding Gell-Mann matrix introduced in section 12.2; the colour parts of the q and $\bar{q}$ wavefunctions are understood to be included in the u and v factors; and $(\not{p}_1 + \not{p}_3)/2p_1 \cdot p_3$ is the virtual quark propagator (cf (L.6) in appendix L of volume 1) before gluon radiation, and similarly for the antiquark. Since the colour parts separate from the Dirac trace parts, we shall ignore them to begin with, and reinstate the result of the colour sum (via problem (14.5)) in the final answer (14.73).

Averaging over $e^{\pm}$ spins and summing over final state quark spins and gluon polarization λ (using (8.171), and noting the discussion after (13.93)), we obtain (problem 14.6)

$$\frac{1}{4}\sum_{\text{spins},\lambda} |\mathcal{M}_{q\bar{q}g}|^2 = \frac{e^4 e_a^2 g_s^2}{Q^4} L^{\mu\nu}(k_1,k_2) H_{\mu\nu}(p_1,p_2,p_3) \tag{14.60}$$

where the lepton tensor is, as usual (equation (8.119)),

$$L^{\mu\nu}(k_1,k_2) = 2(k_1^\mu k_2^\nu + k_1^\nu k_2^\mu - k_1 \cdot k_2 g^{\mu\nu}) \tag{14.61}$$

and the hadron tensor is

$$\begin{aligned} H_{\mu\nu}(p_1,p_2,p_3) \;=\; & \frac{1}{p_1 \cdot p_3}[L_{\mu\nu}(p_2,p_3) - L_{\mu\nu}(p_1,p_1) + L_{\mu\nu}(p_1,p_2)] \\ & + \frac{1}{p_2 \cdot p_3}[L_{\mu\nu}(p_1,p_3) - L_{\mu\nu}(p_2,p_2) \\ & \qquad + L_{\mu\nu}(p_1,p_2)] \\ & + \frac{p_1 \cdot p_2}{(p_1 \cdot p_3)(p_2 \cdot p_3)}[2L_{\mu\nu}(p_1,p_2) + L_{\mu\nu}(p_1,p_3) \\ & \qquad + L_{\mu\nu}(p_2,p_3)] \end{aligned} \tag{14.62}$$

Combining (14.61) and (14.62) allows complete expressions for the five-fold differential cross section to be obtained (Ellis *et al.* 1976).

For the subsequent discussion it will be useful to integrate over the three angles describing the orientation (relative to the beam axis) of the production plane containing the three partons. After this integration, the (doubly differential) cross section is a function of two independent Lorentz invariant variables, which are conveniently taken to be two of the three s_{ij} defined by

$$s_{ij} = (p_i + p_j)^2. \tag{14.63}$$

Since we are considering the massless case $p_i^2 = 0$ throughout, we may also write

$$s_{ij} = 2p_i \cdot p_j. \tag{14.64}$$

These variables are linearly related by

$$2(p_1 \cdot p_2 + p_2 \cdot p_3 + p_3 \cdot p_1) = Q^2 \tag{14.65}$$

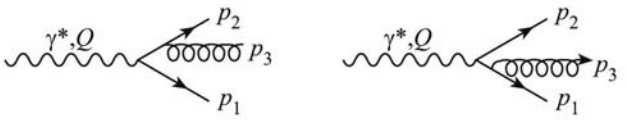


FIGURE 14.16
Virtual photon decaying to $q\bar{q}g$.

as follows from

$$(p_1 + p_2 + p_3)^2 = Q^2 \tag{14.66}$$

and $p_i^2 = 0$. The integration yields (Ellis *et al.* 1976, 1977)

$$\frac{\mathrm{d}^2\sigma}{\mathrm{d}s_{13}\mathrm{d}s_{23}} = \frac{2}{3}\alpha^2 e_a^2 \alpha_s \frac{1}{(Q^2)^3}\left(\frac{s_{13}}{s_{23}} + \frac{s_{23}}{s_{13}} + \frac{2Q^2 s_{12}}{s_{13}s_{23}}\right) \tag{14.67}$$

where $\alpha_s = g_s^2/4\pi$.

We may understand the form of this result in a simple way, as follows. It seems plausible that after integrating over the production angles, the lepton tensor will be proportional to $Q^2 g^{\mu\nu}$, all directional knowledge of the k_i having been lost. Indeed, if we use $-g^{\mu\nu}L_{\mu\nu}(p,p) = 4p \cdot p'$ together with (14.62) we easily find that

$$-\frac{1}{4}g^{\mu\nu}H_{\mu\nu} = \frac{p_1 \cdot p_3}{p_2 \cdot p_3} + \frac{p_2 \cdot p_3}{p_1 \cdot p_3} + \frac{p_1 \cdot p_2 Q^2}{(p_1 \cdot p_3)(p_2 \cdot p_3)} = \frac{s_{13}}{s_{23}} + \frac{s_{23}}{s_{13}} + \frac{2Q^2 s_{12}}{s_{13}s_{23}}, \tag{14.68}$$

exactly the factor appearing in (14.67). In turn, the result may be given a simple physical interpretation. From (7.118) we note that we can replace $-g^{\mu\nu}$ by $\sum_{\lambda'} \epsilon^\mu(\lambda')\epsilon^{\nu *}(\lambda')$ for a virtual photon of polarization λ', the $\lambda' = 0$ state contributing negatively. Thus effectively the result of doing the angular integration is (up to constants and Q^2 factors) to replace the lepton factor $\bar{v}(k_2)\gamma^\mu u(k_1)$ by $-\mathrm{i}\epsilon^\mu(\lambda')$, so that $\mathcal{M}_{q\bar{q}g}$ is proportional to the $\gamma* \to q\bar{q}g$ processes shown in figure 14.16. But these are basically the same amplitudes as the ones we already met in Compton scattering (section 8.6). To compare with section 8.6.3, we convert the initial state fermion (electron/quark) into a final state antifermion (positron/antiquark) by $p \to -p$, and then identify the variables of figure 14.16 with those of figure 8.14 (a) by

$$\begin{aligned} p' \to p_1 \quad k' \to p_3 \qquad -p \to p_2 \qquad & s \to 2p_1 \cdot p_3 = s_{13} \\ t \to 2p_1 \cdot p_2 = s_{12} \qquad & u \to 2p_2 \cdot p_3 = s_{23}. \end{aligned} \tag{14.69}$$

Remembering that in (8.181) the virtual γ had squared 4-momentum $-Q^2$, we see that the Compton '$\sum |\mathcal{M}|^2$' of (8.181) indeed becomes proportional to the factor (14.68), as expected.

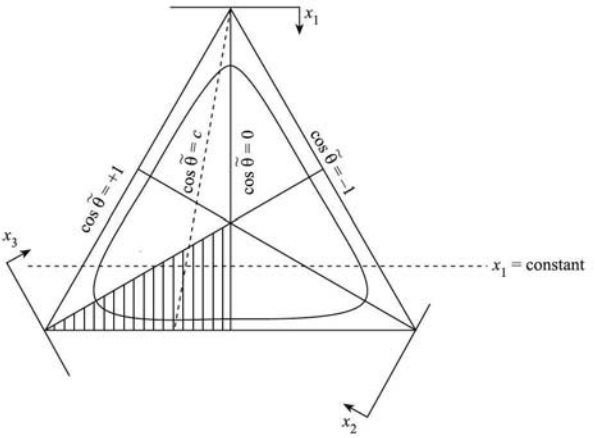


FIGURE 14.17
The kinematically allowed region in (x_i) is the interior of the equilateral triangle.

14.4.2 Soft and collinear divergences

In three-body final states of the type under discussion here it is often convenient to preserve the symmetry between the s_{ij}'s and use *three* (dimensionless) variables x_i defined by

$$s_{23} = Q^2(1 - x_1) \text{ and cyclic permutations.} \tag{14.70}$$

These are related by (14.65), which becomes

$$x_1 + x_2 + x_3 = 2. \tag{14.71}$$

An event with a given value of the set x_i can then be plotted as a point in an equilateral triangle of height 1, as shown in figure 14.17. In order to find the limits of the allowed physical region in this x_i space, we now transform from the overall three-body CMS to the CMS of 2 and 3 (figure 14.18). If $\tilde{\theta}$ is the angle between 1 and 3 in this system, then (problem 14.7)

$$\begin{aligned} x_2 &= (1 - x_1/2) + (x_1/2)\cos\tilde{\theta} \\ x_3 &= (1 - x_1/2) - (x_1/2)\cos\tilde{\theta}. \end{aligned} \tag{14.72}$$

The limits of the physical region are then clearly $\cos\tilde{\theta} = \pm 1$, which correspond to $x_2 = 1$ and $x_3 = 1$. By symmetry, we see that the entire perimeter of the triangle in figure 14.17 is the required boundary: physical events fall anywhere inside the triangle. (This is the massless limit of the classic Dalitz plot, first introduced by Dalitz (1953) for the analysis of $K \to 3\pi$.) Lines of constant $\tilde{\theta}$ are shown in figure 14.17.

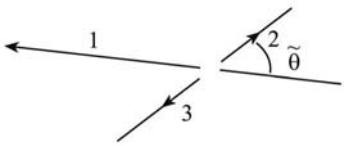


FIGURE 14.18
Definition of $\tilde{\theta}$.

Now consider the distribution provided by the QCD bremsstrahlung process, equation (14.67), which can be written equivalently as

$$\frac{\mathrm{d}^2\sigma}{\mathrm{d}x_1\mathrm{d}x_2} = \sigma_{\mathrm{pt}}\frac{2\alpha_{\mathrm{s}}}{3\pi}\left(\frac{x_1^2+x_2^2}{(1-x_1)(1-x_2)}\right) \tag{14.73}$$

where σ_{pt} is the pointlike $\mathrm{e}^+\mathrm{e}^- \to$ hadrons total cross section of (9.99), and a factor of 4 has been introduced from the colour sum (problem 14.5). The factor in large parentheses is (14.68) written in terms of the x_i (problem 14.8). The most striking feature of (14.73) is that it is *infinite* as x_1 or x_2, or both, tend to 1 – and in such a way that the cross section integrated over x_1 and x_2 diverges logarithmically.

This is a quite different infinity from the ones encountered in the loop integrals of chapters 10 and 11. No integral over an arbitrarily large internal momentum is involved here – the tree amplitude itself is becoming singular on the phase space boundary. We can trace the origin of the singularity back to the denominator factors $(p_1\cdot p_3)^{-1} \sim (1-x_2)^{-1}$ and $(p_2\cdot p_3)^{-1} \sim (1-x_1)^{-1}$ in (14.59). These become zero in two distinct configurations of the gluon momentum:

$$\text{(i)} \quad p_3 \propto p_1 \ \text{ or } p_3 \propto p_2 \ \ (\text{using } p_i^2 = 0) \tag{14.74}$$

$$\text{(ii)} \quad p_3 \to 0 \tag{14.75}$$

which are easily interpreted physically. Condition (i) corresponds to a situation in which the 4-momentum of the gluon is parallel to that of either the quark or the antiquark; this is called a 'collinear divergence' and the configuration is pictured in figure 14.19(a). If we restore the quark masses, $p_1^2 = m_1^2 \neq 0$ and $p_2^2 = m_2^2 \neq 0$, then the factor $(2p_1\cdot p_3)^{-1}$, for example, becomes $((p_1+p_3)^2 - m_1^2)^{-1}$ which only vanishes as $p_3 \to 0$, which is condition (ii). The divergence of type (i) is therefore also termed a 'mass singularity', as it would be absent if the quarks had mass. Condition (ii) corresponds to the emission of a very 'soft' gluon (figure 14.19(b)) and is called a 'soft, or infrared, divergence'. In contrast to this, the gluon momentum p_3 in type (i) does *not* have to be vanishingly small.

FIGURE 14.19
Gluon configurations leading to divergences of equation (14.73): (a) gluon emitted approximately collinear with quark (or antiquark): (b) soft gluon emission. The events are viewed in the overall CMS.

It is apparent from these figures that in either of these two cases the observed final state hadrons, after the fragmentation process, will in fact resemble a *two*-jet configuration. Such events will be found in the regions $x_1 \approx 1$ and/or $x_2 \approx 1$ of the kinematical plot shown in figure 14.17, which correspond to strips adjacent to two of the boundaries of the triangle. Events outside these strips should be essentially three-jet events, corresponding to the emission of a hard, non-collinear gluon. To isolate such events, we must keep away from the boundaries of the triangle (the strip along the third boundary $x_3 = 1$ will not contain a divergence, but will be included in a physical jet measure – see the following section). Thus to order $\alpha^2\alpha_s$ the total annihilation cross section to three jets is given by the integral of (14.73) over a suitably defined inner triangular region in figure 14.17.

Assuming such a separation of three- and two-jet events can be done satisfactorily (see the next section), their ratio carries important information – namely, it should be proportional to α_s. This follows simply from the extra factor of g_s associated with the gluon emissions in figure 14.15. Glossing over a number of technicalities (for which the reader is referred to Ellis, Stirling and Webber 1996, section 3.3), we show in figure 14.20 a compilation of data on the fraction of three-jet events at different e^+e^- annihilation energies. The most remarkable feature of this figure is, of course, that this fraction – and hence α_s – *changes with energy, decreasing as the energy increases.* This is, in fact, direct evidence for asymptotic freedom. A more recent comparison between theory and experiment (the agreement is remarkable) will be presented in the following chapter, section 15.3, after we have introduced the theoretical framework for calculating the energy dependence of α_s.

14.5 Definition of the two-jet cross section in e^+e^- annihilation

As just noted, the integral of (14.73) over the remaining regions of figure 14.17, near the phase-space boundaries, will contribute to the two-jet annihilation cross section – and it is divergent. Clearly this is not a physically acceptable

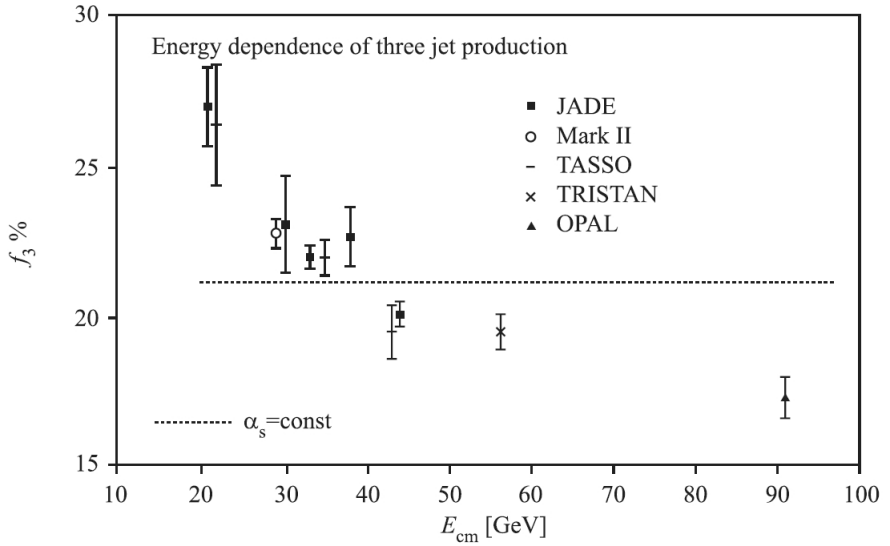


FIGURE 14.20
A compilation of three-jet fractions at different e^+e^- annihilation energies. Adapted from Akrawy *et al.* (OPAL) (1990); figure from R K Ellis, W J Stirling and B R Webber (1996) *QCD and Collider Physics*, courtesy Cambridge University Press.

result: we want a finite two-jet cross section. The cure lies in recognizing that at the order to which we are working, namely $\alpha^2\alpha_s$, other parton-level graphs can contribute. These are the one-gluon loop graphs shown in figure 14.21, which are of order $\alpha\alpha_s$. They turn out to contain exactly the same soft and collinear divergences, this time associated with configurations of virtual momenta inside the loops. In a carefully defined two-jet cross section, these two classes of divergences (one from real gluon emission, the other from virtual gluons) actually cancel.

Let us call the amplitude for the sum of these three graphs F_{vg}, where 'vg' stands for virtual gluon. F_{vg} is the order α_s correction to the original order α parton-level graph of figure 9.17, shown here again in figure 14.22, with amplitude F_γ. The cross section from these contributions is proportional to $|F_\gamma + F_{vg}|^2$. There are three terms in this expression: one of order α^2, from $|F_\gamma|^2$; another of order $\alpha^2\alpha_s^2$, from $|F_{vg}|^2$, which we drop since it is of higher order in α_s; and an *interference* term of order $\alpha^2\alpha_s$, the same as (14.73). Thus the interference term must be included in calculating the two-jet cross section to this order. When it is, the soft and collinear divergences cancel[5]: the resulting two-jet cross section is IRC (infrared and collinear) 'safe'.

[5]The usual ultraviolet divergences in the loop graphs are removed by conventional renormalization.

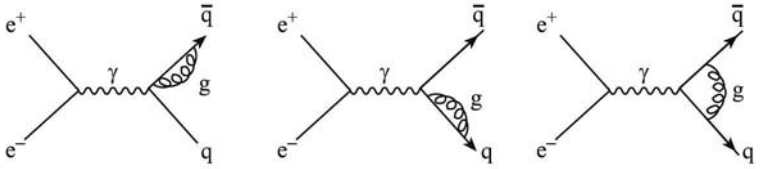


FIGURE 14.21
Virtual gluon corrections to figure 14.20.

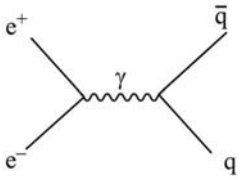


FIGURE 14.22
One-photon annihilation amplitude in $e^+e^- \to \bar{q}q$.

This result was first shown by Sterman and Weinberg (1977), in a paper which initiated the modern treatment of jets within the framework of QCD. They defined the two-jet differential cross section to include those events in which all but a fraction ϵ of the total e^+e^- energy E $(= \sqrt{Q^2})$ is emitted within some pair of oppositely directed cones of half-angle $\delta \ll 1$, lying at an angle θ to the e^+e^- beam line. Including the contributions of real and virtual gluons up to order $\alpha\alpha_s$, the result is (Muta 2010, section 5.4.1)

$$\left(\frac{d\sigma}{d\Omega}\right)_{2-\text{jet}} = \left(\frac{d\sigma}{d\Omega}\right)_{\text{pt}} \left[1 - \frac{4}{3}\frac{\alpha_s}{\pi}\left(3\ln\delta + 4\ln\delta\ln 2\epsilon + \frac{\pi^2}{3} - \frac{5}{2}\right)\right], \quad (14.76)$$

where $(\frac{d\sigma}{d\Omega})_{\text{pt}}$ is the contribution of the lowest-order graph, figure 14.22, which is given by equation (9.102) summed over quark colours and charges; terms of order δ and ϵ, and higher powers, are neglected. It is evident from (14.76) that the jet parameters ϵ and δ serve to control the soft and collinear divergences, which reappear as ϵ and δ tend to zero; they are 'resolution parameters'.

The remarkable cancellation of the soft and collinear divergences between the real and virtual emission processes is actually a general result in QED (recall that in chapter 11 we declined to pursue the problem of such infrared divergences). The Bloch–Nordsieck (1937) theorem states that 'soft' singularities cancel between real and virtual processes when one adds up all final states which are indistinguishable by virtue of the energy resolution of the apparatus. The Kinoshita (1962) Lee and Nauenberg (1964) theorem states, roughly speaking, that mass singularities are absent if one adds up all indis-

tinguishable mass-degenerate states. This is the reason for the finiteness of the Sterman-Weinberg 2-jet cross section, in an analogous QCD case.

Returning to (14.76), it is important to note that the angular distribution of this well-defined two-jet process is given precisely by the lowest-order expression (9.102), just as was hoped in the simple parton model of section 9.5. Of course, the cross section depends on the jet parameters δ and ϵ. The formula (14.76) can be used, for example, to estimate the angular radius of the jets, as a function of E.

Although the Sterman-Weinberg jet definition was historically the first, it is not the only possible one. Another, in some ways simpler, definition (Kramer and Lampe 1987) is directly phrased in terms of the offending denominators s_{13}^{-1} and s_{23}^{-1} in (14.67). Let us introduce the dimensionless jet mass variables

$$y_{ij} = s_{ij}/Q^2 = 2E_iE_j(1-\cos\theta_{ij})/Q^2 \tag{14.77}$$

for any two partons i and j; s_{12} will be included, though no singularity is involved. Here E_i and E_j are the (massless) parton energies, and θ_{ij} is the angle between their 3-momenta, in the overall CMS. Then i and j are defined to be in one jet if y_{ij} is less than some given number y. Note that for small θ_{ij}, $s_{ij} \approx E_iE_j\theta_{ij}^2/Q^2$, so the single parameter y provides effectively both an energy and an angle cut. Clearly this definition is equivalent to a formulation in terms of strips $1 \le x_k < 1-y$ on figure 14.7, as discussed earlier. Including contributions, as before, from figures 14.22, 14.21, and 14.16, the resulting 2-jet cross section is found to be (Kramer and Lampe 1987)

$$\sigma_{2-\text{jet}} = \sigma_{\text{pt}}[1 - \frac{2}{3}\frac{\alpha_{\text{s}}}{\pi}(2\ln^2 y + 3\ln y - 4y\ln y + 1 - \pi^2/3)]. \tag{14.78}$$

Terms of order y were calculated numerically. These include the contribution from the (non-singular) region $y_{12} < y$, where the two quarks are in one jet and the other jet is a pure gluon jet. Plainly the IRC singularities have been eliminated from (14.78), at the cost of the jet mass resolution parameter y. Kramer and Lampe also calculated the order α_{s}^2 corrections to (14.78).

These two ways of regulating the IRC divergences in the 2-jet partonic cross section have each been extensively developed into *jet algorithms*, as we shall briefly discuss in section 14.6.2.

14.6 Further developments

14.6.1 Test of non-Abelian nature of QCD in $e^+e^- \to 4$ jets

We have seen in section 14.3.1 how the colour factors associated with different QCD vertices (problem 14.5) play an important part in determining the relative weights of different parton-level processes. The quark-gluon colour factor

C_F enters into the parton-level three-jet amplitude (14.67), but the triple-gluon vertex is not involved at order α_s. This vertex is an essential feature of non-Abelian gauge theories, being absent in Abelian theories such as QED. A direct measurement of the triple-gluon vertex colour factor, C_A, can be made in the process $e^+e^- \to 4$ jets.

4-jet events originate from the parton-level process $e^+e^- \to q\bar{q}g$ via three mechanisms: the emission of a second bremsstrahlung gluon, splitting of the first gluon into two gluons, and splitting of the first gluon into n_f quark pairs. As problem 14.5 shows, these three types of splitting vertices are characterized in cross sections by the colour factors C_F, C_A and $n_f T_R$, so that the cross section can be written as (Ali and Kramer 2011)

$$\sigma_{4-\text{jet}} = \left(\frac{\alpha_s}{\pi}\right) C_F[C_F\sigma_{b\bar{b}} + C_A\sigma_{gg} + n_f T_R \sigma_{q\bar{q}}]. \tag{14.79}$$

Measurements yield (Abbiendi *et al.* 2001)

$$\begin{aligned} C_A/C_F &= 2.29 \pm 0.06[\text{stat.}] \pm 0.14[\text{syst.}] \\ T_R/C_F &= 0.38 \pm 0.03[\text{stat.}] \pm 0.06[\text{syst.}], \end{aligned} \tag{14.80}$$

in good agreement with the theoretical predictions $C_A/C_F = 9/4$ and $T_R/C_F = 3/8$ in QCD.

14.6.2 Jet algorithms

From the examples already discussed in this chapter, it is clear that jets are an essential element in making comparisons between experimental measurements involving final state particles in detectors, and theoretical calculations at the parton level using perturbative QCD. Conceptually, jets provide a common representation for these two classes of event – those at the detector level, and those at the parton level. For precision comparisons, it is necessary to have a rigorous definition of a jet – a *jet algorithm* – which should be equally applicable at the detector, and at the parton, level. In the more than thirty years that have passed since Sterman and Weinberg's 1977 paper, many jet definitions have been developed and applied. All involve the basic notion of clustering together objects that are in some sense 'near' to each other. Two main classes of jet algorithm may be distinguished: cone algorithms based on proximity in coordinate space, as in the Sterman-Weinberg approach, and used extensively, until recently, at hadron colliders; and sequential recombination algorithms based on proximity in momentum space, as in the jet-mass criterion of Kramer and Lampe (1987), and widely used at e^+e^- and e p colliders. Recent general reviews of jet algorithms are provided by Salam (2010) and by Ali and Kramer (2011); see also Ellis *et al.* (2008), Campbell *et al.* (2007), and Kluth (2006). Here we shall give only a brief introduction to sequential recombination algorithms – all of which are IRC safe – since it seems likely that they will dominate future jet analyses.

The JADE algorithm (Bartel *et al.* 1986, Bethke *et al.* 1988) is a prominent early example of sequential recombination algorithms applied in e^+e^- annihilation reactions. Particles are clustered in a jet iteratively as long as the quantity y_{ij} of (14.77) is less than some prescribed value y_c. If for some pair (i,j), $y_{ij} < y_c$, particles i and j are combined into a compound object (with the resultant 4-momentum, typically), and the process continues by pairing the compound with a new particle k. The procedure stops when all y_{ij} distances are greater than y_c, and the compounds that remain at this stage are the jets, by definition.

One drawback with this scheme is that in higher orders of perturbation theory one meets terms of the form $\alpha_s^2 \ln^{2n} y$ (generalizations of the $\alpha_s \ln^2 y$ term in (14.78)). Such terms can be large enough to invalidate a perturbative approach. Also, it is possible for two soft particles moving in opposite directions to get combined in the early stages of clustering, which runs counter to the intuitive notion of a jet being restricted in angular radius. The k_t-algorithm (Catani *et al.* 1991) avoids these problems by replacing the y_{ij} of (14.77) by

$$y_{ij} = 2\text{min.}[E_i^2, E_j^2](1 - \cos\theta_{ij})/Q^2. \tag{14.81}$$

This amounts to defining 'distance' by the minimum transverse momentum k_t of the particles in nearby collinear pairs. The use of the minimum energy ensures that the distance between two soft, back-to-back particles is larger than that between a soft particle and a hard one that is close to it in angle. The k_t algorithm was widely used at LEP.

The basic idea of the k_t algorithm was extended to hadron colliders (Ellis and Soper 1993, Catani *et al.* 1993), where the total energy of the hard scattering particles is not well defined experimentally. The distance measure y_{ij} is replaced by

$$d_{ij} = \text{min.}[p_{ti}^{2p}, p_{tj}^{2p}][(y_i - y_j)^2 + (\phi_i - \phi_j)^2]/R^2 \tag{14.82}$$

where, for particle i, p_{ti} is the transverse momentum with respect to the (beam) z-axis, y_i is the rapidity along the beam axis (defined by $y_i = \frac{1}{2}\ln[(E_i + p_{zi})/(E_i - p_{zi})]$), ϕ_i is the azimuthal angle in the plane transverse to the beam, and R is a jet parameter. The variables y_i, ϕ_i have the property that they are invariant under boosts along the beam direction. In addition, recombination with the beam jets is controlled by the quantity $d_{ij} = k_{ti}^{2p}$, which is included along with the d_{ij}'s when recombining all the particles into (i) jets with non-zero transverse momentum, and (ii) beam jets. The power parameter p takes the value 1 in the (extended) k_t algorithm, and -1 in the 'anti-k_t' algorithm (Cacciari *et al.* 2008). Whereas the former (and $p = 0$) leads to irregularly shaped jet boundaries, the latter leads to cone-like boundaries. The choice $p = -1$ was made in early LHC analyses.

Problems

14.1

(a) Show that the antisymmetric 3q combination of equation (14.2) is (i) a determinant, and (ii) invariant under the transformation (14.14) for each colour wavefunction.

(b) Suppose that p_α and q_α stand for two $\mathrm{SU}(3)_\mathrm{c}$ colour wavefunctions, transforming under an infinitesimal $\mathrm{SU}(3)_\mathrm{c}$ transformation via

$$p' = (1 + \mathrm{i}\boldsymbol{\eta} \cdot \boldsymbol{\lambda}/2)p,$$

and similarly for q. Consider the antisymmetric combination of their components, given by

$$\begin{pmatrix} p_2q_3 - p_3q_2 \\ p_3q_1 - p_1q_3 \\ p_1q_2 - p_2q_1 \end{pmatrix} \equiv \begin{pmatrix} Q_1 \\ Q_2 \\ Q_3 \end{pmatrix};$$

that is, $Q_\alpha = \epsilon_{\alpha\beta\gamma}p_\beta q_\gamma$. Check that the three components Q_α transform as a $\mathbf{3}^*_\mathrm{c}$, in the particular case for which only the parameters η_1, η_2, η_3 and η_8 are non-zero. [Note: you will need the explicit forms of the $\boldsymbol{\lambda}$ matrices (appendix M); you need to verify the transformation law

$$Q' = (1 - \mathrm{i}\boldsymbol{\eta} \cdot \boldsymbol{\lambda}^*/2)Q.]$$

14.2

(a) Verify that the normally ordered QCD interaction $\bar{\hat{q}}_\mathrm{f}\gamma^\mu \frac{1}{2}\lambda_a \hat{q}_\mathrm{f}\hat{A}_{a\mu}$ is **C**-invariant.

(b) Show that $\lambda_a \hat{F}_{a\mu\nu}$ transforms under **C** according to (14.36).

14.3 Verify that the Lorentz-invariant 'contraction' $\epsilon_{\mu\nu\rho\sigma}\hat{F}^{\mu\nu}\hat{F}^{\rho\sigma}$ of two U(1) (Maxwell) field strength tensors is equal to $8\boldsymbol{E} \cdot \boldsymbol{B}$.

14.4 Verify that the cross section for the exchange of a single massless scalar gluon between two quarks (or between a quark and an antiquark) contains no '$1/\hat{t}^2$' factor.

14.5 This problem is concerned with the evaluation of various '*colour factors*'.

(a) Consider first the colour factor needed for equation (14.73). The 'colour wavefunction' part of the amplitude (14.59) is

$$\sum_c a_c(c_3)\chi^\dagger(c_1)\frac{\lambda_c}{2}\chi(c_2) \tag{14.83}$$

where c_1, c_2 and c_3 label the colour degree of freedom of the quark, antiquark and gluon respectively, and the sum on the index c has been indicated explicitly. The χ's are the colour wavefunctions of the quark and antiquark, and are represented by three-component column vectors; a convenient choice is

$$\chi(r) = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \quad \chi(b) = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \quad \chi(g) = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \tag{14.84}$$

by analogy with the spin wavefunctions of SU(2). The cross section is obtained by forming the modulus squared of (14.83) and summing over the colour labels c_i:

$$\sum_{c,c_1,c_2,c_3} a_c(c_3)\chi_r^*(c_1)\frac{(\lambda_c)_{rs}}{2}\chi_s(c_2)\chi_l^*(c_2)\frac{(\lambda_d)_{lm}}{2}\chi_m(c_1)a_d^*(c_3) \tag{14.85}$$

where summation is understood on the matrix indices on the χ's and λ's, which have been indicated explicitly. In this form the expression is very similar to the *spin* summations considered in chapter 8 (cf equation (8.62)). We proceed to evaluate it as follows:

(i) Show that

$$\sum_{c_2} \chi_s(c_2)\chi_l^*(c_2) = \delta_{sl}.$$

(ii) Assuming the analogous result

$$\sum_{c_3} a_c(c_3)a_d^*(c_3) = \delta_{cd}$$

show that (14.85) becomes

$$\sum_{c=1}^{8} \left(\frac{\lambda_c}{2}\frac{\lambda_c}{2}\right)_{rr},$$

where the (implied) sum on r runs from 1 to 3.

(iii) The expression $\sum_c \frac{\lambda_c}{2}\frac{\lambda_c}{2}$ is just the Casimir operator $\hat{C}_2$ (see section M.5 in appendix M) for SU(3) in the fundamental representation **3**, which from (M.67) has the value $C_F\mathbf{1}_3$, where $\mathbf{1}_3$ is the unit 3×3 matrix, and $C_F = 4/3$. Hence show that the colour factor for (14.73) is 4.

Note that if we averaged over the colours of the initial quark, or considered one particular colour, the colour factor would be C_F.

(b) The colour part for the triple gluon vertex $g_1 \to g_2 + g_3$ is

$$\sum_{c,d,e} a_d^*(c_2) a_e^*(c_3) f_{dec} a_c(c_1).$$

Show that the modulus squared of this, averaged over the initial gluon colours and summed over the final gluon colours, is

$$\frac{1}{8} \sum_{c,d,e} f_{dec} f_{dec},$$

where each of c, d, e runs from 1 to 8. Deduce using (12.84) that this expression can be written as

$$\frac{1}{8} \sum_{e} \left(\sum_{d} \mathrm{G}_d^{(8)} \mathrm{G}_d^{(8)} \right)_{ee},$$

where $\mathrm{G}_d^{(8)}$, $(d = 1 \ldots 8)$ are the 8×8 matrices representing the generators of SU(3) in the **8**-dimensional (adjoint) representation (see section 12.2). The expression $(\sum_d \mathrm{G}_d^{(8)} \mathrm{G}_d^{(8)})$ is the SU(3) Casimir operator $\hat{C}_2$ in the adjoint representation, which from (M.67) has the value $C_A \mathbf{1}_8$, where $\mathbf{1}_8$ is the 8×8 unit matrix, and $C_A = 3$. Hence show that the (averaged, summed) triple gluon vertex colour factor is $C_A = 3$.

(c) The colour part of the $g \to q + \bar{q}$ vertex is

$$\chi_r^*(c_3) (\frac{\lambda_c}{2})_{rs} \chi_s(c_2) a_c(c_1).$$

Show that the modulus squared of this, averaged over the initial gluon colours and summed over the final quark colours is

$$\frac{1}{8} \sum_{c} \left(\frac{\lambda_c}{2} \frac{\lambda_c}{2} \right)_{rr} = \frac{1}{2}.$$

This number is usually denoted by T_R.

14.6 Verify equation (14.60).

14.7 Verify equation (14.72).

14.8 Verify that expression (14.68) becomes the factor in large parentheses in equation (14.73), when expressed in terms of the x_i's.


Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

15

QCD II: Asymptotic Freedom, the Renormalization Group, and Scaling Violations

In the previous chapter we learned that QCD amplitudes contributing to $e^+e^- \to$ jets generally have IRC singularities, but that finite physical cross sections can be obtained by including together kinematically indistinguishable final states. The partial cross sections (for example $\sigma(e^+e^- \to 2 \text{ jets})$) will depend on the IRC cut-off parameter(s). What about the *fully inclusive* process $e^+e^- \to$ hadrons, where all final states are summed over? At order $\alpha\alpha_s$, the parton-level diagrams contributing to this process are the same ones we considered in section 14.5, namely figures 14.16, 14.21 and 14.22. If we denote the amplitudes for these contributions by F_{rg} (for real gluon emission), F_{vg} (for virtual gluon emission) and F_γ for the Born graph, then the partial cross section $\sigma(e^+e^- \to 2 \text{ jets})$ includes $|F_\gamma|^2$, the interference term $2\text{Re}(F_\gamma F^*_{vg})$, and the integral of $|F_{rg}|^2$ over strips near the boundaries of figure 14.17. At this order, the partial cross section $\sigma(e^+e^- \to 3 \text{ jets})$ is given by the integral of $|F_{rg}|^2$ over the remaining (interior) region of figure 14.17. The corresponding total cross section is thus simply the sum of $|F_\gamma|^2, 2\text{Re}(F_\gamma F^*_{vg})$, and the integral of $|F_{rg}|^2$ over the whole of the $x_1 - x_2$ phase space. Clearly the IRC singularities will cancel, as in the 2-jet cross section, and the result will not depend on any IRC cut-off parameter. Indeed, the result is (see for example Muta 2010, section 5.1.2)

$$\sigma(e^+e^- \to \text{hadrons}) = \sigma_{pt}(Q^2)(1 + \alpha_s/\pi). \tag{15.1}$$

This fully inclusive cross section is finite and free of IRC cut-offs.

At first sight, this result might appear satisfactory. It predicts a cross section somewhat greater than σ_{pt}, as is observed in figure 14.1 – from which we might infer that $\alpha_s \sim 0.5$ or less. Assuming the expansion parameter is α_s/π, the implied perturbation series in powers of α_s would seem to be rapidly convergent. However, this is an illusion, which is dispelled as soon as we go to the next order in α_s (i.e. to the order $\alpha^2\alpha_s^2$ in the cross section).

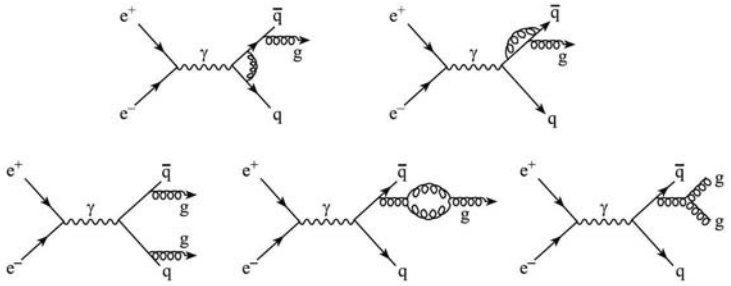


FIGURE 15.1
Some higher-order processes contributing to $e^+e^- \to$ hadrons at the parton level.

15.1 Higher-order QCD corrections to $\sigma(e^+e^- \to \text{hadrons})$: large logarithms

Some typical graphs contributing to this order of the cross section are shown in figure 15.1 (note that, as with the $O(\alpha^2\alpha_s)$ terms, some graphs will contribute via their modulus squared and some via interference terms). The result was obtained numerically by Dine and Saperstein (1979), and analytically by Chetyrkin *et al.* (1979) and by Celmaster and Gonsalves (1980). For our present purposes, the crucial feature of the answer is the appearance of a term

$$\sigma_{\text{pt}}\left[-\beta_0\frac{\alpha_s^2}{\pi}\ln(Q^2/\mu^2)\right] \tag{15.2}$$

where μ is a mass scale (about which we shall shortly have a lot more to say, but which for the moment may be thought of as related in some way to an average quark mass), and the coefficient β_0 is given by

$$\beta_0 = \left(\frac{33-2N_f}{12\pi}\right) \tag{15.3}$$

where N_f is the number of 'active' flavours (e.g. $N_f = 5$ above the $b\bar{b}$ threshold). The term (15.2) raises the following problem. The ratio between it and the $O(\alpha\alpha_s)$ term is clearly

$$-\beta_0\alpha_s\ln(Q^2/\mu^2). \tag{15.4}$$

If we take $N_f = 5, \alpha_s \approx 0.4, \mu \sim 1$ GeV and $Q^2 \sim (10\text{ GeV})^2$, (15.4) is of order 1, and can in no sense be regarded as a small perturbation. Furthermore, the correction (15.4), by itself, would predict large *scaling violations* in this cross

section – that is, large Q^2-dependent departures from the point-like Born cross section, $\sigma_{\rm pt}(Q^2)$. But the data actually follow the point-like prediction very well.

Suppose that, nevertheless, we consider the sum of (15.1) and (15.2), which is

$$\sigma_{\rm pt}[1 + \frac{\alpha_{\rm s}}{\pi}\{1 - \beta_0\alpha_{\rm s}\ln(Q^2/\mu^2)\}]. \tag{15.5}$$

This suggests that one effect, at least, of these higher-order corrections is to convert $\alpha_{\rm s}$ to a Q^2-dependent quantity, namely $\alpha_{\rm s}\{1 - \beta_0\alpha_{\rm s}\ln(Q^2/\mu^2)\}$. We have seen something very like this before, in equation (11.56), for the case of QED. There is, however, one remarkable difference: here the coefficient of the ln is *negative*, whereas that in (11.56) is positive. Apart from this (vital!) difference, however, we can reasonably begin to think in terms of an effective 'Q^2-dependent strong coupling constant $\alpha_{\rm s}(Q^2)$'.

Pressing on with the next order ($\alpha^2\alpha_{\rm s}^3$) terms, we encounter a term (Samuel and Surguladze 1991, Gorishnii *et al.* 1991)

$$\sigma_{\rm pt}\left[\alpha_{\rm s}\beta_0\ln(Q^2/\mu^2)\right]^2\frac{\alpha_{\rm s}}{\pi}, \tag{15.6}$$

and the ratio between this and (15.2) is precisely (15.4) once again! We are now strongly inclined to suspect that we are seeing, in this class of terms, an expansion of the form $(1+x)^{-1} = 1 - x + x^2 - x^3 \ldots$. If true, this would imply that all terms of the form (15.2) and (15.6), and higher, *sum up* to give (cf (11.63))

$$\sigma_{\rm pt}\left[1 + \frac{\alpha_{\rm s}/\pi}{1 + \alpha_{\rm s}\beta_0\ln(Q^2/\mu^2)}\right]. \tag{15.7}$$

The 're-summation' effected by (15.7) has a remarkable effect: the 'dangerous' large logarithms in (15.2) and (15.6) are now effectively in the *denominator* (cf (11.56)), and their effect is such as to *reduce* the effective value of $\alpha_{\rm s}$ as Q^2 increases – exactly the property of *asymptotic freedom*.

We hasten to say that of course this is not how the property was discovered – which was, rather, through the calculations of Politzer (1973) and Gross and Wilczek (1973). Prior to their work, it was widely believed that any quantum field theory would have a running coupling which behaved like that of QED which, as we saw in section 11.5.3, increases for large Q^2 (short distances). Such behaviour would make the scaling violations due to a term like (15.7) even worse. It was therefore a mystery how quantum field theory could account for the small scaling violations seen in the data. The discovery that the running couplings of non-Abelian gauge theories became weaker at large Q^2 opened the way for a quantitative understanding of parton-model scaling, and perturbative QCD corrections to it.

To place the asymptotic freedom calculation in its proper context requires a considerable detour. Referring to our previous discussion, we may ask: are we guaranteed that still-higher-order terms will indeed continue to contain pieces corresponding to the expression of (15.7)? And what exactly is the

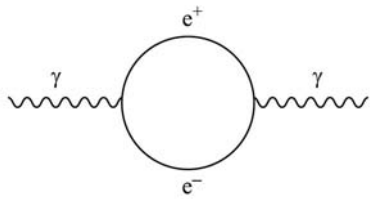


FIGURE 15.2
One-loop vacuum polarization contribution to Z_3.

mass parameter μ? Answering these questions will lead to the important body of ideas going under the name of the 'renormalization group'.

15.2 The renormalization group and related ideas in QED

15.2.1 Where do the large logs come from?

We have taken the title of this section from that of Section 18.1 in Weinberg (1996), which we have found very illuminating, and to which we refer for a more detailed discussion.

As we have just mentioned, the phenomenon of 'large logarithms' arises also in the simpler case of QED. There, however, the factor corresponding to $\alpha_s\beta_0 \sim \frac{1}{4}$ is $\alpha/3\pi \sim 10^{-3}$, so that it is only at quite unrealistically enormous $|q^2|$ values that the corresponding factor $(\alpha/3\pi)\ln(|q^2|/m_e^2)$ (where m_e is the electron mass) becomes of order unity. Nevertheless, the origin of the logarithmic term is essentially the same in both cases, and the technicalities are much simpler for QED (no photon self-interactions, no ghosts). We shall therefore forget about QCD for a while, and concentrate on QED. Indeed, the discussion of renormalization of QED given in chapter 11 will be sufficient to answer the question in the title of this subsection.

For the answer does, in fact, fundamentally have to do with renormalization. Let us go back to the renormalization of the charge in QED. We learned in chapter 11 that the renormalized charge e was given in terms of the 'bare' charge e_0 by the relation $e = e_0(Z_2/Z_1)Z_3^{\frac{1}{2}}$ (see (11.6)), where in fact due to the Ward identity Z_1 and Z_2 are equal (section 11.6), so that only $Z_3^{\frac{1}{2}}$ is needed. To order e^2 in renormalized perturbation theory, including only the e^+e^- loop of figure 15.2, Z_3 is given by (cf (11.31))

$$Z_3^{[2]} = 1 + \Pi_\gamma^{[2]}(0) \tag{15.8}$$

where, from (11.23) and (11.24),

$$\Pi_\gamma^{[2]}(q^2) = 8e^2\mathrm{i}\int_0^1 \mathrm{d}x \int \frac{\mathrm{d}^4k'}{(2\pi)^4}\ \frac{x(1-x)}{(k'^2 - \Delta_\gamma + \mathrm{i}\epsilon)^2} \tag{15.9}$$

and $\Delta_\gamma = m_\mathrm{e}^2 - x(1-x)q^2$ with $q^2 < 0$. We regularize the k' integral by a cut-off Λ as explained in sections 10.3.1 and 10.3.2, obtaining (problem 15.1)

$$\Pi_\gamma^{[2]}(q^2) = -\frac{e^2}{\pi^2}\int_0^1 \mathrm{d}x\ x(1-x)\left\{\ln\left(\frac{\Lambda + \sqrt{\Lambda^2 + \Delta_\gamma}}{\Delta_\gamma^{\frac{1}{2}}}\right) - \frac{\Lambda}{(\Lambda^2 + \Delta_\gamma)^{1/2}}\right\}. \tag{15.10}$$

Setting $q^2 = 0$ and retaining the dominant $\ln\Lambda$ term, we find that

$$\left(Z_3^{[2]}\right)^{\frac{1}{2}} = 1 - \left(\frac{\alpha}{3\pi}\right)\ln(\Lambda/m_\mathrm{e}). \tag{15.11}$$

It is not a coincidence that the coefficient $\alpha/3\pi$ of the ultraviolet divergence is also the coefficient of the $\ln(|q^2|/m_\mathrm{e}^2)$ term in (11.55)–(11.57); we need to understand why.

We first recall how (11.55) was arrived at. It refers to the *renormalized* self-energy part, which is defined by the 'subtracted' form

$$\bar\Pi_\gamma^{[2]}(q^2) = \Pi_\gamma^{[2]}(q^2) - \Pi_\gamma^{[2]}(0). \tag{15.12}$$

In the process of subtraction, the dependence on the cut-off Λ disappears and we are left with

$$\bar\Pi_\gamma^{[2]}(q^2) = -\frac{2\alpha}{\pi}\int_0^1 \mathrm{d}x\ x(1-x)\ln\left[\frac{m_\mathrm{e}^2}{m_\mathrm{e}^2 - q^2x(1-x)}\right] \tag{15.13}$$

as in equation (11.34). For large values of $|q^2|$ this leads to the 'large log' term $(\alpha/3\pi)\ln(|q^2|/m_\mathrm{e}^2)$. Now, in order to form such a term, it is obviously not possible to have just '$\ln|q^2|$' appearing: the argument of the logarithm must be dimensionless, so that some mass scale must be present, to which $|q^2|$ can be compared. In the present case, that mass scale is evidently m_e, which entered via the quantity $\Pi_\gamma^{[2]}(0)$, or equivalently via the renormalization constant $Z_3^{[2]}$ (cf (15.11)). This is the beginning of the answer to our questions.

Why is it m_e that enters into $\Pi_\gamma^{[2]}(0)$ or Z_3? Part of the answer – once again – is of course that a '$\ln\Lambda$' cannot appear in that form, but must be '$\ln(\Lambda/\text{some mass})$'. So we must enquire: what determines the 'some mass'? With this question we have reached the heart of the problem (for the moment). The answer is, in fact, not immediately obvious: it lies in the *prescription used to define the renormalized coupling constant*; this prescription, whatever it is, determines Z_3.

The value (15.8) (or (11.31)) was determined from the requirement that the $O(e^2)$ corrected photon propagator (in $\xi = 1$ gauge) had the simple form

$-\mathrm{i}g_{\mu\nu}/q^2$ as $q^2 \to 0$; that is, as the photon goes on-shell. Now, this is a perfectly 'natural' definition of the renormalized charge – but it is by no means forced upon us. In fact the appearance of a singularity in $Z_3^{[2]}$ as $m_\mathrm{e} \to 0$ suggests that it is inappropriate to the case in which fermion masses are neglected. We could in principle choose a different value of q^2, say $q^2 = -\mu^2$, at which to 'subtract'. Certainly the difference between $\Pi_\gamma^{[2]}(q^2 = 0)$ and $\Pi_\gamma^{[2]}(q^2 = -\mu^2)$ is finite as $\Lambda \to \infty$, so such a redefinition of 'the' renormalized charge would only amount to a finite shift. Nevertheless, even a finite shift is alarming, to those accustomed to a certain 'sanctity' in the value $\alpha = \frac{1}{137}$! We have to concede, however, that if the point of renormalization is to render amplitudes finite by taking certain constants from experiment, then any choice of such constants should be as good as any other – for example, the 'charge' defined at $q^2 = -\mu^2$ rather than at $q^2 = 0$.

Thus there is, actually, a considerable *arbitrariness* in the way renormalization can be done – a fact to which we did not draw attention in our earlier discussions in chapters 10 and 11. Nevertheless, it must somehow be the case that, despite this arbitrariness, *physical results remain the same.* We shall come back to this important point shortly.

15.2.2 Changing the renormalization scale

The recognition that the *renormalization scale* ($-\mu^2$ in this case) is arbitrary suggests a way in which we might exploit the situation, so as to avoid large '$\ln(|q^2|/m_\mathrm{e}^2)$' terms: we renormalize at a *large* value of μ^2! Consider what happens if we define a new $Z_3^{[2]}$ by

$$Z_3^{[2]}(\mu) = 1 + \Pi_\gamma^{[2]}(q^2 = -\mu^2). \tag{15.14}$$

Then for $\mu^2 \gg m_\mathrm{e}^2$, but $\mu^2 \ll \Lambda^2$, we have

$$\left(Z_3^{[2]}(\mu)\right)^{\frac{1}{2}} = 1 - \left(\frac{\alpha}{3\pi}\right) \ln\left(\Lambda/\mu\right), \tag{15.15}$$

and a new renormalized self-energy

$$\begin{aligned} \bar{\Pi}_\gamma^{[2]}(q^2, \mu) &= \Pi_\gamma^{[2]}(q^2) - \Pi_\gamma^{[2]}(q^2 = -\mu^2) \\ &= -\frac{e^2}{2\pi^2} \int_0^1 \mathrm{d}x \; x(1-x) \ln\left[\frac{m_\mathrm{e}^2 + \mu^2 x(1-x)}{m_\mathrm{e}^2 - q^2 x(1-x)}\right]. \end{aligned} \tag{15.16}$$

For μ^2 and $-q^2$ both $\gg m_\mathrm{e}^2$, the logarithm is now $\ln(|q^2|/\mu^2)$ which is small when $|q^2|$ is of order μ^2. It seems, therefore, that with this different renormalization prescription we have 'tamed' the large logarithms.

However, we have forgotten that, for consistency, the 'e' we should now be using is the one defined, in terms of e_0, via

$$e_\mu = \left(Z_3^{[2]}(\mu)\right)^{\frac{1}{2}} e_0 = \left(1 - \frac{\alpha}{3\pi} \ln(\Lambda/\mu)\right) e_0 \tag{15.17}$$

rather than

$$e = \left(Z_3^{[2]}\right)^{\frac{1}{2}} e_0 = \left(1 - \frac{\alpha}{3\pi}\ln(\Lambda/m_\mathrm{e})\right) e_0, \tag{15.18}$$

working always to one-loop order with an $\mathrm{e}^+\mathrm{e}^-$ loop. The relation between e_μ and e is then

$$e_\mu = \frac{\left(1 - \frac{\alpha}{3\pi}\ln(\Lambda/\mu)\right)}{\left(1 - \frac{\alpha}{3\pi}\ln(\Lambda/m_\mathrm{e})\right)} e \approx \left(1 + \frac{\alpha}{3\pi}\ln(\mu/m_\mathrm{e})\right) e \tag{15.19}$$

to leading order in α. Equation (15.19) indeed represents, as anticipated, a finite shift from 'e' to 'e_μ', but the problem with it is that a 'large log' has resurfaced in the form of $\ln(\mu/m_\mathrm{e})$ (remember that our idea was to take $\mu^2 \gg m_\mathrm{e}^2$). Although the numerical coefficient of the log in (15.19) is certainly small, a similar procedure applied to QCD will involve the larger coefficient $\beta_0\alpha_\mathrm{s}$ as in (15.5), and the correction analogous to (15.19) will be of order 1, invalidating the approach.

We have to be more subtle. Instead of making one jump from m_e^2 to a large value μ^2, we need to proceed in stages. We can calculate e_μ from e as long as μ is not too different from m_e. Then we can proceed to $e_{\mu'}$ for μ' not too different from μ, and so on. Rather than thinking of such a process in discrete stages $m_\mathrm{e} \to \mu \to \mu' \to \ldots$, it is more convenient to consider infinitesimal steps – that is, we regard $e_{\mu'}$ at the scale μ' as being a continuous function of e_μ at scale μ, and of whatever other dimensionless variables exist in the problem (since the e's are themselves dimensionless). In the present case, these other variables are μ'/μ and m_e/μ, so that $e_{\mu'}$ must have the form

$$e_{\mu'} = E(e_\mu, \mu'/\mu, m_\mathrm{e}/\mu). \tag{15.20}$$

Differentiating (15.20) with respect to μ' and letting $\mu' = \mu$ we obtain

$$\mu\frac{\mathrm{d}e_\mu}{\mathrm{d}\mu} = \beta_\mathrm{em}(e_\mu, m_\mathrm{e}/\mu) \tag{15.21}$$

where

$$\beta_\mathrm{em}(e_\mu, m_\mathrm{e}/\mu) = \left[\frac{\partial}{\partial z}E(e_\mu, z, m_\mathrm{e}/\mu)\right]_{z=1}. \tag{15.22}$$

For $\mu \gg m_\mathrm{e}$ equation (15.21) reduces to

$$\mu\frac{\mathrm{d}e_\mu}{\mathrm{d}\mu} = \beta_\mathrm{em}(e_\mu, 0) \equiv \beta_\mathrm{em}(e_\mu), \tag{15.23}$$

which is a form of *Callan–Symanzik equation* (Callan 1970, Symanzik 1970); it governs the change of the coupling constant e_μ as the renormalization scale μ changes.

To this one-loop order, it is easy to calculate the crucial quantity $\beta_{\rm em}(e_\mu)$. Returning to (15.17), we may write the bare coupling e_0 as

$$\begin{aligned} e_0 &= e_\mu \left(1 - \frac{\alpha}{3\pi} \ln(\Lambda/\mu)\right)^{-1} \\ &\approx e_\mu \left(1 + \frac{\alpha}{3\pi} \ln(\Lambda/\mu)\right) \\ &\approx e_\mu \left(1 + \frac{\alpha_\mu}{3\pi} \ln(\Lambda/\mu)\right) \end{aligned} \tag{15.24}$$

where the last step follows from the fact that e and e_μ differ by $O(e^3)$, which would be a higher-order correction to (15.24). Now the unrenormalized coupling is certainly independent of μ. Hence, differentiating (15.24) with respect to μ at fixed e_0, we find

$$\left.\frac{{\rm d}e_\mu}{{\rm d}\mu}\right|_{e_0} - \frac{e_\mu \alpha_\mu}{3\pi\mu} - \ln(\Lambda/\mu) \cdot \frac{e_\mu^2}{4\pi^2} \left.\frac{{\rm d}e_\mu}{{\rm d}\mu}\right|_{e_0} = 0. \tag{15.25}$$

Working to order e_μ^3 we can drop the last term in (15.25), obtaining finally (to one-loop order)

$$\mu \left.\frac{{\rm d}e_\mu}{{\rm d}\mu}\right|_{e_0} = \frac{e_\mu^3}{12\pi^2} \quad \left(\equiv \beta_{\rm em}^{[2]}(e_\mu)\right). \tag{15.26}$$

We can now integrate equation (15.26) to obtain e_μ at an arbitrary scale μ, in terms of its value at some scale $\mu = M$, chosen in practice large enough so that for variable scales μ greater than M we can neglect $m_{\rm e}$ compared with μ, but small enough so that $\ln(M/m_{\rm e})$ terms do not invalidate the perturbation theory calculation of e_M from e. The solution of (15.26) is then (problem 15.2)

$$\ln(\mu/M) = 6\pi^2 \left(\frac{1}{e_M^2} - \frac{1}{e_\mu^2}\right) \tag{15.27}$$

or equivalently

$$e_\mu^2 = \frac{e_M^2}{1 - \frac{e_M^2}{12\pi^2} \ln(\mu^2/M^2)}, \tag{15.28}$$

which is

$$\alpha_\mu = \frac{\alpha_M}{1 - \frac{\alpha_M}{3\pi} \ln(\mu^2/M^2)} \tag{15.29}$$

where $\alpha = e^2/4\pi$. The crucial point is that the 'large log' is now in the *denominator* (and has coefficient $\alpha_M/3\pi$!). We note that the general solution of (15.23) may be written as

$$\ln(\mu/M) = \int_{e_M}^{e_\mu} \frac{{\rm d}e}{\beta_{\rm em}(e)}. \tag{15.30}$$

We have made progress in understanding how the coupling changes as the renormalization scale changes, and how 'large logarithmic' change as in (15.19) can be brought under control via (15.29). The final piece in the puzzle

is to understand how this can help us with the large $-q^2$ behaviour of our cross section, the problem we originally started from.

15.2.3 The RGE and large $-q^2$ behaviour in QED

To see the connection we need to implement the fundamental requirement, stated at the end of section 15.2.2, that predictions for physically measurable quantities must *not* depend on the renormalization scale μ. Consider, for example, our annihilation cross section σ for $e^+e^- \to$ hadrons, pretending that the one-loop corrections we are interested in are those due to QED rather than QCD. We need to work in the spacelike region, so as to be consistent with all the foregoing discussion. To make this clear, we shall now denote the 4-momentum of the virtual photon by q rather than Q, and take $q^2 < 0$ as in sections 15.2.1 and 15.2.2. Bearing in mind the way we used the 'dimensionless-ness' of the e's in (15.20), let us focus on the dimensionless ratio $\sigma/\sigma_{\rm pt} \equiv S$. Neglecting all masses, S can only be a function of the dimensionless ratio $|q^2|/\mu^2$ and of e_μ:

$$S = S(|q^2|/\mu^2, e_\mu). \tag{15.31}$$

But S must ultimately have no μ dependence. It follows that *the μ^2 dependence arising via the $|q^2|/\mu^2$ argument must cancel that associated with e_μ.* This is why the μ^2-dependence of e_μ controls the $|q^2|$ dependence of S, and hence of σ. In symbols, this condition is represented by the equation

$$\left(\left.\frac{\partial}{\partial \mu}\right|_{e_\mu} + \left.\frac{\mathrm{d}e_\mu}{\mathrm{d}\mu}\right|_{e_0} \frac{\partial}{\partial e_\mu} \right) S\left(|q^2|/\mu^2, e_\mu\right) = 0, \tag{15.32}$$

or

$$\left(\mu \left.\frac{\partial}{\partial \mu}\right|_{e_\mu} + \beta_{\rm em}(e_\mu) \frac{\partial}{\partial e_\mu} \right) S\left(|q|^2/\mu^2, e_\mu\right) = 0. \tag{15.33}$$

Equation (15.33) is referred to as 'the renormalization group equation (RGE) for S'. The terminology goes back to Stueckelberg and Peterman (1953), who were the first to discuss the freedom associated with the choice of renormalization scale. The 'group' connotation is a trifle obscure – but all it really amounts to is the idea that if we do one infinitesimal shift in μ^2, and then another, the result will be a third such shift; in other words, it is a kind of 'translation group'. It was, however, Gell-Mann and Low (1954) who realized how equation (15.33) could be used to calculate the large $|q^2|$ behaviour of S, as we now explain.

It is convenient to work in terms of μ^2 and α rather than μ and e. Equation (15.33) is then

$$\left(\mu^2 \left.\frac{\partial}{\partial \mu^2}\right|_{\alpha_\mu} + \beta_{\rm em}(\alpha_\mu) \frac{\partial}{\partial \alpha_\mu} \right) S\left(|q^2|/\mu^2, \alpha_\mu\right) = 0, \tag{15.34}$$

where $\beta_{\rm em}(\alpha_\mu)$ is defined by

$$\beta_{\rm em}(\alpha_\mu) \equiv \mu^2 \frac{\partial \alpha_\mu}{\partial \mu^2}\bigg|_{e_0}. \tag{15.35}$$

From (15.35) and (15.26) we deduce that, to the one-loop order to which we are working,

$$\beta^{[2]}_{\rm em}(\alpha_\mu) = \frac{e_\mu}{4\pi}\beta^{[2]}_{\rm em}(e_\mu) = \frac{\alpha_\mu^2}{3\pi}. \tag{15.36}$$

Now introduce the important variable

$$t = \ln(|q^2|/\mu^2). \tag{15.37}$$

Equation (15.34) then becomes

$$\left[-\frac{\partial}{\partial t} + \beta_{\rm em}(\alpha_\mu)\frac{\partial}{\partial \alpha_\mu}\right] S\left(e^t, \alpha_\mu\right) = 0. \tag{15.38}$$

This is a first-order differential equation which can be solved by implicitly defining a new function – the *running coupling* $\alpha(|q^2|)$ – as follows (compare (15.30):

$$t = \int_{\alpha_\mu}^{\alpha(|q^2|)} \frac{{\rm d}\alpha}{\beta_{\rm em}(\alpha)}. \tag{15.39}$$

To see how this helps, we have to recall how to differentiate an integral with respect to one of its limits – or, more generally, the formulae

$$\frac{\partial}{\partial a}\int^{f(a)} g(x){\rm d}x = g\left(f(a)\right)\frac{\partial f}{\partial a}. \tag{15.40}$$

First, let us differentiate (15.39) with respect to t at fixed α_μ; we obtain

$$1 = \frac{1}{\beta_{\rm em}(\alpha(|q^2|))}\ \frac{\partial \alpha(|q^2|)}{\partial t}. \tag{15.41}$$

Next, differentiate (15.39) with respect to α_μ at fixed t (note that $\alpha(|q^2|)$ will depend on μ and hence on α_μ); we obtain

$$0 = \frac{\partial \alpha(|q^2|)}{\partial \alpha_\mu}\ \frac{1}{\beta_{\rm em}(\alpha(|q^2|))} - \frac{1}{\beta_{\rm em}(\alpha_\mu)} \tag{15.42}$$

the minus sign coming from the fact that α_μ is the lower limit in (15.39). From (15.41) and (15.42) we find

$$\left[-\frac{\partial}{\partial t} + \beta_{\rm em}(\alpha_\mu)\frac{\partial}{\partial \alpha_\mu}\right] \alpha(|q^2|) = 0. \tag{15.43}$$

It follows that $S(1, \alpha(|q^2|))$ is a solution of (15.38).

This is a remarkable result. It shows that all the dependence of S on the (momentum)^2 variable $|q^2|$ enters through that of the running coupling $\alpha(|q^2|)$. Of course, this result is only valid in a regime of $-q^2$ which is much greater than all quantities with dimension (mass)^2 – for example the squares of all particle masses, which do not appear in (15.31). This is why the technique applies only at 'high' $-q^2$. The result implies that if we can calculate $S(1, \alpha_\mu)$ (i.e. S at the point $q^2 = -\mu^2$) at some definite order in perturbation theory, then replacing α_μ by $\alpha(|q^2|)$ will allow us to predict the q^2-dependence (at large $-q^2$). All we need to do is solve (15.39). Indeed, for QED with one e^+e^- loop we have seen that $\beta^{[2]}_{\rm em}(\alpha) = \alpha^2/3\pi$. Hence integrating (15.39) we obtain

$$\alpha(|q^2|) = \frac{\alpha_\mu}{1 - \frac{\alpha_\mu}{3\pi}t} = \frac{\alpha_\mu}{1 - \frac{\alpha_\mu}{3\pi}\ln(|q^2|/\mu^2)}. \tag{15.44}$$

This is almost exactly the formula we proposed in (11.57), on plausibility grounds.[1]

Suppose now that the leading QED perturbative contribution to $S(1, \alpha_\mu)$ is $S_1\alpha_\mu$. Then the terms contained in $S(1, \alpha(|q^2|))$ in this approximation can be found by expanding in powers of α_μ:

$$\begin{aligned} S(1, \alpha(|q^2|)) &\approx 1 + S_1\alpha(|q^2|) = 1 + S_1\alpha_\mu \left[1 - \frac{\alpha_\mu}{3\pi}t\right]^{-1} \\ &= 1 + S_1\alpha_\mu \left[1 + \frac{\alpha_\mu t}{3\pi} + \left(\frac{\alpha_\mu t}{3\pi}\right)^2 + \ldots\right], \end{aligned} \tag{15.45}$$

where $t = \ln(|q^2|/\mu^2)$. The next-higher-order calculation of $S(1, \alpha_\mu)$ would be $S_2\alpha_\mu^2$, say, which generates the terms

$$S_2\alpha^2(|q^2|) = S_2\alpha_\mu^2 \left[1 + \frac{2\alpha_\mu t}{3\pi} + \ldots\right]. \tag{15.46}$$

Comparing (15.45) and (15.46) we see that each power of the large log factor appearing in (15.46) comes with one more power of α_μ than in (15.45). Provided α_μ is small, then, the *leading* terms in $t, t^2, \ldots$ are contained in (15.45). It is in this sense that replacing $S(1, \alpha_\mu)$ by $S(1, \alpha(|q^2|))$ sums all 'leading log terms'.

In fact, of course, the one-loop (and higher) corrections to S in which we are really interested are those due to QCD, rather than QED, corrections. But the logic is exactly the same. The leading ($O(\alpha_s)$) perturbative contribution to $S = \sigma/\sigma_{\rm pt}$ at $q^2 = -\mu^2$ is given in (15.1) as $\alpha_s(\mu^2)/\pi$. It follows that the 'leading log corrections' at high $-q^2$ are summed up by replacing this expression by $\alpha_s(|q^2|)/\pi$, where the running $\alpha_s(|q^2|)$ is determined by solving (15.39) with the QCD analogue of (15.36) – to which we now turn.

[1] The difference has to do, of course, with the different renormalization prescriptions. Eq (11.57) is written in terms of an 'α' defined at $q^2 = 0$, and without neglect of m_e.

15.3 Back to QCD: asymptotic freedom

15.3.1 One loop calculation

The reader will of course have realized, some time back, that the quantity β_0 introduced in (15.3) must be precisely the coefficient of α_s^2 in the one-loop contribution to the β-function of QCD defined by

$$\beta_{\rm s} = \mu^2 \frac{\partial \alpha_{\rm s}}{\partial \mu^2}\bigg|_{\text{fixed bare } \alpha_{\rm s}}; \tag{15.47}$$

that is to say,

$$\beta_{\rm s}(\text{one loop}) = -\beta_0 \alpha_{\rm s}^2 \tag{15.48}$$

with

$$\beta_0 = \frac{33 - 2N_{\rm f}}{12\pi}. \tag{15.49}$$

For $N_{\rm f} \leq 16$ the quantity β_0 is *positive*, so that the sign of (15.48)) is opposite to that of the QED analogue, equation (15.36). Correspondingly, (15.44) is replaced by

$$\alpha_{\rm s}(|q^2|) = \frac{\alpha_{\rm s}(\mu^2)}{[1 + \alpha_{\rm s}(\mu^2)\beta_0 \ln(Q^2/\mu^2)]}, \tag{15.50}$$

where $Q^2 = |q^2|$.[2] Then replacing $\alpha_{\rm s}$ in (15.1) by (15.50) leads to (15.7).

Thus in QCD the strong coupling runs in the opposite way to QED, becoming smaller at large values of Q^2 (or small distances) – the property of asymptotic freedom. The justly famous result (15.49) was first obtained by Politzer (1973), Gross and Wilczek (1973), and 't Hooft. 't Hooft's result, announced at a conference in Marseilles in 1972, was not published. The published calculation of Politzer and of Gross and Wilczek quickly attracted enormous interest, because it immediately offered a way to understand how the successful parton model could be reconciled with the undoubtedly very strong binding forces between quarks. The resolution, we now understand, lies in quite subtle properties of renormalized quantum field theory, involving first the exposure of 'large logarithms', then their re-summation in terms of the running coupling, and of course the crucial sign of the β-function. Not only did the result (15.49) explain the success of the parton model: it also, we repeat, opened the prospect of performing reliable perturbative calculations in a *strongly* interacting theory, at least at high Q^2. For example, at sufficiently high Q^2, we can reliably compute the β function in perturbation theory. The result of Politzer and of Gross and Wilczek, when combined with

[2] Except that in (15.50) $\alpha_{\rm s}$ is evaluated at large *spacelike* values of q^2, whereas in (15.7) it is wanted at large *timelike* values. Readers troubled by this may consult Peskin and Schroeder (1995) section 18.5. The difficulty is evaded in the approach of section 15.6 below.

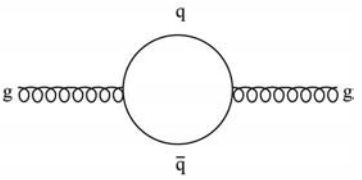


FIGURE 15.3
$q\bar{q}$ vacuum polarization correction to the gluon propagator.

the motivations for a colour SU(3) group discussed in the previous chapter, led rapidly to the general acceptance of QCD as the theory of strong interactions, a conclusion reinforced by the demonstration by Coleman and Gross (1973) that no theory without Yang-Mills fields possessed the property of asymptotic freedom.

In section 11.5.3 we gave the conventional physical interpretation of the way in which the running of the QED coupling tends to *increase* its value at distances short enough to probe inside the screening provided by e^+e^- pairs ($|q|^{-1} \ll m_e^{-1}$). This vacuum polarization screening effect is also present in (15.49) via the term $-\frac{2N_f}{12\pi}$, the value of which can be quite easily understood. It arises from the '$q\bar{q}$' vacuum polarization diagram of figure 15.3, which is precisely analogous to the e^+e^- diagram used to calculate $\bar{\Pi}_\gamma^{[2]}(q^2)$ in QED. The only new feature in figure 15.3 is the presence of the $\frac{\lambda}{2}$-matrices at each vertex. If 'a' and 'b' are the colour labels of the ingoing and outgoing gluons, the $\frac{\lambda}{2}$-matrix factors must be

$$\sum_{\alpha,\beta=1}^{3} \left(\frac{\lambda_a}{2}\right)_{\alpha\beta} \left(\frac{\lambda_b}{2}\right)_{\beta\alpha} \tag{15.51}$$

since there are no free quark indices (of type α, β) on the external legs of the diagram. It is simple to check that (15.51) has the value $\frac{1}{2}\delta_{ab}$ (this is, in fact, the way the λ's are conventionally normalized). Hence for one quark flavour we expect '$\alpha/3\pi$' to be replaced by '$\alpha_s/6\pi$', in agreement with the second term in (15.49).

The all-important, positive, first term must therefore be due to the gluons. The one-loop graphs contributing to the calculation of β_0 are shown in figure 15.4. They include figure 15.3, of course, but there are also, characteristically, graphs involving the gluon self-coupling which is absent in QED, and also (in covariant gauges) ghost loops. We do not want to enter into the details of the calculation of $\beta(\alpha_s)$ here (they are given in Peskin and Schroeder 1995, chapter 16, for example), but it would be nice to have a simple intuitive picture of the 'antiscreening' result in terms of the gluon interactions, say. Unfortunately no fully satisfactory simple explanation exists, though the reader may be inter-

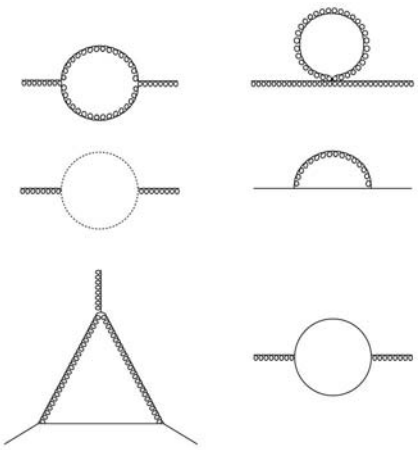


FIGURE 15.4
Graphs contributing to the one-loop β function in QCD. The curly line represents a gluon, a dotted line a ghost (see section 13.3.3) and a straight line a quark.

ested to consult Hughes (1980, 1981) and Nielsen (1981) for a 'paramagnetic' type of explanation, rather than a 'dielectric' one.

Returning to (15.50), we note that the equation effectively provides a prediction of α_s at any scale Q^2, given its value at a particular scale $Q^2 = \mu^2$, which must be taken from experiment. The reference scale is now normally taken to be the Z^0 mass; the value $\alpha_s(m_Z^2)$ then plays the role in QCD that $\alpha \sim 1/137$ does in QED.

Despite appearances, equation (15.50) does not really involve two parameters – after all, (15.47) is only a first-order differential equation. By introducing

$$\ln \Lambda_{\rm QCD}^2 = \ln \mu^2 - 1/(\beta_0 \alpha_s(\mu^2)), \tag{15.52}$$

equation (15.50) can be rewritten (problem 15.3) as

$$\alpha_s(Q^2) = \frac{1}{\beta_0 \ln(Q^2/\Lambda_{\rm QCD}^2)}. \tag{15.53}$$

Equation (15.53) is equivalent to (cf (15.30))

$$\ln\left(Q^2/\Lambda_{\rm QCD}^2\right) = \int_{\alpha_s(Q^2)}^{\infty} \frac{{\rm d}\alpha_s}{\beta_s(\text{one loop})} \tag{15.54}$$

with $\beta_s(\text{one loop}) = -\beta_0 \alpha_s^2$. $\Lambda_{\rm QCD}$ is therefore an integration constant, representing the scale at which α_s would diverge to infinity (if we extended our calculation beyond its perturbative domain of validity). More usefully, $\Lambda_{\rm QCD}$

is a measure of the scale at which α_s really does become 'strong'. The extraction of a value of Λ_{QCD} is a somewhat complicated matter, as we shall briefly indicate in the following section, but a typical value is in the region of 200 MeV. Note that this is a distance scale of order $(200\ \text{MeV})^{-1} \sim 1$ fm, just about the size of a hadron – a satisfactory connection.

15.3.2 Higher-order calculations, and experimental comparison

So far we have discussed only the 'one-loop' calculation of $\beta(\alpha_s)$. The general perturbative expansion for β_s can be written as

$$\beta_s(\alpha_s) = -\beta_0\alpha_s^2 - \beta_1\alpha_s^3 - \beta_2\alpha_s^4 + \dots \tag{15.55}$$

where β_0 is the one-loop coefficient given in (15.49), β_1 is the two-loop coefficient, and so on. β_1 was calculated by Caswell (1974) and Jones (1974), and has the value

$$\beta_1 = \frac{153 - 19N_f}{24\pi^2}. \tag{15.56}$$

The three-loop coefficient β_2, obtained by Tarasov *et al.* (1980) and by Larin and Vermaseren (1993), is

$$\beta_2 = \frac{77139 - 15099N_f + 325N_f^2}{3456\pi^2}. \tag{15.57}$$

The four-loop coefficient β_3 was calculated by van Ritbergen *et al.* (1997) and by Czakon (2005); we shall not give it here. A technical point to note is that while β_0 and β_1 are independent of the scheme adopted for renormalization (see appendix O), the higher-order coefficients do depend on it; the value (15.57) is in the widely used $\overline{\text{MS}}$ scheme. Likewise, Λ_{QCD} will be scheme-dependent (see appendix O), and the value $\Lambda_{\overline{\text{MS}}}$ will be used here (the 'QCD' now being understood).

Only in the one-loop approximation for β_s can an analytic solution of (15.47) be obtained. However, a useful approximate solution can be found iteratively, as follows. Consider the two-loop version of (15.54), namely

$$\ln(Q^2/\Lambda^2_{\overline{\text{MS}}}) = -\int \frac{\mathrm{d}\alpha_s}{\beta_0\alpha_s^2 + \beta_1\alpha_s^3}. \tag{15.58}$$

Expanding the denominator and integrating gives

$$\ln(Q^2/\Lambda^2_{\overline{\text{MS}}}) = \frac{1}{\beta_0\alpha_s} + \frac{b_1}{\beta_0}\ln\alpha_s + C, \tag{15.59}$$

where $b_1 = \beta_1/\beta_0$ and C is a constant. In the $\overline{\text{MS}}$ scheme, C is given by $C = (b_1/\beta_0)\ln\beta_0$. Then the equation for α_s is

$$L = \frac{1}{\beta_0\alpha_s} + \frac{b_1}{\beta_0}\ln\beta_0\alpha_s, \tag{15.60}$$

where we have defined $L = \ln(Q^2/\Lambda^2_{\overline{\rm MS}})$. In first approximation, one sets b_1 to zero and finds $\alpha_{\rm s} = (1/\beta_0 L)$ as before. To obtain the next approximation, we set $\alpha_{\rm s} = (1/\beta_0 L)$ in the b_1 term of (15.60), and solve for $\alpha_{\rm s}$ to first order in b_1. This gives (problem 15.4 (a))

$$\alpha_{\rm s} = \frac{1}{\beta_0 L} - \frac{1}{\beta_0^3 L^2}\beta_1 \ln L. \tag{15.61}$$

Problem 15.4 (b) carries the calculation to the three-loop stage.

The current world average value of $\alpha_{\rm s}(m_{\rm Z}^2)$ is (Bethke 2009)

$$\alpha_{\rm s}(m_{\rm Z}^2) = 0.1184 \pm 0.0007. \tag{15.62}$$

The remarkable precision of this number represents extraordinary consistency among the many methods used to determine it[3], which include deep inelastic scattering, electroweak fits, $\rm e^+e^- \to$ jets, and lattice calculations (see chapter 16). If (15.62) is used to determine $\Lambda_{\overline{\rm MS}}$ from (15.61), one finds $\Lambda_{\overline{\rm MS}} = 231$ MeV; using the 3-loop formula of problem 15.4 (b) gives $\Lambda_{\overline{\rm MS}} = 213$ MeV (Bethke 2009).

These values of $\Lambda_{\overline{\rm MS}}$ are for $N_{\rm f} = 5$, appropriate for the $\rm Z^0$ mass region, well above the b threshold. As Q^2 runs to smaller values, and a quark mass threshold is crossed, $N_{\rm f}$ changes by one unit, and so correspondingly do the coefficients $\beta_0, \beta_1, \ldots$. Physical quantities must however be continuous across a quark threshold. This requires that the values of $\alpha_{\rm s}$ above and below that threshold satisfy certain matching conditions (Rodrigo and Santamaria 1993, Bernreuther and Wetzel 1982, Chetyrkin *et al.* 1997). These are satisfied by allowing $\Lambda_{\overline{\rm MS}}$ to depend on $N_{\rm f}$. At one and two loop order, the matching condition is simply $\alpha_{\rm s}^{(N_{\rm f}-1)} = \alpha_{\rm s}^{(N_{\rm f})}$, which can be straightforwardly implemented in terms of $\Lambda_{\overline{\rm MS}}^{(N_{\rm f}-1)}$ and $\Lambda_{\overline{\rm MS}}^{(N_{\rm f}-1)}$. In higher orders the matching conditions contain additional terms, which are required at $(n-1)$-loop order for an n-loop calculation of $\alpha_{\rm s}$.

Figure 15.5 shows a summary (Bethke 2009) of measurements of $\alpha_{\rm s}$ as a function of the energy scale Q, compared with the QCD prediction. The latter is evaluated in 4-loop approximation, using 3-loop threshold matching conditions at the masses $m_{\rm c} = 1.5$ GeV and $m_{\rm b} = 4.7$ GeV. The agreement is perfect, a triumph for both experiment and theory.

15.4 $\sigma(\rm e^+e^- \to$ hadrons) revisited

We may now return to the physical process which originally motivated this extensive detour. The perturbative corrections to $\sigma_{\rm pt}(Q^2)$ are expressed as a

[3]With the exception of a long-standing systematic difference: results from structure functions prefer a smaller value of $\alpha_{\rm s}(m_{\rm Z}^2)$ than most of the others.

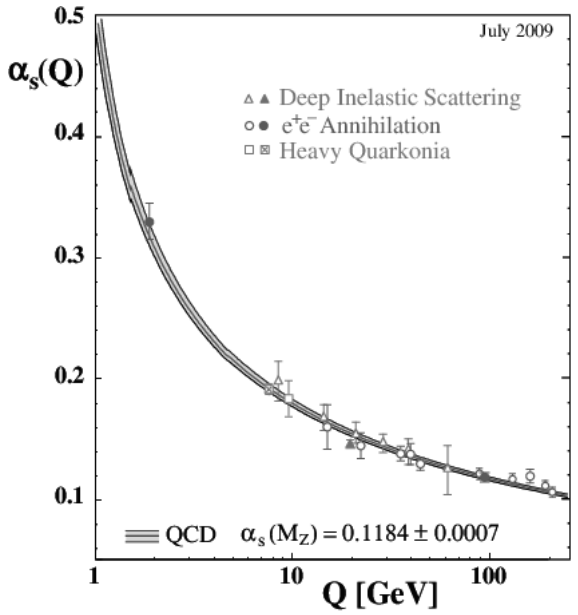


FIGURE 15.5
Comparison between measurements of α_s and the theoretical prediction, as a function of the energy scale Q (Bethke 2009). (See color plate I.)

power series in α_s,

$$\sigma(e^+e^- \to \text{hadrons}) = \sigma_{pt}(Q^2) \left[1 + \sum_{n=1}^{\infty} c_n(Q^2/\mu^2) \left(\frac{\alpha_s(\mu^2)}{\pi}\right)^n\right], \quad (15.63)$$

where μ is the renormalization scale. (A similar expansion can be written for many other physical quantities too.) The coefficients from c_2 onwards depend on the renormalization scheme (see appendix O), and are usually quoted in the $\overline{\text{MS}}$ scheme. c_1 is the leading order (LO) coefficient, and we already know that $c_1 = 1$ from (15.1). c_2 is the next-to-leading (NLO) coefficient; $c_2(1)$ was calculated by Dine and Sapirstein (1979), Chetyrkin *et al.* (1979) and by Celmaster and Gonsalves (1980), and has the value $1.9857 - 0.1152N_f$. The next-to-next-to-leading (NNLO) coefficient $c_3(1)$ was calculated by Samuel and Surguladze (1991) and by Gorishnii *et al.* (1991), and is equal to -12.8 for five flavours. The N3LO coefficient $c_4(1)$ (which requires the evaluation of some twenty thousand diagrams) may be found in Baikov *et al.* (2008) and Baikov *et al.* (2009).

The physical cross section $\sigma(e^+e^- \to \text{hadrons})$ must be independent of the renormalization scale μ^2, and this would also be true of the series in (15.63) if an infinite number of terms were kept: the μ^2-dependence of the coefficients

$c_n(Q^2/\mu^2)$ would cancel that of $\alpha_s(\mu^2)$. This requirement can be imposed order by order in α_s to fix the μ^2-dependence of the coefficients, and is a direct way of applying the RGE idea. Consider, for example, truncating the series at the $n = 2$ stage:

$$\sigma(e^+e^- \to \text{hadrons}) \approx \sigma_{pt}(Q^2)\left(1 + \frac{\alpha_s(\mu^2)}{\pi} + c_2(Q^2/\mu^2)(\alpha_s(\mu^2)/\pi)^2\right). \tag{15.64}$$

Differentiating with respect to μ^2 and setting the result to zero we obtain

$$\mu^2 \frac{dc_2}{d\mu^2} = -\frac{\pi\beta(\alpha_s(\mu^2))}{(\alpha_s(\mu^2))^2} \tag{15.65}$$

where an $O(\alpha_s^3)$ term has been dropped. Substituting the one-loop result (15.48) – as is consistent to this order – we find

$$c_2(Q^2/\mu^2) = c_2(1) - \pi\beta_0 \ln(Q^2/\mu^2). \tag{15.66}$$

The second term on the right-hand side of (15.66) gives the contribution identified in (15.2).

In practice only a finite number of terms $n = N$ will be available, and a μ^2-dependence will remain, which implies an uncertainty in the prediction of the cross section (and similar physical observables), due to the arbitrariness of the scale choice. This uncertainty will be of the same order as the neglected terms, i.e. of order α_s^{N+1}. Thus the scale dependence of a QCD prediction gives a measure of the uncertainties due to neglected terms. For $\sigma(e^+e^- \to \text{hadrons})$ the choice of scale $\mu^2 = Q^2$ is usually made, so as to avoid large logarithms in relations such as (15.66).

Before proceeding to our second main application of the RGE, scaling violations in deep inelastic scattering, it is necessary to take another detour, to enlarge our understanding of the scope of the RGE.

15.5 A more general form of the RGE: anomalous dimensions and running masses

The reader may have wondered why, for QCD, all the graphs of figure 15.6 are needed, whereas for QED we got away with only figure 11.3. The reason for the simplification in QED was the equality between the renormalization constants Z_1 and Z_2, which therefore cancelled out in the relation between the renormalized and bare charges e and e_0, as briefly stated before equation (15.8) (this equality was discussed in section 11.6). We recall that Z_1 is the field strength renormalization factor for the charged fermion in QED, and Z_1 is the vertex part renormalization constant; their relation to the counter terms

was given in equation (11.7). For QCD, although gauge invariance does imply generalizations of the Ward identity used to prove $Z_1 = Z_2$ (Taylor 1971, Slavnov 1972), the consequence is no longer the simple relation '$Z_1 = Z_2$' in this case, due essentially to the ghost contributions. In order to see what change $Z_1 \neq Z_2$ would make, let us return to the one-loop calculation of β for QED, pretending that $Z_1 \neq Z_2$. We have

$$e_0 = \frac{Z_1}{Z_2} Z_3^{-\frac{1}{2}} e_\mu \tag{15.67}$$

where, because we are renormalizing at scale μ, all the Z_i's depend on μ (as in (15.15)), but we shall now not indicate this explicitly. Taking logs and differentiating with respect to μ at constant e_0, we obtain

$$\mu \frac{\mathrm{d}}{\mathrm{d}\mu}\bigg|_{e_0} \ln Z_1 - \mu \frac{\mathrm{d}}{\mathrm{d}\mu}\bigg|_{e_0} \ln Z_2 - \frac{1}{2}\mu \frac{\mathrm{d}}{\mathrm{d}\mu}\bigg|_{e_0} \ln Z_3 + \frac{\mu}{e_\mu} \frac{\mathrm{d}e_\mu}{\mathrm{d}\mu}\bigg|_{e_0} = 0. \tag{15.68}$$

Hence

$$\beta(e_\mu) \equiv \mu \frac{\mathrm{d}e_\mu}{\mathrm{d}\mu}\bigg|_{e_0} = e_\mu \gamma_3 + 2e_\mu \gamma_2 - e_\mu \mu \frac{\mathrm{d}}{\mathrm{d}\mu} \ln Z_1, \tag{15.69}$$

where

$$\gamma_2 \equiv \frac{1}{2}\mu \frac{\mathrm{d}}{\mathrm{d}\mu}\bigg|_{e_0} \ln Z_2, \quad \gamma_3 = \frac{1}{2}\mu \frac{\mathrm{d}}{\mathrm{d}\mu}\bigg|_{e_0} \ln Z_3. \tag{15.70}$$

To leading order in e_μ, the γ_3 term in (15.70) reproduces (15.26) when (15.15) is used for Z_3, the other two terms in (15.68) cancelling via $Z_1 = Z_2$. So if, as in the case of QCD, Z_1 is not equal to Z_2, we need to introduce the contributions from loops determining the fermion field strength renormalization factor, as well as those related to the vertex parts (together with appropriate ghost loops), in addition to the vacuum polarization loop associated in the Z_3.

Quantities such as γ_2 and γ_3 have an interesting and important significance, which we shall illustrate in the case of γ_2 for QED. Z_2 enters into the relation between the propagator of the bare fermion $\langle\Omega|T(\hat{\psi}_0(x)\hat{\bar{\psi}}_0(0))|\Omega\rangle$ and the renormalized one, via (cf (11.2))

$$\langle\Omega|T(\hat{\bar{\psi}}(x)\hat{\psi}(0)|\Omega\rangle = \frac{1}{Z_2}\langle\Omega|T(\hat{\bar{\psi}}_0(x)\hat{\psi}_0(0))|\Omega\rangle, \tag{15.71}$$

where (cf section 10.1.3) $|\Omega\rangle$ is the vacuum of the interacting theory. The Fourier transform of (15.71) is, of course, the Feynman propagator:

$$\tilde{S}'_F(q^2) = \int \mathrm{d}^4x e^{\mathrm{i}q\cdot x}\langle\Omega|T(\hat{\bar{\psi}}(x)\hat{\psi}(0))|\Omega\rangle. \tag{15.72}$$

Suppose we now ask: what is the large $-q^2$ behaviour of (15.72) for space-like q^2, with $-q^2 \gg m^2$ where m is the fermion mass? This sounds very similar to the question answered in 15.2.3 for the quantity $S(|q^2|/\mu^2, e_\mu)$. However,

the latter was dimensionless whereas (recalling that $\hat{\psi}$ has mass dimension $\frac{3}{2}$) $\tilde{S}'_F(q^2)$ has dimension M^{-1}. This dimensionality is, of course, just what a propagator of the free-field form $i/(\not{q} - m)$ would provide.

Accordingly, we extract this $(\not{q})^{-1}$ factor (compare $\sigma/\sigma_{\rm pt}$) and consider the dimensionless ratio $\tilde{R}'_F(|q^2|/\mu^2, \alpha_\mu) = \not{q}\tilde{S}'_F(q^2)$. We might guess that, just as for $S(|q^2|/\mu^2, \alpha_\mu)$, to get the leading large $|q^2|$ behaviour we will need to calculate $\tilde{R}'_F$ to some order in α_μ, and then replace α_μ by $\alpha(|q^2|/\mu^2)$. But this is not quite all. The factor Z_2 in (15.71) will – as noted above – depend on the renormalization scale μ, just as Z_3 of (15.15) did. Thus when we change μ, the normalization of the $\hat{\psi}$'s will change via the $Z_2^{\frac{1}{2}}$ factors – of course by a finite amount here – and we must include this change when writing down the analogue of (15.33) for this case (i.e. the condition that the 'total change, on changing μ, is zero'). The required equation is

$$\left[\mu^2 \frac{\partial}{\partial \mu^2}\bigg|_{\alpha_\mu} + \beta(\alpha_\mu)\frac{\partial}{\partial \alpha_\mu} + \gamma_2(\alpha_\mu)\right] \tilde{R}'_F(|q^2|/\mu^2, \alpha_\mu) = 0. \tag{15.73}$$

The solution of (15.73) is somewhat more complicated than that of (15.33). We can gain insight into the essential difference caused by the presence of γ_2 by considering the special case $\beta(\alpha_\mu) = 0$. In this case, we easily find

$$\tilde{R}'_F(|q^2|/\mu^2, \alpha_\mu) \propto (\mu^2)^{-\gamma_2(\alpha_\mu)}. \tag{15.74}$$

But since $\tilde{R}'_F$ can only depend on μ via $|q^2|/\mu^2$, we learn that if $\beta = 0$ then the large $|q^2|$ behaviour of $\tilde{R}'_F$ is given by $(|q^2|/\mu^2)^{\gamma_2}$ – or, in other words, that at large $|q^2|$

$$\tilde{S}'_F(|q^2|/\mu^2, \alpha_\mu) \propto \frac{1}{\not{q}}\left(\frac{|q^2|}{\mu^2}\right)^{\gamma_2(\alpha_\mu)}. \tag{15.75}$$

Thus, *at a zero of the β-function*, $\tilde{S}'_F$ has an 'anomalous'power law dependence on $|q^2|$ (i.e. in addition to the obvious $\not{q}^{-1}$ factor), which is controlled by the parameter γ_2. The latter is called the 'anomalous dimension' of the fermion field, since its presence effectively means that the $|q^2|$ behaviour of $\tilde{S}'_F$ is not determined by its 'normal' dimensionality M^{-1}. The behaviour (15.75) is often referred to as 'scaling with anomalous dimension', meaning that if we multiply $|q^2|$ by a scale factor λ, then $\tilde{S}'_F$ is multiplied by $\lambda^{\gamma_2(\alpha_\mu)-1}$ rather than just λ^{-1}. Anomalous dimensions turn out to play a vital role in the theory of critical phenomena – they are, in fact, closely related to 'critical exponents' (see section 16.4.3, and Peskin and Schroeder 1995, chapter 13). Scaling with anomalous dimensions is also exactly what occurs in deep inelastic scattering of leptons from nucleons, as we shall see in section 15.6.

The full solution of (15.73) for $\beta \neq 0$ is elegantly discussed in Coleman (1985), chapter 3; see also Peskin and Schroeder (1995) section 12.3. We quote

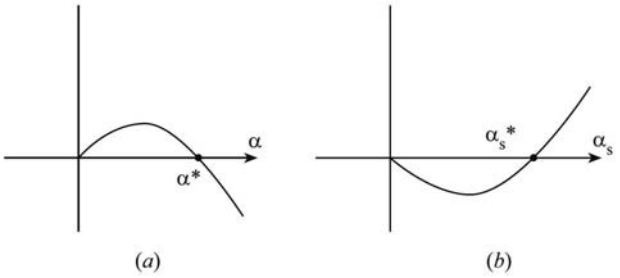


FIGURE 15.6
Possible behaviour of β functions. (a) The slope is positive near the origin (as in QED), and negative near $\alpha = \alpha^*$. (b) The slope is negative at the origin (as in QCD), and positive near $\alpha_s = \alpha_s^*$.

it here:

$$\tilde{R}'_F(|q^2|/\mu^2), \alpha_\mu), = \tilde{R}'_F(1, \alpha(|q^2|/\mu^2)) \exp\left\{\int_0^t \mathrm{d}t' \gamma_2(\alpha(t'))\right\}. \qquad (15.76)$$

The first factor is the expected one from section 15.2.3; the second results from the addition of the γ_2 term in (15.73). Suppose now that $\beta(\alpha)$ has a zero at some point α^*, in the vicinity of which $\beta(\alpha) \approx -B(\alpha - \alpha^*)$ with $B > 0$. Then, near this point the evolution of α is given by (cf (15.39))

$$\ln(|q^2|/\mu^2) = \int_{\alpha_\mu}^{\alpha(|q^2|)} = \frac{\mathrm{d}\alpha}{-B(\alpha - \alpha^*)}, \qquad (15.77)$$

which implies

$$\alpha(|q^2|) = \alpha^* + \text{constant } \times (\mu^2/|q^2|)^B. \qquad (15.78)$$

Thus asymptotically for large $|q^2|$, the coupling will evolve to the '*fixed point*' α^*. In this case, at sufficiently large $-q^2$, the integral in (15.76) can be evaluated by setting $\alpha(t') = \alpha^*$, and $\tilde{R}'_F$ will scale with an anomalous dimension $\gamma_2(\alpha^*)$ determined by the fixed point value of α. The behaviour of such an α is shown in figure 15.6(a). We emphasize that there is no reason to believe that the QED β function actually does behave like this.

The point α^* in figure 15.6(a) is called an ultraviolet-stable fixed point: α 'flows' towards it at large $|q^2|$. In the case of QCD, the β function starts out negative, so that the corresponding behaviour (with a zero at a $\alpha_s^* \neq 0$) would look like that shown in figure 15.6(b). In this case, the reader can check (problem 15.5) that α_s^* is reached in the infrared limit $q^2 \to 0$, and so α_s^* is called an infrared-stable fixed point. Clearly it is the slope of β near the fixed point that determines whether it is u-v or i-r stable. This applies equally to a fixed point at the origin, so that QED is i-r stable at $\alpha = 0$ while QCD is u-v stable at $\alpha_s = 0$.

We must now point out to the reader an error in the foregoing analysis, in the case of a gauge theory. The quantity Z_2 is not gauge invariant in QED (or QCD), and hence γ_2 depends on the choice of gauge. This is really no surprise, because the full fermion propagator itself is not gauge invariant (the free-field propagator is gauge invariant, of course). What ultimately matters is that the complete physical amplitude for any process, at a given order of α, be gauge invariant. Thus the analysis given above really only applies – in this simple form – to non-gauge theories, such as the ABC model of chapter 6, or to gauge-invariant quantities.

This is an appropriate point at which to consider the treatment of quark masses in the RGE-based approach. Up to now we have simply assumed that the relevant $|q^2|$ is very much greater than all quark masses, the latter therefore being neglected. While this may be adequate for the light quarks u, d, s, it seems surely a progressively worse assumption for c, b and t. However, in thinking about how to re-introduce the quark masses into our formalism, we are at once faced with a difficulty: how are they to be defined? For an unconfined particle, such as a lepton, it seems natural to define 'the' mass as the position of the pole of the propagator (i.e. the 'on-shell' value $p^2 = m^2$), a definition we followed in chapters 10 and 11. Significantly, renormalization is required (in the shape of a mass counter-term) to achieve a pole at the 'right' physical mass m, in this sense. But this prescription is inherently perturbative, and cannot be used for a confined particle, which never 'escapes' beyond the range of the non-perturbative confining forces, and whose propagator can therefore never approach the form $\sim (\not{p} - m)^{-1}$ of a free particle.

Our present perspective on renormalization suggests an obvious way forward. Just as there was, in principle, no *necessity* to define the QED coupling parameter e via an on-shell prescription, so here a mass parameter in the Lagrangian can be defined in any way we find convenient; all that is necessary is that it should be possible to determine its value from some measurable quantity (for example, quark masses from lattice QCD predictions of hadron masses). Effectively, we are regarding the 'm' in a term such as $-m\bar{\hat{\psi}}(x)\hat{\psi}(x)$ as a 'coupling constant' having mass dimension 1 (and, after all, the ABC coupling itself had mass dimension 1). Incidentally, the operator $\bar{\hat{\psi}}(x)\hat{\psi}(x)$ *is* gauge invariant, as is any such *local* operator. Taking this point of view, it is clear that a renormalization scale will be involved in such a general definition of mass, and we must expect to see our mass parameters 'evolve' with this scale, just as the gauge (or other) couplings do. In turn, this will get translated into a $|q^2|$-dependence of the mass parameters, just as for $\alpha(|q^2|)$ and $\alpha_s(|q^2|)$.

The RGE in such a scheme now takes the form

$$\left[\mu^2\frac{\partial}{\partial\mu^2} + \beta(\alpha_s)\frac{\partial}{\partial\alpha_s} + \sum_i \gamma_i(\alpha_s) + \gamma_m(\alpha_s)m\frac{\partial}{\partial m}\right] R(|q^2|/\mu^2, \alpha_s, m/|q|) = 0 \tag{15.79}$$

where the partial derivatives are taken at fixed values of the other two vari-

ables. Here the γ_i are the anomalous dimensions relevant to the quantity R, and γ_m is an analogous 'anomalous mass dimension', arising from finite shifts in the mass parameter when the scale μ^2 is changed. Just as with the solution (15.76) of (15.73), the solution of (15.79) is given in terms of a 'running mass' $m(|q^2|)$. Formally, we can think of γ_m in (15.79) as analogous to $\beta(\alpha_s)$ and $\ln m$ as analogous to α_s. Then equation (15.41) for the running α_s,

$$\frac{\partial \alpha_s(|q^2|)}{\partial t} = \beta(\alpha_s(|q^2|)) \tag{15.80}$$

where $t = \ln(|q^2|/\mu^2)$, becomes

$$\frac{\partial(\ln m(|q^2|))}{\partial t} = \gamma_m(\alpha_s(|q|^2)). \tag{15.81}$$

Equation (15.81) has the solution

$$m(|q^2|) = m(\mu^2) \exp \int_{\ln \mu^2}^{\ln |q^2|} \mathrm{d} \ln |q'^2|\ \gamma_m(\alpha_s(|q'^2|). \tag{15.82}$$

To one-loop order in QCD, $\gamma_m(\alpha_s)$ turns out to be $-\frac{1}{\pi}\alpha_s$ (Peskin and Schroeder 1995, section 18.1). Inserting the one-loop solution for α_s in the form (15.53), we find

$$m(|q^2|) = m(\mu^2) \left[\frac{\ln(\mu^2/\Lambda^2)}{\ln(|q^2|/\Lambda^2)} \right]^{\frac{1}{\pi\beta_0}}, \tag{15.83}$$

where $(\pi\beta_0)^{-1} = 12/(33 - 2N_f)$. Thus the quark masses decrease logarithmically as $|q^2|$ increases, rather like $\alpha_s(|q^2|)$. It follows that, in general, quark mass effects are suppressed both by explicit $m^2/|q^2|$ factors, and by the logarithmic decrease given by (15.83). Further discussion of the treatment of quark masses is contained in Ellis, Stirling and Webber (1996), section 2.4; see also the review by Manohar and Sachrajda in Nakamura *et al.* 2010.

15.6 QCD corrections to the parton model predictions for deep inelastic scattering: scaling violations

As we saw in section 9.2, the parton model provides a simple intuitive explanation for the experimental observation that the nucleon structure functions in deep inelastic scattering depend, to a good first approximation, only on the dimensionless ratio $x = Q^2/2M\nu$, rather than on Q^2 and ν separately; this behaviour is referred to as 'scaling'. Here M is the nucleon mass, and Q^2 and ν are defined in (9.7) and (9.8). In this section we shall show how QCD

corrections to the simple parton model, calculated using RGE techniques, predict observable violations of scaling in deep inelastic scattering. As we shall see, comparison between the theoretical predictions and experimental measurements provides strong evidence for the correctness of QCD as the theory of nucleonic constituents.

15.6.1 Uncancelled mass singularities at order α_s.

The free parton model amplitudes we considered in chapter 9 for deep inelastic lepton-nucleon scattering were of the form shown in figure 15.7 (cf figure 9.4). The obvious first QCD corrections will be due to real gluon emission by either the initial or final quark, as shown in figure 15.8, but to these we must add the one-loop virtual gluon processes of figure 15.9 in order (see below) to get rid of infrared divergences similar to those encountered in section 14.4.2, and also the diagram of figure 15.10, corresponding to the presence of gluons in the nucleon. To simplify matters, we shall consider what is called a 'non-singlet structure function' $F_2^{\rm NS}$, such as $F_2^{\rm ep} - F_2^{\rm en}$ in which the (flavour) singlet gluon contribution cancels out, leaving only the diagrams of figures 15.8 and 15.9.

We now want to perform, for these diagrams, calculations analogous to those of section 9.2, which enabled us to find the e-N structure functions νW_2 and MW_1 from the simple parton process of figure 15.7. There are two problems here: one is to find the parton level W's corresponding to figure 15.8 (leaving aside figure 15.9 for the moment) – cf equations (9.29) and (9.30) in the case of the free parton diagram figure 15.7; the other is to relate these parton W's to observed nucleon W's via an integration over momentum fractions. In section 9.2 we solved the first problem by explicitly calculating the parton level $\mathrm{d}^2\sigma^i/\mathrm{d}Q^2\mathrm{d}\nu$ and picking off the associated νW_2^i, W_1^i. In principle, the same can be done here, starting from the five-fold differential cross section for our $\mathrm{e}^- + \mathrm{q} \to \mathrm{e}^- + \mathrm{q} + \mathrm{g}$ process. However, a simpler – if somewhat heuristic – way is available. We note from (9.46) that in general $F_1 = MW_1$ is given by the transverse virtual photon cross section

$$W_1 = \sigma_T/(4\pi^2\alpha/K) = \frac{1}{2}\sum_{\lambda=\pm1} \epsilon_\mu^*(\lambda)\epsilon_\nu(\lambda)W^{\mu\nu} \tag{15.84}$$

where $W^{\mu\nu}$ was defined in (9.3). Further, the Callan–Gross relation is still true (the photon only interacts with the charged partons, which are quarks with spin $\frac{1}{2}$ and charge e_i), and so

$$F_2/x = 2F_1 = 2MW_1 = \sigma_T/(4\pi^2\alpha/2MK). \tag{15.85}$$

These formulae are valid for both parton and proton W_1's and $W^{\mu\nu}$'s, with appropriate changes for parton masses $\hat{M}$. Hence the parton level $2\hat{F}_1$ for figure 15.8 is just the transverse photon cross section as calculated from the graphs of figure 15.11, divided by the factor $4\pi^2\alpha/2\hat{M}\hat{K}$, where as usual '$\hat{}$' denotes kinematic quantities in the corresponding parton process. This cross section,

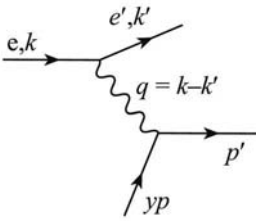


FIGURE 15.7
Electron-quark scattering via one-photon exchange.

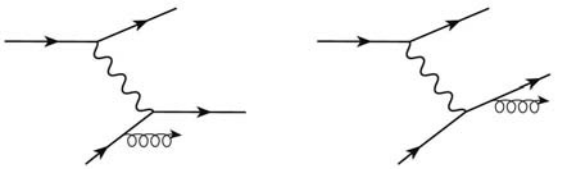


FIGURE 15.8
Electron-quark scattering with single-gluon emission

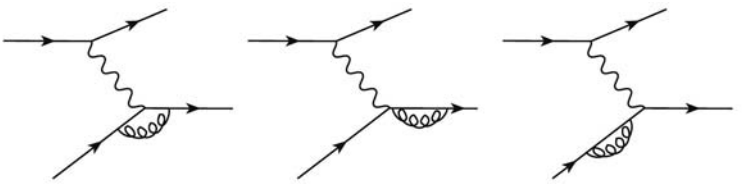


FIGURE 15.9
Virtual single-gluon corrections to figure 15.7.

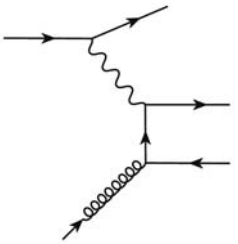


FIGURE 15.10
Electron-gluon scattering with $\bar{q}q$ production.

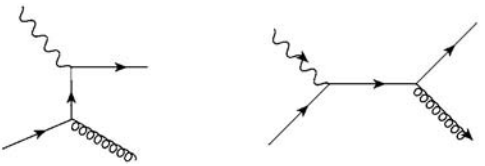


FIGURE 15.11
Virtual photon processes entering into figure 15.8.

however, is – apart from a colour factor – just the virtual Compton cross section calculated in section 8.6. Also, taking the same (Hand) convention for the individual photon flux factors,

$$2\hat{M}\hat{K} = \hat{s}. \tag{15.86}$$

Thus for the parton processes of figure 15.9,

$$\begin{aligned} 2\hat{F}_1 &= \hat{\sigma}_T/(4\pi^2\alpha/2\hat{M}\hat{K}) \\ &= \frac{\hat{s}}{4\pi^2\alpha}\int_{-1}^{1} \mathrm{d}\cos\theta \cdot \frac{4}{3} \cdot \frac{\pi {e_i}^2 \alpha\alpha_{\mathrm{s}}(\mu^2)}{\hat{s}} \left(-\frac{\hat{t}}{\hat{s}} - \frac{\hat{s}}{\hat{t}} + \frac{2\hat{u}Q^2}{\hat{s}\hat{t}}\right) \end{aligned} \tag{15.87}$$

where, in going from (8.181) to (15.87), we have inserted a colour factor $\frac{4}{3}$ (problem 14.5 (a)), renamed the variables $\hat{t} \to \hat{u}, \hat{u} \to \hat{t}$ in accordance with figure 15.11, and replaced α^2 by ${e_i}^2\alpha\alpha_{\mathrm{s}}(\mu^2)$.

Before proceeding with (15.87), it is helpful to consider the other part of the calculation – namely the relation between the nucleon F_1 and the parton $\hat{F}_1$. We mimic the discussion of section 9.2, but with one significant difference: the quark 'taken' from the proton still has momentum fraction y (momentum yp), but now its longitudinal momentum must be degraded in the final state due to the gluon bremsstrahlung process we are calculating. Let us call the quark momentum after gluon emission zyp (figure 15.12). Then, assuming as in section 9.2 that it stays on-shell, we have

$$q^2 + 2zyq \cdot p = 0 \tag{15.88}$$

or

$$x = yz, \qquad x = Q^2/2q \cdot p, \qquad q^2 = -Q^2 \tag{15.89}$$

and we can write (cf (9.31))

$$\frac{F_2}{x} = 2F_1 = \sum_i \int_0^1 \mathrm{d}y f_i(y) \int_0^1 \mathrm{d}z \; 2\hat{F}_1^i \delta(x - yz) \tag{15.90}$$

where the $f_i(y)$ are the parton distribution functions introduced in section 9.2 (we often call them $q(x)$ or $g(x)$ as the case may be) for parton type i, and

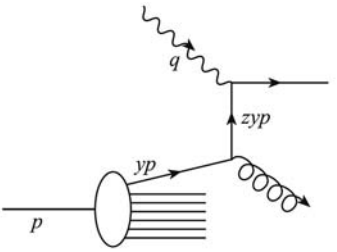


FIGURE 15.12
The first process of figure 15.11, viewed as a contribution to e^--nucleon scattering.

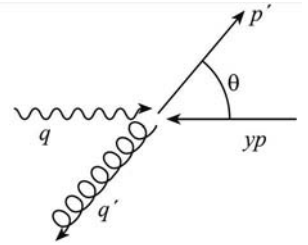


FIGURE 15.13
Kinematics for the parton process of figure 15.12.

the sum is over contributing partons. The reader may enjoy checking that (15.90) does reduce to (9.34) for free partons by showing that in that case $2\hat{F}_1^i = e_i^2\delta(1-z)$ (see Halzen and Martin 1984, section 10.3, for help), so that $2F_1^{\text{free}} = \sum_i e_i^2 f_i(x)$.

To proceed further with the calculation (i.e. of (15.87) inserted into (15.90)), we need to look at the kinematics of the $\gamma q \to qg$ process, in the CMS. Referring to figure 15.13, we let k, k' be the magnitudes of the CMS momenta $\boldsymbol{k}, \boldsymbol{k}'$. Then

$$\begin{aligned} \hat{s} &= 4k'^2 = (yp+q)^2 = Q^2(1-z)/z, \qquad z = Q^2/(\hat{s}+Q^2) \\ \hat{t} &= (q-p')^2 = -2kk'(1-\cos\theta) = -Q^2(1-c)/2z, \qquad c = \cos\theta \\ \hat{u} &= (q-q')^2 = -2kk'(1+\cos\theta) = -Q^2(1+c)/2z. \end{aligned} \tag{15.91}$$

We now note that in the integral (15.87) for $\hat{F}_1$, when we integrate over $c = \cos\theta$, we shall obtain an infinite result

$$\sim \int^1 \frac{\mathrm{d}c}{1-c} \tag{15.92}$$

associated with the vanishing of $\hat{t}$ in the 'forward' direction (i.e. when q and p'

are parallel). This is a divergence of the 'collinear' type, in the terminology of section 14.4.2 – or, as there, a 'mass singularity', occurring in the zero quark mass limit. If we simply replace the propagator factor $\hat{t}^{-1} = [(q-p')^2]^{-1}$ by $[(q-p')^2 - m^2]^{-1}$, where m is a quark mass, then (15.92) becomes

$$\sim \int^1 \frac{\mathrm{d}c}{(1+2m^2z/Q^2)-c} \tag{15.93}$$

which will produce a factor of the form $\ln(Q^2/m^2)$ as $m^2 \to 0$. Thus m regulates the divergence. We have here an uncancelled mass singularity, and it *violates scaling*. This crucial physical result is present in the lowest-order QCD correction to the parton model, in this case. As we are learning, such logarithmic violations of scaling are a characteristic feature of all QCD corrections to the free (scaling) parton model.

We may calculate the coefficient of the $\ln Q^2$ term by retaining in (15.87) only the terms proportional to $\hat{t}^{-1}$:

$$2\hat{F}_1^i \approx e_i^2 \int_{-1}^1 \frac{\mathrm{d}c}{1-c}\left(\frac{\alpha_\mathrm{s}(\mu^2)}{2\pi}\cdot\frac{4}{3}\cdot\frac{1+z^2}{1-z}\right) \tag{15.94}$$

and so, for just one quark species, this QCD correction contributes (from (15.90)) a term

$$\frac{e_i^2\alpha_\mathrm{s}(\mu^2)}{2\pi}\int_x^1 \frac{\mathrm{d}y}{y}q(y)\left\{\hat{P}_{\mathrm{qq}}(x/y)\ln(Q^2/m^2)+C(x/y)\right\} \tag{15.95}$$

to $2F_1$, where

$$\hat{P}_{\mathrm{qq}}(z) = \frac{4}{3}\left(\frac{1+z^2}{1-z}\right), \tag{15.96}$$

and $C(x/y)$ has no mass singularity.

Our result so far is therefore that the 'free' quark distribution function $q(x)$, which depended only on the scaling variable x, becomes modified to

$$q(x) + \frac{\alpha_\mathrm{s}(\mu^2)}{2\pi}\int_x^1 \frac{\mathrm{d}y}{y}q(y)\left\{\hat{P}_{\mathrm{qq}}(x/y)\ln\left(Q^2/m^2\right)+C(x/y)\right\} \tag{15.97}$$

$$= q(x) + \frac{\alpha_\mathrm{s}(\mu^2)}{2\pi}\int_0^1 \mathrm{d}y\int_0^1 \mathrm{d}z\delta(zy-x)q(y)\{\hat{P}_{\mathrm{qq}}(z)\ln(Q^2/m^2) + C(z)\} \tag{15.98}$$

due to lowest-order gluon radiation. Clearly, this corrected distribution function violates scaling because of the $\ln Q^2$ term. But the result as it stands cannot represent a well-controlled approximation, since it contains divergences as $z \to 1$ and as $m^2 \to 0$.

We postpone discussion of the mass divergence until the next section. The divergence as $z \to 1$ is a standard infrared divergence (the quark momentum

yzp after gluon emission becomes equal to the quark momentum *yp* before emission), and we expect that it can be cured by including the virtual gluon diagrams of figure 15.9, as indicated at the start of the section (and as was done analogously in the case of e^+e^- annihilation). This has been verified explicitly by Kim and Schilcher (1978) and by Altarelli *et al.* (1978 a, b; 1979). Alternatively, we follow the procedure of Altarelli and Parisi (1977). First we regulate the divergence as $z \to 1$ by defining a regulated function $1/(1-z)_+$ such that

$$\int_0^1 \frac{f(z)}{(1-z)_+}\mathrm{d}z = \int_0^1 \frac{f(z)-f(1)}{(1-z)}\mathrm{d}z = \int_0^1 \ln(1-z)\frac{\mathrm{d}f(z)}{\mathrm{d}z}\mathrm{d}z, \qquad (15.99)$$

where $f(z)$ is any test function sufficiently regular at the end points. Now the gluon loops which will cancel the i-r divergence only contribute at $z \to 1$, in leading log approximation. Thus the i-r finite version of $\hat{P}_{\mathrm{qq}}$ has the form

$$P_{\mathrm{qq}}(z) = \frac{4}{3}\frac{1+z^2}{(1-z)_+} + A\delta(1-z). \qquad (15.100)$$

The coefficient A is determined by the physical requirement that the net number of quarks (i.e. the number of quarks minus the number of antiquarks) does not vary with Q^2. From (15.98) this implies

$$\int_0^1 P_{\mathrm{qq}}(z)\mathrm{d}z = 0. \qquad (15.101)$$

Inserting (15.100) into (15.101), and using (15.99), we find (problem 15.6)

$$A = 2, \qquad (15.102)$$

so that

$$P_{\mathrm{qq}}(z) = \frac{4}{3}\frac{(1+z^2)}{(1-z)_+} + 2\delta(1-z). \qquad (15.103)$$

The function P_{qq} is called a 'splitting function', and it has an important physical interpretation. The quantity $\alpha_{\mathrm{s}}(\mu^2)/(2\pi)\ P_{\mathrm{qq}}(z)$ is, for $z < 1$, the probability that, to first order in α_{s}, a quark having radiated a gluon is left with a fraction z of its original momentum. Similar functions arise in QED in connection with what is called the 'equivalent photon approximation' (Weizsäcker 1934, Williams 1934, Chen and Zerwas 1975). The application of these techniques to QCD corrections to the free parton model is due to Altarelli and Parisi (1977), who thereby opened the way to this simpler and more physical way of understanding scaling violations, which had previously been discussed mainly within the rather technical operator product formalism (Wilson 1969).

We must now find some way of making sense, physically, of the uncancelled mass divergence in (15.97).

15.6.2 Factorization, and the order α_s DGLAP equation

The key is to realize that when two partons are in the collinear configuration their relative momentum is very small, and hence the interaction between them is very strong, beyond the reach of a perturbative calculation. This suggests that we should absorb such uncalculable effects into a modified distribution function $q(x, \mu_F^2)$ given by

$$q(x, \mu_F^2) = q(x) + \frac{\alpha_s(\mu^2)}{2\pi} \int_x^1 \frac{dy}{y} q(y) P_{qq}(x/y) \left\{ \ln(\mu_F^2/m^2) + C(x/y) \right\} \tag{15.104}$$

which we have to take from experiment. Note that we have also absorbed the non-singular term $C(x/y)$ into $q(x, \mu_F^2)$. In terms of this quantity, then, we have

$$F_2(x, Q^2) \equiv e_i^2 x q(x, Q^2) \tag{15.105}$$

$$= x e_i^2 \int_x^1 \frac{dy}{y} q(y, \mu_F^2) \left\{ \delta(1 - x/y) + \frac{\alpha_s(\mu^2)}{2\pi} P_{qq}(x/y) \ln(Q^2/\mu_F^2) \right\} \tag{15.106}$$

to this order in α_s, and for one quark type.

This procedure is, of course, very reminiscent of ultraviolet renormalization, in which u-v divergences are controlled by similarly importing some quantities from experiment. In this example, we have essentially made use of the simple fact that

$$\ln(Q^2/m^2) = \ln(Q^2/\mu_F^2) + \ln(\mu_F^2/m^2). \tag{15.107}$$

The arbitrary scale μ_F is analogous to renormalization scale μ (which we have retained in $\alpha_s(\mu^2)$), and is here referred to as a 'factorization scale'. It is the scale entering into the separation in (15.107), between one (uncalculable) factor which depends on the i-r parameter m but not on Q^2, and the other (calculable) factor which depends on Q^2. The scale μ_F can be thought of as one which separates the perturbative short-distance physics from the non-perturbative long-distance physics. Thus partons emitted at small transverse momenta $< \mu_F$ (i.e. approximately collinear processes) should be considered as part of the hadron structure, and are absorbed into $q(x, \mu_F^2)$. Partons emitted at large transverse momenta contribute to the short-distance (calculable) part of the cross section. Just as for the renormalization scale, the more terms that can be included in the perturbative contributions to the mass-singular terms, the weaker the dependence on μ_F will be. We have demonstrated the possibility of factorization only to $O(\alpha_s)$, but proofs to all orders in perturbation theory exist; reviews are provided by Collins and Soper (1987, 1988).

Returning now to (15.106), the reader can guess what is coming next: we shall impose the condition that the physical quantity $F_2(x, Q^2)$ must be independent of the choice of factorization scale μ_F^2. Differentiating (15.106)

partially with respect to $\mu_{\rm F}^2$, and setting the result to zero, we obtain (to order $\alpha_{\rm s}$ on the right-hand side)

$$\mu_{\rm F}^2 \frac{\partial q(x,\mu_{\rm F}^2)}{\partial \mu_{\rm F}^2} = \frac{\alpha_{\rm s}(\mu^2)}{2\pi} \int_x^1 \frac{{\rm d}y}{y} P_{\rm qq}(x/y) q(y,\mu_{\rm F}^2). \tag{15.108}$$

This equation is the analogue of equation (15.35) describing the running of the coupling $\alpha_{\rm s}$ with μ^2, and is a fundamental equation in the theory of perturbative applications of QCD. It is called the DGLAP equation, after Dokshitzer (1977), Gribov and Lipatov (1972), and Altarelli and Parisi (1977). The above derivation is not rigorous: a more sophisticated treatment (Georgi and Politzer 1974, Gross and Wilczek 1974) confirms the result and extends it to higher orders.

Equation (15.108) shows that, although perturbation theory cannot be used to calculate the distribution function $q(x,\mu_{\rm F}^2)$ at any particular value $\mu_{\rm F}^2 = \mu_0^2$, it can be used to predict how the distribution *changes* (or 'evolves') as $\mu_{\rm F}^2$ varies. (We recall from (15.105) that $q(x,\mu_0^2)$ can be found experimentally via $xq(x,\mu_0^2) = 2F_2(x,Q^2=\mu_0^2)/e_i^2$.) As in the case of $\sigma(\mathrm{e^+e^-} \to \text{hadrons})$ and the scale μ^2, the choice of factorization scale is arbitrary, and would cancel from physical quantities if all powers in the perturbation series were included. Truncating at N terms results in an ambiguity of order $\alpha_{\rm s}^{(N+1)}$. In deep inelastic predictions, the standard choice for scales is $\mu^2 = \mu_{\rm F}^2 = Q^2$.

The way the non-singlet distribution changes can be understood qualitatively as follows. The change in the distribution for a quark with momentum fraction x, which absorbs the virtual photon, is given by the integral over y of the corresponding distribution for a quark with momentum fraction y, which radiated away (via a gluon) a fraction x/y of its momentum with probability $(\alpha_{\rm s}/2\pi)P_{\rm qq}(x/y)$. This probability is high for large momentum fractions: high-momentum quarks lose momentum by radiating gluons. Thus there is a predicted tendency for the distribution function $q(x,\mu^2)$ to get smaller at large x as μ^2 increases, and larger at small x (due to the build-up of slower partons), while maintaining the integral of the distribution over x as a constant. The effect is illustrated qualitatively in figure 15.14. In addition, the radiated gluons produce more $q\bar{q}$ pairs at small x. Thus the nucleon may be pictured as having more and more constituents, all contributing to its total momentum, as its structure is probed on ever smaller distance (larger μ^2) scales.

In general, the right-hand side of (15.108) will have to be supplemented by terms (calculable from figure 15.10) in which quarks are generated from the gluon distribution; the equations must then be closed by a corresponding one describing the evolution of the gluon distributions (Altarelli 1982). In the now commonly used notation, this generalization of (15.108) reads

$$\mu_{\rm F}^2 \frac{\partial f_{i/{\rm p}}(x,\mu_{\rm F}^2)}{\partial \mu_{\rm F}^2} = \sum_{j={\rm q,g}} \frac{\alpha_{\rm s}(\mu_{\rm F}^2)}{2\pi} \int_x^1 \frac{{\rm d}y}{y} P_{i\leftarrow j}^{(1)}(x/y) f_{j/{\rm p}}(y,\mu_{\rm F}^2), \tag{15.109}$$

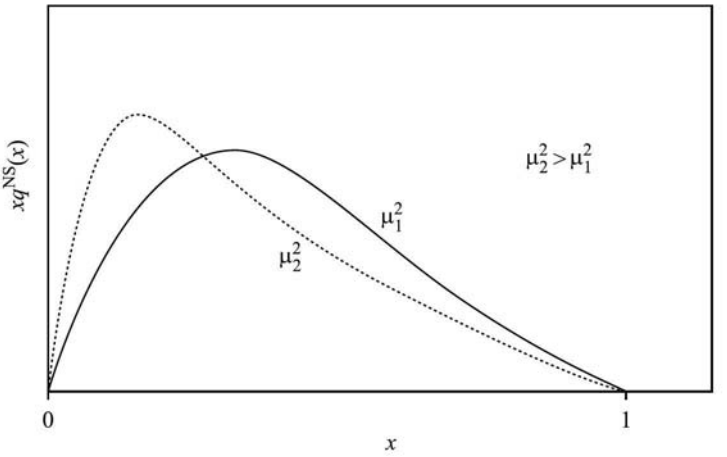


FIGURE 15.14
Evolution of the distribution function with μ^2.

where the sum is over quark types q and gluons g, $P^{(1)}_{i\leftarrow j}$ is the $j \to i$ splitting function to this order, and $f_{i/\mathrm{p}}$ is the parton distribution function for partons of type i in the proton. In our previous notation, $P^{(1)}_{\mathrm{q}\leftarrow\mathrm{q}}(x/y) = P_{\mathrm{qq}}(x/y)$, and $f_{\mathrm{q/p}}(x, \mu_{\mathrm{F}}^2) = q(x, \mu_{\mathrm{F}}^2)$. The other splitting functions may be found in Altarelli (1982).

Both the splitting functions and expression (15.106) for $F_2(x, Q^2)$ can be extended to higher orders in α_{s}. Thus the perturbative expansion (15.106) becomes

$$F_2(x, Q^2) = x \sum_{n=0}^{\infty} \frac{\alpha_{\mathrm{s}}^n(\mu_{\mathrm{F}}^2)}{(2\pi)^n} \sum_{i=\mathrm{q,g}} \int_x^1 \frac{\mathrm{d}z}{z} C^{(n)}_{2,i}(z, Q^2, \mu_{\mathrm{F}}^2) f_{i/\mathrm{p}}(x/z, \mu_{\mathrm{F}}^2), \quad (15.110)$$

where we have chosen $\mu = \mu_{\mathrm{F}}$. The expansion (15.110) is analogous to (15.63), and as in that case the coefficient functions will depend on μ_{F}^2 in such a way that, order by order, the μ_{F}^2 dependence will cancel. At zeroth order the coefficients are the μ_{F}^2-independent free parton ones, $C^{(0)}_{2,\mathrm{q}} = e_{\mathrm{q}}^2 \delta(1-z)$ and $C^{(0)}_{2,\mathrm{g}} = 0$. In most cases the coefficients have been calculated up to order α_{s}^2 (Nakamura *et al.* 2010).

We ought also to mention that there are in principle non-perturbative corrections to both (15.63) and (15.110), which are of order $(\Lambda^2_{\overline{\mathrm{MS}}}/Q^2)^2$ and $(\Lambda^2_{\overline{\mathrm{MS}}}/Q^2)$ respectively.

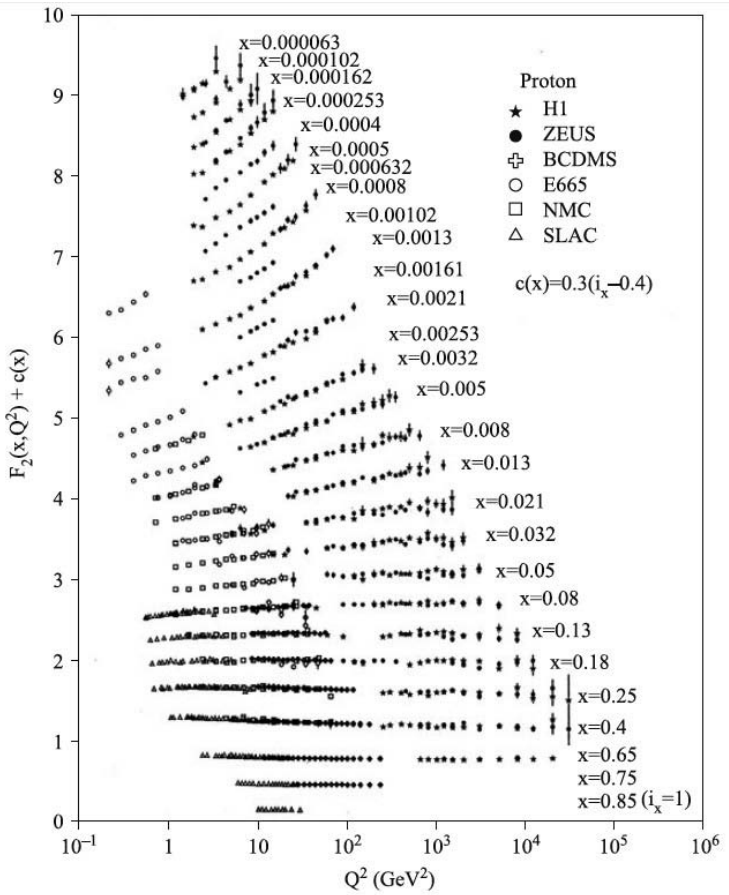


FIGURE 15.15
Q^2-dependence of the proton structure function F_2^{p} for various fixed x values (Hagiwara *et al.* 2002). i_x is a number depending on the x-bin, ranging from $i_x = 1$ ($x = 0.85$) to $i_x = 28$ ($x = 0.000063$). Figure reprinted with permission from K Hagiwara *et al. Phys.Rev.* D **66** 010001 (2002). Copyright 2002 by the American Physical Society.

15.6.3 Comparison with experiment

Data on nucleon structure functions do indeed show the trend described in the previous section. Figure 15.15 shows the Q^2-dependence of the proton structure function $F_2^{\mathrm{p}}(x, Q^2) = \sum e_i^2 x f_{i/\mathrm{p}}(x, Q^2)$ for various fixed x values, as compiled by B. Foster, A.D. Martin and M.G. Vincter for the 2002 Particle Data Group review (Hagiwara *et al.* 2002). Clearly at larger x ($x \geq 0.13$) the function gets smaller as Q^2 increases, while at smaller x it increases.

Fits to the data have been made in various ways. One (theoretically convenient) way is to consider 'moments' (Mellin transforms) of the structure functions, defined by

$$M_{\rm q}^n(t) = \int_0^1 {\rm d}x x^{n-1} q(x,t), \tag{15.111}$$

where we have taken $\mu^2 = \mu_{\rm F}^2$ and introduced the variable $t = \ln \mu^2$. Taking moments of both sides of (15.108) and interchanging the order of the x and y integrations, we find

$$\frac{{\rm d}M_{\rm q}^n(t)}{{\rm d}t} = \frac{\alpha_{\rm s}(t)}{2\pi} \int_0^1 {\rm d}y y^{n-1} q(y,t) \int_0^y \frac{{\rm d}x}{y} (x/y)^{n-1} P_{\rm qq}(x/y). \tag{15.112}$$

Changing the variable to $z = x/y$ in the second integral, and defining[4]

$$\gamma_{\rm qq}^n = 4 \int_0^1 {\rm d}z z^{n-1} P_{\rm qq}(z), \tag{15.113}$$

we obtain

$$\frac{{\rm d}M_{\rm q}^n(t)}{{\rm d}t} = \frac{\alpha_{\rm s}(t)}{8\pi} \gamma_{\rm qq}^n M_{\rm q}^n(t). \tag{15.114}$$

Thus the integral in (15.108) – which is of convolution type – has been reduced to product form by this transformation. Now we also know from (15.47) and (15.48) that

$$\frac{{\rm d}\alpha_{\rm s}}{{\rm d}t} = -\beta_0 \alpha_{\rm s}^2 \tag{15.115}$$

with $\beta_0 = (33 - 2N_{\rm f})/12\pi$ as usual, to this (one-loop) order. Thus (15.114) becomes

$$\frac{{\rm d}\ln M_{\rm q}^n}{{\rm d}\ln \alpha_{\rm s}} = -\frac{\gamma_{\rm qq}^n}{8\pi\beta_0} = -d_{\rm qq}^n, \text{ say.} \tag{15.116}$$

The solution to (15.116) is easily found to be

$$M_{\rm q}^n(t) = M_{\rm q}^n(t_0) \left(\frac{\alpha_{\rm s}(t_0)}{\alpha_{\rm s}(t)} \right)^{d_{\rm qq}^n}. \tag{15.117}$$

Applying the prescription (15.99) to γ_n, we find (problem 15.9)

$$\gamma_{\rm qq}^n = -\frac{8}{3} \left[1 - \frac{2}{n(n+1)} + 4 \sum_{j=2}^{n} \frac{1}{j} \right] \tag{15.118}$$

[4]The notation is not chosen accidentally: the γ's are indeed anomalous dimensions of certain operators which appear in Wilson's operator product approach to scaling violations (Wilson 1969); interested readers may pursue this with Peskin and Schroeder 1995, chapter 18.

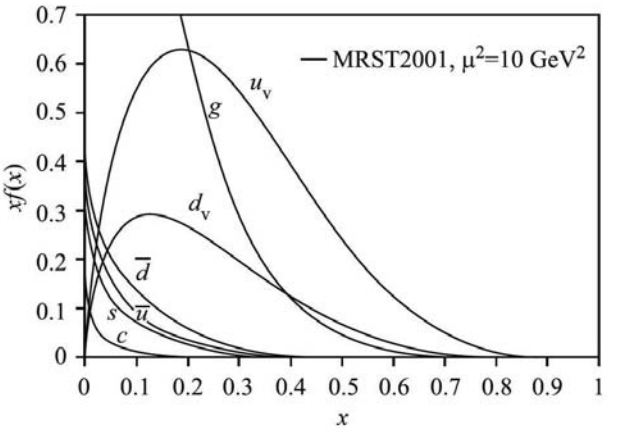


FIGURE 15.16
Distributions of x times the unpolarized parton distributions $f(x, \mu^2)$ (where $f = u_v, d_v, \bar{u}, \bar{d}, s, c, g$) using the MRST2001 parametrization (Martin *et al.* 2002) at a scale $\mu^2 = 10\text{GeV}^2$. Figure reprinted with permission from K Hagiwara *et al. Phys. Rev.* D **66** 010001 (2002). Copyright 2002 by the American Physical Society.

and then

$$d^n_{\text{qq}} = \frac{4}{33 - 2N_{\text{f}}} \left[1 - \frac{2}{n(n+1)} + 4 \sum_{j=2}^{n} \frac{1}{j} \right]. \tag{15.119}$$

We emphasize again that all the foregoing analysis is directly relevant only to distributions in which the flavour singlet gluon distributions do not contribute to the evolution equations. In the more general case, analogous splitting functions $P_{\text{qg}}, P_{\text{gq}}$ and P_{gg} will enter, folded appropriately with the gluon distribution function $g(x, t)$, together with the related quantities $\gamma^n_{\text{qg}}, \gamma^n_{\text{gq}}$ and γ^n_{gg}. Equation (15.108) is then replaced by a 2×2 matrix equation for the evolution of the quark and gluon moments M^n_{q} and M^n_{g}.

Returning to (15.117), one way of testing it is to plot the logarithm of one moment, $\ln M^n_{\text{q}}$, versus the logarithm of another, $\ln M^m_{\text{q}}$, for different n, m values. A more direct procedure, applicable to the non-singlet case too of course, is to choose a reference point μ^2_0 and parametrize the parton distribution functions $f_i(x, t_0)$ in some way. These may then be evolved numerically, via the DGLAP equations, to the desired scale. Figure 15.16 shows a typical set of distributions at $\mu^2 = 10$ GeV^2 (Martin *et al.* 2002). A global numerical fit is then performed to determine the best values of the parameters, including the parameter $\Lambda_{\overline{\text{MS}}}$ which enters into $\alpha_{\text{s}}(t)$. An example of such a fit, due to Martin *et al.* (1994), is shown in figure 15.17.

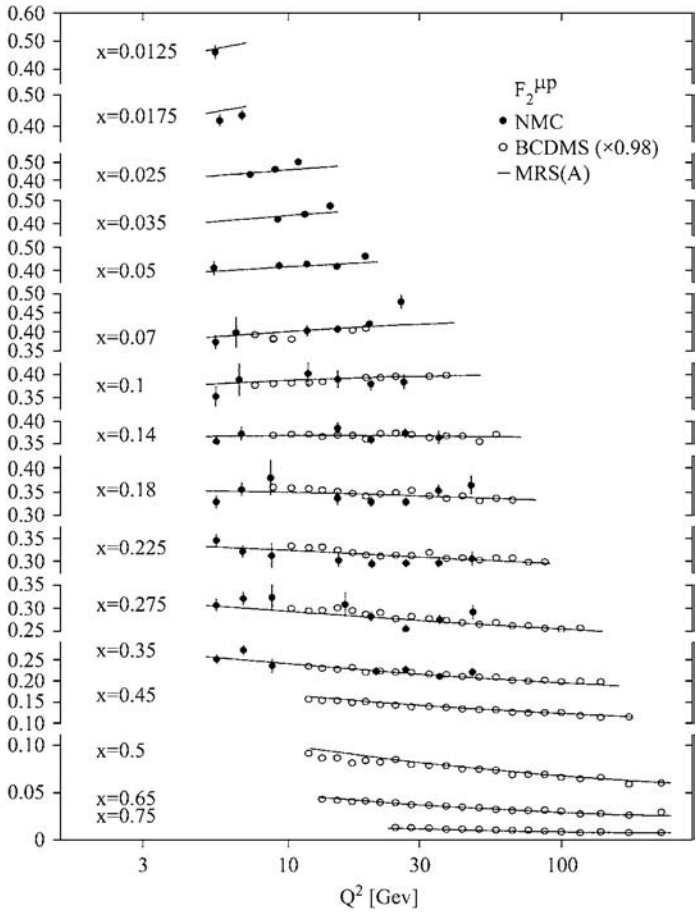


FIGURE 15.17
Data on the structure function F_2 in muon-proton deep inelastic scattering, from BCDMS (Benvenuti *et al.* 1989) and NMC (Amaudruz *et al.* 1992). The curves are QCD fits (Martin *et al.* 1994) as described in the text. Figure reprinted with permission from A D Martin *et al. Phys. Rev.* D **50** 6734 (1994). Copyright 1994 by the American Physical Society.

It may be worth pausing to reflect on how far our understanding of *structure* has developed, via quantum field theory, from the simple 'fixed number of constituents' models which are useful in atomic and nuclear physics. When nucleons are probed on finer and finer scales, more and more partons (gluons, $q\bar{q}$ pairs) appear, in a way quantitatively predicted by QCD. The precise experimental confirmation of these predictions (and many others, as discussed by Ellis, Stirling and Webber 1996, for example) constitutes a remarkable vote of confidence, by Nature, in relativistic quantum field theory.

Problems

15.1 Verify equation (15.10).

15.2 Verify equation (15.27).

15.3 Check that (15.50) can be rewritten as (15.53).

15.4 (a) Verify (15.61). (b) Show that the next term in the expansion (15.60) is

$$\frac{(b_2 - b_1^2)}{\beta_0}\alpha_s$$

where $b_2 = \beta_2/\beta_0$. By iteratively solving the resulting modified equation (15.60), show that the corresponding correction to (15.61) is

$$+\frac{1}{\beta_0^3 L^3}[b_1^2(\ln^2 L^2 - \ln L - 1) + b_2].$$

15.5 Verify that for the type of behaviour of the β function shown in figure 15.7(b), α_s^* is reached as $q^2 \to 0$.

15.6 Verify equation (15.102).

15.7 Check that the electromagnetic charge e has dimension $(\text{mass})^{\epsilon/2}$ in $d = 4 - \epsilon$ dimensions.

15.8 Verify equation (O.20) in appendix O.

15.9 Verify equation (15.118).


Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

16

Lattice Field Theory, and the Renormalization Group Revisited

16.1 Introduction

Throughout this book, thus far, we have relied on perturbation theory as the calculational tool, justifying its use in the case of QCD by the smallness of the coupling constant at short distances; note, however, that this result itself required the summation of an infinite series of perturbative terms. As remarked in section 15.3, the concomitant of asymptotic freedom is that α_s really does become strong at small Q^2, or at long distances of order $\Lambda_{\overline{\rm MS}}^{-1} \sim 1$ fm. Here we have no prospect of getting useful results from perturbation theory: it is the *non-perturbative regime*. But this is precisely the regime in which quarks bind together to form hadrons. If QCD is indeed the true theory of the interaction between quarks, then it should be able to explain, ultimately, the vast amount of data that exists in low energy hadronic physics. For example: what are the masses of mesons and baryons? Are there novel colourless states such as glueballs? Is $\mathrm{SU}(2)_\mathrm{f}$ or $\mathrm{SU}(3)_\mathrm{f}$ chiral symmetry spontaneously broken? What is the form of the effective interquark potential? What are the hadronic form factors, in electromagnetic (chapter 9) or weak (chapter 20) processes?

After more than 30 years of theoretical development, and machine advances, numerical simulations of *lattice QCD* are now yielding precise answers to many of these questions, thereby helping to establish QCD as the correct theory of the strong interactions of quarks, and also providing reliable input needed for the discovery of new physics. Lattice QCD is a highly mature field, and many technical details are beyond our scope. Rather, in this chapter we aim to give an elementary introduction to lattice field theory in general, including some important insights that it generates concerning the renormalization group. We return to QCD in the final section, with some illustrative results.

In thinking about how to formulate a non-perturbative approach to quantum field theory, several questions immediately arise. First of all, how can we regulate the ultraviolet divergences, and thus define the theory, if we cannot get to grips with them via the specific divergent integrals supplied by perturbation theory? We need to be able to regulate the divergences in a way which does not rely on their appearance in the Feynman graphs of pertur-

bation theory. As Wilson (1974, 1975) was the first to propose, one quite natural non-perturbative way of regulating ultraviolet divergences is to approximate continuous space-time by a discrete lattice of points. Such a lattice will introduce a minimum distance – namely the lattice spacing 'a' between neighbouring points. Since no two points can ever be closer than a, there is now a corresponding maximum momentum $\Lambda = \pi/a$ (see following equation (16.6)) in the lattice version of the theory. Thus the theory is automatically ultraviolet finite from the start, without presupposing the existence of any perturbative expansion; renormalization questions will, however, enter when we consider the a dependence of our parameters. As long as the lattice spacing is much smaller than the physical size of the hadrons one is studying, the lattice version of the theory should be a good approximation. Of course, Lorentz invariance is sacrificed in such an approach, and replaced by some form of hypercubic symmetry; we must hope that for small enough a this will not matter. We shall discuss how simple field theories are 'discretized' in the next section; scalar fields, fermion fields, and gauge fields each require their own prescriptions.

Next, we must ask how a discretized quantum field theory can be formulated in a way suitable for numerical computation. Any formalism based on non-commuting *operators* seems to be ruled out, since it is hard to see how they could be numerically simulated. Indeed, the same would be true of ordinary quantum mechanics. Fortunately a formulation does exist which avoids operators: Feynman's *sum over paths* approach, which was briefly mentioned in section 5.2.2. This method is the essential starting point for the lattice approach to quantum field theory, and it will be introduced in section 16.3. The sum over paths approach does not involve quantum operators, but fermions still have to be accommodated somehow. The way this is done is briefly described in section 16.3: see also appendix P.

It turns out that this formulation enables direct contact to be made between quantum field theory and *statistical mechanics*, as we shall discuss in section 16.3.3. This relationship has proved to be extremely fruitful, allowing physical insights and numerical techniques to pass from one subject to the other, in a way that has been very beneficial to both. In section 16.4 we make a worthwhile detour to explore the physics of renormalization and of the RGE from a lattice/statistical mechanics perspective, before returning to QCD in section 16.5.

16.2 Discretization

16.2.1 Scalar fields

We start by considering a simple field theory involving a scalar field ϕ. Postponing until section 16.3 the question of exactly how we shall use it, we assume

that we shall still want to formulate the theory in terms of an action of the form

$$S = \int \mathrm{d}^4x \; \mathcal{L}(\phi, \boldsymbol{\nabla}\phi, \dot{\phi}). \tag{16.1}$$

It seems plausible that it might be advantageous to treat space and time as symmetrically as possible, from the start, by formulating the theory in 'Euclidean' space, instead of Minkowskian, by introducing $t = -\mathrm{i}\tau$; further motivation for doing this will be provided in section 16.3. In that case, the action (16.1) becomes

$$S \;\; \rightarrow \;\; -\mathrm{i}\int \mathrm{d}^3\boldsymbol{x}\mathrm{d}\tau \; \mathcal{L}(\phi, \boldsymbol{\nabla}\phi, \mathrm{i}\frac{\partial\phi}{\partial\tau}) \tag{16.2}$$

$$\equiv \;\; \mathrm{i}\int \mathrm{d}^3\boldsymbol{x}\mathrm{d}\tau \; \mathcal{L}_{\mathrm{E}} \equiv \mathrm{i}S_{\mathrm{E}}. \tag{16.3}$$

A typical free scalar action is then

$$S_{\mathrm{E}}(\phi) = \frac{1}{2}\int \mathrm{d}^3\boldsymbol{x}\mathrm{d}\tau \left[(\partial_\tau\phi)^2 + (\boldsymbol{\nabla}\phi)^2 + m^2\phi^2\right]. \tag{16.4}$$

We now represent all of space-time by a finite-volume 'hypercube'. For example, we may have N_1 lattice points along the x-axis, so that a field $\phi(x)$ is replaced by the N_1 numbers $\phi(n_1 a)$ with $n_1 = 0, 1, \ldots N_1 - 1$. We write $L = N_1 a$ for the length of the cube side. In this notation, integrals and differentials are replaced by the finite sums and difference expressions

$$\int \mathrm{d}x \to a\sum_{n_1}, \qquad \frac{\partial\phi}{\partial x} \to \frac{1}{a}[\phi(n_1+1) - \phi(n_1)], \tag{16.5}$$

so that a typical integral (in one dimension) becomes

$$\int \mathrm{d}x \left(\frac{\partial\phi}{\partial x}\right)^2 \to a\sum_{n_1}\frac{1}{a^2}\left[\phi(n_1+1) - \phi(n_1)\right]^2. \tag{16.6}$$

As in all our previous work, we can alternatively consider a formulation in momentum space, which will also be discretized. It is convenient to impose periodic boundary conditions such that $\phi(x) = \phi(x + L)$. Then the allowed k-values may be taken to be $k_{\nu_1} = 2\pi\nu_1/L$ with $\nu_1 = -N_1/2+1, \ldots 0, \ldots N_1/2$ (we take N_1 to be even). It follows that the maximum allowed magnitude of the momentum is then π/a, indicating that a^{-1} is (as anticipated) playing the role of our earlier momentum cut-off Λ. We then write

$$\phi(n_1) = \sum_{\nu_1}\frac{1}{(N_1 a)^{\frac{1}{2}}}\mathrm{e}^{\mathrm{i}2\pi\nu_1 n_1/N_1}\tilde{\phi}(\nu_1), \tag{16.7}$$

which has the inverse

$$\tilde{\phi}(\nu_1) = \left(\frac{a}{N_1}\right)^{\frac{1}{2}}\sum_{n_1}\mathrm{e}^{-\mathrm{i}2\pi\nu_1 n_1/N_1}\phi(n_1), \tag{16.8}$$

since (problem 16.1)

$$\frac{1}{N_1}\sum_{n_1=0}^{N_1-1} \mathrm{e}^{\mathrm{i}2\pi n_1(\nu_1-\nu_2)/N_1} = \delta_{\nu_1,\nu_2}. \tag{16.9}$$

Equation (16.9) is a discrete version of the δ-function relation given in (E.25) of volume 1. A one-dimensional version of the mass term in (16.4) then becomes (problem 16.2)

$$\frac{1}{2}\int \mathrm{d}x\ m^2\phi(x)^2 \to \frac{1}{2}m^2\sum_{\nu_1}\tilde{\phi}(\nu_1)\tilde{\phi}(-\nu_1), \tag{16.10}$$

while

$$\frac{1}{2}\int \mathrm{d}x \left(\frac{\partial\phi}{\mathrm{d}x}\right)^2 \quad\to\quad \frac{2}{a^2}\sum_{\nu_1}\tilde{\phi}(\nu_1)\sin^2\left(\frac{\pi\nu_1}{N_1}\right)\tilde{\phi}(-\nu_1) \tag{16.11}$$

$$= \quad \frac{1}{2a^2}\sum_{k_{\nu_1}}\tilde{\phi}(k_{\nu_1})4\sin^2\left(\frac{k_{\nu_1}a}{2}\right)\tilde{\phi}(-k_{\nu_1}). \tag{16.12}$$

Thus a one-dimensional version of the free action (16.4) is

$$\frac{1}{2}\sum_{k_{\nu_1}}\tilde{\phi}(k_{\nu_1})\left[\frac{4\sin^2(k_{\nu_1}a/2)}{a^2}+m^2\right]\tilde{\phi}(-k_{\nu_1}). \tag{16.13}$$

In the continuum case, (16.13) would be replaced by

$$\frac{1}{2}\int\frac{\mathrm{d}k}{2\pi}\tilde{\phi}(k)\left[k^2+m^2\right]\tilde{\phi}(-k) \tag{16.14}$$

as usual, which implies that the propagator in the discrete case is proportional to

$$\left[\frac{4\sin^2(k_{\nu_1}a/2)}{a^2}+m^2\right]^{-1} \tag{16.15}$$

rather than to $\left[k^2+m^2\right]^{-1}$ (remember we are in one-dimensional Euclidean space). The two expressions do coincide in the continuum limit $a\to 0$. The manipulations we have been going through will be easily recognized by readers familiar with the theory of lattice vibrations and phonons, and lead to a satisfactory discretization of scalar fields. For Dirac fields the matter is not so straightforward.

16.2.2 Dirac fields

The first obvious problem has already been mentioned: how are we to represent such entirely non-classical objects, which obey anticommutation relations? This is part of the wider problem of representing field operators in

a form suitable for numerical simulation, which we defer until section 16.3. There is, however, a quite separate problem which arises when we try to repeat for the Dirac field the discretization used for the scalar field.

First note that the Euclidean Dirac matrices $\gamma_\mu^{\rm E}$ are related to the usual Minkowski ones $\gamma_\mu^{\rm M}$ by $\gamma_{1,2,3}^{\rm E} \equiv -{\rm i}\gamma_{1,2,3}^{\rm M}, \gamma_4^{\rm E} \equiv -{\rm i}\gamma_4^{\rm M} \equiv \gamma_0^{\rm M}$. They satisfy $\{\gamma_\mu^{\rm E}, \gamma_\nu^{\rm E}\} = 2\delta_{\mu\nu}$ for $\mu = 1, 2, 3, 4$. The Euclidean Dirac Lagrangian is then $\bar\psi(x)\left[\gamma_\mu^{\rm E}\partial_\mu + m\right]\psi(x)$, which should be written now in Hermitean form

$$m\bar\psi(x)\psi(x) + \frac{1}{2}\left\{\bar\psi(x)\gamma_\mu^{\rm E}\partial_\mu\psi(x) - (\partial_\mu\bar\psi(x))\gamma_\mu^{\rm E}\psi(x)\right\}. \tag{16.16}$$

The corresponding 'one-dimensional' discretized action is then

$$\begin{aligned} a\sum_{n_1} m\bar\psi(n_1)\psi(n_1) \quad &+ \quad \frac{a}{2}\left\{\sum_{n_1}\bar\psi(n_1)\gamma_1^E\left[\frac{\psi(n_1+1)-\psi(n_1)}{a}\right]\right. \\ &- \quad \left.\sum_{n_1}\left(\frac{\bar\psi(n_1+1)-\bar\psi(n_1)}{a}\right)\gamma_1^{\rm E}\psi(n_1)\right\} \end{aligned} \tag{16.17}$$

$$= a\sum_{n_1}\left\{m\bar\psi(n_1)\psi(n_1) + \frac{1}{2a}\left[\bar\psi(n_1)\gamma_1^{\rm E}\psi(n_1+1) - \bar\psi(n_1+1)\gamma_1^{\rm E}\psi(n_1)\right]\right\}. \tag{16.18}$$

In momentum space this becomes (problem 16.3)

$$\sum_{k_{\nu_1}}\bar{\tilde\psi}(k_{\nu_1})\left[{\rm i}\gamma_1^{\rm E}\,\frac{\sin(k_{\nu_1}a)}{a} + m\right]\tilde\psi(k_{\nu_1}), \tag{16.19}$$

and the inverse propagator is $\left[{\rm i}\gamma_1^{\rm E}\,\frac{\sin(k_{\nu_1}a)}{a} + m\right]$. Thus the propagator itself is

$$\left[m - {\rm i}\gamma_1^{\rm E}\,\frac{\sin(k_{\nu_1}a)}{a}\right] \Big/ \left[m^2 + \frac{\sin^2(k_{\nu_1}a)}{a^2}\right]. \tag{16.20}$$

But here there is a problem: in addition to the correct continuum limit ($a \to 0$) found at $k_{\nu_1} \to 0$, an alternative finite $a \to 0$ limit is found at $k_{\nu_1} \to \pi/a$ (consider expanding $a^{-1}\sin[(\pi/a - \delta)a]$ for small δ). Thus two modes survive as $a \to 0$, a phenomenon known as the 'fermion doubling problem'. Actually in four dimensions there are *sixteen* such corners of the hypercube, so we have far too many degenerate lattice copies (which are called different 'tastes', to distinguish them from the real quark flavours).

Various solutions to this problem have been proposed. Wilson (1975), for example, suggested adding the extra term

$$-\frac{1}{2a}r\sum_{n_1}\bar\psi(n_1)[\psi(n_1+1) + \psi(n_1-1) - 2\psi(n_1)] \tag{16.21}$$

to the fermion action in this one-dimensional case, where r is dimensionless. Evidently this is a *second* difference, and it would correspond to the term

$$-\frac{1}{2}ra\int \mathrm{d}^3\boldsymbol{x}\mathrm{d}\tau\ \bar{\psi}(x)(\partial_\tau^2+\boldsymbol{\nabla}^2)\psi(x) \tag{16.22}$$

in the four-dimensional continuum action. Note the presence of the lattice spacing 'a' in (16.22), which ensures its disappearance as $a\to 0$. The higher-derivative term $\bar{\psi}(\partial_\tau^2+\boldsymbol{\nabla}^2)\psi$ has mass dimension 5, and therefore requires a coupling constant with mass dimension -1, i.e. a length in units $\hbar=c=1$; it is, in fact, a non-renormalizable term. However, if we recall the discussion of section 11.8 in volume 1, we would expect it to be suppressed at low momenta much less than the cut-off π/a. Hence it is natural to see a coupling proportional to a appearing in (16.22). (We shall see in section 16.5.3 how renormalization group ideas provide a different perspective on such non-renormalizable interactions, classifying them as 'irrelevant').

How does the extra term (16.21) help the doubling problem? One easily finds that it changes the (one-dimensional) inverse propagator to

$$\left[\mathrm{i}\gamma_1^{\mathrm{E}}\ \frac{\sin(k_{\nu_1}a)}{a}+m\right]+\frac{r}{a}(1-\cos(k_{\nu_1}a)). \tag{16.23}$$

By considering the expansion of the cosine near $k_{\nu_1}\approx 0$ it can be seen that the second term disappears in the continuum limit, as expected. However, for $k_{\nu_1}\approx\pi/a$ it gives a large term of order $\frac{1}{a}$ which adds to the mass m, effectively banishing the 'doubled' state to a very high mass, far from the physical spectrum.

Unfortunately there is a price to pay. The problem is that, as we learned in section 12.3.2, the QCD lagrangian has an exact chiral symmetry for massless quarks. To the extent that m_{u} and m_{d} (and m_{s}, but less so) are small on a hadronic scale such as $\Lambda_{\overline{\mathrm{MS}}}$, we expect chiral symmetry to have important physical consequences. These will indeed be explored in chapter 18. For the moment, we note merely that it is important for lattice-based QCD calculations to be able to deal correctly with the light quarks. Now we cannot simply choose the bare Lagrangian mass parameters to be small, and leave it at that. In any interacting theory, renormalization effects will cause shifts in these masses. In a chirally symmetric theory, or one which is chirally symmetric as a fermion mass goes to zero, such a mass shift is proportional to the fermion mass itself; in particular it does not simply add to the mass. We drew attention to this fact in the case of the electron mass renormalization in QED, in section 11.2. So in chirally symmetric theories, mass renormalizations are 'protected', in this sense. But the modification (16.21), while avoiding physical fermion doublers, breaks chiral symmetry badly. This can easily be seen by noting (see (12.154) for example) that the crucial property required for chiral symmetry to hold is

$$\gamma_5\not{D}+\not{D}\gamma_5=0, \tag{16.24}$$

where $\not{D}$ is the $SU(3)_c$-covariant Dirac derivative. Any addition to $\not{D}$ which is proportional to the unit 4×4 matrix will violate (16.24), and hence break chiral symmetry. The Lagrangian mass m itself is of this form, and it breaks chiral symmetry, but 'softly' – i.e. in a way that disappears as m goes to zero (thereby preserving the symmetry in this limit). The Wilson addition (16.21) also breaks chiral symmetry, but it remains there even as $m \to 0$: it is a 'hard' breaking.

This means that in the theory with the Wilson modification (i.e. with 'Wilson fermions') fermion mass renormalization will not be protected by the chiral symmetry, so that large additive renormalizations are possible. This will require repeated fine-tunings of the bare mass parameters, to bring them down to the desired small values. And it turns out that this seriously lengthens the computing time.

Another approach ('staggered fermions') was suggested by Kogut and Susskind (1975), Banks *et al.* (1976), and Susskind (1977). This essentially involves distributing the 4 spin degrees of freedom of the Dirac field across different lattice sites (we shall not need the details). At each site there is now a *one*-component fermion, with the colour degrees of freedom, which speeds the calculations. The 16-fold 'doubling' degeneracy can be re-arranged as four different tastes of 4-component fermions, while retaining enough chiral symmetry to forbid additive mass renormalizations.

Since the different components of the staggered Dirac field now live on different sites, they will experience slightly different gauge field interactions. (These are of course local in the continuum limit, but the point remains true after discretization, as we shall see in the following section.) These interactions will mix fields of different tastes, causing new problems, but they can be suppressed by adding further terms to the action. There is still the 4-fold degeneracy to get rid of, but a trick is available for that, as we shall explain in section 16.3.

One might wonder if a lattice theory with fermions could be formulated such that it both avoids doublers and preserves chiral symmetry. For quite a long time it was believed that this was not possible – a conclusion which was essentially the content of the Nielsen–Ninomaya theorem (Nielsen and Ninomaya 1981a, b, c). But more recently a way was found to formulate chiral gauge theories with fermions satisfactorily on the lattice at finite spacing a. The key is to replace the condition (16.24) by the Ginsparg–Wilson (1982) relation

$$\gamma_5 \not{D} + \not{D}\gamma_5 = a \not{D}\gamma_5 \not{D} \; . \tag{16.25}$$

This relation implies (Lüscher 1998) that the associated action has an exact symmetry, with infinitesimal variations proportional to

$$\delta\psi = \gamma_5 \left(1 - \frac{1}{2} a \not{D}\right) \psi \tag{16.26}$$

$$\delta\overline{\psi} = \overline{\psi} \left(1 - \frac{1}{2} a \not{D}\right) \gamma_5 \; . \tag{16.27}$$

The symmetry under (16.26)–(16.27), which is proportional to the infinitesimal version of (12.152) as $a \to 0$, provides a lattice theory with all the fundamental symmetry properties of continuum chiral gauge theories (Hasenfratz *et al.* 1998). Finding an operator which satisfies (16.25) is, however, not so easy – but that problem has now been solved, indeed in three different ways: Kaplan's 'domain wall' fermions (Kaplan 1992); 'classically perfect fermions' (Hasenfratz and Niedermayer 1994); and overlap fermions (Narayanan and Neuberger 1993a, b, 1994, 1995). Unfortunately all these proposals are computationally more expensive than the Wilson or staggered fermion alternatives.

16.2.3 Gauge fields

Having explored the discretization of actions for free scalars and Dirac fermions, we must now think about how to implement gauge invariance on the lattice. In the usual (continuum) case, we saw in chapter 13 how this was implemented by replacing ordinary derivatives by *covariant derivatives*, the geometrical significance of which (in terms of parallel transport) is discussed in appendix N. It is very instructive to see how the same ideas arise naturally in the lattice case.

We illustrate the idea in the simple case of the Abelian U(1) theory, QED. Consider, for example, a charged scalar field $\phi(x)$, with charge e. To construct a gauge-invariant current, for example, we replaced $\phi^\dagger \partial_\mu \phi$ by $\phi^\dagger(\partial_\mu + \mathrm{i}eA_\mu)\phi$, so we ask: what is the discrete analogue of this? The term $\phi^\dagger(x)\frac{\partial}{\partial x}\phi(x)$ becomes, as we have seen,

$$\phi^\dagger(n_1)\frac{1}{a}[\phi(n_1+1) - \phi(n_1)a] \tag{16.28}$$

in one dimension. We do *not* expect (16.28) by itself to be gauge invariant, and it is easy to check that it is not. Under a gauge transformation for the continuous case, we have

$$\phi(x) \to \mathrm{e}^{\mathrm{i}e\theta(x)}\phi(x), A(x) \to A(x) + \frac{\mathrm{d}\theta(x)}{\mathrm{d}x}; \tag{16.29}$$

then $\phi^\dagger(x)\phi(y)$ transforms by

$$\phi^\dagger(x)\phi(y) \to \mathrm{e}^{-\mathrm{i}e[\theta(x)-\theta(y)]}\phi^\dagger(x)\phi(y), \tag{16.30}$$

and is clearly not invariant. The essential reason is that this operator involves the fields at two *different* points, and so the term $\phi^\dagger(n_1)\phi(n_1+1)$ in (16.28) will not be gauge invariant either. The discussion in appendix N prepares us for this: we are trying to compare two 'vectors' (here, fields) at two different points, when the 'coordinate axes' are changing as we move about. We need to parallel transport one field to the same point as the other, before they can be properly compared. The solution (N.18) shows us how to do this. Consider the quantity

$$\mathcal{O}(x,y) = \phi^\dagger(x)\exp[\mathrm{i}e\int_y^x A\mathrm{d}x']\phi(y). \tag{16.31}$$

Under the gauge transformation (16.29), $\mathcal{O}(x,y)$ transforms by

$$\mathcal{O}(x,y) \to \phi^\dagger(x)\mathrm{e}^{-\mathrm{i}e\theta(x)}\mathrm{e}^{\{\mathrm{i}e\int_y^x A\mathrm{d}x' + \mathrm{i}e[\theta(x)-\theta(y)]\}}\mathrm{e}^{\mathrm{i}e\theta(y)}\phi(y) = \mathcal{O}(x,y), \tag{16.32}$$

and it is therefore gauge invariant. The familiar 'covariant derivative' rule can be recovered by letting $y = x + \mathrm{d}x$ for infinitesimal $\mathrm{d}x$, and by considering the gauge-invariant quantity

$$\lim_{\mathrm{d}x\to 0}\left[\frac{\mathcal{O}(x,x+\mathrm{d}x)-\mathcal{O}(x,x)}{\mathrm{d}x}\right]. \tag{16.33}$$

Evaluating (16.33) one finds (problem 16.4) the result

$$\phi^\dagger(x)\left(\frac{\mathrm{d}}{\mathrm{d}x} - \mathrm{i}eA\right)\phi(x) \tag{16.34}$$

$$\equiv \phi^\dagger(x)D_x\phi(x) \tag{16.35}$$

with the usual definition of the covariant derivative. In the discrete case, we merely keep the finite version of (16.31), and replace $\phi^\dagger(n_1)\phi(n_1+1)$ in (16.28) by the gauge invariant quantity

$$\phi^\dagger(n_1)U(n_1,n_1+1)\phi(n_1+1), \tag{16.36}$$

where the *link variable* U is defined by

$$U(n_1,n_1+1) = \exp\left[\mathrm{i}e\int_{(n_1+1)a}^{n_1 a} A\mathrm{d}x'\right]. \tag{16.37}$$

Note that

$$U(n_1,n_1+1) \to \exp[-\mathrm{i}eA(n_1)a] \tag{16.38}$$

in the small a limit.

Similarly, the free Dirac term $\bar\psi(n_1)\gamma_1^{\mathrm{E}}\psi(n_1+1) - \bar\psi(n_1+1)\gamma_1^{\mathrm{E}}\psi(n_1)$ in (16.18) is replaced by the gauge-invariant term

$$\bar\psi(n_1)\gamma_1^{\mathrm{E}}U(n_1,n_1+1)\psi(n_1+1) - \bar\psi(n_1+1)\gamma_1^{\mathrm{E}}U(n_1+1,n_1)\psi(n_1). \tag{16.39}$$

The generalization to more dimensions is straightforward. In the non-Abelian $SU(2)$ or $SU(3)$ case, 'eA' in (16.38) is replaced by $gt^aA^a(n_1)$ where the t's are the appropriate matrices, as in the continuum form of the covariant derivative. A link variable $U(n_2,n_1)$ may be drawn as in figure 16.1. Note that the order of the arguments is significant: $U(n_2,n_1) = U^{-1}(n_1,n_2) = U^\dagger(n_1,n_2)$ from (16.38), which is why the link carries an arrow.

Thus gauge invariant discretized derivatives of charged fields can be constructed. What about the Maxwell action for the U(1) gauge field? This does not exist in only one dimension ($\partial_\mu A_\nu - \partial_\nu A_\mu$ cannot be formed), so let us move into two. Again, our discussion of the geometrical significance of $F_{\mu\nu}$ as

FIGURE 16.1
Link variable $U(n_2; n_1)$ in one dimension.

a curvature guides us to the answer. Consider the product $U_\Box$ of link variables around a square path (figure 16.2) of side a (reading from the right):

$$\begin{aligned} U_\Box &= U(n_x, n_y; n_x, n_{y+1})U(n_x, n_{y+1}; n_{x+1}, n_{y+1}) \\ &\quad \times U(n_{x+1}, n_{y+1}; n_{x+1}, n_y)U(n_{x+1}, n_y; n_x, n_y). \end{aligned} \tag{16.40}$$

It is straightforward to verify, first, that $U_\Box$ is gauge invariant. Under a gauge transformation, the link $U(n_{x+1}, n_y; n_x, n_y)$, for example, transforms by a factor (cf equation (16.32))

$$\exp\{\mathrm{i}e[\theta(n_{x+1}, n_y) - \theta(n_x, n_y)]\}, \tag{16.41}$$

and similarly for the three other links in $U_\Box$. In this Abelian case the exponentials contain no matrices, and the accumulated phase factors cancel out, verifying the gauge invariance. Next, let us see how to recover the Maxwell action. Adding the exponentials again, we can write

$$\begin{aligned} U_\Box &\equiv \exp\{-\mathrm{i}eaA_y(n_x, n_y) - \mathrm{i}eaA_x(n_x, n_y + 1) \\ &\quad + \mathrm{i}eaA_y(n_x + 1, n_y) + \mathrm{i}eaA_x(n_x, n_y)\} \qquad (16.42) \\ &= \exp\left\{-\mathrm{i}ea^2\left[\frac{A_x(n_x, n_y + 1) - A_x(n_x, n_y)}{a}\right]\right. \\ &\quad \left. + \mathrm{i}ea^2\left[\frac{A_y(n_x + 1, n_y) - A_y(n_x, n_y)}{a}\right]\right\} \qquad (16.43) \\ &= \exp\left\{+\mathrm{i}ea^2\left(\frac{\partial A_y}{\partial x} - \frac{\partial A_x}{\partial y}\right)\right\}, \qquad (16.44) \end{aligned}$$

using the derivative definition of (16.5). For small 'a' we may expand the exponential in (16.44). We also take the real part to remove the imaginary terms, leading to

$$\sum_\Box (1 - \mathrm{Re}\ U_\Box) \to \frac{1}{2}\sum_\Box e^2 a^4 (F_{xy})^2, \tag{16.45}$$

where $F_{xy} = \frac{\partial A_y}{\partial x} - \frac{\partial A_x}{\partial y}$ as usual. To relate this to the continuum limit we must note that we sum over each such *plaquette* with only one definite orientation, so that the sum over plaquettes is equivalent to half of the entire

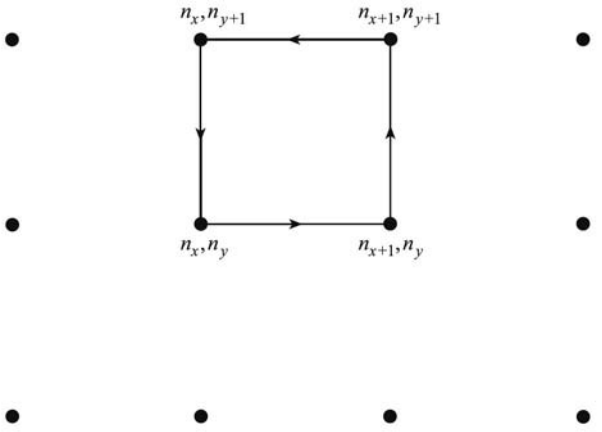


FIGURE 16.2
A simple plaquette in two dimensions.

sum. Thus

$$\sum_{\square}(1 - \mathrm{Re}\ U_{\square}) \rightarrow \frac{1}{4}\sum_{n_1,n_2} e^2 a^4 F_{xy}^2 \rightarrow e^2 a^2 \int\int \frac{1}{4} F_{xy}^2 \mathrm{d}x\mathrm{d}y. \tag{16.46}$$

(Note that in two dimensions 'e' has dimensions of mass.) In four dimensions similar manipulations lead to the form

$$S_{\mathrm{E}} = \frac{1}{e^2}\sum_{\square}(1 - \mathrm{Re}\ U_{\square}) \rightarrow \frac{1}{4}\int \mathrm{d}^3\boldsymbol{x}\mathrm{d}\tau F_{\mu\nu}^2 \tag{16.47}$$

for the lattice action, as required. In the non-Abelian case, as noted above, 'eA' is replaced by '$gt \cdot A$'; for SU(3), the analogue of (16.47) is

$$S_g = \frac{2}{g^2}\sum_{\square} \mathrm{Tr}(1 - \mathrm{Re}\ U_{\square}), \tag{16.48}$$

where the trace is over the SU(3) matrices.

16.3 Representation of quantum amplitudes

So (with some suitable fermionic action) we have a gauge-invariant 'classical' field theory defined on a lattice, with a suitable continuum limit. (Actually,

the $a \to 0$ limit of the quantum theory is, as we shall see in section 16.5, more subtle than the naive replacements (16.5) because of renormalization issues, as should be no surprise to the reader by now). However, we have not yet considered how we are going to turn this classical lattice theory into a quantum one. The fact that the calculations are mostly going to have to be done numerically seems at once to require a formulation that avoids non-commuting operators. This is precisely what is provided by Feynman's *sum over paths* formulation of quantum mechanics and of quantum field theory, and it is therefore an essential element in the lattice approach to quantum field theory. In this section we give a brief introduction to this formalism, starting with quantum mechanics.

16.3.1 Quantum mechanics

In section 5.2.2 we stated that in this approach the amplitude for a quantum system, described by a Lagrangian L depending on one degree of freedom $q(t)$, to pass from a state in which $q = q^{\mathrm{i}}$ at $t = t_{\mathrm{i}}$ to a state in which $q = q^{\mathrm{f}}$ at time $t = t_{\mathrm{f}}$, is proportional to (with $\hbar = 1$)

$$\sum_{\text{all paths } q(t)} \exp\left(\mathrm{i}\int_{t_{\mathrm{i}}}^{t_{\mathrm{f}}} L(q(t),\dot{q}(t))\mathrm{d}t\right), \tag{16.49}$$

where $q(t_{\mathrm{i}}) = q^{\mathrm{i}}$, and $q(t_{\mathrm{f}}) = q^{\mathrm{f}}$. We shall now provide some justification for this assertion.

We begin by recalling how, in ordinary quantum mechanics, state vectors and observables are related in the Schrödinger and Heisenberg pictures (see appendix I of volume 1). Let $\hat{q}$ be the canonical coordinate operator in the Schrödinger picture, with an associated complete set of eigenvectors $|q\rangle$ such that

$$\hat{q}|q\rangle = q|q\rangle \ . \tag{16.50}$$

The corresponding Heisenberg operator $\hat{q}_{\mathrm{H}}(t)$is defined by

$$\hat{q}_{\mathrm{H}}(t) = \mathrm{e}^{\mathrm{i}\hat{H}(t-t_0)}\hat{q}\mathrm{e}^{-\mathrm{i}\hat{H}(t-t_0)} \tag{16.51}$$

where $\hat{H}$ is the Hamiltonian, and t_0 is the (arbitrary) time at which the two pictures coincide. Now define the Heisenberg picture state $|q_t\rangle_{\mathrm{H}}$ by

$$|q_t\rangle_{\mathrm{H}} = \mathrm{e}^{\mathrm{i}\hat{H}(t-t_0)}|q\rangle \ . \tag{16.52}$$

We then easily obtain from (16.50)–(16.52) the result

$$\hat{q}_{\mathrm{H}}(t)|q_t\rangle_{\mathrm{H}} = q|q_t\rangle_{\mathrm{H}} \ , \tag{16.53}$$

which shows that $|q_t\rangle_{\mathrm{H}}$ is the (Heisenberg picture) state which at time t is an eigenstate of $\hat{q}_{\mathrm{H}}(t)$ with eigenvalue q. Consider now the quantity

$${}_{\mathrm{H}}\langle q^{\mathrm{f}}_{t_{\mathrm{f}}}|q^{\mathrm{i}}_{t_{\mathrm{i}}}\rangle_{\mathrm{H}} \tag{16.54}$$

which is, indeed, the amplitude for the system described by $\hat{H}$ to go from $q^{\rm i}$ at $t_{\rm i}$ to $q^{\rm f}$ at $t_{\rm f}$. Using (16.52) we can write

$$ {}_{\rm H}\langle q^{\rm f}_{t_{\rm f}}|q^{\rm i}_{t_{\rm i}}\rangle_{\rm H} = \langle q^{\rm f}|{\rm e}^{-{\rm i}\hat{H}(t_{\rm f}-t_{\rm i})}|q^{\rm i}\rangle \; ; \tag{16.55} $$

we want to understand how (16.55) can be represented as (16.49).

We shall demonstrate the connection explicitly for the special case of a free particle, for which

$$ \hat{H} = \frac{\hat{p}^2}{2m} \; . \tag{16.56} $$

For this case, we can evaluate (16.55) directly as follows. Inserting a complete set of momentum eigenstates, we obtain[1]

$$ \begin{aligned} \langle q^{\rm f}|{\rm e}^{-{\rm i}\hat{H}(t_{\rm f}-t_{\rm i})}|q^{\rm i}\rangle &= \int_{-\infty}^{\infty} \langle q^{\rm f}|p\rangle\langle p|{\rm e}^{-{\rm i}\hat{H}(t_{\rm f}-t_{\rm i})}|q^{\rm i}\rangle \, {\rm d}p \\ &= \frac{1}{2\pi}\int_{-\infty}^{\infty} {\rm e}^{{\rm i}pq^{\rm f}}{\rm e}^{-{\rm i}p^2(t_{\rm f}-t_{\rm i})/2m}{\rm e}^{-{\rm i}pq^{\rm i}} \, {\rm d}p \\ &= \frac{1}{2\pi}\int_{-\infty}^{\infty} \exp\left\{-{\rm i}\left[\frac{p^2(t_{\rm f}-t_{\rm i})}{2m} - p(q^{\rm f}-q^{\rm i})\right]\right\} {\rm d}p \; . \end{aligned} \tag{16.57} $$

To evaluate the integral, we complete the square via the steps

$$ \begin{aligned} \frac{p^2(t_{\rm f}-t_{\rm i})}{2m} - p(q^{\rm f}-q^{\rm i}) &= \left(\frac{t_{\rm f}-t_{\rm i}}{2m}\right)\left[p^2 - \frac{2mp(q^{\rm f}-q^{\rm i})}{t_{\rm f}-t_{\rm i}}\right] \\ &= \left(\frac{t_{\rm f}-t_{\rm i}}{2m}\right)\left\{\left[p - \frac{m(q^{\rm f}-q^{\rm i})}{t_{\rm f}-t_{\rm i}}\right]^2 - \frac{m^2(q^{\rm f}-q^{\rm i})^2}{(t_{\rm f}-t_{\rm i})^2}\right\} \\ &= \left(\frac{t_{\rm f}-t_{\rm i}}{2m}\right)p'^2 - \frac{m(q^{\rm f}-q^{\rm i})^2}{2(t_{\rm f}-t_{\rm i})} \; , \end{aligned} \tag{16.58} $$

where

$$ p' = p - \frac{m(q^{\rm f}-q^{\rm i})}{t_{\rm f}-t_{\rm i}} \; . \tag{16.59} $$

We then shift the integration variable in (16.57) to p', and obtain

$$ \langle q^{\rm f}|{\rm e}^{-{\rm i}\hat{H}(t_{\rm f}-t_{\rm i})}|q^{\rm i}\rangle = \frac{1}{2\pi}\exp\left[{\rm i}\frac{m(q^{\rm f}-q^{\rm i})^2}{2(t_{\rm f}-t_{\rm i})}\right]\int_{-\infty}^{\infty} {\rm d}p' \, \exp\left[-\frac{{\rm i}(t_{\rm f}-t_{\rm i})p'^2}{2m}\right] \; . \tag{16.60} $$

As it stands, the integral in (16.60) is not well-defined, being rapidly oscillatory for large p'. However, it is at this point that the motivation for passing to

[1] Remember that $\langle q|p\rangle$ is the q-space wavefunction of a state with definite momentum p, and is therefore a plane wave; we are using the normalization of equation (E.26) in volume 1.

'Euclidean' space-time arises. If we make the replacement $t \to -\mathrm{i}\tau$, (16.60) becomes

$$\langle q^{\mathrm{f}}|\mathrm{e}^{-\hat{H}(\tau_{\mathrm{f}}-\tau_{\mathrm{i}})}|q^{\mathrm{i}}\rangle = \frac{1}{2\pi}\exp\left[-\frac{m(q^{\mathrm{f}}-q^{\mathrm{i}})^2}{2(\tau_{\mathrm{f}}-\tau_{\mathrm{i}})}\right]\int_{-\infty}^{\infty}\mathrm{d}p'\,\exp\left[-\frac{(\tau_{\mathrm{f}}-\tau_{\mathrm{i}})p'^2}{2m}\right] \tag{16.61}$$

and the integral is a simple convergent Gaussian. Using the result

$$\int_{-\infty}^{\infty}\mathrm{d}\xi\mathrm{e}^{-b\xi^2} = \sqrt{\frac{\pi}{b}} \tag{16.62}$$

we finally obtain

$$\langle q^{\mathrm{f}}|\mathrm{e}^{-\hat{H}(\tau_{\mathrm{f}}-\tau_{\mathrm{i}})}|q^{\mathrm{i}}\rangle = \left[\frac{m}{2\pi(\tau_{\mathrm{f}}-\tau_{\mathrm{i}})}\right]^{\frac{1}{2}}\exp\left[-\frac{m(q^{\mathrm{f}}-q^{\mathrm{i}})^2}{2(\tau_{\mathrm{f}}-\tau_{\mathrm{i}})}\right]\,. \tag{16.63}$$

We must now understand how the result (16.63) can be represented in the form (16.49). In Euclidean space, (16.49) is

$$\sum_{\text{paths}}\exp\left(-\int_{\tau_{\mathrm{i}}}^{\tau_{\mathrm{f}}}\frac{1}{2}m\left(\frac{\mathrm{d}q}{\mathrm{d}\tau}\right)^2\mathrm{d}\tau\right) \tag{16.64}$$

in the free-particle case. We interpret the τ integral in terms of a discretization procedure, similar to that introduced in section 16.2. We split the interval $\tau_{\mathrm{f}}-\tau_{\mathrm{i}}$ into N segments each of size ϵ, as shown in figure 16.3. The τ-integral in (16.64) becomes the sum

$$m\sum_{j=1}^{N}\frac{(q^j-q^{j-1})^2}{2\epsilon}\,, \tag{16.65}$$

and the 'sum over paths', in going from $q^0 \equiv q^{\mathrm{i}}$ at τ_{i} to $q^N \equiv q^{\mathrm{f}}$ at τ_{f}, is now interpreted as a multiple integral over all the intermediate positions $q^1, q^2, \ldots, q^{N-1}$ which paths can pass through at 'times' $\tau_1, \tau_2, \ldots, \tau_{N-1}$:

$$\frac{1}{A(\epsilon)}\int\int\ldots\int\exp\left[-m\sum_{j=1}^{N}\frac{(q^j-q^{j-1})^2}{2\epsilon}\right]\frac{\mathrm{d}q^1}{A(\epsilon)}\frac{\mathrm{d}q^2}{A(\epsilon)}\ldots\frac{\mathrm{d}q^{N-1}}{A(\epsilon)}, \tag{16.66}$$

where $A(\epsilon)$ is a normalizing factor, depending on ϵ, which is to be determined.

The integrals in (16.66) are all of Gaussian form, and since the integral of a Gaussian is again a Gaussian (cf the manipulations leading from (16.57) to (16.60), but without the 'i' in the exponents), we may perform all the integrations analytically. We follow the method of Feynman and Hibbs (1965), section 3.1. Consider the integral over q^1:

$$I^1 \equiv \int\exp\left\{-\frac{m}{2\epsilon}\left[(q^2-q^1)^2+(q^1-q^{\mathrm{i}})^2\right]\right\}\,\mathrm{d}q^1\,. \tag{16.67}$$

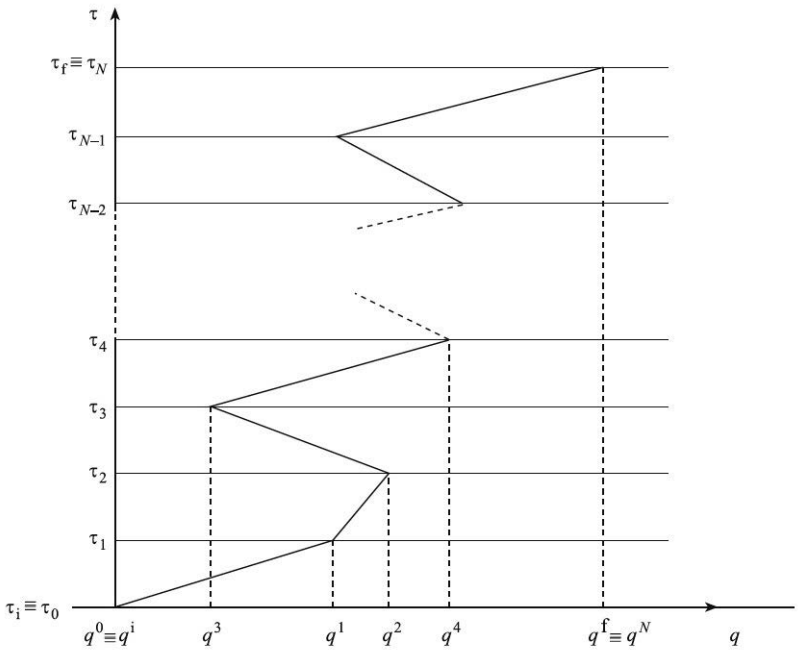


FIGURE 16.3
A 'path' from $q^0 \equiv q^{\mathrm{i}}$ at τ_{i} to $q^N \equiv q^{\mathrm{f}}$ at τ_{f}, via the intermediate positions $q^1, q^2, \ldots, q^{N-1}$ at $\tau_1, \tau_2, \ldots, \tau_{N-1}$.

This can be evaluated by completing the square, shifting the integration variable, and using (16.62), to obtain (problem 16.5)

$$I^1 = \left(\frac{\pi\epsilon}{m}\right)^{\frac{1}{2}} \exp\left[\frac{-m}{4\epsilon}(q^2 - q^{\mathrm{i}})^2\right] . \tag{16.68}$$

Now the procedure may be repeated for the q^2 integral

$$I^2 \equiv \int \exp\left\{-\frac{m}{4\epsilon}(q^2 - q^{\mathrm{i}})^2 - \frac{m}{2\epsilon}(q^3 - q^2)^2\right\} \mathrm{d}q^2 , \tag{16.69}$$

which yields (problem 16.5)

$$I^2 = \left(\frac{4\pi\epsilon}{3m}\right)^{\frac{1}{2}} \exp\left[\frac{-m}{6\epsilon}(q^3 - q^{\mathrm{i}})^2\right] . \tag{16.70}$$

As far as the exponential factors in (16.63) in (16.64) are concerned, the pattern is now clear: after $n-1$ steps we shall have an exponential factor

$$\exp\left[-m(q^n - q^{\mathrm{i}})^2/(2n\epsilon)\right] . \tag{16.71}$$

Hence, after $N-1$ steps we shall have a factor

$$\exp\left[-m(q^{\rm f}-q^{\rm i})^2/2(\tau_{\rm f}-\tau_{\rm i})\right] \ , \tag{16.72}$$

remembering that $q^N \equiv q^{\rm f}$ and that $\tau_{\rm f}-\tau_{\rm i} = N\epsilon$. So we have recovered the correct exponential factor of (16.63), and all that remains is to choose $A(\epsilon)$ in (16.66) so as to produce the same normalization as (16.63).

The required $A(\epsilon)$ is

$$A(\epsilon) = \sqrt{\frac{2\pi\epsilon}{m}} \ , \tag{16.73}$$

as we now verify. For the first (q^1) integration, the formula (16.66) contains two factors of $A^{-1}(\epsilon)$, so that the result (16.68) becomes

$$\begin{aligned} \frac{1}{\left[A(\epsilon)\right]^2} I^1 &= \frac{m}{2\pi\epsilon}\left(\frac{\pi\epsilon}{m}\right)^{\frac{1}{2}} \exp\left[-\frac{m}{4\epsilon}(q^2-q^{\rm i})^2\right] \\ &= \left(\frac{m}{2\pi 2\epsilon}\right)^{\frac{1}{2}} \exp\left[-\frac{m}{4\epsilon}(q^2-q^{\rm i})^2\right] \ . \end{aligned} \tag{16.74}$$

For the second (q^2) integration, the accumulated constant factor is

$$\frac{1}{A(\epsilon)}\left(\frac{m}{2\pi 2\epsilon}\right)^{\frac{1}{2}}\left(\frac{4\pi\epsilon}{3m}\right)^{\frac{1}{2}} = \left(\frac{m}{2\pi 3\epsilon}\right)^{\frac{1}{2}} \ . \tag{16.75}$$

Proceeding in this way, one can convince oneself that after $N-1$ steps, the accumulated constant is

$$\left(\frac{m}{2\pi N\epsilon}\right)^{\frac{1}{2}} = \left[\frac{m}{2\pi(\tau_{\rm f}-\tau_{\rm i})}\right]^{\frac{1}{2}} \ , \tag{16.76}$$

as in (16.63).

The equivalence of (16.63) and (16.64) (in the sense $\epsilon \to 0$) is therefore established for the free-particle case. More general cases are discussed in Feynman and Hibbs (1965) chapter 5, and in Peskin and Schroeder (1995) chapter 9. The conventional notation for the path-integral amplitude is

$$\langle q^{\rm f}|{\rm e}^{-\hat{H}(\tau_{\rm f}-\tau_{\rm i})}|q^{\rm i}\rangle = \int \mathcal{D}q(\tau){\rm e}^{-\int_{\tau_{\rm i}}^{\tau_{\rm f}} L\ {\rm d}\tau} \ , \tag{16.77}$$

where the right-hand side of (16.77) is interpreted in the sense of (16.66).

We now proceed to discuss further aspects of the path-integral formulation. Consider the (Euclideanized) amplitude $\langle q^{\rm f}|e^{-\hat{H}(\tau_{\rm f}-\tau_{\rm i})}|q^{\rm i}\rangle$, and insert a complete set of energy eigenstates $|n\rangle$ such that $\hat{H}|n\rangle = E_n|n\rangle$:

$$\langle q^{\rm f}|{\rm e}^{-\hat{H}(\tau_{\rm f}-\tau_{\rm i})}|q^{\rm i}\rangle = \sum_n \langle q^{\rm f}|n\rangle\langle n|q^{\rm i}\rangle {\rm e}^{-E_n(\tau_{\rm f}-\tau_{\rm i})} \ . \tag{16.78}$$

Equation (16.78) shows that if we take the limits $\tau_{\rm i} \to -\infty$, $\tau_{\rm f} \to \infty$, then the

state of lowest energy E_0 (the ground state) provides the dominant contribution. Thus, in this limit, our amplitude will represent the process in which the system begins in its ground state $|\Omega\rangle$ at $\tau_{\rm i} \to -\infty$, with $q = q^{\rm i}$, and ends in $|\Omega\rangle$ at $\tau_{\rm f} \to \infty$, with $q = q^{\rm f}$.

How do we represent propagators in this formalism? Consider the expression (somewhat analogous to a field theory propagator)

$$G_{\rm fi}(t_a, t_b) \equiv \langle q^{\rm f}_{t_{\rm f}} | T\left\{\hat{q}_{\rm H}(t_a)\hat{q}_{\rm H}(t_b)\right\} | q^{\rm i}_{t_{\rm i}} \rangle \ , \tag{16.79}$$

where T is the usual time-ordering operator. Using (16.51) and (16.52), (16.79) can be written, for $t_b > t_a$, as

$$G_{\rm fi}(t_a, t_b) = \langle q^{\rm f} | {\rm e}^{-{\rm i}\hat{H}(t_{\rm f}-t_b)} \hat{q} {\rm e}^{-{\rm i}\hat{H}(t_b-t_a)} \hat{q} {\rm e}^{-{\rm i}\hat{H}(t_a-t_{\rm i})} | q^{\rm i} \rangle \ . \tag{16.80}$$

Inserting a complete set of states and Euclideanizing, (16.80) becomes

$$\begin{aligned} G_{\rm fi}(t_a, t_b) \quad &= \quad \int {\rm d}q^a {\rm d}q^b \ q^a q^b \langle q^{\rm f} | {\rm e}^{-\hat{H}(\tau_{\rm f}-\tau_b)} | q^b \rangle \\ &\quad \times \ \langle q^b | {\rm e}^{-\hat{H}(\tau_b-\tau_a)} | q^a \rangle \langle q^a | {\rm e}^{-\hat{H}(\tau_a-\tau_{\rm i})} | q^{\rm i} \rangle \ . \end{aligned} \tag{16.81}$$

Now, each of the three matrix elements has a discretized representation of the form (16.63), with say $N_1 - 1$ variables in the interval $(\tau_a, \tau_{\rm i})$, $N_2 - 1$ in (τ_b, τ_a) and $N_3 - 1$ in $(\tau_{\rm f}, \tau_b)$. Each such representation carries one 'surplus' factor of $[A(\epsilon)]^{-1}$, making an overall factor of $[A(\epsilon)]^{-3}$. Two of these factors can be associated with the ${\rm d}q^a {\rm d}q^b$ integration in (16.81), so that we have a total of $N_1 + N_2 + N_3 - 1$ properly normalized integrations, and one 'surplus' factor $[A(\epsilon)]^{-1}$ as in (16.66). If we now identify $q(\tau_a) \equiv q^a$, $q(\tau_n) \equiv q^b$, it follows that (16.81) is simply

$$\int \mathcal{D}q(\tau) q(\tau_a) q(\tau_b) {\rm e}^{-\int_{\tau_{\rm i}}^{\tau_{\rm f}} L \, {\rm d}\tau} \ . \tag{16.82}$$

In obtaining (16.82), we took the case $\tau_b > \tau_a$. Suppose alternatively that $\tau_a > \tau_b$. Then the order of τ_a and τ_b inside the interval $(\tau_{\rm i}, \tau_{\rm f})$ is simply reversed, but since q^a and q^b in (16.81), or $q(\tau_a)$ and $q(\tau_b)$ in (16.82), are ordinary (commuting) numbers, the formula (16.82) is unaltered, and actually does represent the matrix element (16.79) of the time-ordered product.

16.3.2 Quantum field theory

The generalizations of these results to the field theory case are intuitively clear. For example, in the case of a single scalar field $\phi(\mathbf{x})$, we expect the analogue of (16.82) to be (cf (16.4))

$$\int \mathcal{D}\phi(x) \, \phi(x_a) \phi(x_b) \exp\left[- \int_{\tau_{\rm i}}^{\tau_{\rm f}} \mathcal{L}_{\rm E}(\phi, \nabla\phi, \partial_\tau \phi) \, {\rm d}^4 x_{\rm E} \right] \ , \tag{16.83}$$

where

$$\mathrm{d}^4x_\mathrm{E} = \mathrm{d}^3\boldsymbol{x}\mathrm{d}\tau, \tag{16.84}$$

and the boundary conditions are given by $\phi(\mathbf{x},\tau_\mathrm{i}) = \phi^\mathrm{i}(x)$, $\phi(\mathbf{x},\tau_\mathrm{f}) = \phi^\mathrm{f}(x)$, $\phi(\mathbf{x},\tau_a) = \phi^a(x)$ and $\phi(\mathbf{x},\tau_b) = \phi^b(x)$, say. In (16.83), we have to understand that a *four*-dimensional discretization of Euclidean space-time is implied, the fields being Fourier-analyzed by four-dimensional generalizations of expressions such as (16.7). Just as in (16.79)–(16.82), (16.83) is equal to

$$\langle\phi^\mathrm{f}(x)|\mathrm{e}^{-\hat{H}\tau_\mathrm{f}}T\left\{\hat{\phi}_\mathrm{H}(x_a)\hat{\phi}_\mathrm{H}(x_b)\right\}\mathrm{e}^{-\hat{H}\tau_\mathrm{i}}|\phi^\mathrm{i}(x)\rangle\ . \tag{16.85}$$

Taking the limits $\tau_\mathrm{i}\to-\infty$, $\tau_\mathrm{f}\to\infty$ will project out the configuration of lowest energy, as discussed after (16.78), which in this case is the (interacting) vacuum state $|\Omega\rangle$. Thus in this limit the surviving part of (16.85) is

$$\langle\phi^\mathrm{f}(x)|\Omega\rangle\mathrm{e}^{-E_\Omega\tau}\langle\Omega|T\left\{\hat{\phi}_\mathrm{H}(x_a)\hat{\phi}_\mathrm{H}(x_b)\right\}|\Omega\rangle\mathrm{e}^{-E_\Omega\tau}\langle\Omega|\phi^\mathrm{i}(x)\rangle \tag{16.86}$$

with $\tau\to\infty$. The exponential and overlap factors can be removed by dividing by the same quantity as (16.85) but without the additional fields $\phi(x_a)$ and $\phi(x_b)$. In this way, we obtain the formula for the *field theory propagator* in four-dimensional Euclidean space:

$$\langle\Omega|T\left\{\hat{\phi}_\mathrm{H}(x_a)\hat{\phi}_\mathrm{H}(x_b)\right\}|\Omega\rangle = \lim_{\tau\to\infty}\frac{\int\mathcal{D}\phi\ \phi(x_a)\phi(x_b)\exp[-\int_{-\tau}^{\tau}\mathcal{L}_\mathrm{E}\mathrm{d}^4x_\mathrm{E}]}{\int\mathcal{D}\phi\exp[-\int_{-\tau}^{\tau}\mathcal{L}_\mathrm{E}\mathrm{d}^4x_\mathrm{E}]}\ . \tag{16.87}$$

Vacuum expectation values of time-ordered products of more fields will simply have more factors of ϕ on both sides.

Perturbation theory can be developed in this formalism also. Suppose $\mathcal{L}_\mathrm{E} = \mathcal{L}_\mathrm{E}^0 + \mathcal{L}_\mathrm{E}^\mathrm{int}$, where $\mathcal{L}_\mathrm{E}^0$ describes a free scalar field and $\mathcal{L}_\mathrm{E}^\mathrm{int}$ is an interaction, for example $\lambda\phi^4$. Then, assuming λ is small, the exponential in (16.87) can be expressed as

$$\exp\left[-\int\mathrm{d}^4x_\mathrm{E}\,(\mathcal{L}_\mathrm{E}^0+\mathcal{L}_\mathrm{E}^\mathrm{int})\right] = \left(\exp-\int\mathrm{d}^4x_\mathrm{E}\,\mathcal{L}_\mathrm{E}^0\right)\left(1-\lambda\int\mathrm{d}^4x_\mathrm{E}\phi^4+\ldots\right) \tag{16.88}$$

and both numerator and denominator of (16.87) may be expressed as vevs of products of free fields. Compact techniques exist for analyzing this formulation of perturbation theory (Ryder 1985, chapter 6, Peskin & Schroeder 1995, chapter 9), and one finds exactly the same 'Feynman rules' as in the canonical (operator) approach.

In the case of gauge theories, we can easily imagine a formula similar to (16.87) for the gauge field propagator, in which the integral is carried out over all gauge fields $A_\mu(x)$ (in the $U(1)$ case, for example). But we already know from chapter 7 (or from chapter 13 in the non-Abelian case) that we shall not be able to construct a well-defined perturbation theory in this way, since the gauge field propagator will not exist unless we 'fix the gauge' by imposing

some constraint, such as the Lorentz gauge condition. Such constraints can be imposed on the corresponding path integral, and indeed this was the route followed by Faddeev and Popov (1967) in first obtaining the Feynman rules for non-Abelian gauge theories, as mentioned in section 13.5.3.

In the discrete case, the appropriate integration variables are the link variables $U(l_i)$ where l_i is the i^{th} link. They are elements of the relevant gauge group – for example $U(n_1, n_1+1)$ of (16.3.1) is an element of $U(1)$. In the case of the unitary groups, such elements typically have the form (cf (12.35)) $\sim$ exp(i Hermitean matrix), where the 'Hermitean matrix' can be parametrized in some convenient way – for example, as in (12.31) for $SU(2)$. In all these cases, the variables in the parametrization of U vary over some bounded domain (they are essentially 'angle-type' variables, as in the simple $U(1)$ case), and so, with a finite number of lattice points, the integral over the link variables is well-defined without gauge-fixing. The integration measure for the link variables can be chosen so as to be gauge invariant, and hence provided the action is gauge invariant, the formalism provides well-defined expressions, independently of perturbation theory, for vevs of gauge invariant quantities.

There remains one more conceptual problem to be addressed in this approach: namely, how are we to deal with fermions? It seems that we must introduce new variables which, though not quantum field operators, must nevertheless *anti*commute with each other. Such 'classical' anticommuting variables are called *Grassmann variables*, and are briefly described in appendix P. Further details are contained in Ryder (1985) and in Peskin and Schroeder (1995) section 9.5). For our purposes, the important point is that the fermion Lagrangian is *bilinear* in the (Grassmann) fermion fields ψ, the fermionic action for one flavour having the form

$$S_{\psi_{\mathrm{f}}} = \int \mathrm{d}^4 x_{\mathrm{E}}\, \bar{\psi}_{\mathrm{f}} M_{\mathrm{f}}(U) \psi_{\mathrm{f}}, \tag{16.89}$$

where M_{f} is a matrix representing the Dirac operator $\mathrm{i}\not{D} - m_{\mathrm{f}}$ in its discretized and Euclideanized form. This means that in a typical fermionic amplitude of the form (cf the denominator of (16.87))

$$Z_{\psi_{\mathrm{f}}} = \int \mathcal{D}\bar{\psi}_{\mathrm{f}} \mathcal{D}\psi_{\mathrm{f}} \exp[-S_{\psi_{\mathrm{f}}}]\,, \tag{16.90}$$

one has essentially an integral of Gaussian type (albeit with Grassmann variables), which can actually be performed analytically[2]. The result is simply $\det[M_{\mathrm{f}}(U)]$, the determinant of the Dirac operator matrix. For N_{f} flavours, this easily generalizes to

$$\prod_{\mathrm{f}=1}^{N_{\mathrm{f}}} \det M_{\mathrm{f}}(U). \tag{16.91}$$

[2]See appendix P.

Now we may write

$$\prod_{\mathrm{f}=1}^{N_\mathrm{f}} \det M_\mathrm{f}(U) = \exp\left[\sum_\mathrm{f} \ln\det M_\mathrm{f}(U)\right], \tag{16.92}$$

so that the effect of N_f fermions is to contribute an additional term

$$S_\mathrm{eff}(U) = -\sum_\mathrm{f} \ln\det[M_\mathrm{f}(U)] \tag{16.93}$$

to the gluonic action. But although formally correct, this fermionic contribution is computationally very time-consuming to include. Until the mid-1990s it could not be done, and instead calculations were made using the *quenched approximation*, in which the determinant is set equal to a constant independent of the link variables U. This is equivalent to the neglect of closed fermion loops in a Feynman graph approach, i.e. no vacuum polarization insertions on virtual gluon lines. Vacuum polarization amplitudes typically behave as q^2/m_f^2 for $q^2 \ll m_\mathrm{f}^2$, where q is the momentum flowing into the loop (see equation (11.39), for example, in the case of QED). The quenched approximation is therefore poorer for the light quarks u, d and s.

By the later 1990s it was possible to include the determinant provided the quark masses were not too small: the computation slowed down seriously for light quark masses. So calculations were done for unphysically large values of $m_\mathrm{u}, m_\mathrm{d}$ and m_s, and the results extrapolated towards the physical values.

Beginning in the early 2000s, however, more precise calculations with substantially lighter quark masses became possible, using the staggered fermion formulation discussed in section 16.2.2. It will be recalled that this saves a factor of four in the number of degrees of freedom. But there is still the remaining problem of the four unwanted additional 'tastes'. If these tastes are degenerate, as they would be in the continuum limit, then we can use the simple trick of replacing $S_\mathrm{eff}(U)$ by $\frac{1}{4}S_\mathrm{eff}(U)$, which means that we take the fourth root of the staggered fermion determinant. The true physical (non-degenerate) quark flavour multiplicity still remains, of course, and we arrive at

$$S_\mathrm{eff,stag.} = -\ln\det\{M_{\mathrm{stag.\ u}}(U) M_{\mathrm{stag.\ d}}(U) M_{\mathrm{stag.\ s}}(U)\}^{1/4}. \tag{16.94}$$

Unfortunately, things are not so simple away from the continuum limit, at finite lattice spacing a. Bernard, Golterman and Shamir (2006) pointed out that the quantity

$$\{\det M_\mathrm{stag.}(U)\}^{1/4} \tag{16.95}$$

cannot be represented by a local single-taste theory except in the continuum limit: at finite a, it represents a non-local single-taste action. Locality is a very fundamental property of all successful quantum field theories, and its recovery from (16.95) in the limit $a \to 0$ is not obvious. We refer to Sharpe (2006) for a full discussion, and further references. Meanwhile, as we shall see in section 16.6, some of the currently (in 2011) most accurate published results in lattice QCD are using staggered fermions with the 'rooting' procedure.

16.3.3 Connection with statistical mechanics

Not the least advantage of the path integral formulation of quantum field theory (especially in its lattice form) is that it enables a highly suggestive connection to be set up between quantum field theory and statistical mechanics. We introduce this connection now, by way of a preliminary to the discussion of renormalization in the following section.

The connection is made via the fundamental quantity of equilibrium statistical mechanics, the *partition function* Z defined by

$$Z = \sum_{\text{configurations}} \exp\left(-\frac{H}{k_{\mathrm{B}}T}\right), \tag{16.96}$$

which is simply the 'sum over states' (or configurations) of the relevant degrees of freedom, with the Boltzmann weighting factor. H is the classical Hamiltonian evaluated for each configuration. Consider, for comparison, the denominator in (16.87), namely

$$Z_\phi = \int \mathcal{D}\phi \ \exp(-S_{\mathrm{E}}), \tag{16.97}$$

where

$$S_{\mathrm{E}} = \int \mathrm{d}^4x_{\mathrm{E}}\mathcal{L}_{\mathrm{E}} = \int \mathrm{d}^4x_{\mathrm{E}} \left\{\frac{1}{2}(\partial_\tau\phi)^2 + \frac{1}{2}(\boldsymbol{\nabla}\phi)^2 + \frac{1}{2}m^2\phi^2 + \lambda\phi^4\right\} \tag{16.98}$$

in the case of a single scalar field with mass m and self-interaction $\lambda\phi^4$. The Euclideanized Lagrangian density $\mathcal{L}_{\mathrm{E}}$ is like an energy density: it is bounded from below, and increases when the field has large magnitude or has large gradients in τ or $\boldsymbol{x}$. The factor $\exp(-S_{\mathrm{E}})$ is then a sensible statistical weight for the fluctuations in ϕ, and Z_ϕ may be interpreted as the partition function for a system described by the field degree of freedom ϕ, but of course in *four* 'spatial' dimensions.

The parallel becomes perhaps even stronger when we discretize space-time. In an Ising model (see the following section), the Hamiltonian has the form

$$H = -J\sum_n s_n s_{n+1}\ , \tag{16.99}$$

where J is a constant, and the sum is over lattice sites n, the system variables taking the values ±1. When (16.99) is inserted into (16.96), we arrive at something very reminiscent of the $\phi(n_1)\phi(n_1+1)$ term in (16.6). Naturally, the effective 'Hamiltonian' is not quite the same – though we may note that Wilson (1971b) argued that in the case of a ϕ^4 interaction the parameters can be chosen so as to make the values $\phi = \pm1$ the most heavily weighted in S_{E}. Statistical mechanics does, of course, deal in three spatial dimensions, not the four of our Euclideanized space-time. Nevertheless, it is remarkable that

quantum field theory in three spatial dimensions appears to have such a close relationship to equilibrium statistical mechanics in four spatial dimensions.

One insight we may draw from this connection is that, in the case of pure gauge actions (16.47) or (16.48), the gauge coupling is seen to be analogous to an inverse temperature, by comparison with (16.96). One is led to wonder whether something like *transitions between different 'phases'* exist, as coupling constants (or other parameters) vary – and, indeed, such changes of 'phase' can occur.

A second point is somewhat related to this. In statistical mechanics, an important quantity is the *correlation length* ξ, which for a spin system may be defined via the *spin-spin correlation function*

$$G(\boldsymbol{x}) = \langle s(\boldsymbol{x}) s(\boldsymbol{0}) \rangle = \sum_{\text{all } s(\boldsymbol{x})} s(\boldsymbol{x}) s(\boldsymbol{0}) \mathrm{e}^{-H/k_{\mathrm{B}}T} \,, \tag{16.100}$$

where we are once more reverting to a continuous $\boldsymbol{x}$ variable. For large $|\boldsymbol{x}|$, this takes the form

$$G(\boldsymbol{x}) \propto \frac{1}{|\boldsymbol{x}|} \exp\left(\frac{-|\boldsymbol{x}|}{\xi(T)}\right) \,. \tag{16.101}$$

The Fourier transform of this (in the continuum limit) is

$$\tilde{G}(\boldsymbol{k}^2) \propto \left(\boldsymbol{k}^2 + \xi^{-2}(T)\right)^{-1} \,, \tag{16.102}$$

as we learned in section 1.3.3. Comparing (16.100) with (16.87), it is clear that (16.100) is proportional to the propagator (or Green function) for the field $s(\boldsymbol{x})$; (16.102) then shows that $\xi^{-1}(T)$ is playing the role of a mass term m. Now, near a critical point for a statistical system, correlations exist over very large scales ξ compared to the inter-atomic spacing a; in fact, at the critical point $\xi(T_{\mathrm{c}}) \sim L$, where L is the size of the system. In the quantum field theory, as indicated earlier, we may regard a^{-1} as playing a role analogous to a momentum cut-off Λ, so the regime $\xi \gg a$ is equivalent to $m \ll \Lambda$, as was indeed always our assumption. Thus studying a quantum field theory this way is analogous to studying a four-dimensional statistical system near a critical point. This shows rather clearly why it is not going to be easy: correlations over all scales will have to be included. At this point, we are naturally led to the consideration of *renormalization* in the lattice formulation.

16.4 Renormalization, and the renormalization group, on the lattice

16.4.1 Introduction

In the continuum formulation which we have used elsewhere in this book, fluctuations over short distances of order Λ^{-1} generally lead to divergences

in the limit $\Lambda \to \infty$, which are controlled (in a renormalizable theory) by the procedure of renormalization. Such divergent fluctuations turn out, in fact, to affect a renormalizable theory only through the values of some of its parameters and, if these parameters are taken from experiment, all other quantities become finite, even as $\Lambda \to \infty$. This latter assertion is not easy to prove, and indeed is quite surprising. However, this is by no means all there is to renormalization theory: we have seen the power of 'renormalization group' ideas in making testable predictions for QCD. Nevertheless, the methods of chapter 15 were rather formal, and the reader may well feel the need of a more physical picture of what is going on. Such a picture was provided by Wilson (1971a) (see also Wilson and Kogut 1974), using the 'lattice + path integral' approach. Another important advantage of this formalism is, therefore, precisely the way in which, thanks to Wilson's work, it provides access to a more intuitive way of understanding renormalization theory. The aim of this section is to give a brief introduction to Wilson's ideas, so as to illuminate the formal treatment of the previous chapter.

In the 'lattice + path integral' approach to quantum field theory, the degrees of freedom involved are the values of the field(s) at each lattice site, as we have seen. Quantum amplitudes are formed by integrating suitable quantities over all values of these degrees of freedom, as in (16.87) for example. From this point of view, it should be possible to examine specifically how the 'short distance' or 'high momentum' degrees of freedom affect the result. In fact, the idea suggests itself that we might be able to perform explicitly the integration (or summation) over those degrees of freedom located near the cutoff Λ in momentum space, or separated by only a lattice site or two in co-ordinate space. If we can do this, the result may be compared with the theory as originally formulated, to see how this 'integration over short-distance degrees of freedom' affects the physical predictions of the theory. Having done this once, we can imagine doing it again – and indeed *iterating* the process, until eventually we arrive at some kind of 'effective theory' describing physics in terms of 'long-distance' degrees of freedom.

There are several aspects of such a programme which invite comment. First, the process of 'integrating out' short-distance degrees of freedom will obviously *reduce* the number of effective degrees of freedom, which is necessarily very large in the case $\xi \gg a$, as envisaged above. Thus it must be a step in the right direction. Secondly, the above sketch of the 'integrating out' procedure suggests that, at any given stage of the integration, we shall be considering the system as described by parameters (including masses and couplings) *appropriate to that scale*, which is of course strongly reminiscent of RGE ideas. And thirdly, we may perhaps anticipate that the result of this 'integrating out' will be not only to render the parameters of the theory scale-dependent, but also, in general, to introduce new kinds of *effective interactions* into the theory. We now consider some simple examples which we hope will illustrate these points.

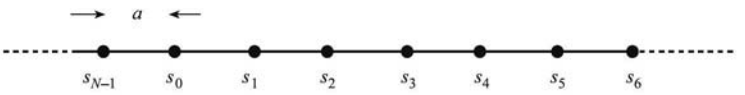


FIGURE 16.4
A portion of the one-dimensional lattice of spins in the Ising model.

16.4.2 Two one-dimensional examples

Consider first a simple one-dimensional Ising model with Hamiltonian (16.99) and partition function

$$Z = \sum_{\{s_n\}} \exp\left[K \sum_{n=0}^{N-1} s_n s_{n+1}\right], \tag{16.103}$$

where $K = J/(k_{\mathrm{B}}T) > 0$. In (16.103) all the s_n variables take the values ± 1 and the 'sum over $\{s_n\}$' means that all possible configurations of the N variables $s_0, s_1, s_2, \ldots, s_{N-1}$ are to be included. The spin s_n is located at the lattice site na, and we shall (implicitly) be assuming the periodic boundary condition $s_n = s_{N+n}$. Figure 16.4 shows a portion of the one-dimensional lattice with the spins on the sites, each site being separated by the lattice constant a. Thus, for the portion $\{s_{N-1}, s_0, \ldots s_4\}$ we are evaluating

$$\sum_{s_{N-1},s_0,s_1,s_2,s_3,s_4} \exp[K(s_{N-1}s_0 + s_0s_1 + s_1s_2 + s_2s_3 + s_3s_4)]\,. \tag{16.104}$$

Now suppose we want to describe the system in terms of a 'coarser' lattice, with lattice spacing $2a$, and corresponding new spin variables s'_n. There are many ways we could choose to describe the s'_n, but here we shall only consider a very simple one (Kadanoff 1977) in which each s'_n is simply identified with the s_n at the corresponding site (see figure 16.5). For the portion of the lattice under consideration, then, (16.104) becomes

$$\sum_{s_{N-1},s'_0,s_1,s'_1,s_3,s'_2} \exp\left[K\left(s_{N-1}s'_0 + s'_0s_1 + s_1s'_1 + s'_1s_3 + s_3s'_2\right)\right]. \tag{16.105}$$

If we can now perform the sums over s_1 and s_3 in (16.105), we shall end up (for this portion) with an expression involving the 'effective' spin variables s'_0, s'_1 and s'_2, situated twice as far apart as the original ones, and therefore providing a more 'coarse grained' description of the system. Summing over s_1 and s_3 corresponds to 'integrating out' two short-distance degrees of freedom as discussed earlier.

In fact, these sums are easy to do. Consider the quantity $\exp(Ks'_0s_1)$, expanded as a power series:

$$\exp(Ks'_0s_1) = 1 + Ks'_0s_1 + \frac{K^2}{2!} + \frac{K^3}{3!}(s'_0s_1) + \ldots \tag{16.106}$$

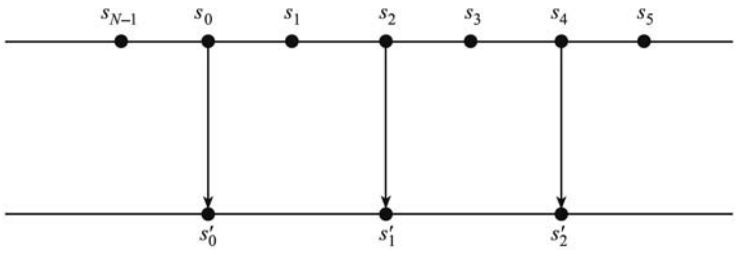


FIGURE 16.5
A 'coarsening' transformation applied to the lattice portion shown in figure 16.4. The new (primed) spin variables are situated twice as far apart as the original (unprimed) ones.

where we have used $(s'_0 s_1)^2 = 1$. It follows that

$$\exp(K s'_0 s_1) = \cosh K \, (1 + s'_0 s_1 \tanh K) \,, \tag{16.107}$$

and similarly

$$\exp(K s_1 s'_1) = \cosh K \, (1 + s_1 s'_1 \tanh K) \,. \tag{16.108}$$

Thus the sum over s_1 is

$$\sum_{s_1 = \pm 1} \cosh^2 K \left(1 + s'_0 s_1 \tanh K + s_1 s'_1 \tanh K + s'_0 s'_1 \tanh^2 K\right) . \tag{16.109}$$

Clearly, the terms linear in s_1 vanish after summing, and the s_1 sum becomes just

$$2 \cosh^2 K \left(1 + s'_0 s'_1 \tanh^2 K\right) . \tag{16.110}$$

Remarkably, (16.110) contains a new 'nearest-neighbour' interaction, $s'_0 s'_1$, just like the original one in (16.103), but with an *altered coupling* (and a different spin-independent piece). In fact, we can write (16.110) in the standard form

$$\exp \left[g_1(K) + K' s'_0 s'_1 \right] \tag{16.111}$$

and then use (16.107) to set

$$\tanh K' = \tanh^2 K \tag{16.112}$$

and identify

$$g_1(K) = \ln \left(\frac{2 \cosh^2 K}{\cosh K'} \right) . \tag{16.113}$$

Exactly the same steps can be followed through for the sum on s_3 in (16.105), and indeed for *all* the sums over the 'integrated out' spins. The upshot is that, apart from the accumulated spin-independent part, the new partition

function, defined on a lattice of size $2a$, has the same form as the old one, but with a new coupling K' related to the old one K by (16.112).

Equation (16.112) is an example of a *renormalization transformation*: the number of degrees of freedom has been halved, the lattice spacing has doubled, and the coupling K has been renormalized to K'.

It is clear that we could apply the same procedure to the new Hamiltonian, introducing a coupling K'' which is related to K' , and thence to K by

$$\tanh K'' = (\tanh K')^2 = (\tanh K)^4. \tag{16.114}$$

This is equivalent to *iterating* the renormalization transformation; after n iterations, the effective lattice constant is $2^n a$, and the effective coupling is given by

$$\tanh K^{(n)} = (\tanh K)^n. \tag{16.115}$$

The successive values $K', K'', \ldots$ of the coupling under these iterations can be regarded as a '*flow*' in the (one-dimensional) space of K-values: a *renormalization flow*.

Of particular interest is a point (or points) K^* such that

$$\tanh K^* = \tanh^2 K^*. \tag{16.116}$$

This is called a *fixed point* of the renormalization tranformation. At such a point in K-space, changing the scale by a factor of 2 (or 2^n for that matter) will make no difference, which means that the system must be in some sense ordered. Remembering that $K = J/(k_{\mathrm{B}}T)$, we see that $K = K^*$ when the temperature is 'tuned' to the value $T = T^* = J/(k_{\mathrm{B}}K^*)$. Such a T^* would be the temperature of a *critical point* for the thermodynamics of the system, corresponding to the onset of ordering. In the present case, the only fixed points are $K^* = \infty$ and $K^* = 0$. Thus there is no critical point at a non-zero T^*, and hence no transition to an ordered phase. However, we may describe the behaviour as $T \to 0$ as 'quasi-critical'. For large K, we may use

$$\tanh K \simeq 1 - 2\mathrm{e}^{-2K} \tag{16.117}$$

to write (16.115) as

$$K^{(n)} = K - \frac{1}{2}\ln n, \tag{16.118}$$

which shows that K^n changes only very slowly (logarithmically) under iterations when in the vicinity of a very large value of K, so that this is 'almost' a fixed point.

We may represent the flow of K, under the renormalization transformation (16.115), as in figure 16.6. Note that the flow is away from the quasi-fixed point at $K^* = \infty$ ($T = 0$) and towards the (non-interacting) fixed point at $K^* = 0$.

A renormalization transformation which has a fixed point at a finite (neither zero nor infinite) value of the coupling is clearly of greater interest, since

FIGURE 16.6
'Renormalization flow': the arrows show the direction of flow of the coupling K as the lattice constant is increased. The starred values are fixed points.

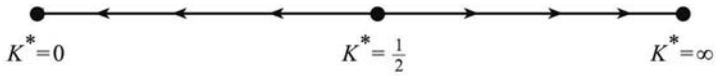


FIGURE 16.7
The renormalization flow for the transformation (16.120).

this will correspond to a critical point at a finite temperature. A simple such example given by Kadanoff (1977) is the transformation

$$K' = \frac{1}{2}(2K)^2 \tag{16.119}$$

for a doubling of the effective lattice size, or

$$K^{(n)} = \frac{1}{2}(2K)^n \tag{16.120}$$

for n such iterations. The model leading to (16.120) involves fermions in one dimension, but the details are irrelevant to our purpose here. The renormalization transformation (16.120) has three fixed points: $K^* = 0$, $K^* = \infty$ and the finite point $K^* = \frac{1}{2}$. The renormalization flow is shown in figure 16.7.

The striking feature of this flow is that the motion is always away from the finite fixed point, under successive iterations. This may be understood by recalling that at the fixed point (which is a critical point for the statistical system) the correlation length ξ must be infinite (as $L \to \infty$). As we iterate away from this point, ξ decreases and we leave the fixed (or critical) point. For this model, ξ is given by Kadanoff (1977) as

$$\xi = \frac{a}{|\ln 2K|} \tag{16.121}$$

which indeed goes to infinity at $K = \frac{1}{2}$.

16.4.3 Connections with particle physics

Let us now begin to think about how all this may relate to the treatment of the renormalization group in particle physics, as given in the previous chapter.

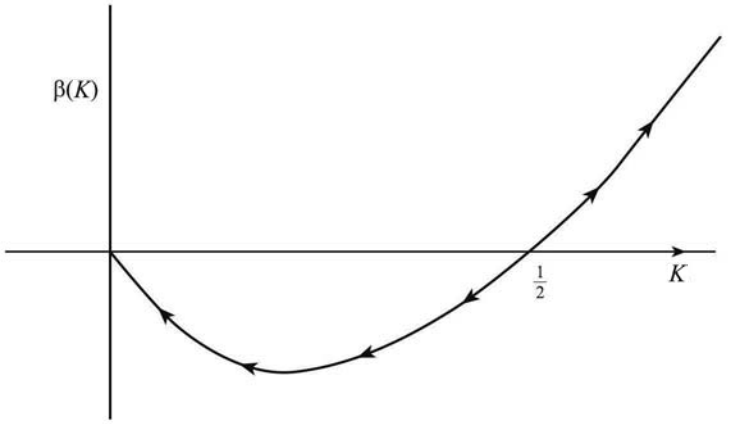


FIGURE 16.8
The β-function of (16.124); the arrows indicate increasing f.

First, we need to consider a continuous change of scale, say by a factor of f. In the present model, the transformation (16.120) then becomes

$$K(fa) = \frac{1}{2}(2K(a))^f. \tag{16.122}$$

Differentiating (16.122) with respect to f, we find

$$f\frac{\mathrm{d}K(fa)}{\mathrm{d}f} = K(fa)\ln\left[2K(fa)\right]. \tag{16.123}$$

We may reasonably call (16.123) a renormalization group equation, describing the 'running' of $K(fa)$ with the scale f, analogous to the RGE's for α and α_s considered in chapter 15. In this case, the β-function is

$$\beta(K) = K\ln(2K), \tag{16.124}$$

which is sketched in figure 16.8. The zero of β is indeed at the fixed (critical) point $K = \frac{1}{2}$, and this is an infrared unstable fixed point, the flow being away from it as f increases.

The foregoing is exactly analogous to the discussion in section 15.5: see in particular figure 15.6 and the related discussion. Note, however, that in the present case we are considering rescalings in *position* space, not momentum space. Since momenta are measured in units of a^{-1}, it is clear that scaling a by f is the same as scaling k by $f^{-1} = t$, say. This will produce a change in sign in $\mathrm{d}K/\mathrm{d}t$ relative to $\mathrm{d}K/\mathrm{d}f$, and accounts for the fact that $K = \frac{1}{2}$ is an infrared unstable fixed point in figure 16.8, while α_s^* is an infrared stable fixed point in figure 15.6(b). Allowing for the change in sign, figure 16.8 is quite analogous to figure 15.6(a).

We have emphasized that, at a critical point, and in the continuum limit, the correlation length $\xi \to \infty$, or equivalently the mass parameter (cf (16.102)) $m = \xi^{-1} \to 0$. In this case, the Fourier transform of the spin-spin correlation function should behave as

$$\tilde{G}(\mathbf{k}^2) \propto \frac{1}{\mathbf{k}^2}. \tag{16.125}$$

This is indeed the $\mathbf{k}^2$-dependence of the propagator of a free, massless scalar particle, but – as we learned for the fermion propagator in section 15.5 – it is no longer true in an interacting theory. In the interacting case, (16.125) generally becomes modified to

$$\tilde{G}(\mathbf{k}^2) \propto \frac{1}{(\mathbf{k}^2)^{1-\frac{\eta}{2}}}, \tag{16.126}$$

or equivalently

$$G(\mathbf{x}) \propto \frac{1}{|\mathbf{x}|^{1+\eta}} \tag{16.127}$$

in three spatial dimensions, and in the continuum limit. Thus, at a critical point, the spin-spin correlation function exhibits scaling under the transformation $\mathbf{x}' = f\mathbf{x}$, but it is not free-field scaling. Comparing (16.126) with (15.75), we see that $\eta/2$ is precisely the *anomalous dimension* of the field $s(\mathbf{x})$, so – just as in section 15.5 – we have an example of scaling with anomalous dimension. In the statistical mechanics case, η is a *critical exponent*, one of a number of such quantities characterizing the critical behaviour of a system. In general, η will depend on the coupling constant $\eta(K)$: at a non-trivial fixed point, η will be evaluated at the fixed point value K^*, $\eta(K^*)$. Enormous progress was made in the theory of critical phenomena when the powerful methods of quantum field theory were applied to calculate critical exponents (see for example Peskin & Schroeder 1995, chapter 13, and Binney *et al.* 1992).

In our discussion so far, we have only considered simple models with just one 'coupling constant', so that diagrams of renormalization flow were one-dimensional. Generally, of course, Hamiltonians will consist of several terms, and the behaviour of all their coefficients will need to be considered under a renormalization transformation. The general analysis of renormalization flow in multi-dimensional coupling space was given by Wegner (1972). In simple terms, the coefficients show one of three types of behaviour under renormalization transformations such that $a \to fa$, characterized by their behaviour in the vicinity of a fixed point: (i) the difference from the fixed point value grows as f increases, so that the system moves away from the fixed point (as in the single-coupling examples considered earlier); (ii) the difference decreases as f increases, so the system moves towards the fixed point; (iii) there is no change in the value of the coupling as f changes. The corresponding coefficients are called, respectively, (i) *relevant*, (ii) *irrelevant* and (iii) *marginal* couplings; the terminology is also frequently applied to the operators in the Hamiltonians

themselves. The intuitive meaning of 'irrelevant' is clear enough: the system will head towards a fixed point as $f \to \infty$ whatever the initial values of the irrelevant couplings. The critical behaviour of the system will therefore be independent of the number and type of all irrelevant couplings, and will be determined by the relatively few (in general) marginal and relevant couplings. Thus *all* systems which flow close to the fixed point will display the same critical exponents determined by the dynamics of these few couplings. This explains the property of *universality* observed in the physics of phase transitions, whereby many apparently quite different physical systems are described (in the vicinity of their critical points) by the same critical exponents.

Additional terms in the Hamiltonian are, in fact, generally introduced following a renormalization transformation. In the quantum field case, we may expect that renormalization transformations associated with $a \to fa$, and iterations thereof, will in general lead to an effective theory involving all possible couplings allowed by whatever symmetries are assumed to be relevant. Thus, if we start with a typical 'ϕ^4' scalar theory as given by (16.98), we shall expect to generate all possible couplings involving ϕ and its derivatives. At first sight, this may seem disturbing: after all, the original theory (in four dimensions) is a renormalizable one, but an interaction such as $A\phi^6$ is *not* renormalizable according to the criterion given in section 11.8 (in four dimensions ϕ has mass dimension unity, so that A must have mass dimension -2). It is, however, essential to remember that in this 'Wilsonian' approach to renormalization, summations over momenta appearing in loops do not, after one iteration $a \to fa$, run up to the original cut-off value π/a, but only up to the lower cut-off π/fa. The additional interactions compensate for this change.

In fact, we shall now see how the coefficients of non-renormalizable interactions correspond precisely to *irrelevant* couplings in Wilson's approach, so that their effect becomes negligible as we iterate to scales much larger than a. We consider continuous changes of scale characterized by a factor f, and we discuss a theory with only a single scalar field ϕ for simplicity. Imagine, therefore, that we have integrated out, in (16.97), those components of $\phi(\mathbf{x})$ with $a < |\mathbf{x}| < fa$. We will be left with a functional integral of the form (16.97), but with $\phi(\mathbf{x})$ restricted to $|\mathbf{x}| > fa$, and with additional interaction terms in the action. In order to interpret the result in Wilson's terms, we must rewrite it so that it has the same general form as the original Z_ϕ of (16.97). A simple way to do this is to rescale distances by

$$\mathbf{x}' = \frac{\mathbf{x}}{f} \tag{16.128}$$

so that the functional integral is now over $\phi(\mathbf{x}')$ with $|\mathbf{x}'| > a$, as in (16.97). We now *define* the fixed point of the renormalization transformation to be that in which all the terms in the action are *zero*, except the 'kinetic' piece; this is the 'free-field' fixed point. Thus, we require the kinetic action to be

unchanged:

$$\begin{aligned}\int \mathrm{d}^4 x_{\mathrm{E}}\ (\partial_\mu \phi)^2 &= \int \mathrm{d}^4 x'_{\mathrm{E}} (\partial'_\mu \phi')^2 \\ &= \int \frac{1}{f^2} \mathrm{d}^4 x_{\mathrm{E}} (\partial_\mu \phi')^2 \ , \end{aligned} \tag{16.129}$$

from which it follows that $\phi' = f\phi$. Consider now a term of the form $A\phi^6$:

$$A \int \mathrm{d}^4 x_{\mathrm{E}}\ \phi^6 = \frac{A}{f^2} \int \mathrm{d}^4 x'_{\mathrm{E}}\ \phi'^6 \ . \tag{16.130}$$

(16.130) shows that the 'new' A' is related to the old one by $A' = \frac{A}{f^2}$, and in particular that, as f increases, A' decreases and is therefore an *irrelevant* coupling, tending to zero as we reach large scales. But such an interaction is precisely a non-renormalizable one (in four dimensions), according to the criterion of section 11.8. The mass dimension of ϕ is unity, and hence that of A must be -2 so that the action is dimensionless; couplings with negative mass dimensions correspond to non-renormalizable interactions. The reader may verify the generality of this result for any interaction with p powers of ϕ, and q derivatives of ϕ.

However, the mass term $m^2\phi^2$ behaves differently:

$$m^2 \int \mathrm{d}^4 x_{\mathrm{E}}\ \phi^2 = m^2 f^2 \int \mathrm{d}^4 x'_{\mathrm{E}}\ \phi'^2 \tag{16.131}$$

showing that $m'^2 = m^2 f^2$ and the 'coupling' m^2 is *relevant*, since it grows with f^2. Such a term has positive mass dimension, and corresponds to a 'super-renormalizable' interaction. Finally, the $\lambda\phi^4$ interaction transforms as

$$\lambda \int \mathrm{d}^4 x_{\mathrm{E}}\ \phi^4 = \lambda \int \mathrm{d}^4 x'_{\mathrm{E}}\ \phi'^4 \tag{16.132}$$

and so $\lambda' = \lambda$. The coupling is *marginal*, which may correspond (though not necessarily) to a renormalizable interaction. To find out if such couplings increase or decrease with f, we have to include higher-order loop corrections. The foregoing analysis in terms of the suppression of non-renormalizable interactions by powers of f^{-1} parallels precisely the similar one in section 11.8. We saw that such terms were suppressed at low energies by factors of E/Λ, where Λ is the cut-off scale beyond which the theory is supposed to fail on physical grounds (e.g. Λ might be the Planck mass). The result is that as we renormalize, in Wilson's sense, down to much lower energy scales, the non-renormalizable terms disappear and we are left with an effective renormalizable theory. This is the field theory analogue of 'universality'.

These ideas have an important application in lattice QCD. One of the reasons for systematic inaccuracies in lattice computations is that the continuum is being simulated by a lattice of finite spacing. Symanzik (1983) showed

that corrections to continuum theory results stemming from finite lattice spacing could be diminished systematically by the use of lattice actions that also include suitable irrelevant terms. This procedure is routinely adopted in accurate lattice calculations with 'Symanzik-improved' actions.

One further word should be said about terms such as '$m^2\phi^2$' (which arise in the Higgs sector of the Standard Model, for instance). As we have seen, m^2 scales by $m'^2 = m^2 f^2$, which is a rapid growth with f. If we imagine starting at a very high scale, such as 10^{15} TeV and flowing down to 1 TeV, then the 'initial' value of m will have to be very finely 'tuned' in order to end up with a mass of order 1 TeV. Thus, in this picture, it seems unnatural to have scalar particles with masses much less than the physical cut-off scale, unless some symmetry principle 'protects' their light masses. We shall return to this problem in section 22.8.1.

We now return to lattice QCD, with a brief survey of some of the impressive results now being obtained numerically.

16.5 Lattice QCD

16.5.1 Introduction, and the continuum limit

Let us begin by considering some numbers. The lattice must be large enough so that the spatial dimension R of the object we wish to describe – say the size of a hadron – fits comfortably inside it, otherwise the result will be subject to 'finite size effects' as the hypercube side length L is varied. We also need $R \gg a$, or else the granularity of the lattice resolution will become apparent. Further, as indicated earlier, we expect the mass m (which is of order R^{-1}) to be very much less than a^{-1}. Thus ideally we need

$$a \ll R \sim 1/m \ll L = Na \tag{16.133}$$

so that N must be large. For example, if $N = 64$ and $a \sim 0.1$fm the condition (16.133) would be reasonably satisfied by a light hadron mass. But remember that each field at each lattice point is an independent degree of freedom: dealing with integrals such as (16.87) presents a formidable numerical challenge.

Ignoring any statistical inaccuracy, the results will depend on the parameters $g_{\rm L}$ and N, where $g_{\rm L}$ is the bare lattice gauge coupling (we assume for simplicity that the quarks are massless). Despite the fact that $g_{\rm L}$ is dimensionless, we shall now see that its value actually controls the physical size of the lattice spacing a, as a result of renormalization effects. The computed mass of a hadron M, say, must be related to the only quantity with mass dimension, a^{-1}, by a relation of the form

$$M = \frac{1}{a} f(g_{\rm L}). \tag{16.134}$$

Thus in approaching the continuum limit $a \to 0$, we shall also have to change

$g_{\rm L}$ suitably, so as to ensure that M remains finite. This is, of course, quite analogous to saying that, in a renormalizable theory, the bare parameters of the theory depend on the momentum cut-off Λ in such a way that, as $\Lambda \to \infty$, finite values are obtained for the corresponding physical parameters (see the last paragraph of section 10.1.2, for example). In practice, of course, the extent to which the lattice 'a' can really be taken to be very small is severely limited by the computational resources available – that is, essentially, by the number of mesh points N.

Equation (16.134) should therefore really read

$$M = \frac{1}{a} f\left(g_{\rm L}(a)\right) \ . \tag{16.135}$$

As $a \to 0$, M should be finite and independent of a. However, we know that the behaviour of $g_{\rm L}(a)$ at small scales is in fact calculable in perturbation theory, thanks to the asymptotic freedom of QCD. This will allow us to determine the form of $f(g_{\rm L})$, up to a constant, and lead to an interesting prediction for M (equations (16.141)–(16.142)).

Differentiating (16.135) we find

$$0 = \frac{{\rm d}M}{{\rm d}a} = -\frac{1}{a^2} f\left(g_{\rm L}(a)\right) + \frac{1}{a}\frac{{\rm d}f}{{\rm d}g_{\rm L}}\frac{{\rm d}g_{\rm L}(a)}{{\rm d}a} \ , \tag{16.136}$$

so that

$$\left(a\frac{dg_{\rm L}(a)}{{\rm d}a}\right)\frac{{\rm d}f}{{\rm d}g_{\rm L}} = f\left(g_{\rm L}(a)\right) \ . \tag{16.137}$$

Meanwhile, the scale dependence of $g_{\rm L}$ is given (to one loop order) by

$$a\frac{{\rm d}g_{\rm L}(a)}{{\rm d}a} = \frac{\beta_0}{4\pi} g_{\rm L}^3(a) \ , \tag{16.138}$$

where the sign is the opposite of (15.47) since $a \sim \mu^{-1}$ is the relevant scale parameter here (compare the comments after equation (16.124)). The integration of (16.138) requires, as usual, a dimensionful constant of integration (cf (15.53)):

$$\frac{g_{\rm L}^2(a)}{4\pi} = \frac{1}{\beta_0 \ln(1/a^2\Lambda_{\rm L}^2)} \ . \tag{16.139}$$

Equation (16.139) shows that $g_{\rm L}(a)$ tends logarithmically to zero as $a \to 0$, as we expect from asymptotic freedom. $\Lambda_{\rm L}$ can be regarded as a lattice equivalent of the continuum $\Lambda_{\overline{\rm MS}}$, and it is defined (at one loop order) by

$$\Lambda_{\rm L} \equiv \lim_{g_{\rm L}\to 0} \frac{1}{a} \exp\left(-\frac{2\pi}{\beta_0 g_{\rm L}^2}\right) \ . \tag{16.140}$$

Equation (16.140) may also be read as showing that the lattice spacing a must go exponentially to zero as $g_{\rm L}$ tends to zero. Higher-order corrections can of course be included.

In a similar way, integrating (16.137) using (16.138) gives, in (16.134),

$$M = \text{constant} \times \left[\frac{1}{a}\exp\left(-\frac{2\pi}{\beta_0 g_{\rm L}^2}\right)\right] \tag{16.141}$$

$$= \text{constant} \times \Lambda_{\rm L} \,. \tag{16.142}$$

Equation (16.141) is known as *asymptotic scaling*: it predicts how any physical mass, expressed in lattice units a^{-1}, should vary as a function of $g_{\rm L}$. The form (16.142) is remarkable, as it implies that all calculated masses must be proportional, in the continuum limit $a \to 0$, to the same universal scale factor $\Lambda_{\rm L}$.

How are masses calculated on the lattice? The principle is very similar to the way in which the ground state was selected out as $\tau_{\rm i} \to -\infty$, $\tau_{\rm f} \to +\infty$ in (16.78). Consider a correlation function for a scalar field, for simplicity:

$$C(\tau) = \langle\Omega|\phi(\mathbf{x}=0,\tau)\phi(0)|\Omega\rangle$$

$$= \sum_n |\langle\Omega|\phi(0)|n\rangle|^2 \, {\rm e}^{-E_n\tau} \,. \tag{16.143}$$

As $\tau \to \infty$, the term with the minimum value of E_n, namely $E_n = M_\phi$, will survive; M_ϕ can be measured from a fit to the exponential fall-off as a function of τ.

The behaviour predicted by (16.141) and (16.142) can be tested in actual calculations. A quantity such as the ρ meson mass is calculated via a correlation function of the form (16.143), the result being expressed in terms of a certain number of lattice units a^{-1} at a certain value of $g_{\rm L}$. By comparison with the known ρ mass, a^{-1} can be converted to GeV. Then the calculation is repeated for a different $g_{\rm L}$ value and the new a^{-1} (GeV) extracted. A plot of $\ln[a^{-1}({\rm GeV})]$ versus $1/g_{\rm L}^2$ should then give a straight line with slope $2\pi/\beta_0$ and intercept $\ln\Lambda_{\rm L}$. Figure 16.9 shows such a plot, taken from Ellis *et al.* (1996), from which it appears that the calculations are indeed being performed close to the continuum limit. The value of $\Lambda_{\rm L}$ has been adjusted to fit the numerical data, and has the value $\Lambda_{\rm L} = 1.74$ MeV in this case. This may seem alarmingly far from the kind of value expected for $\Lambda_{\rm QCD}$, but we must remember that the renormalization schemes involved in the two cases are quite different. In fact, we may expect $\Lambda_{\rm QCD} \approx 50\Lambda_{\rm L}$ (Montvay and Munster (1994), section 5.1.6).

16.5.2 The static $\mathrm{q\bar{q}}$ potential

The calculations of m_ρ represented in figure 16.9 were done in the quenched approximation. As a first example of a calculation with dynamical (unquenched) fermions we show in figure 16.10 a lattice calculation of the static $\mathrm{q\bar{q}}$ potential (Allton *et al.* 2002, UKQCD Collaboration), using two degenerate flavours of

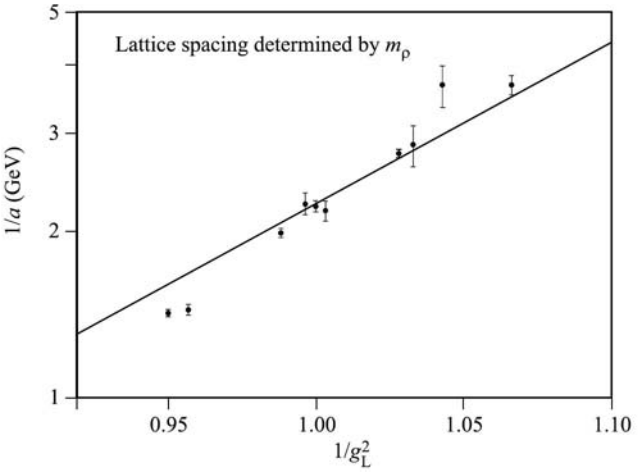


FIGURE 16.9
$\ln(a^{-1}$ in GeV) plotted against $1/g_L^2$; figure from R K Ellis, W J Stirling and B R Webber (1996) *QCD and Collider Physics*, courtesy Cambridge University Press, as adapted from Allton (1995).

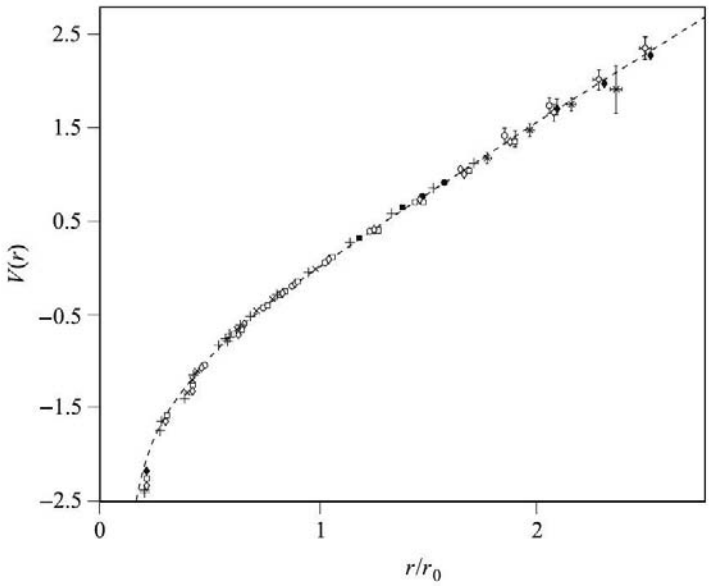


FIGURE 16.10
The static QCD potential, expressed in units of r_0. The broken curve is the functional form (16.147). Figure reprinted with permission from C R Allton *et al.* (UKQCD Collaboration) *Phys. Rev.* D **65** 054502 (2002). Copyright 2002 by the American Physical Society.

dynamical quarks[3] on a $16^3 \times 32$ lattice. As usual, one dimensionful quantity has to be fixed in order to set the scale. In the present case this has been done via the scale parameter r_0 of Sommer (1994), defined by

$$r_0^2 \left. \frac{\mathrm{d}V}{\mathrm{d}r} \right|_{r=r_0} = 1.65 \; . \tag{16.144}$$

Applying (16.144) to the Cornell (Eichten *et al.* 1980) or Richardson (1979) phenomenological potentials gives $r_0 \simeq 0.49$ fm, conveniently in the range which is well-determined by $c\bar{c}$ and $b\bar{b}$ data. The data are well described by the expression

$$V(r) = V_0 + \sigma r - \frac{A}{r} \; , \tag{16.145}$$

where in accordance with (16.144)

$$\sigma = \frac{(1.65 - A)}{r_0^2} \; , \tag{16.146}$$

and where V_0 has been chosen such that $V(r_0) = 0$. Thus (16.145) becomes

$$r_0 V(r) = (1.65 - A) \left(\frac{r}{r_0} - 1 \right) - A \left(\frac{r_0}{r} - 1 \right) \; . \tag{16.147}$$

This is – up to a constant – exactly the functional form mentioned in chapter 1, equation (1.33). The quantity $\sqrt{\sigma}$ (there called b) is referred to as the 'string tension', and has a value of about 465 MeV in the present calculations. Phenomenological models suggest a value of around 440 MeV (Eichten *et al.* 1980). The parameter A is found to have a value of about 0.3. In lowest-order perturbation theory, and in the continuum limit, A would be given by one-gluon exchange as

$$A = \frac{4}{3} \alpha_{\mathrm{s}}(\mu) \tag{16.148}$$

where μ is some energy scale. This would give $\alpha_{\mathrm{s}} \simeq 0.22$, a reasonable value for $\mu \simeq 3$ GeV. Interestingly, the form (16.147) is predicted by the 'universal bosonic string model' (Lüscher *et al.* 1980, Lüscher 1981), in which A has the 'universal' value $\frac{\pi}{12} \simeq 0.26$.

The existence of the linearly rising term with $\sigma > 0$ is a signal for confinement, since – if the potential maintained this form – it would cost an infinite amount of energy to separate a quark and an antiquark. But at some point, enough energy will be stored in the 'string' to create a $q\bar{q}$ pair from the vacuum: the string then breaks, and the two $q\bar{q}$ pairs form mesons. There is no evidence for string breaking in figure 16.10, but we must note that the largest distance probed is only about 1.3 fm.

[3]Comparison with matched data in the quenched approximation revealed very little difference, in this case.

16.5.3 Calculation of $\alpha(M_Z^2)$

Our second example of a precision lattice calculation with dynamical quarks is the determination of $\alpha_s(M_Z^2)$ by Davies *et al.* (2008) (HPQCD Collaboration). The reported value is

$$\alpha_s(M_Z^2) = 0.1183(8). \tag{16.149}$$

The accuracy of this result is extremely impressive, and it implies that this determination is an important ingredient in the world average value quoted in (15.62). It is worth sketching some of the elements that went into this landmark calculation.

The work used 12 gluon configurations from the MILC collaboration (Aubin *et al.* 2004), and built on a joint effort by several groups (see Davies *et al.* (HPQCD, UKQCD, MILC, and Fermilab collaborations) 2004). Vacuum polarization effects from all three light quarks u, d and s were included, using a Symanzik-improved staggered-quark discretization, with rooting. The effects of c and b quarks were incorporated using perturbation theory. The strange quark mass was physical, while the u and d quark mass (set to be the same) was three times too large, but small enough for chiral perturbation theory (see chapter 18) to be reliable for extrapolating to the physical mass.

There were 5 parameters: $m_u = m_d, m_s, m_c, m_b$ and the bare QCD coupling g_L (or equivalently the lattice spacing a). The mass parameters were tuned to reproduce experimentally measured values of $m_\pi^2, 2m_K^2 - m_\pi^2, m_D$ and m_Υ respectively. The lattice spacing was adjusted to make the $\Upsilon - \Upsilon'$ mass difference agree with experiment (Gray *et al.* 2005). With the free parameters all determined, the simulation accurately reproduced QCD, and predictions for physical quantities could proceed. *En passant*, we show in figure 16.11 results obtained (Davies *et al.* 2004), divided by experimental results, for nine different quantities, with and without quark vacuum polarization (left and right panels respectively). The values on the left deviate from experiment by as much as $10\% - 15\%$; those on the right agree with experiment to within systematic and statistical errors of 3% or less.

To extract a value of the coupling constant, the general strategy is to calculate (with the tuned simulation) a non-perturbative numerical value for a short-distance quantity, for which perturbation theory should be reliable. Then, by comparing the numerically computed value to the known perturbative expansion, a value of the coupling constant can be found.

In this case, the quantities calculated were vacuum expectation values of small Wilson loop operators W_{mn} (and related quantities) where

$$W_{mn} \equiv \frac{1}{3}\langle 0|\mathrm{Re\,Tr\,P}\exp[-\mathrm{i}g_L \int_{nm} A \cdot \mathrm{d}x]|0\rangle, \tag{16.150}$$

where P denotes path ordering, $A_\mu = \boldsymbol{\lambda}/2 \cdot \boldsymbol{A}_\mu$ is the QCD (matrix-valued) vector potential, and the integral is over a closed $ma \times na$ rectangular path, not necessarily planar. The 1×1 Wilson loop is just the vev of the simple plaquette operator $U_\Box$ of section 16.2.3.

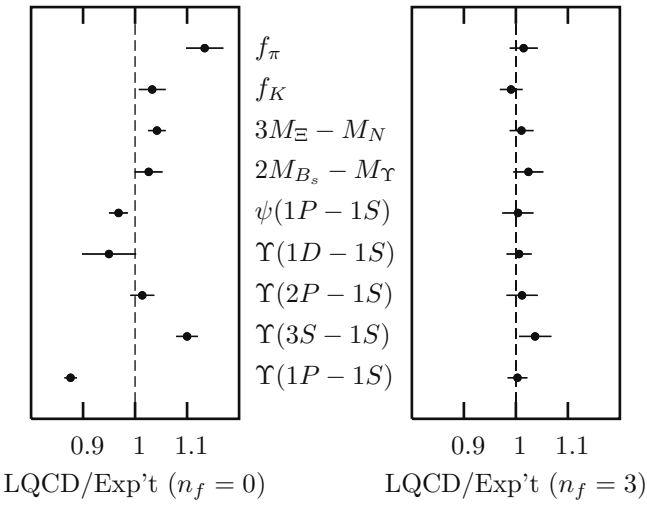


FIGURE 16.11
Lattice QCD results divided by experimental results for nine different quantities, with and without quark vacuum polarization (left and right panels, respectively). Figure reprinted with permission from C T H Davies *et al.* (HPQCD Collaboration) *Phys. Rev. Lett.* **92** 022001 (2004). Copyright 2004 by the American Physical Society.

In order to compare the numerical evaluation of (16.150) with perturbation theory, one has to decide what is a suitable expansion parameter. It was shown by Lepage and Mackenzie (1993) that the obvious first choice, the bare lattice coupling constant, is generally a poor one due to renormalization effects, even for short distance quantities. Instead, a renormalized coupling should be used – but this raises the questions of what renormalization *scheme* to adopt, and what *scale* at which to evaluate the (now running) coupling. In the present case, the scheme proposed by Brodsky, Lepage and Mackenzie (1983) was followed. It is defined in terms of the heavy quark potential $V(q)$, and is called the 'V-scheme'. The strong coupling in the V-scheme is defined by

$$V(q) = -\frac{4}{3}\frac{4\pi\alpha_{\rm V}(q)}{q^2} \tag{16.151}$$

with no higher-order corrections.

The numerically calculated short-distance quantities $Y^{(r)}$ are therefore to be expanded as the series

$$Y^{(r)} = \sum_{n=1}^{\infty} c_n^{(r)} \alpha_{\rm V}^n(d^{(r)}/a), \tag{16.152}$$

where $c_n^{(r)}$ and $d^{(r)}$ are dimensionless constants independent of the lattice spacing a, but dependent on the particular $Y^{(r)}$, and $\alpha_{\rm V}(d^{(r)}/a)$ is the running QCD coupling in the V-scheme, with $N_{\rm f} = 3$ light quark flavours. The perturbative coefficients $c_n^{(r)}$ for the various Y's were computed using Feynman diagrams, for $n \leq 3$, for the same quark and gluon actions which were used to create the sets of gluon field configurations employed in the numerical evaluation of the Y's. The renormalization scale $d^{(r)}/a$ varies for each short-distance quantity, being chosen according to the Lepage-Mackenzie (1993) prescription (or in some cases a more robust procedure due to Hornbostel, Lepage and Morningstar (2003)).

There were 22 $Y^{(r)}$'s, each of which was analyzed separately, fitting the expansion (16.152) to the 12 values of that Y calculated using the 12 gluon configurations. In the simplest terms, the result of each such fit would be the value of $\alpha_{\rm V}$ at a particular scale, which was chosen to be $\alpha_{\rm V}(7.5\ {\rm GeV})$. The values required at the scales $\alpha_{\rm V}(d^{(r)}/a_i)$ were found by numerically integrating the evolution equation (at four-loop order) for $\alpha_{\rm V}$; here a_i is the lattice spacing for each configuration (there were 6 different spacings). In fact, the fitting was more sophisticated, including further parameters related to various corrections; the interested reader can consult Davies *et al.* (2008) for the details. Having obtained $\alpha_{\rm V}(7.5\ {\rm GeV})$, this was then converted to the $\overline{\rm MS}$ scheme, using the relation (Brodsky, Lepage and Mackenzie 1983)

$$\alpha_{\rm V}(\mu) = \alpha_{\overline{\rm MS}}({\rm e}^{-5/6}\mu). \tag{16.153}$$

Finally, the resultant $\alpha_{\overline{\rm MS}}$ was evolved to $M_{\rm Z}^2$. The value (16.149) is the final result after performing a weighted average over the 22 separate determinations. A full discussion of the error estimate, which includes finite lattice spacing, finite lattice volume, and chiral extrapolation uncertainties, is given in Davies *et al.* (2008).

16.5.4 Hadron masses

For our last example of a precise lattice QCD calculation, it is appropriate to consider the mass spectrum of light hadrons. After all, protons and neutrons account for nearly all the mass of ordinary matter, and 95% of their mass is the result of QCD interactions. It has long been a fundamental challenge to predict hadron masses accurately from QCD.

As one example of such calculations, we show in figure 16.12 the light hadron spectrum of QCD as reported by Dürr *et al.* (2008). Horizontal lines and bands are the experimental values (which have been isospin-averaged) with their decay widths. The solid circles are the predicted values. Vertical error bars represent combined statistical and systematic error estimates. The masses of the π, K and Ξ have no error bars, because they have been used to set the values of $m_{\rm u} = m_{\rm d}$, $m_{\rm s}$ and the overall scale, respectively. Once again, the agreement with experiment is very impressive.

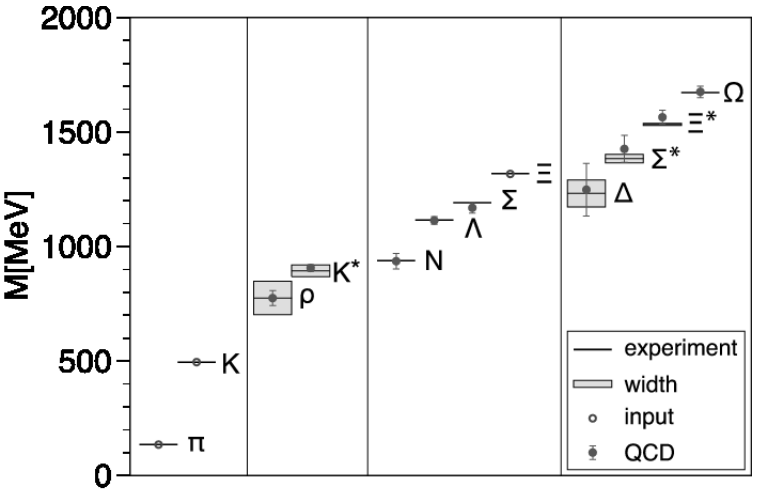


FIGURE 16.12
The light hadron spectrum of QCD, from Dürr *et al.* (2008). (See color plate II.)

These calculations used a Symanzik-improved gauge action (Lüscher and Weisz 1985), and 2+1 flavours of light dynamical Wilson fermions, with various improvements (Morningstar and Peardon 2004). The physical scale was set either by fitting to the mass of the Ξ, or to the mass of the Ω; the two ways gave consistent results. Pion masses in the range (approximately) 800 MeV to 190 MeV were used to extrapolate to the physical value, with lattice sizes approximately four times the inverse pion mass. A particular type of finite-volume effect arises in the case of strongly decaying resonant states: a procedure for reconstructing the infinite-volume resonance mass, given by Lüscher (1986, 1991a, 1991b), was followed here. This was satisfactory, except for the ρ and Δ at the lightest pion mass point, which was omitted from the extrapolation for these two channels. For further details, and additional references, we refer the reader to the supplementary material to Dürr *et al.* (2008) provided online.

We have been able to give only a brief introduction into what is now, almost forty years after its initial inception by Wilson (1974), the highly mature field of lattice QCD. A great deal of effort has gone into ingenious and subtle improvements to the lattice action, to the numerical algorithms, and to the treatment of fermions – to name a few of the issues. Lattice QCD is now a major part of particle physics. From the perspective of this chapter and the previous one, we can confidently say that, both in the short-distance (perturbative) regime, and in the long-distance (non-perturbative) regime, QCD is established as the correct theory of the strong interactions of quarks, beyond reasonable doubt.

Problems

16.1 Verify equation (16.9).

16.2 Verify equation (16.10).

16.3 Show that the momentum space version of (16.18) is (16.19).

16.4 Use (16.31) in (16.33) to verify (16.34).

16.5 Verify (16.68) and (16.70).

16.6 In a modified one-dimensional Ising model, spin variables s_n at sites labelled by $n = 1, 2, 3, \ldots N$ take the values $s_n = \pm 1$, and the energy of each spin configuration is

$$E = -\sum_{n=1}^{N-1} J_n s_n s_{n+1} \ ,$$

where all the constants J_n are positive. Show that the partition function Z_N is given by

$$Z_N = 2 \prod_{n=1}^{N-1} (2 \cosh K_n) \ ,$$

where $K_n = J_n / k_{\mathrm{B}} T$. Hence calculate the entropy for the particular case in which all the J_n's are equal to J and $N \gg 1$, and discuss the behaviour of the entropy in the limits $T \to \infty$ and $T \to 0$.

Let 'p' denote a particular site such that $1 \ll p \ll N$. Show that the average value $\langle s_p s_{p+1} \rangle$ of the product $s_p s_{p+1}$ is given by

$$\langle s_p s_{p+1} \rangle = \frac{1}{Z_N} \frac{\partial Z_N}{\partial K_p}.$$

Show further that

$$\langle s_p s_{p+j} \rangle = \frac{1}{Z_N} \frac{\partial^j Z_N}{\partial K_p \partial K_{p+1} \ldots \partial K_{p+j-1}}.$$

Hence show that in the case $J_1 = J_2 = \ldots = J_N = J$,

$$\langle s_p s_{p+j} \rangle = \mathrm{e}^{-ja/\xi} \ ,$$

where

$$\xi = -a / [\ln(\tanh K)] \ ,$$

and $K = J / k_{\mathrm{B}} T$. Discuss the physical meaning of ξ, considering the $T \to \infty$ and $T \to 0$ limits explicitly.


Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

Part VII

Spontaneously Broken Symmetry


Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

17

Spontaneously Broken Global Symmetry

Previous chapters have introduced the non-Abelian symmetries SU(2) and SU(3) in both global and local forms, and we have seen how they may be applied to describe such typical physical phenomena as particle multiplets, and massless gauge fields. Remarkably enough, however, these symmetries are also applied, in the Standard Model, in two cases where the physical phenomena appear to be very different. Consider the following two questions: (i) Why are there no signs in the baryonic spectrum, such as parity doublets in particular, of the global chiral symmetry introduced in section 12.3.2? (ii) How can weak interactions be described by a local non-Abelian gauge theory when we know the mediating gauge field quanta are not massless? The answers to these questions each involve the same fundamental idea, which is a crucial component of the Standard Model, and perhaps also of theories which go beyond it. This is the idea that a symmetry can be 'spontaneously broken', or 'hidden'. By contrast, the symmetries considered hitherto may be termed 'manifest symmetries'.

The physical consequences of spontaneous symmetry breaking turn out to be rather different in the global and local cases. However, the essentials for a theoretical understanding of the phenomenon are contained in the simpler global case, which we consider in this chapter. The application to spontaneously broken chiral symmetry will be treated in chapter 18, and spontaneously broken local symmetry will be discussed in chapter 19, and applied in chapter 22.

17.1 Introduction

We begin by considering, in response to question (i) above, what could go wrong with the argument for symmetry multiplets that we gave in chapter 12. To understand this, we must use the field theory formulation of section 12.3, in which the generators of the symmetry are Hermitian field operators, and the states are created by operators acting on the vacuum. Thus consider two states $|A\rangle$, $|B\rangle$[1] :

$$|A\rangle = \hat{\phi}_A^\dagger|0\rangle, \qquad |B\rangle = \hat{\phi}_B^\dagger|0\rangle \tag{17.1}$$

[1]We now revert to the ordinary notation $|0\rangle$ for the vacuum state, rather than $|\Omega\rangle$, but it must be borne in mind that $|0\rangle$ is the full (interacting) vacuum.

where $\hat{\phi}_A^\dagger$ and $\hat{\phi}_B^\dagger$ are related to each other by (cf (12.100))

$$[\hat{Q}, \hat{\phi}_A^\dagger] = \hat{\phi}_B^\dagger \tag{17.2}$$

for some generator $\hat{Q}$ of a symmetry group, such that

$$[\hat{Q}, \hat{H}] = 0. \tag{17.3}$$

(17.2) is equivalent to

$$\hat{U}\hat{\phi}_A^\dagger \hat{U}^{-1} \approx \hat{\phi}_A^\dagger + \mathrm{i}\epsilon\hat{\phi}_B^\dagger \tag{17.4}$$

for an infinitesimal transformation $\hat{U} \approx 1 + \mathrm{i}\epsilon\hat{Q}$. Thus $\hat{\phi}_A^\dagger$ is 'rotated' into $\hat{\phi}_B^\dagger$ by $\hat{U}$, and the operators will create states related by the symmetry transformation. We want to see what are the assumptions necessary to prove that

$$E_A = E_B, \quad \text{where} \quad \hat{H}|A\rangle = E_A|A\rangle \quad \text{and} \quad \hat{H}|B\rangle = E_B|B\rangle. \tag{17.5}$$

We have

$$E_B|B\rangle = \hat{H}|B\rangle = \hat{H}\hat{\phi}_B^\dagger|0\rangle = \hat{H}(\hat{Q}\hat{\phi}_A^\dagger - \hat{\phi}_A^\dagger\hat{Q})|0\rangle. \tag{17.6}$$

Now if

$$\hat{Q}|0\rangle = 0 \tag{17.7}$$

we can rewrite the right-hand side of (17.6) as

$$\begin{aligned} \hat{H}\hat{Q}\hat{\phi}_A^\dagger|0\rangle &= \hat{Q}\hat{H}\hat{\phi}_A^\dagger|0\rangle \quad \text{using (17.3)} \quad = \hat{Q}\hat{H}|A\rangle = E_A\hat{Q}|A\rangle \\ &= E_A\hat{Q}\hat{\phi}_A^\dagger|0\rangle = E_A(\hat{\phi}_B^\dagger + \hat{\phi}_A^\dagger\hat{Q})|0\rangle \quad \text{using (17.2)} \\ &= E_A|B\rangle \quad \text{if (17.7) holds;} \end{aligned} \tag{17.8}$$

whence, comparing (17.8) with (17.6), we see that

$$E_A = E_B \qquad \text{if (17.7) holds.} \tag{17.9}$$

Remembering that $\hat{U} = \exp(\mathrm{i}\alpha\hat{Q})$, we see that (17.7) is equivalent to

$$|0\rangle' \equiv \hat{U}|0\rangle = |0\rangle. \tag{17.10}$$

Thus a multiplet structure will emerge provided that the vacuum is left invariant under the symmetry transformation. *The 'spontaneously broken symmetry' situation arises in the contrary case – that is, when the vacuum is not invariant under the symmetry,* which is to say when

$$\hat{Q}|0\rangle \neq 0. \tag{17.11}$$

In this case, the argument for the existence of symmetry multiplets breaks down, and although the Hamiltonian or Lagrangian may exhibit a non-Abelian

symmetry, this will not be manifested in the form of multiplets of mass-degenerate particles.

The preceding italicized sentence does correctly define what is meant by a spontaneously broken symmetry in field theory, but there is another way of thinking about it which is somewhat less abstract though also less rigorous. The basic condition is $\hat{Q}|0\rangle \neq 0$, and it seems tempting to infer that, in this case, the application of $\hat{Q}$ to the vacuum gives, not zero, but *another possible vacuum*, $|0\rangle'$. Thus we have the physically suggestive idea of 'degenerate vacua' (they must be degenerate since $[\hat{Q}, H] = 0$). We shall see in a moment why this notion, though intuitively helpful, is not rigorous.

It would seem, in any case, that the properties of the *vacuum* are all-important, so we begin our discussion with a somewhat formal, but nonetheless fundamental, theorem about the quantum field vacuum.

17.2 The Fabri–Picasso theorem

Suppose that a given Lagrangian $\hat{\mathcal{L}}$ is invariant under some one-parameter continuous global internal symmetry with a conserved Noether current $\hat{j}^\mu$, such that $\partial_\mu \hat{j}^\mu = 0$. The associated 'charge' is the Hermitian operator $\hat{Q} = \int \hat{j}^0 \mathrm{d}^3\boldsymbol{x}$, and $\dot{\hat{Q}} = 0$. We have hitherto assumed that the transformations of such a U(1) group are representable in the space of physical states by unitary operations $\hat{U}(\lambda) = \exp \mathrm{i}\lambda\hat{Q}$ for arbitrary λ, with the vacuum invariant under $\hat{U}$, so that $\hat{Q}|0\rangle = 0$. Fabri and Picasso (1966) showed that there are actually *two* possibilities:

(i) $\hat{Q}|0\rangle = 0$, and $|0\rangle$ is an eigenstate of $\hat{Q}$ with eigenvalue 0, so that $|0\rangle$ is invariant under $\hat{U}$(i.e. $\hat{U}|0\rangle = |0\rangle$);
or

(ii) $\hat{Q}|0\rangle$ does not exist in the space (its norm is infinite).

The statement (ii) is technically more correct than the more intuitive statements '$\hat{Q}|0\rangle \neq 0$' or '$\hat{U}|0\rangle = |0\rangle'$', suggested above.

To prove this result, consider the vacuum matrix element $\langle 0|\hat{j}^0(x)\hat{Q}|0\rangle$. From translation invariance, implemented by the unitary operator[2] $\hat{U}(x) = \exp \mathrm{i}\hat{P} \cdot x$ (where $\hat{P}^\mu$ is the 4-momentum operator) we obtain

$$\begin{aligned} \langle 0|\hat{j}^0(x)\hat{Q}|0\rangle &= \langle 0|\mathrm{e}^{\mathrm{i}\hat{P}\cdot x}\hat{j}^0(0)\mathrm{e}^{-\mathrm{i}\hat{P}\cdot x}\hat{Q}|0\rangle \\ &= \langle 0|\mathrm{e}^{\mathrm{i}\hat{P}\cdot x}\hat{j}^0(0)\hat{Q}\mathrm{e}^{-\mathrm{i}\hat{P}\cdot x}|0\rangle \end{aligned}$$

[2]If this seems unfamiliar, it may be regarded as the 4-dimensional generalization of the transformation (I.7) in appendix I of volume 1, from Schrödinger picture operators at $t = 0$ to Heisenberg operators at $t \neq 0$.

where the second line follows from

$$[\hat{P}^\mu, \hat{Q}] = 0 \tag{17.12}$$

since $\hat{Q}$ is an *internal* symmetry. But the vacuum is an eigenstate of $\hat{P}^\mu$ with eigenvalue zero, and so

$$\langle 0|\hat{j}^0(x)\hat{Q}|0\rangle = \langle 0|\hat{j}^0(0)\hat{Q}|0\rangle \tag{17.13}$$

which states that the matrix element we started from is in fact independent of x. Now consider the norm of $\hat{Q}|0\rangle$:

$$\langle 0|\hat{Q}\hat{Q}|0\rangle = \int \mathrm{d}^3\boldsymbol{x}\langle 0|\hat{j}^0(x)\hat{Q}|0\rangle \tag{17.14}$$

$$= \int \mathrm{d}^3\boldsymbol{x}\langle 0|\hat{j}^0(0)\hat{Q}|0\rangle, \tag{17.15}$$

which must diverge in the infinite volume limit, unless $\hat{Q}|0\rangle = 0$. Thus either $\hat{Q}|0\rangle = 0$ or $\hat{Q}|0\rangle$ has infinite norm. The foregoing can be easily generalized to non-Abelian symmetry operators $\hat{T}_i$.

Remarkably enough, the argument can also, in a sense, be reversed. Coleman (1966) proved that if an operator

$$\hat{Q}(t) = \int \mathrm{d}^3\boldsymbol{x}\hat{j}^0(x) \tag{17.16}$$

is the spatial integral of the $\mu = 0$ component of a 4-vector (but *not assumed* to be conserved), and if it annihilates the vacuum

$$\hat{Q}(t)|0\rangle = 0, \tag{17.17}$$

then in fact $\partial_\mu \hat{j}^\mu = 0$, $\hat{Q}$ is independent of t, and the symmetry is unitarily implementable by operators $\hat{U} = \exp(\mathrm{i}\lambda\hat{Q})$.

We might now simply proceed to the chiral symmetry application. We believe, however, that the concept of spontaneous symmetry breaking is so important to particle physics that a more extended discussion is amply justified. In particular, there are crucial insights to be gained by considering the analogous phenomenon in condensed matter physics. After a brief look at the ferromagnet, we shall describe the Bogoliubov model for the ground state of a superfluid, which provides an important physical example of a spontaneously broken global Abelian U(1) symmetry. We shall see that the excitations away from the ground state are *massless modes* and we shall learn, via Goldstone's theorem, that such modes are an inevitable result of spontaneously breaking a global symmetry. Next, we shall introduce the 'Goldstone model' which is the simplest example of a spontaneously broken global U(1) symmetry, involving just one complex scalar field. The generalization of this to the non-Abelian case will draw us in the direction of the Higgs sector of the Standard Model.

Returning to condensed matter systems, we introduce the BCS ground state for a superconductor, in a way which builds on the Bogoliubov model of a superfluid. We are then prepared for the application, in chapter 18, to spontaneous chiral symmetry breaking (question (i) above), following Nambu's profound analogy with one aspect of superconductivity. In chapter 19 we shall see how a different aspect of superconductivity provides a model for the answer to question (ii) above.

17.3 Spontaneously broken symmetry in condensed matter physics

17.3.1 The ferromagnet

We have seen that everything depends on the properties of the vacuum state. An essential aid to understanding hidden symmetry in quantum field theory is provided by Nambu's (1960) remarkable insight that the *vacuum* state of a quantum field theory is analogous to the ground state of an interacting many-body system. It is the state of lowest energy – the equilibrium state, given the kinetic and potential energies as specified in the Hamiltonian. Now the ground state of a complicated system (for example, one involving interacting fields) may well have unsuspected properties – which may, indeed, be very hard to predict from the Hamiltonian. But we can postulate (even if we cannot yet prove) properties of the quantum field theory vacuum $|0\rangle$ which are analogous to those of the ground states of many physically interesting many-body systems – such as superfluids and superconductors, to name two with which we shall be principally concerned.

Now it is generally the case, in quantum mechanics, that the ground state of any system described by a Hamiltonian is non-degenerate. Sometimes we may meet systems in which apparently more than one state has the same lowest energy eigenvalue. Yet in fact none of these states will be the true ground state: tunnelling will take place between the various degenerate states, and the true ground state will turn out to be a unique linear superposition of them. This is, in fact, the only possibility for systems of finite spatial extent, though in practice a state which is not the true ground state may have an extremely long lifetime. However, in the case of fields (extending presumably throughout all space), the Fabri–Picasso theorem shows that there is an alternative possibility, which is often described as involving a 'degenerate ground state' – a term we shall now elucidate. In case (a) of the theorem, the ground state is unique. For, suppose that several ground states $|0,a\rangle, |0,b\rangle, \ldots$ existed, with the symmetry unitarily implemented. Then one ground state will be related to another by

$$|0,a\rangle = \mathrm{e}^{\mathrm{i}\lambda\hat{Q}}|0,b\rangle \tag{17.18}$$

for some λ. However, in case (a) the charge annihilates a ground state, and so all of them are really identical. In case (b), on the other hand, we cannot write (17.18) – since $\hat{Q}|0\rangle$ does not exist – and we do have the possibility of many degenerate ground states. In simple models one can verify that these alternative ground states are all orthogonal to each other, in the infinite volume limit – or perhaps more physically, the limit in which the number of degrees of freedom becomes infinite. And each member of every 'tower' of excited states, built on these alternative ground states, is also orthogonal to all the members of other towers. But any single tower must constitute a complete space of states. It follows that states in different towers belong to *different* complete spaces of states, that is to different – and inequivalent – 'worlds', each one built on one of the possible orthogonal ground states.

At first sight, a familiar example of these ideas seems to be that of a ferromagnet, below its Curie temperature $T_{\rm C}$. Consider an 'ideal Heisenberg ferromagnet' with N atoms each of spin 1/2, described by a Hamiltonian of Heisenberg exchange form $H_S = -J \sum \hat{\boldsymbol{S}}_i \cdot \hat{\boldsymbol{S}}_j$, where i and j label the atomic sites. This Hamiltonian is invariant under spatial rotations, since it only depends on the dot product of the spin operators. Such rotations are implemented by unitary operators $\exp(\mathrm{i}\hat{\boldsymbol{S}} \cdot \boldsymbol{\alpha})$ where $\hat{\boldsymbol{S}} = \sum_i \hat{\boldsymbol{S}}_i$, and spins at different sites are assumed to commute. As usual with angular momentum in quantum mechanics, the eigenstates of H_S are labelled by the eigenvalues of total squared spin, and of one component of spin, say of $\hat{S}_z = \sum_i \hat{S}_{iz}$. The quantum mechanical ground state of H_S is an eigenstate with total spin quantum number $S = N/2$, and this state is $(2 \cdot N/2 + 1) = (N + 1)-$ fold degenerate, allowing for all the possible eigenvalues $(N/2, N/2 - 1, \ldots - N/2)$ of $\hat{S}_z$ for this value of S. We are free to choose any one of these degenerate states as 'the' ground state, say the state with eigenvalue $S_z = N/2$.

It is clear that the ground state is not invariant under the spin-rotation symmetry of H_S, which would require the eigenvalues $S = S_z = 0$. Furthermore, this ground state is degenerate. So two important features of what we have so far learned to expect of a spontaneously broken symmetry are present – namely, 'the ground state is not invariant under the symmetry of the Hamiltonian', and 'the ground state is degenerate'. However, it has to be emphasized that this ferromagnetic ground state does, in fact, respect the symmetry of H_S, in the sense that it belongs to an irreducible representation of the symmetry group: the unusual feature is that it is not the 'trivial' (singlet) representation, as would be the case for an invariant ground state. The spontaneous symmetry breaking which is the true model for particle physics is that in which a many body ground state is *not* an eigenstate (trivial or otherwise) of the symmetry operators of the Hamiltonian: rather it is a superposition of such eigenstates. We shall explore this for the superfluid and the superconductor in due course.

Nevertheless, there are some useful insights to be gained from the ferromagnet. First, consider two ground states differing by a spin rotation. In the first, the spins are all aligned along the 3-axis, say, and in the second along

the axis $\hat{\boldsymbol{n}} = (0, \sin\alpha, \cos\alpha)$. Thus the first ground state is

$$\chi_0 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}_1 \begin{pmatrix} 1 \\ 0 \end{pmatrix}_2 \cdots \begin{pmatrix} 1 \\ 0 \end{pmatrix}_N \qquad \text{(N products)} \tag{17.19}$$

while the second is (cf (4.31), (4.32))

$$\chi_0^{(\alpha)} = \begin{pmatrix} \cos\alpha/2 \\ \mathrm{i}\sin\alpha/2 \end{pmatrix}_1 \cdots \begin{pmatrix} \cos\alpha/2 \\ \mathrm{i}\sin\alpha/2 \end{pmatrix}_N . \tag{17.20}$$

The scalar product of (17.19) and (17.20) is $(\cos\alpha/2)^N$, which goes to zero as $N \to \infty$. Thus any two such 'rotated ground states' are indeed orthogonal in the infinite volume (or infinite number of degrees of freedom) limit.

We may also enquire about the excited states built on one such ground state, say the one with $\hat{S}_z$ eigenvalue $N/2$. Suppose for simplicity that the magnet is one-dimensional (but the spins have all three components). Consider the state $\chi_n = \hat{S}_{n-}\chi_0$ where $\hat{S}_{n-}$ is the spin lowering operator $\hat{S}_{n-} = (\hat{S}_{nx} - \mathrm{i}\hat{S}_{ny})$ at site n, such that

$$\hat{S}_{n-} \begin{pmatrix} 1 \\ 0 \end{pmatrix}_n = \begin{pmatrix} 0 \\ 1 \end{pmatrix}_n ; \tag{17.21}$$

so $\hat{S}_{n-}\chi_0$ differs from the ground state χ_0 by having the spin at site n flipped. The action of $\hat{H}_S$ on χ_n can be found by writing

$$\sum_{i\neq j} \hat{\boldsymbol{S}}_i \cdot \hat{\boldsymbol{S}}_j = \sum_{i\neq j} \frac{1}{2}(\hat{S}_{i-}\hat{S}_{j+} + \hat{S}_{j-}\hat{S}_{i+}) + \hat{S}_{iz}\hat{S_{jz}} \tag{17.22}$$

(remembering that spins on different sites commute), where $\hat{S}_{i+} = \hat{S}_{ix} + \mathrm{i}\hat{S}_{iy}$. Since all $\hat{S}_{i+}$ operators give zero on a spin 'up' state, the only non-zero contributions from the first (bracketed) term in (17.22) come from terms in which either $\hat{S}_{i+}$ or $\hat{S}_{j+}$ act on the 'down' spin at n, so as to restore it to 'up'. The 'partner' operator $\hat{S}_{i-}$ (or $\hat{S}_{j-}$) then simply lowers the spin at i (or j), leading to the result

$$\sum_{i\neq j} \frac{1}{2}(\hat{S}_{i-}\hat{S}_{j+} + \hat{S}_{j-}\hat{S}_{i+})\chi_n = \sum_{i\neq n} \chi_i . \tag{17.23}$$

Thus the state χ_n is not an eigenstate of $\hat{H}_S$. However, a little more work shows that the superpostitions

$$\tilde{\chi}_q = \frac{1}{\sqrt{N}} \sum_n \mathrm{e}^{\mathrm{i}qna}\chi_n \tag{17.24}$$

are eigenstates; here q is one of the discretized wavenumbers produced by appropriate boundary conditions, as is usual in one-dimensional 'chain' problems. The states (17.24) represent *spin waves*, and they have the important

feature that for low q (long wavelength) their frequency ω tends to zero with q (actually $\omega \propto q^2$). In this respect, therefore, they behave like massless particles when quantized – and this is another feature we should expect when a symmetry is spontaneously broken.

The ferromagnet gives us one more useful insight. We have been assuming that one particular ground state (e.g. the one with $S_z = N/2$) has been somehow 'chosen'. But what does the choosing? The answer to this is clear enough in the (perfectly realistic) case in which the Hamiltonian $\hat{H}_S$ is supplemented by a term $-g\mu B \sum_i \hat{S}_{iz}$, representing the effect of an applied field B directed along the z-axis. This term will indeed ensure that the ground state is unique, and has $S_z = N/2$. Consider now the two limits $B \to 0$ and $N \to \infty$, both at finite temperature. When $B \to 0$ at finite N, the $N+1$ different S_z eigenstates become degenerate, and we have an ensemble in which each enters with an equal weight; there is therefore no loss of symmetry, even as $N \to \infty$ (but only *after* $B \to 0$). On the other hand, if $N \to \infty$ at finite $B \neq 0$, the single state with $S_z = N/2$ will be selected out as the unique ground state and this asymmetric situation will persist even in the limit $B \to 0$. In a (classical) mean field theory approximation we suppose that an 'internal field' is 'spontaneously generated', which is aligned with the external B and survives even as $B \to 0$, thus 'spontaneously' breaking the symmetry.

The ferromagnet therefore provides an easily pictured system exhibiting many of the features associated with spontaneous symmetry breaking; most importantly, it strongly suggests that what is really characteristic about the phenonenon is that it entails 'spontaneous ordering'.[3] Generally such ordering occurs below some characteristic 'critical temperature', $T_{\rm C}$. The field which develops a non-zero equilibrium value below $T_{\rm C}$ is called an 'order parameter'. This concept forms the basis of Landau's theory of second-order phase transitions (see for example chapter XIV of Landau and Lifshitz 1980).

We now turn to an example much more closely analogous to the particle physics applications: the superfluid.

17.3.2 The Bogoliubov superfluid

Consider the non-relativistic Hamiltonian (in the Schrödinger picture)

$$\begin{aligned} \hat{H} &= \frac{1}{2m}\int \mathrm{d}^3\boldsymbol{x}\,\boldsymbol{\nabla}\hat{\phi}^\dagger\cdot\boldsymbol{\nabla}\hat{\phi} \\ &\quad + \frac{1}{2}\int\int \mathrm{d}^3\boldsymbol{x}\,\mathrm{d}^3\boldsymbol{y}\, v(|\boldsymbol{x}-\boldsymbol{y}|)\hat{\phi}^\dagger(\boldsymbol{x})\hat{\phi}^\dagger(\boldsymbol{y})\hat{\phi}(\boldsymbol{y})\hat{\phi}(\boldsymbol{x}) \end{aligned} \qquad (17.25)$$

where $\hat{\phi}^\dagger(\boldsymbol{x})$ creates a boson of mass m at position $\boldsymbol{x}$. This $\hat{H}$ describes identical bosons interacting via a potential v, which is assumed to be weak (see, for example, Schiff 1968 section 55, or Parry 1973 chapter 1). We note

[3] It is worth pausing to reflect on the idea that *ordering* is associated with *symmetry breaking*.

at once that $\hat{H}$ is invariant under the global U(1) symmetry

$$\hat{\phi}(\boldsymbol{x}) \to \hat{\phi}'(\boldsymbol{x}) = \mathrm{e}^{-\mathrm{i}\alpha}\hat{\phi}(\boldsymbol{x}), \tag{17.26}$$

the generator being the conserved number operator

$$\hat{N} = \int \hat{\phi}^\dagger \hat{\phi}\ \mathrm{d}^3\boldsymbol{x} \tag{17.27}$$

which obeys $[\hat{N}, \hat{H}] = 0$. Our ultimate concern will be with the way this symmetry is 'spontaneously broken' in the superfluid ground state. Naturally, since this is an Abelian, rather than a non-Abelian, symmetry the physics will not involve any (hidden) multiplet structure. But the nature of the 'symmetry breaking ground state' in this U(1) case (and in the BCS model of section 17.7) will serve as a physical model for non-Abelian cases also.

We begin by re-writing $\hat{H}$ in terms of mode creation and annihilation operators in the usual way. We expand $\hat{\phi}(\boldsymbol{x})$ as a superposition of solutions of the $v = 0$ problem, which are plane waves quantized in a large cube of volume Ω:

$$\hat{\phi}(\boldsymbol{x}) = \frac{1}{\Omega^{\frac{1}{2}}} \sum_{\boldsymbol{k}} \hat{a}_{\boldsymbol{k}} \mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{x}} \tag{17.28}$$

where $\hat{a}_{\boldsymbol{k}}|0\rangle = 0$, $\hat{a}^\dagger_{\boldsymbol{k}}|0\rangle$ is a one-particle state, and $[\hat{a}_{\boldsymbol{k}}, \hat{a}^\dagger_{\boldsymbol{k}'}] = \delta_{\boldsymbol{k},\boldsymbol{k}'}$, with all other commutators vanishing. We impose periodic boundary conditions at the cube faces, and the free particle energies are $\epsilon_k = \boldsymbol{k}^2/2m$. Inserting (17.28) into (17.25) leads (problem 17.1) to

$$\hat{H} = \sum_{\boldsymbol{k}} \epsilon_k \hat{a}^\dagger_{\boldsymbol{k}} a_{\boldsymbol{k}} + \frac{1}{2\Omega} \sum_{\Delta} \bar{v}(|\boldsymbol{k}_1 - \boldsymbol{k}'_1|) \hat{a}^\dagger_{\boldsymbol{k}_1} \hat{a}^\dagger_{\boldsymbol{k}_2} \hat{a}_{\boldsymbol{k}'_2} \hat{a}_{\boldsymbol{k}'_1} \Delta(\boldsymbol{k}_1 + \boldsymbol{k}_2 - \boldsymbol{k}'_1 - \boldsymbol{k}'_2) \tag{17.29}$$

where the sum is over all momenta $\boldsymbol{k}_1, \boldsymbol{k}_2, \boldsymbol{k}'_1, \boldsymbol{k}'_2$ subject to the conservation law imposed by the Δ function:

$$\Delta(\boldsymbol{k}) = 1 \qquad \text{if } \boldsymbol{k} = \boldsymbol{0} \tag{17.30}$$

$$= 0 \qquad \text{if } \boldsymbol{k} \neq \boldsymbol{0}. \tag{17.31}$$

The interaction term in (17.29) is easily visualized as in figure 17.1. A pair of particles in states $\boldsymbol{k}'_1, \boldsymbol{k}'_2$ is scattered (conserving momentum) to a pair in states $\boldsymbol{k}_1, \boldsymbol{k}_2$ via the Fourier transform of v:

$$\bar{v}(|\boldsymbol{k}|) = \int v(\boldsymbol{r}) \mathrm{e}^{-\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{r}} \mathrm{d}^3\boldsymbol{r}. \tag{17.32}$$

Now, below the superfluid transition temperature T_S, we know that in the limit as $v \to 0$ the ground state has all the particles 'condensed' into the lowest energy state, which has $\boldsymbol{k} = \boldsymbol{0}$. Thus the ground state will be proportional to

$$|N, 0\rangle = (\hat{a}^\dagger_0)^N |0\rangle. \tag{17.33}$$

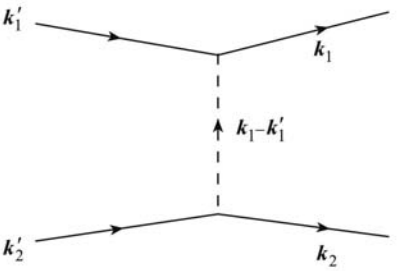


FIGURE 17.1
The interaction term in (17.29).

When a weak repulsive v is included, it is reasonable to hope that most of the particles remain in the condensate, only relatively few being excited to states with $\boldsymbol{k} \neq \boldsymbol{0}$. Let N_0 be the number of particles with $\boldsymbol{k} = \boldsymbol{0}$, where by assumption $N_0 \approx N$. We now consider the limit N (and N_0) $\to \infty$ and $\Omega \to \infty$ such that the density $\rho = N/\Omega$ (and $\rho_0 = N_0/\Omega$) stays constant. Bogoliubov (1947) argued that in this limit we may effectively replace both $\hat{a}_0$ and $\hat{a}_0^\dagger$ in the second term in (17.29) by the number $N_0^{1/2}$. This amounts to saying that in the commutator

$$\frac{\hat{a}_0}{\Omega^{1/2}}\frac{\hat{a}_0^\dagger}{\Omega^{1/2}} - \frac{\hat{a}^\dagger}{\Omega^{1/2}}\frac{\hat{a}_0}{\Omega^{1/2}} = \frac{1}{\Omega} \tag{17.34}$$

the two terms on the left-hand side are each of order N_0/Ω and hence finite, while their difference may be neglected as $\Omega \to \infty$. Replacing $\hat{a}_0$ and $\hat{a}_0^\dagger$ by $N_0^{1/2}$ leads (problem 17.2) to the following approximate form for $\hat{H}$:

$$\begin{aligned}\hat{H} \approx \hat{H}_{\mathrm{B}} \quad \equiv \quad & {\sum_{\boldsymbol{k}}}' \hat{a}_{\boldsymbol{k}}^\dagger \hat{a}_{\boldsymbol{k}} E_k + \frac{1}{2}\frac{N^2}{\Omega}\bar{v}(0) \\ & + \frac{1}{2}{\sum_{\boldsymbol{k}}}' \frac{N}{\Omega}\bar{v}(|\boldsymbol{k}|)[\hat{a}_{\boldsymbol{k}}^\dagger \hat{a}_{-\boldsymbol{k}}^\dagger + \hat{a}_{\boldsymbol{k}}\hat{a}_{-\boldsymbol{k}}],\end{aligned} \tag{17.35}$$

where

$$E_k = \epsilon_k + \frac{N}{\Omega}\bar{v}(|\boldsymbol{k}|), \tag{17.36}$$

primed summations do not include $\boldsymbol{k} = \boldsymbol{0}$, and terms which tend to zero as $\Omega \to \infty$ have been dropped (thus, N_0 has been replaced by N).

The most immediately striking feature of (17.35), as compared with $\hat{H}$ of (17.29), is that $\hat{H}_{\mathrm{B}}$ does not conserve the U(1) (number) symmetry (17.26) while $\hat{H}$ does: it is easy to see that for (17.26) to be a good symmetry, the number of $\hat{a}$'s must equal the number of $\hat{a}^\dagger$'s in every term. Thus the ground state of $\hat{H}_{\mathrm{B}}$, $|\mathrm{ground}\rangle_{\mathrm{B}}$, cannot be expected to be an eigenstate of the number

operator. However, it is important to be clear that the number non-conserving aspect of (17.35) is of a completely different kind, conceptually, from that which would be associated with a (hypothetical) '*explicit*' number violating term in the original Hamiltonian – for example, the addition of a term of the form '$\hat{a}^\dagger\hat{a}\hat{a}$'. In arriving at (17.35), we effectively replaced (17.28) by

$$\hat{\phi}_{\mathrm{B}}(\boldsymbol{x}) = \rho_0^{1/2} + \frac{1}{\Omega^{1/2}} \sum_{\boldsymbol{k}\neq\boldsymbol{0}} \hat{a}_{\boldsymbol{k}} \mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{x}} \tag{17.37}$$

where $\rho_0 = N_0/\Omega, N_0 \approx N$, and N_0/Ω remains finite as $\Omega \to \infty$. The limit is crucial here: it enables us to picture the condensate N_0 as providing an infinite reservoir of particles, with which excitations away from the ground state can exchange particle number. From this point of view, a number non-conserving ground state may appear more reasonable. The ultimate test, of course, is whether such a state is a good approximation to the true ground state, for a large but finite system.

What is $|\mathrm{ground}\rangle_{\mathrm{B}}$? Remarkably, $\hat{H}_{\mathrm{B}}$ can be exactly diagonalized by means of the *Bogoliubov quasiparticle operators* (for $\boldsymbol{k} \neq \boldsymbol{0}$)

$$\hat{\alpha}_{\boldsymbol{k}} = f_k \hat{a}_{\boldsymbol{k}} + g_k \hat{a}^\dagger_{-\boldsymbol{k}}, \quad \hat{\alpha}^\dagger_{\boldsymbol{k}} = f_k \hat{a}^\dagger_{\boldsymbol{k}} + g_k \hat{a}^\dagger_{-\boldsymbol{k}} \tag{17.38}$$

where f_k and g_k are real functions of $k = |\boldsymbol{k}|$. We must again at once draw attention to the fact that this transformation does not respect the symmetry (17.26) either, since $\hat{a}_{\boldsymbol{k}} \to \mathrm{e}^{-\mathrm{i}\alpha}\hat{a}_{\boldsymbol{k}}$ while $\hat{a}^\dagger_{-\boldsymbol{k}} \to \mathrm{e}^{+\mathrm{i}\alpha}\hat{a}^\dagger_{-\boldsymbol{k}}$. In fact, the operators $\hat{\alpha}^\dagger_{\boldsymbol{k}}$ will turn out to be precisely *creation operators for quasiparticles* which exchange particle number with the ground state.

The commutator of $\hat{\alpha}_{\boldsymbol{k}}$ and $\hat{\alpha}^\dagger_{\boldsymbol{k}}$ is easily evaluated:

$$[\hat{\alpha}_{\boldsymbol{k}}, \hat{\alpha}^\dagger_{\boldsymbol{k}}] = f_k^2 - g_k^2, \tag{17.39}$$

while two $\hat{\alpha}$'s or two $\hat{\alpha}^\dagger$'s commute. We choose f_k and g_k such that $f_k^2 - g_k^2 = 1$, so that the $\hat{a}$'s and the $\hat{\alpha}$'s have the same (bosonic) commutation relations, and the transformation (17.38) is 'canonical'. A convenient choice is $f_k = \cosh\theta_k, g_k = \sinh\theta_k$. We now assert that $\hat{H}_{\mathrm{B}}$ can be written in the form

$$\hat{H}_{\mathrm{B}} = {\sum_{\boldsymbol{k}}}' \omega_k \hat{\alpha}^\dagger_{\boldsymbol{k}} \hat{\alpha}_{\boldsymbol{k}} + \beta \tag{17.40}$$

for certain ω_k and β. Equation (17.40) implies, of course, that the eigenvalues of $\hat{H}_{\mathrm{B}}$ are $\beta + \sum_{\boldsymbol{k}}(n + 1/2)\omega_k$, and that $\hat{\alpha}^\dagger_{\boldsymbol{k}}$ acts as the creation operator for the quasiparticle of energy ω_k, as just anticipated.

We verify (17.40) slightly indirectly. We note first that it implies that

$$[\hat{H}_{\mathrm{B}}, \hat{\alpha}^\dagger_{\boldsymbol{l}}] = \omega_l \hat{\alpha}^\dagger_{\boldsymbol{l}}. \tag{17.41}$$

Substituting for $\hat{\alpha}^\dagger_{\boldsymbol{l}}$ from (17.38), we require

$$[\hat{H}_{\rm B}, \cosh\theta_l\, \hat{a}^\dagger_{\boldsymbol{l}} + \sinh\theta_l\, \hat{a}_{\_\boldsymbol{l}}] = \omega_l(\cosh\theta_l\, \hat{a}^\dagger_{\boldsymbol{l}} + \sinh\theta_l\, \hat{a}_{\_\boldsymbol{l}}), \tag{17.42}$$

which must hold as an identity in the $\hat{a}$'s and $\hat{a}^\dagger$'s. Using the expression (17.35) for $\hat{H}_{\rm B}$, and some patient work with the commutation relations (problem 17.3), one finds

$$(\omega_l - E_l)\cosh\theta_l \quad + \quad \frac{N}{\Omega}\bar{v}(|\boldsymbol{l}|)\sinh\theta_l = 0 \tag{17.43}$$

$$\frac{N}{\Omega}\bar{v}(|\boldsymbol{l}|)\cosh\theta_l \quad - \quad (\omega_l + E_l)\sinh\theta_l = 0. \tag{17.44}$$

For consistency, therefore, we require

$$E_l^2 - \omega_l^2 - \left(\frac{N}{\Omega^2}\right)^2 (\bar{v}(|\boldsymbol{l}|))^2 = 0, \tag{17.45}$$

or (recalling the definitions of E_l and ϵ_l)

$$\omega_l = \left[\frac{\boldsymbol{l}^2}{2m}\left(\frac{\boldsymbol{l}^2}{2m} + 2\rho\bar{v}(|\boldsymbol{l}|)\right)\right]^{1/2} \tag{17.46}$$

where $\rho = N/\Omega$. The value of $\tanh\theta_l$ is then determined via either of (17.43), (17.44).

Equation (17.46) is an important result, giving the frequency as a function of the momentum (or wavenumber); it is an example of a 'dispersion relation'. At the risk of stating the obvious, let us emphasize that equation (17.40) tells us that the original system of interacting bosons is equivalent (under the approximations made) to a system of non-interacting quasiparticles, whose frequency ω_l is related to wavenumber by (17.46). These are the true modes of the system. Let us consider this dispersion relation.

First of all, in the non-interacting case $\bar{v} = 0$, we recover the usual frequency-wavenumber relation for a massive non-relativistic particle, $\omega_l = \boldsymbol{l}^2/2m$. But if $\bar{v}(0) \neq 0$, the behaviour at small $\boldsymbol{l}$ is very different: $\omega_l \approx c_{\rm s}|\boldsymbol{l}|$, where $c_{\rm s} = (\rho\bar{v}(0)/m)^{1/2}$. This dispersion relation is characteristic of a massless mode, but in this case it is sound rather than light, with speed of sound $c_{\rm s}$. The spectrum is therefore phonon-like, not (non-relativistic) particle-like. The two behaviours can be easily distinguished experimentally, by measuring the low-temperature specific heat: in three dimensions, for $\omega_l \sim \boldsymbol{l}^2$ it goes to zero as $T^{3/2}$, whereas for $\omega_l \sim |\boldsymbol{l}|$ it goes as T^3. The latter behaviour is observed in superfluids. At large values of $|\boldsymbol{l}|$, however, ω_l behaves essentially like $\boldsymbol{l}^2/2m$ and the spectrum returns to the 'particle-like' one of massive bosons. Thus (17.46) interpolates between phonon-like behaviour at small $|\boldsymbol{l}|$ and particle-like behaviour at large $|\boldsymbol{l}|$.

There is still more to be learned from (17.46). If, in fact, $\bar{v}(|\boldsymbol{l}|) \sim 1/\boldsymbol{l}^2$, then $\omega_l \to$ constant as $|\boldsymbol{l}| \to 0$, and the spectrum would *not* be phonon-like. Indeed, if $\bar{v}(|\boldsymbol{l}|) \sim e^2/\boldsymbol{l}^2$, then $\omega_l \sim |e|(\rho/m)^{1/2}$ for small $|\boldsymbol{l}|$, which is just the 'plasma frequency' $\omega_{\rm p}$. In particle physics terms, this would be analogous to a dispersion relation of the form $\omega_l \sim (\omega_{\rm p}^2 + \boldsymbol{l}^2)^{1/2}$, which describes a particle with mass $\omega_{\rm p}$. Such a $\bar{v}$ is, of course, Colombic (the Fourier transform of $e^2/|\boldsymbol{x}|$), indicating that *in the case of such a long-range force the frequency spectrum acquires a mass-gap.* This will be the topic of chapter 19.

Having discussed the spectrum of quasiparticle excitations, let us now concentrate on the ground state in this model. From (17.40), it is clear that it is defined as the state $|\text{ground}\rangle_{\rm B}$ such that

$$\hat{\alpha}_{\boldsymbol{k}}|\text{ground}\rangle_{\rm B} = 0 \qquad \text{for all } \boldsymbol{k} \neq \boldsymbol{0}; \tag{17.47}$$

i.e. as the state with no non-zero-momentum quasiparticles in it. This is a complicated state in terms of the original $\hat{a}_{\boldsymbol{k}}$ and $\hat{a}^\dagger_{\boldsymbol{k}}$ operators, but we can give a formal expression for it, as follows. Since the $\hat{\alpha}$'s and $\hat{a}$'s are related by a canonical transformation, there must exist a unitary operator $\hat{U}_{\rm B}$ such that

$$\hat{\alpha}_{\boldsymbol{k}} = \hat{U}_{\rm B}\hat{a}_{\boldsymbol{k}}\hat{U}_{\rm B}^{-1}, \qquad \hat{a}_{\boldsymbol{k}} = \hat{U}_{\rm B}^{-1}\hat{\alpha}_{\boldsymbol{k}}\hat{U}_{\rm B}. \tag{17.48}$$

Now we know that $\hat{a}_{\boldsymbol{k}}|0\rangle = 0$. Hence it follows that

$$\hat{\alpha}_{\boldsymbol{k}}\hat{U}_{\rm B}|0\rangle = 0, \tag{17.49}$$

and we can identify $|\text{ground}\rangle_{\rm B}$ with $\hat{U}_{\rm B}|0\rangle$. In problem 17.4, $\hat{U}_{\rm B}$ is evaluated for an $\hat{H}_{\rm B}$ consisting of a single $\boldsymbol{k}$-mode only, in which case the operator effecting the transformation analogous to (17.48) is $\hat{U}_1 = \exp[\theta(\hat{a}\hat{a} - \hat{a}^\dagger\hat{a}^\dagger)/2]$ where θ replaces θ_k in this case. This generalizes (in the form of products of such operators) to the full $\hat{H}_{\rm B}$ case, but we shall not need the detailed result; an analogous result for the BCS ground state is discussed more fully in section 17.7. The important point is the following. It is clear from expanding the exponentials that $\hat{U}_{\rm B}$ creates a state in which the number of a-quanta (i.e. the original bosons) *is not fixed.* Thus unlike the simple non-interacting ground state $|N, 0\rangle$ of (17.33), $|\text{ground}\rangle_{\rm B} = \hat{U}_{\rm B}|0\rangle$ does not have a fixed number of particles in it: that is to say, it is *not* an eigenstate of the symmetry operator $\hat{N}$, as anticipated in the comment following (17.36). This is just the situation alluded to in the paragraph before equation (17.19), in our discussion of the ferromagnet.

Consider now the expectation value of $\hat{\phi}(\boldsymbol{x})$ in any state of definite particle number – that is, in an eigenstate of the symmetry operator $\hat{N}$. It is easy to see that this must vanish (remember that $\hat{\phi}$ destroys a boson, and so $\hat{\phi}|N\rangle$ is proportional to $|N-1\rangle$, which is orthogonal to $|N\rangle$). On the other hand, this is *not* true of $\hat{\phi}_{\rm B}(\boldsymbol{x})$: for example, in the non-interacting ground state (17.33), we have

$$\langle N, 0|\hat{\phi}_{\rm B}(\boldsymbol{x})|N, 0\rangle = \rho_0^{1/2}. \tag{17.50}$$

Furthermore, using the inverse of (17.38)

$$\hat{a}_{\boldsymbol{k}} = \cosh\theta_k \hat{\alpha}_{\boldsymbol{k}} - \sinh\theta_k \hat{\alpha}^\dagger_{-\boldsymbol{k}} \tag{17.51}$$

together with (17.47), we find the similar result:

$${}_{\rm B}\langle \text{ground}|\hat{\phi}_{\rm B}(\boldsymbol{x})|\text{ground}\rangle_{\rm B} = \rho_0^{1/2}. \tag{17.52}$$

The question is now how to generalize (17.50) or (17.52) to the complete $\hat{\phi}(\boldsymbol{x})$ and the true ground state $|\text{ground}\rangle$, in the limit $N, \Omega \to \infty$ with fixed N/Ω. We make the *assumption* that

$$\langle \text{ground}|\hat{\phi}(\boldsymbol{x})|\text{ground}\rangle \neq 0; \tag{17.53}$$

that is, we abstract from the Bogoliubov model the crucial feature that *the field acquires a non-zero expectation value in the ground state*, in the infinite volume limit.

We are now at the heart of spontaneous symmetry breaking in field theory. Condition (17.53) has the form of an 'ordering' condition: it is analogous to the non-zero value of the total spin in the ferromagnetic case, but in (17.53) – we must again emphasize – $|\text{ground}\rangle$ is *not* an eigenstate of the symmetry operator $\hat{N}$; if it were, (17.53) would vanish, as we have just seen. Recalling the association 'quantum vacuum $\leftrightarrow$ many body ground state' we expect that the occurrence of a non-zero vacuum expectation value (vev) for an operator transforming non-trivially under a symmetry operator will be the key requirement for spontaneous symmetry breaking in field theory. Such operators are generically called *order parameters*. In the next section we show how this requirement necessitates one (or more) massless modes, via Goldstone's theorem (1961).

Before leaving the superfluid, we examine (17.37) and (17.52) in another way, which is only rigorous for a finite system but is nevertheless very suggestive. Since the original $\hat{H}$ has a U(1) symmetry under which $\hat{\phi}$ transforms to $\hat{\phi}' = \exp(-\mathrm{i}\alpha)\hat{\phi}$, we should be at liberty to replace (17.37) by

$$\hat{\phi}'_{\rm B} = \mathrm{e}^{-\mathrm{i}\alpha}\rho_0^{1/2} + \frac{1}{\Omega^{1/2}} \sum_{\boldsymbol{k}\neq\boldsymbol{0}} \hat{a}_{\boldsymbol{k}} \mathrm{e}^{-\mathrm{i}\alpha} \mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{x}}. \tag{17.54}$$

But in that case our condition (17.52) becomes

$${}_{\rm B}\langle \text{ground}|\hat{\phi}'_{\rm B}|\text{ground}\rangle_{\rm B} = \mathrm{e}^{-\mathrm{i}\alpha}{}_{\rm B}\langle \text{ground}|\hat{\phi}_{\rm B}|\text{ground}\rangle_{\rm B}. \tag{17.55}$$

Now $\hat{\phi}' = \hat{U}_\alpha \hat{\phi} \hat{U}_\alpha^{-1}$ where $\hat{U}_\alpha = \exp(\mathrm{i}\alpha\hat{N})$. Hence (17.55) may be written as

$${}_{\rm B}\langle \text{ground}|\hat{U}_\alpha \hat{\phi} \hat{U}_\alpha^{-1}|\text{ground}\rangle_{\rm B} = \mathrm{e}^{-\mathrm{i}\alpha}{}_{\rm B}\langle \text{ground}|\hat{\phi}_{\rm B}|\text{ground}\rangle_{\rm B}. \tag{17.56}$$

If $|\text{ground}\rangle_{\rm B}$ were an eigenstate of $\hat{N}$ with eigenvalue N, say, then the $\hat{U}_\alpha$ factors in (17.56) would become just $\mathrm{e}^{\mathrm{i}\alpha N} \cdot \mathrm{e}^{-\mathrm{i}\alpha N}$ and would cancel out, leaving a

contradiction. Instead, however, knowing that $|\text{ground}\rangle_{\text{B}}$ is not an eigenstate of $\hat{N}$, we can regard $\hat{U}_\alpha^{-1}|\text{ground}\rangle_{\text{B}}$ as an 'alternative ground state' $|\text{ground}, \alpha\rangle_{\text{B}}$ such that

$$ {}_{\text{B}}\langle\text{ground}, \alpha|\hat{\phi}|\text{ground}, \alpha\rangle_{\text{B}} = \mathrm{e}^{-\mathrm{i}\alpha}\,{}_{\text{B}}\langle\text{ground}|\hat{\phi}_{\text{B}}|\text{ground}\rangle_{\text{B}}, \tag{17.57} $$

the original choice (17.52) corresponding to $\alpha = 0$. There are infinitely many such ground states since α is a continuous parameter. No physical consequence follows from choosing one rather than another, but we do have to choose one, thus 'spontaneously' breaking the symmetry. In choosing say $\alpha = 0$, we are deciding (arbitrarily) to pick the ground state such that ${}_{\text{B}}\langle\text{ground}|\hat{\phi}|\text{ground}\rangle_{\text{B}}$ is aligned in the 'real' direction. By hypothesis, a similar situation obtains for the true ground state. None of the states $|\text{ground}, \alpha\rangle$ is an eigenstate for $\hat{N}$: instead, they are certain coherent superpositions of states with different eigenvalues N, such that the expectation value of $\hat{\phi}$ has a definite phase.

17.4 Goldstone's theorem

We return to quantum field theory proper, and show following Goldstone (1961) (see also Goldstone, Salam and Weinberg 1962) how in case (b) of the Fabri–Picasso theorem massless particles will necessarily be present. Whether these particles will actually be observable depends, however, on whether the theory also contains gauge fields. In this chapter we are concerned solely with global symmetries, and gauge fields are absent; the local symmetry case is treated in chapter 19.

Suppose, then, that we have a Lagrangian $\hat{\mathcal{L}}$ with a continuous symmetry generated by a charge $\hat{Q}$, which is independent of time, and is the space integral of the $\mu = 0$ component of a conserved Noether current:

$$ \hat{Q} = \int \hat{j}_0(x)\,\mathrm{d}^3\boldsymbol{x}. \tag{17.58} $$

We consider the case in which the vacuum of this theory is not invariant, i.e. is not annihilated by $\hat{Q}$.

Suppose $\hat{\phi}(y)$ is some field operator which is not invariant under the continuous symmetry in question, and consider the vacuum expectation value

$$ \langle 0|[\hat{Q}, \hat{\phi}(y)]|0\rangle. \tag{17.59} $$

Just as in equation (17.13), translation invariance implies that this vev is, in fact, independent of y, and we may set $y = 0$. If $\hat{Q}$ were to annihilate $|0\rangle$, the expression (17.18) would clearly vanish: we investigate the consequences of it *not* vanishing. Since $\hat{\phi}$ is not invariant under $\hat{Q}$, the commutator in (17.59) will give some other field, call it $\hat{\phi}'(y)$; thus the hallmark of the hidden symmetry

situation is the existence of some field (here $\hat{\phi}'(y)$) with *non-vanishing vacuum expectation value*, just as in (17.53).

From (17.58), we can write (17.59) as

$$0 \neq \langle 0|\hat{\phi}'(y)|0\rangle \tag{17.60}$$

$$= \langle 0|[\int \mathrm{d}^3\boldsymbol{x}\hat{j}_0(x), \hat{\phi}(y)]|0\rangle. \tag{17.61}$$

Since, by assumption, $\partial_\mu \hat{j}^\mu = 0$, we have as usual

$$\frac{\partial}{\partial x^0}\int \mathrm{d}^3\boldsymbol{x}\hat{j}_0(x) + \int \mathrm{d}^3\boldsymbol{x}\,\mathrm{div}\hat{\boldsymbol{j}}(\mathrm{x}) = 0, \tag{17.62}$$

whence

$$\frac{\partial}{\partial x^0}\int \mathrm{d}^3\boldsymbol{x}\langle 0|[\hat{j}_0(x), \hat{\phi}(y)]|0\rangle = -\int \mathrm{d}^3\boldsymbol{x}\langle 0|[\mathrm{div}\hat{\boldsymbol{j}}(x), \hat{\phi}(y)]|0\rangle \tag{17.63}$$

$$= -\int \mathrm{d}\boldsymbol{S}\cdot\langle 0|[\hat{\boldsymbol{j}}(x), \hat{\phi}(y)]|0\rangle. \tag{17.64}$$

If the surface integral vanishes in (17.64), (17.61) will be independent of x_0. The commutator in (17.64) involves local operators separated by a very large space-like interval, and therefore the vanishing of (17.64) would seem to be unproblematic. Indeed so it is – with the exception of the case in which the symmetry is local and gauge fields are present. A detailed analysis of exactly how this changes the argument being presented here will take us too far afield at this point, and the reader is referred to Guralnik *et al.* (1968) and Bernstein (1974). We shall treat the 'spontaneously broken' gauge theory case in chapter 19, but in less formal terms.

Let us now see how the independence of (17.61) on x_0 leads to the necessity for a massless particle in the spectrum. Inserting a complete set of states in (17.61), we obtain

$$0 \neq \int \mathrm{d}^3\boldsymbol{x}\sum_n\{\langle 0|\hat{j}_0(x)|n\rangle\langle n|\hat{\phi}(y)|0\rangle - \langle 0|\hat{\phi}(y)|n\rangle\langle n|\hat{j}_0(x)|0\rangle\} \tag{17.65}$$

$$= \int \mathrm{d}^3\boldsymbol{x}\sum_n\{\langle 0|\hat{j}_0(0)|n\rangle\langle n|\hat{\phi}(y)|0\rangle \mathrm{e}^{-\mathrm{i}p_n\cdot x} - \langle 0|\hat{\phi}(y)|n\rangle\langle n|\hat{j}_0(0)|0\rangle \mathrm{e}^{\mathrm{i}p_n\cdot x}\} \tag{17.66}$$

using translation invariance, with p_n the 4-momentum eigenvalue of the state $|n\rangle$. Performing the spatial integral on the right-hand side we find (omitting the irrelevant $(2\pi)^3$)

$$0 \neq \sum_n \delta^3(\boldsymbol{p}_n)[\langle 0|\hat{j}_0(0)|n\rangle\langle n|\hat{\phi}(y)|0\rangle \mathrm{e}^{\mathrm{i}p_{n0}x_0} - \langle 0|\hat{\phi}(y)|n\rangle\langle n|\hat{j}_0(0)|0\rangle \mathrm{e}^{-\mathrm{i}p_{n0}x_0}]. \tag{17.67}$$

But this expression is independent of x_0. *Massive* states $|n\rangle$ will produce explicit x_0-dependent factors $\mathrm{e}^{\pm \mathrm{i}M_n x_0}$ ($p_{n0} \to M_n$ as the δ-function constrains $\boldsymbol{p}_n = \mathbf{0}$), hence the matrix elements of $\hat{j}_0$ between $|0\rangle$ and such a massive state must *vanish*, and such states contribute zero to (17.67). Equally, if we take $|n\rangle = |0\rangle$, (17.67) vanishes identically. But it has been assumed to be not zero. Hence *some* state or states must exist among $|n\rangle$ such that $\langle 0|j_0|n\rangle \neq 0$ and yet (17.67) is independent of x_0. The only possibility is states whose energy p_{n0} goes to zero as their 3-momentum does (from $\delta^3(\boldsymbol{p}_n)$). Such states are, of course, massless; they are called generically *Goldstone modes*. Thus the existence of a non-vanishing vacuum expectation value for a field, in a theory with a continuous symmetry, appears to lead inevitably to the necessity of having a massless particle, or particles, in the theory. This is the Goldstone result.

The superfluid provided us with an explicit model exhibiting the crucial non-zero expectation value $\langle \mathrm{ground}|\hat{\phi}|\mathrm{ground}\rangle \neq 0$, in which the now expected massless mode emerged dynamically. We now discuss a simpler, relativistic model, in which the symmetry breaking is brought about more 'by hand' – that is, by choosing a parameter in the Lagrangian appropriately. Although in a sense less 'dynamical' than the Bogoliubov superfluid (or the BCS superconductor, to be discussed shortly) this *Goldstone model* does provide a very simple example of the phenomenon of spontaneous symmetry breaking in field theory.

17.5 Spontaneously broken global U(1) symmetry: the Goldstone model

We consider, following Goldstone (1961), a complex scalar field $\hat{\phi}$ as in section 7.1, with

$$\hat{\phi} = \frac{1}{\sqrt{2}}(\hat{\phi}_1 - \mathrm{i}\hat{\phi}_2), \qquad \hat{\phi}^\dagger = \frac{1}{\sqrt{2}}(\hat{\phi}_1 + \mathrm{i}\hat{\phi}_2), \tag{17.68}$$

described by the Lagrangian

$$\hat{\mathcal{L}}_{\mathrm{G}} = (\partial_\mu \hat{\phi}^\dagger)(\partial^\mu \hat{\phi}) - \hat{V}(\hat{\phi}). \tag{17.69}$$

We begin by considering the 'normal' case in which the potential has the form

$$\hat{V} = \hat{V}_{\mathrm{S}} \equiv \frac{1}{4}\lambda(\hat{\phi}^\dagger\hat{\phi})^2 + \mu^2\hat{\phi}^\dagger\hat{\phi} \tag{17.70}$$

with $\mu^2, \lambda > 0$. The Hamiltonian density is then

$$\hat{\mathcal{H}}_{\mathrm{G}} = \dot{\hat{\phi}}^\dagger\dot{\hat{\phi}} + \boldsymbol{\nabla}\hat{\phi}^\dagger \cdot \boldsymbol{\nabla}\hat{\phi} + \hat{V}(\hat{\phi}). \tag{17.71}$$

Clearly $\hat{\mathcal{L}}_{\rm G}$ is invariant under the global U(1) symmetry

$$\hat{\phi} \to \hat{\phi}' = e^{-{\rm i}\alpha}\hat{\phi}, \tag{17.72}$$

the generator being $\hat{N}_\phi$ of (7.23). We shall see how this symmetry may be 'spontaneously broken'.

We know that everything depends on the nature of the ground state of this field system – that is, the vacuum of the quantum field theory. In general, it is a difficult, non-perturbative, problem to find the ground state (or a good approximation to it – witness the superfluid). But we can make some progress by first considering the theory *classically*. It is clear that the absolute minimum of the classical Hamiltonian $\mathcal{H}_{\rm G}$ is reached for

(i) $\phi =$ constant, which reduces the $\dot{\phi}$ and $\boldsymbol{\nabla}\phi$ terms to zero;

(ii) $\phi = \phi_0$ where ϕ_0 is the minimum of the classical version of the potential, V.

For $V = V_{\rm S}$ as in (17.70) but without the hats, and with λ and μ^2 both positive, the minimum of $V_{\rm S}$ is clearly at $\phi = 0$, and is unique. In the quantum theory, we expect to treat small oscillations of the field about this minimum as approximately harmonic, leading to the usual quantized modes. To implement this, we expand $\hat{\phi}$ about the classical minimum at $\phi = 0$, writing as usual

$$\hat{\phi} = \int \frac{{\rm d}^3\boldsymbol{k}}{(2\pi)^3\sqrt{2\omega}}[\hat{a}(k){\rm e}^{-{\rm i}k\cdot x} + b^\dagger(k){\rm e}^{{\rm i}k.x}] \tag{17.73}$$

where the plane waves are solutions of the 'free' ($\lambda = 0$) problem. For $\lambda = 0$ the Lagrangian is simply

$$\hat{\mathcal{L}}_{\rm free} = \partial_\mu\hat{\phi}^\dagger\partial^\mu\hat{\phi} - \mu^2\hat{\phi}^\dagger\hat{\phi}, \tag{17.74}$$

which represents a complex scalar field, consisting of two degrees of freedom, each with the same mass μ (see section 7.1). Thus in (17.73) $\omega = (\boldsymbol{k}^2+\mu^2)^{1/2}$, and the vacuum is defined by

$$\hat{a}(k)|0\rangle = \hat{b}(k)|0\rangle = 0, \tag{17.75}$$

and so clearly

$$\langle 0|\hat{\phi}|0\rangle = 0. \tag{17.76}$$

It seems reasonable to interpret quantum field average values as corresponding to classical field values, and on this interpretation (17.76) is consistent with the fact that the classical minimum energy configuration has $\phi = 0$.

Consider now the case in which the classical minimum is *not* at $\phi = 0$. This can be achieved by altering the sign of μ^2 in (17.70) 'by hand', so that the classical potential is now the 'symmetry breaking' one

$$V = V_{\rm SB} \equiv \frac{1}{4}\lambda(\phi^\dagger\phi)^2 - \mu^2\phi^\dagger\phi. \tag{17.77}$$

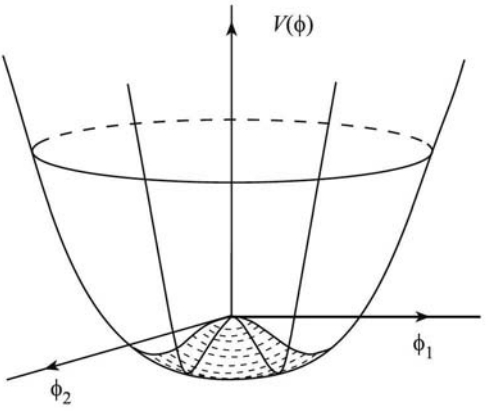


FIGURE 17.2
The classical potential $V_{\rm SB}$ of (17.77).

This is sketched versus ϕ_1 and ϕ_2 in figure 17.2. This time, although the origin $\phi_1 = \phi_2 = 0$ is a stationary point, it is an (unstable) maximum rather than a minimum. The minimum of $V_{\rm SB}$ occurs when

$$(\phi^\dagger \phi) = \frac{2\mu^2}{\lambda}, \tag{17.78}$$

or alternatively when

$$\phi_1^2 + \phi_2^2 = \frac{4\mu^2}{\lambda} \equiv v^2 \tag{17.79}$$

where

$$v = \frac{2|\mu|}{\lambda^{1/2}}. \tag{17.80}$$

The condition (17.79) can also be written as

$$|\phi| = v/\sqrt{2}. \tag{17.81}$$

To have a clearer picture, it is helpful to introduce the 'polar' variables $\rho(x)$ and $\theta(x)$ via

$$\phi(x) = (\rho(x)/\sqrt{2}) \exp(\mathrm{i}\theta(x)/v) \tag{17.82}$$

where for convenience the v is inserted so that θ has the same dimension (mass) as ρ and ϕ. The minimum condition (17.81) therefore represents the circle $\rho = v$; any point on this circle, at any value of θ, represents a possible classical ground state – and it is clear that they are (infinitely) degenerate.

Before proceeding further, we briefly outline a condensed matter analogue of (17.77) and (17.81) which may help in understanding the change in sign of the parameter μ^2. Consider the free energy F of a ferromagnet as a function

of the magnetization $\boldsymbol{M}$ at temperature T, and make an expansion of the form

$$F \approx F_0(T) + \mu^2(T)\boldsymbol{M}^2 + \frac{\lambda}{4}\boldsymbol{M}^4 + \cdots \tag{17.83}$$

valid for weak and slowly varying magnetization. If the parameter μ^2 is positive, it is clear that F has a simple 'bowl' shape as a function of $|\boldsymbol{M}|$, with a minimum at $|\boldsymbol{M}| = 0$. This is the case for T greater than the ferromagnetic transition temperature $T_{\rm C}$. However, if one assumes that $\mu^2(T)$ changes sign at $T_{\rm C}$, becoming negative for $T < T_{\rm C}$, then F will now resemble a vertical section of figure 17.2, the minimum being at $|\boldsymbol{M}| \neq 0$. Any direction of $\boldsymbol{M}$ is possible (only $|\boldsymbol{M}|$ is specified); but the system must choose one particular direction (e.g. via the influence of a very weak external field, as discussed in section 17.3.1), and when it does so the rotational invariance exhibited by F of (17.83) is lost. This symmetry has been broken 'spontaneously' – though this is still only a classical analogue. Nevertheless, the model is essentially the Landau mean field theory of ferromagnetism, and suggests that we should think of the 'symmetric' and 'broken symmetry' situations as different phases of the same system. It may also be the case in particle physics, that parameters such as μ^2 change sign as a function of T, or some other variable, thereby effectively precipitating a phase change.

If we maintain the idea that the vacuum expectation value of the quantum field should equal the ground state value of the classical field, the vacuum in this $\mu^2 < 0$ case must therefore be $|0\rangle_{\rm B}$ such that ${}_{\rm B}\langle 0|\hat{\phi}|0\rangle_{\rm B}$ does not vanish, in contrast to (17.76). It is clear that this is exactly the situation met in the superfluid (but 'B' here will stand for 'broken symmetry'), and is moreover the condition for the existence of massless (Goldstone) modes. Let us see how they emerge in this model.

In quantum field theory, particles are thought of as excitations from a ground state, which is the vacuum. Figure 17.2 strongly suggests that if we want a sensible quantum interpretation of a theory with the potential (17.77), we had better expand the fields about a point on the circle of minima, about which stable oscillations are likely, rather than about the obviously unstable point $\hat{\phi} = 0$. Let us pick the point $\rho = v$, $\theta = 0$ in the classical case. We might well guess that 'radial' oscillations in $\hat{\rho}$ would correspond to a conventional massive field (having a parabolic restoring potential), while 'angle' oscillations in $\hat{\theta}$ – which pass through all the degenerate vacuua – have no restoring force and are massless. Accordingly, we set

$$\hat{\phi}(x) = \frac{1}{\sqrt{2}}(v + \hat{h}(x)) \exp(-\mathrm{i}\hat{\theta}(x)/v) \tag{17.84}$$

and find (problem 17.5) that $\hat{\mathcal{L}}_{\rm G}$ (with $\hat{V} = \hat{V}_{\rm SB}$ of (17.77) with hats on) becomes

$$\hat{\mathcal{L}}_{\rm G} \quad = \quad \frac{1}{2}\partial_\mu \hat{h} \partial^\mu \hat{h} - \mu^2 \hat{h}^2 + \frac{1}{2}\partial_\mu \hat{\theta} \partial^\mu \hat{\theta} + \mu^4/\lambda$$

$$+\frac{\hat{h}}{v}\partial_\mu\hat{\theta}\partial^\mu\hat{\theta}+\frac{1}{2}\frac{\hat{h}^2}{v^2}\partial_\mu\hat{\theta}\partial^\mu\hat{\theta}-\frac{\lambda}{16}v\hat{h}^3-\frac{\lambda}{16}\hat{h}^4, \qquad (17.85)$$

Equation (17.85) is very important. First of all, the first line shows that the particle spectrum in the 'spontaneously broken' case is dramatically different from that in the normal case: instead of two degrees of freedom with the same mass μ, one (the θ-mode) is massless, and the other (the h-mode) has a mass of $\sqrt{2}\mu$. We expect the vacuum $|0\rangle_{\rm B}$ to be annihilated by the mode operators $\hat{a}_h$ and $\hat{a}_\theta$ for these fields. This implies, however, that

$${}_{\rm B}\langle 0|\hat{\phi}|0\rangle_{\rm B} = v/\sqrt{2} \qquad (17.86)$$

which is consistent with our interpretation of the vacuum expectation value (vev) as the classical minimum, and with the occurrence of massless modes. (The constant term in (17.85), which does not affect equations of motion, merely reflects the fact that the minimum value of $V_{\rm SB}$ is $-\mu^4/\lambda$.) The ansatz (17.84) and the non-zero vev (17.86) may be compared with (17.37) and (17.52), respectively, in the superfluid case.

Secondly, the second line of equation (17.85) shows that only the *derivative* of the $\hat{\theta}$ field appears in the interaction terms, whereas this is not true of the $\hat{h}$ field. Indeed, the Lagrangian for the θ-mode cannot have any dependence on a *constant* value of $\hat{\theta}$, since this could be transformed away by a global U(1) transformation (17.72), which is a symmetry of the theory, and under which $\hat{\theta} \to \hat{\theta} + v\alpha$. This will be an important point to remember when we consider effective Lagrangians for Goldstone modes in section 18.3.

Goldstone's model, then, contains much of the essence of spontaneous symmetry breaking in field theory: a non-zero vacuum value of a field which is not an invariant under the symmetry group, zero mass bosons, and massive excitations in a direction in field space which is 'orthogonal' to the degenerate ground states. However, it has to be noted that the triggering mechanism for the symmetry breaking ($\mu^2 \to -\mu^2$) has to be put in by hand, in contrast to the – admittedly approximate, but more 'dynamical' – Bogoliubov approach. The Goldstone model, in short, is essentially phenomenological.

As in the case of the superfluid, we may perfectly well choose a vacuum corresponding to a classical ground state with non-zero θ, say $\theta = -v\alpha$. Then

$${}_{\rm B}\langle 0,\alpha|\hat{\phi}|0,\alpha\rangle_{\rm B} = {\rm e}^{-{\rm i}\alpha}\frac{v}{\sqrt{2}} \qquad (17.87)$$

$$= {\rm e}^{-{\rm i}\alpha}\,{}_{\rm B}\langle 0|\hat{\phi}|0\rangle_{\rm B}, \qquad (17.88)$$

as in (17.57). But we know (see (7.27) and (7.28)) that

$${\rm e}^{-{\rm i}\alpha}\hat{\phi} = \hat{\phi}' = \hat{U}_\alpha\hat{\phi}\hat{U}_\alpha^{-1} \qquad (17.89)$$

where

$$\hat{U}_\alpha = {\rm e}^{{\rm i}\alpha\hat{N}_\phi}. \qquad (17.90)$$

So (17.88) becomes

$$ {}_{\rm B}\langle 0,\alpha|\hat{\phi}|0,\alpha\rangle_{\rm B} = {}_{\rm B}\langle 0|\hat{U}_\alpha\hat{\phi}\hat{U}_\alpha^{-1}|0\rangle_{\rm B} \tag{17.91} $$

and we may interpret $\hat{U}_\alpha^{-1}|0\rangle_{\rm B}$ as the 'alternative vacuum' $|0,\alpha\rangle_{\rm B}$ (this argument is, as usual, not valid in the infinite volume limit where $\hat{N}_\phi$ fails to exist).

It is interesting to find out what happens to the symmetry current corresponding to the invariance (17.72), in the 'broken symmetry' case. This current is given in (7.23) which we write again here in slightly different notation:

$$ \hat{j}^\mu_\phi = {\rm i}(\hat{\phi}^\dagger\partial^\mu\hat{\phi} - (\partial^\mu\hat{\phi})^\dagger\hat{\phi}), \tag{17.92} $$

normal ordering being understood. Written in terms of the $\hat{h}$ and $\hat{\theta}$ of (17.84), $\hat{j}^\mu_\phi$ becomes

$$ \hat{j}^\mu_\phi = v\partial^\mu\hat{\theta} + 2\hat{h}\partial^\mu\hat{\theta} + \hat{h}^2\partial^\mu\hat{\theta}/v. \tag{17.93} $$

The term involving just the *single* field $\hat{\theta}$ is very remarkable: it tells us that there is a non-zero matrix element of the form

$$ {}_{\rm B}\langle 0|\hat{j}^\mu_\phi(x)|\theta,p\rangle = -{\rm i}p^\mu v{\rm e}^{-{\rm i}p\cdot x} \tag{17.94} $$

where $|\theta,p\rangle$ stands for the state with one θ-quantum (Goldstone boson), with momentum p^μ. This is easily seen by writing the usual normal mode expansion for $\hat{\theta}$, and using the standard bosonic commutation relations for $\hat{a}_\theta(k), \hat{a}^\dagger_\theta(k')$. In words, (17.94) asserts that, *when the symmetry is spontaneously broken, the symmetry current connects the vacuum to a state with one Goldstone quantum, with an amplitude which is proportional to the symmetry breaking vacuum expectation value* v, and which vanishes as the 4-momentum goes to zero. The matrix element (17.94), with $x = 0$, is precisely of the type that was shown to be non-zero in the proof of the Goldstone theorem, after (17.67). Note also that (17.94) is consistent with $\partial_\mu\hat{j}^\mu_\phi = 0$ only if $p^2 = 0$, as is required for the massless θ.

We are now ready to generalize the Abelian U(1) model to the (global) non-Abelian case.

17.6 Spontaneously broken global non-Abelian symmetry

We can illustrate the essential features by considering a particular example, which in fact forms part of the Higgs sector of the Standard Model. We consider an SU(2) doublet, but this time not of fermions as in section 12.3,

but of bosons:

$$\hat{\phi} = \begin{pmatrix} \hat{\phi}^+ \\ \hat{\phi}^0 \end{pmatrix} \equiv \begin{pmatrix} \frac{1}{\sqrt{2}}(\phi_1 + \mathrm{i}\phi_2) \\ \frac{1}{\sqrt{2}}(\phi_3 + \mathrm{i}\phi_4) \end{pmatrix} \tag{17.95}$$

where the complex scalar field $\hat{\phi}^+$ destroys positively charged particles and creates negatively charged ones, and the complex scalar field $\hat{\phi}^0$ destroys neutral particles and creates neutral antiparticles. As we shall see in a moment, the Lagrangian we shall use has an additional U(1) symmetry, so that the full symmetry is SU(2) × U(1). This U(1) symmetry leads to a conserved quantum number which we call y. We associate the physical charge Q with the eigenvalue t_3 of the SU(2) generator $\hat{t}_3$, and with y, via

$$Q = e(t_3 + y/2) \tag{17.96}$$

so that $y(\phi^+) = 1 = y(\phi^0)$. Thus ϕ^+ and ϕ^0 can be thought of as analogous to the hadronic iso-doublet $(\mathrm{K}^+, \mathrm{K}^0)$.

The Lagrangian we choose is a simple generalization of (17.69) and (17.77):

$$\hat{\mathcal{L}}_\Phi = (\partial_\mu \hat{\phi}^\dagger)(\partial^\mu \hat{\phi}) + \mu^2 \hat{\phi}^\dagger \hat{\phi} - \frac{\lambda}{4}(\hat{\phi}^\dagger \hat{\phi})^2 \tag{17.97}$$

which has the 'spontaneous symmetry breaking' choice of sign for the parameter μ^2. Plainly, for the 'normal' sign of μ^2, in which '$+\mu^2 \hat{\phi}^\dagger \hat{\phi}$' is replaced by '$-\mu^2 \hat{\phi}^\dagger \hat{\phi}$', with μ^2 positive in both cases, the free ($\lambda = 0$) part would describe a complex doublet, with four degrees of freedom, each with the same mass μ. Let us see what happens in the broken symmetry case.

For the Lagrangian (17.97) with $\mu^2 > 0$, the minimum of the classical potential is at the point

$$(\phi^\dagger \phi)_{\min} = 2\mu^2/\lambda \equiv v^2/2. \tag{17.98}$$

As in the U(1) case, we interpret (17.98) as a condition on the vev of $\hat{\phi}^\dagger \hat{\phi}$,

$$\langle 0|\hat{\phi}^\dagger \hat{\phi}|0\rangle = v^2/2. \tag{17.99}$$

Before proceeding we note that (17.97) is invariant under global SU(2) transformations

$$\hat{\phi} \to \hat{\phi}' = \exp(-\mathrm{i}\boldsymbol{\alpha} \cdot \boldsymbol{\tau}/2)\hat{\phi} \tag{17.100}$$

but also under a separate global U(1) transformation

$$\hat{\phi} \to \hat{\phi}' = \exp(-\mathrm{i}\alpha)\hat{\phi} \tag{17.101}$$

where α is to be distinguished from $\boldsymbol{\alpha} \equiv (\alpha_1, \alpha_2, \alpha_3)$. The symmetry is then referred to as SU(2) × U(1), which is the symmetry of the electroweak sector of the Standard Model, except that in that case it is a *local* symmetry.

As before, in order to get a sensible particle spectrum we must expand the fields $\hat{\phi}$ not about $\hat{\phi} = 0$ but about a point satisfying the stable ground state

(vacuum) condition (17.98). That is, we need to define '$\langle 0|\hat{\phi}|0\rangle$' and expand about it, as in (17.84). In the present case, however, the situation is more complicated than (17.84) since the complex doublet (17.95) contains four real fields as indicated in (17.95), and (17.98) becomes

$$\langle 0|\hat{\phi}_1^2 + \hat{\phi}_2^2 + \hat{\phi}_3^2 + \hat{\phi}_4^2|0\rangle = v^2. \tag{17.102}$$

It is evident that we have a lot of freedom in choosing the $\langle 0|\hat{\phi}_i|0\rangle$ so that (17.102) holds, and it is not at first obvious what an appropriate generalization of (17.84) and (17.85) might be.

Furthermore, in this more complicated (non-Abelian) situation a qualitatively new feature can arise: it may happen that the chosen condition $\langle 0|\hat{\phi}_i|0\rangle \neq 0$ is *invariant* under some subset of the allowed symmetry transformations. This would effectively mean that this particular choice of the vacuum state respected that subset of symmetries, which would therefore not be 'spontaneously broken' after all. Since each broken symmetry is associated with a massless Goldstone boson, we would then get fewer of these bosons than expected. Just this happens (by design) in the present case.

Suppose, then, that we could choose the $\langle 0|\hat{\phi}_i|0\rangle$ so as to break this SU(2) $\times$ U(1) symmetry completely: we would then expect four massless fields. Actually, however, it is not possible to make such a choice. An analogy may make this point clearer. Suppose we were considering just SU(2), and the field '$\hat{\phi}$' was an SU(2)-triplet, $\hat{\boldsymbol{\phi}}$. Then we could always write $\langle 0|\hat{\boldsymbol{\phi}}|0\rangle = v\boldsymbol{n}$ where $\boldsymbol{n}$ is a unit vector; but this form is invariant under rotations about the $\boldsymbol{n}$-axis, irrespective of where that points. In the present case, by using the freedom of global SU(2) $\times$ U(1) phase changes, an arbitrary $\langle 0|\hat{\phi}|0\rangle$ can be brought to the form

$$\langle 0|\hat{\phi}|0\rangle = \begin{pmatrix} 0 \\ v/\sqrt{2} \end{pmatrix}. \tag{17.103}$$

In considering what symmetries are respected or broken by (17.103), it is easiest to look at infinitesimal transformations. It is then clear that the particular transformation

$$\delta\hat{\phi} = -\mathrm{i}\epsilon(1+\tau_3)\hat{\phi} \tag{17.104}$$

(which is a combination of (17.101) and the 'third component' of (17.100)) is still a symmetry of (17.103) since

$$(1+\tau_3)\begin{pmatrix} 0 \\ v/\sqrt{2} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \tag{17.105}$$

so that

$$\langle 0|\phi|0\rangle = \langle 0|\phi + \delta\phi|0\rangle; \tag{17.106}$$

we say that 'the vacuum is invariant under (17.104)', and when we look at the spectrum of oscillations about that vacuum we expect to find only three massless bosons, not four.

Oscillations about (17.103) are conveniently parametrized by

$$\hat{\phi} = \exp(-\mathrm{i}\hat{\boldsymbol{\theta}}(x)\cdot\boldsymbol{\tau}/2v)\begin{pmatrix} 0 \\ \frac{1}{\sqrt{2}}(v+\hat{H}(x)) \end{pmatrix}, \tag{17.107}$$

which is to be compared with (17.84). Inserting (17.107) into (17.97) (see problem 17.6) we easily find that no mass term is generated for the θ fields, while the H field piece is

$$\hat{\mathcal{L}}_H = \frac{1}{2}\partial_\mu\hat{H}\partial^\mu\hat{H} - \mu^2\hat{H}^2 + \text{interactions} \tag{17.108}$$

just as in (17.85), showing that $m_H = \sqrt{2}\mu$.

Let us now note carefully that whereas in the 'normal symmetry' case with the opposite sign for the μ^2 term in (17.97), the free-particle spectrum consisted of a degenerate doublet of four degrees of freedom all with the same mass μ, in the 'spontaneously broken' case no such doublet structure is seen: instead, there is one massive scalar field, and three massless scalar fields. The number of degrees of freedom is the same in each case, but the physical spectrum is completely different.

In the application of this to the electroweak sector of the Standard Model, the SU(2) × U(1) symmetry will be 'gauged' (i.e. made local), which is easily done by replacing the ordinary derivatives in (17.97) by suitable covariant ones. We shall see in chapter 19 that the result, with the choice (17.107), will be to end up with three *massive* gauge fields (those mediating the weak interactions) and one *massless* gauge field (the photon). We may summarize this (anticipated) result by saying, then, that when a spontaneously broken non-Abelian symmetry is gauged, those gauge fields corresponding to symmetries that are broken by the choice of $\langle 0|\hat{\phi}|0\rangle$ acquire a mass, while those that correspond to symmetries that are respected by $\langle 0|\hat{\phi}|0\rangle$ do not. Exactly how this happens will be the subject of chapter 19.

We end this chapter by considering a second important example of spontaneous symmetry breaking in condensed matter physics, as a preliminary to our discussion of chiral symmetry breaking in the following chapter.

17.7 The BCS superconducting ground state

We shall not attempt to provide a self-contained treatment of the Bardeen–Cooper–Schrieffer (1957) – or BCS – theory; rather, we wish simply to focus on one aspect of the theory, namely the occurrence of an *energy gap* separating the ground state from the lowest excited levels of the fermionic energy spectrum. The existence of such a gap is a fundamental ingredient of the theory of superconductivity; in the following chapter we shall see how Nambu (1960)

interpreted a chiral symmetry breaking fermionic mass term as an analogous 'gap'. We emphasize at the outset that we shall here not treat electromagnetic interactions in the superconducting state, leaving that topic for chapter 19.

Our discussion will deliberately have some similarity to that of section 17.3.2. In the present case, of course, we shall be dealing with fermions – namely electrons – rather than the bosons of a superfluid. Nevertheless, we shall see that a similar kind of 'condensation' occurs in the superconductor too. Naturally, such a phenomenon can only occur for bosons. Thus an essential element in the BCS theory is the identification of a mechanism whereby pairs of electrons become correlated, the behaviour of which may have some similarity to that of bosons. Now, direct Coulomb interaction between a pair of electrons is repulsive, and it remains so despite the screening that occurs in a solid. But the positively charged ions do provide sources of attraction for the electrons, and may be used as intermediaries (via 'electron-phonon interactions') to promote an effective attraction between electrons in certain circumstances. At this point we recall the characteristic feature of a weakly interacting gas of electrons at zero temperature: thanks to the Exclusion Principle, the electrons populate single particle energy levels up to some maximum energy $E_{\rm F}$ (the Fermi energy), whose value is fixed by the electron density. It turns out (see for example Kittel 1987, chapter 8) that electron–electron scattering, mediated by phonon exchange, leads to an effective attraction between two electrons whose energies ϵ_k lie in a thin band $E_{\rm F} - \omega_{\rm D} < \epsilon_k < E_{\rm F} + \omega_{\rm D}$ around $E_{\rm F}$, where $\omega_{\rm D}$ is the Debye frequency associated with lattice vibrations. Cooper (1956) was the first to observe that the Fermi 'sea' was unstable with respect to the formation of bound pairs, in the presence of an attractive interaction. What this means is that the energy of the system can be lowered by exciting a pair of electrons above $E_{\rm F}$, which then become bound to a state with a total energy less than $2E_{\rm F}$. This instability modifies the Fermi sea in a fundamental way: a sort of 'condensate' of pairs is created around the Fermi energy, and we need a many-body formalism to handle the situation.

For simplicity we shall consider pairs of equal and opposite momentum $\boldsymbol{k}$, so their total momentum is zero. It can also be argued that the effective attraction will be greater when the spins are antiparallel, but the spin will not be indicated explicitly in what follows: '$\boldsymbol{k}$' will stand for '$\boldsymbol{k}$ with spin up', and '$-\boldsymbol{k}$' for '$-\boldsymbol{k}$ with spin down'. With this by way of motivation, we thus arrive at the *BCS reduced Hamiltonian*

$$\hat{H}_{\rm BCS} = \sum_{\boldsymbol{k}} \epsilon_k \hat{c}^\dagger_{\boldsymbol{k}} \hat{c}_{\boldsymbol{k}} - V \sum_{\boldsymbol{k},\boldsymbol{k}'} \hat{c}^\dagger_{\boldsymbol{k}'} \hat{c}^\dagger_{-\boldsymbol{k}'} \hat{c}_{-\boldsymbol{k}} \hat{c}_{\boldsymbol{k}} \tag{17.109}$$

which is the starting point of our discussion. In (17.109), the $\hat{c}$'s are fermionic operators obeying the usual anticommutation relations, and the ground state is such that $\hat{c}_{\boldsymbol{k}}|0\rangle = 0$. The sum is over states lying near $E_{\rm F}$, as above, and the single particle energies ϵ_k are measured relative to $E_{\rm F}$. The constant V (with the minus sign in front) represents a simplified form of the effective electron–electron attraction. Note that, in the non-interacting ($V = 0$) part,

$\hat{c}^\dagger_{\boldsymbol{k}}\hat{c}_{\boldsymbol{k}}$ is the number operator for the electrons, which because of the Pauli Principle has eigenvalues 0 or 1; this term is of course completely analogous to (7.55), and sums the single particle energies ϵ_k for each occupied level.

We immediately note that $\hat{H}_{\rm BCS}$ is invariant under the global U(1) transformation

$$\hat{c}_{\boldsymbol{k}} \to \hat{c}'_{\boldsymbol{k}} = e^{-{\rm i}\alpha}\hat{c}_{\boldsymbol{k}} \tag{17.110}$$

for all $\boldsymbol{k}$, which is equivalent to $\hat{\psi}'(\boldsymbol{x}) = {\rm e}^{-{\rm i}\alpha}\hat{\psi}(\boldsymbol{x})$ for the electron field operator at $\boldsymbol{x}$. Thus fermion number is conserved by $\hat{H}_{\rm BCS}$. However, just as for the superfluid, we shall see that the BCS ground state does not respect the symmetry.

We follow Bogoliubov (1958) and Bogoliubov *et al.* (1959) (see also Valatin 1958), and make a canonical transformation on the operators $\hat{c}_{\boldsymbol{k}}$, $\hat{c}^\dagger_{-\boldsymbol{k}}$ similar to the one employed for the superfluid problem in (17.38), as motivated by the 'pair condensate' picture. We set

$$\begin{aligned} \hat{\beta}_{\boldsymbol{k}} &= u_k\hat{c}_{\boldsymbol{k}} - v_k\hat{c}^\dagger_{-\boldsymbol{k}}, \qquad \beta^\dagger_{\boldsymbol{k}} = u_k\hat{c}^\dagger_{\boldsymbol{k}} - v_k\hat{c}_{-\boldsymbol{k}} \\ \hat{\beta}_{-\boldsymbol{k}} &= u_k\hat{c}_{-\boldsymbol{k}} + v_k\hat{c}^\dagger_{\boldsymbol{k}}, \qquad \beta^\dagger_{-\boldsymbol{k}} = u_k\hat{c}^\dagger_{-\boldsymbol{k}} + v_k\hat{c}_{\boldsymbol{k}} \end{aligned} \tag{17.111}$$

where u_k and v_k are real, depend only on $k = |\boldsymbol{k}|$, and are chosen so as to preserve *anti*commutation relations for the β's. This last condition implies (problem 17.7)

$$u_k^2 + v_k^2 = 1 \tag{17.112}$$

so that we may conveniently set

$$u_k = \cos\theta_k, v_k = \sin\theta_k. \tag{17.113}$$

Just as in the superfluid case, the transformations (17.111) only make sense in the context of a number non-conserving ground state, since they do not respect the symmetry (17.110). Although $\hat{H}_{\rm BCS}$ of (17.109) is number conserving, we shall shortly make a crucial number non-conserving approximation.

We seek a diagonalization of (17.109), analogous to (17.40), in terms of the mode operators $\hat{\beta}$ and $\hat{\beta}^\dagger$:

$$\hat{H}_{\rm BCS} = \sum_{\boldsymbol{k}} \omega_k(\hat{\beta}^\dagger_{\boldsymbol{k}}\hat{\beta}_{\boldsymbol{k}} + \hat{\beta}^\dagger_{-\boldsymbol{k}}\hat{\beta}_{-\boldsymbol{k}}) + \gamma. \tag{17.114}$$

It is easy to check (problem 17.8) that the form (17.114) implies

$$[\hat{H}_{\rm BCS}, \hat{\beta}^\dagger_{\boldsymbol{l}}] = \omega_l\hat{\beta}^\dagger_{\boldsymbol{l}} \tag{17.115}$$

as in (17.41), despite the fact that the operators obey *anti*commutation relations. Equation (17.115) then implies that the ω_k are the energies of states created by the *quasiparticle operators* $\hat{\beta}^\dagger_{\boldsymbol{k}}$ and $\hat{\beta}^\dagger_{-\boldsymbol{k}}$, the ground state being defined by

$$\hat{\beta}_{\boldsymbol{k}}|{\rm ground}\rangle_{\rm BCS} = \hat{\beta}_{-\boldsymbol{k}}|{\rm ground}\rangle_{\rm BCS} = 0. \tag{17.116}$$

Substituting for $\hat{\beta}^\dagger_{\boldsymbol{l}}$ in (17.115) from (17.111) we therefore require

$$[\hat{H}_{\rm BCS}, \cos\theta_l\, \hat{c}^\dagger_{\boldsymbol{l}} - \sin\theta_l\, \hat{c}_{-\boldsymbol{l}}] = \omega_l(\cos\theta_l\, \hat{c}^\dagger_{\boldsymbol{l}} - \sin\theta_l\, \hat{c}_{-\boldsymbol{l}}), \tag{17.117}$$

which must hold as an identity in the $\hat{c}_{\boldsymbol{l}}$'s and $\hat{c}^\dagger_{\boldsymbol{l}}$'s. Evaluating (17.117) one obtains (problem 17.9)

$$(\omega_l - \epsilon_l)\cos\theta_l - V\sin\theta_l \textstyle\sum_{\boldsymbol{k}} \hat{c}_{-\boldsymbol{k}}\hat{c}_{\boldsymbol{k}} = 0 \tag{17.118}$$

$$-V\cos\theta_l \textstyle\sum_{\boldsymbol{k}} \hat{c}^\dagger_{\boldsymbol{k}}\hat{c}^\dagger_{-\boldsymbol{k}} + (\omega_l + \epsilon_l)\sin\theta_l = 0. \tag{17.119}$$

It is at this point that we make the crucial 'condensate' assumption: we replace the *operator* expressions $\sum_{\boldsymbol{k}} \hat{c}_{-\boldsymbol{k}}\hat{c}_{\boldsymbol{k}}$ and $\sum_{\boldsymbol{k}} \hat{c}^\dagger_{\boldsymbol{k}}\hat{c}^\dagger_{-\boldsymbol{k}}$ by their average values, which are *assumed to be non-zero in the ground state*. Since these operators carry fermion number ± 2, it is clear that this assumption is only valid if the ground state does not, in fact, have a definitive number of particles – just as in the superfluid case. We accordingly make the replacements

$$V\textstyle\sum_{\boldsymbol{k}} \hat{c}_{-\boldsymbol{k}}\hat{c}_{\boldsymbol{k}} \to V\,{}_{\rm BCS}\langle{\rm ground}|\sum_{\boldsymbol{k}} \hat{c}_{-\boldsymbol{k}}\hat{c}_{\boldsymbol{k}}|{\rm ground}\rangle_{\rm BCS} \equiv \Delta \tag{17.120}$$

$$V\textstyle\sum_{\boldsymbol{k}} \hat{c}^\dagger_{\boldsymbol{k}}\hat{c}_{-\boldsymbol{k}}\dagger \to V\,{}_{\rm BCS}\langle{\rm ground}|\sum_{\boldsymbol{k}} \hat{c}^\dagger_{\boldsymbol{k}}\hat{c}^\dagger_{-\boldsymbol{k}}|{\rm ground}\rangle_{\rm BCS} \equiv \Delta^* \tag{17.121}$$

In that case, equations (17.118) and (17.119) become

$$\omega_l\cos\theta_l = \epsilon_l\cos\theta_l + \Delta\sin\theta_l \tag{17.122}$$

$$\omega_l\sin\theta_l = -\epsilon_l\sin\theta_l + \Delta^*\cos\theta_l \tag{17.123}$$

which are consistent if

$$\omega_l = \pm[\epsilon_l^2 + |\Delta|^2]^{1/2}. \tag{17.124}$$

Equation (17.124) is the fundamental result at this stage. Recalling that ϵ_l is measured relative to $E_{\rm F}$, we see that it implies that all excited states are separated from $E_{\rm F}$ by a finite amount, namely $|\Delta|$.

In interpreting (17.124) we must however be careful to reckon energies for an excited state as relative to a BCS state having the same number of pairs, if we consider experimental probes which do not inject or remove electrons. Thus relative to a component of $|{\rm ground}\rangle_{\rm BCS}$ with N pairs, we may consider the excitation of two particles above a BCS state with $N-1$ pairs. The minimum energy for this to be possible is $2|\Delta|$. It is this quantity which is usually called the *energy gap*. Such an excited state is represented by $\beta^\dagger_{\boldsymbol{k}}\beta^\dagger_{-\boldsymbol{k}}|{\rm ground}\rangle_{\rm BCS}$.

We shall need the expressions for $\cos\theta_l$ and $\sin\theta_l$ which may be obtained as follows. Squaring (17.122), and taking Δ now to be real and equal to $|\Delta|$, we obtain

$$|\Delta|^2(\cos^2\theta_l - \sin^2\theta_l) = 2\epsilon_l|\Delta|\cos\theta_l\sin\theta_l, \tag{17.125}$$

which leads to

$$\tan 2\theta_l = |\Delta|/\epsilon_l \tag{17.126}$$

and then

$$\cos\theta_l = \left[\frac{1}{2}\left(1+\frac{\epsilon_l}{\omega_l}\right)\right]^{1/2}, \qquad \sin\theta_l = \left[\frac{1}{2}\left(1-\frac{\epsilon_l}{\omega_l}\right)\right]^{1/2}. \tag{17.127}$$

All our experience to date indicates that the choice 'Δ = real' amounts to a choice of phase for the ground state value:

$$V\,{}_{\rm BCS}\langle{\rm ground}|\sum_{\boldsymbol{k}}\hat{c}_{-\boldsymbol{k}}c_{\boldsymbol{k}}|{\rm ground}\rangle_{\rm BCS} = |\Delta|. \tag{17.128}$$

By making use of the U(1) symmetry (17.110), other phases for Δ are equally possible.

The condition (17.128) has, of course, the by now anticipated form for a spontaneously broken U(1) symmetry, and we must therefore expect the occurrence of a massless mode (which we do not demonstrate here). However, we may now recall that the electrons are charged, so that when electromagnetic interactions are included in the superconducting state, we have to allow the α in (17.110) to become a local function of $\boldsymbol{x}$. At the same time, the massless photon field will enter. Remarkably, we shall learn in chapter 19 that the expected massless (Goldstone) mode is, in this case, not observed: instead, that degree of freedom is incorporated into the gauge field, rendering it massive. As we shall see, this is the physics of the Meissner effect in a superconductor, and that of the 'Higgs mechanism' in the Standard Model. Thus in the (charged) BCS model, both a fermion mass and a gauge boson mass are dynamically generated.

An explicit formula for Δ can be found by using the definition (17.120), together with the expression for $\hat{c}_{\boldsymbol{k}}$ found by inverting (17.111):

$$\hat{c}_{\boldsymbol{k}} = \cos\theta_k\,\hat{\beta}_{\boldsymbol{k}} + \sin\theta_k\,\hat{\beta}^\dagger_{-\boldsymbol{k}}. \tag{17.129}$$

This gives, using (17.120) and (17.129),

$$\begin{aligned}
|\Delta| &= V\,{}_{\rm BCS}\langle{\rm ground}|\sum_{\boldsymbol{k}}(\cos\theta_k\hat{\beta}_{-\boldsymbol{k}} + \sin\theta_k\,\hat{\beta}^\dagger_{\boldsymbol{k}}) \\
&\qquad \times(\cos\theta_k\,\hat{\beta}_{\boldsymbol{k}} + \sin\theta_k\,\hat{\beta}^\dagger_{-\boldsymbol{k}})|{\rm ground}\rangle_{\rm BCS} \\
&= V\,{}_{\rm BCS}\langle{\rm ground}|\sum_{\boldsymbol{k}}\cos\theta_k\sin\theta_k\hat{\beta}_{-\boldsymbol{k}}\hat{\beta}^\dagger_{-\boldsymbol{k}}|{\rm ground}\rangle_{\rm BCS}, \\
&= V\sum_{\boldsymbol{k}}\frac{|\Delta|}{2[\epsilon_k^2+|\Delta|^2]^{1/2}}.
\end{aligned} \tag{17.130}$$

The sum in (17.130) is only over the small band $E_{\rm F}-\omega_{\rm D} < \epsilon_k < E_{\rm F}+\omega_{\rm D}$ over which the effective electron–electron attraction operates. Replacing the

sum by an integral, we obtain the *gap equation*

$$\begin{aligned} 1 &= \frac{1}{2}V \cdot N_{\mathrm{F}} \int_{-\omega_{\mathrm{D}}}^{\omega_{\mathrm{D}}} \frac{\mathrm{d}\epsilon}{[\epsilon^2 + |\Delta|^2]^{\frac{1}{2}}} \\ &= VN_{\mathrm{F}} \sinh^{-1}(\omega_{\mathrm{D}}/|\Delta|) \end{aligned} \tag{17.131}$$

where N_{F} is the density of states at the Fermi level. Equation (17.131) yields

$$|\Delta| = \frac{\omega_{\mathrm{D}}}{\sinh(1/VN_{\mathrm{F}})} \approx 2\omega_{\mathrm{D}} \mathrm{e}^{-1/VN_{\mathrm{F}}} \tag{17.132}$$

for $VN_{\mathrm{F}} \ll 1$. This is the celebrated BCS solution for the gap parameter $|\Delta|$. Perhaps the most significant thing to note about it, for our purpose, is that the expression for $|\Delta|$ is not an analytic function of the dimensionless interaction parameter VN_{F} (it cannot be expanded as a power series in this quantity), and so no perturbative treatment starting from a normal ground state could reach this result. The estimate (17.132) is in reasonably good agreement with experiment, and may be refined.

The explicit form of the ground state in this model can be found by a method similar to the one indicated in section 17.3.2 for the superfluid. Since the transformation from the $\hat{c}$'s to the $\hat{\beta}$'s is canonical, there must exist a unitary operator which effects it via (compare (17.48))

$$\hat{U}_{\mathrm{BCS}}\, \hat{c}_{\boldsymbol{k}}\, \hat{U}_{\mathrm{BCS}}^{\dagger} = \hat{\beta}_{\boldsymbol{k}}, \quad \hat{U}_{\mathrm{BCS}}\, \hat{c}^{\dagger}_{-\boldsymbol{k}}\, \hat{U}_{\mathrm{BCS}}^{\dagger} = \hat{\beta}^{\dagger}_{-\boldsymbol{k}}. \tag{17.133}$$

The operator $\hat{U}_{\mathrm{BCS}}$ is (Blatt 1964 section V.4, Yosida 1958, and compare problem 17.4)

$$\hat{U}_{\mathrm{BCS}} = \prod_{\boldsymbol{k}} \exp[\theta_k(\hat{c}^{\dagger}_{\boldsymbol{k}}\hat{c}^{\dagger}_{-\boldsymbol{k}} - \hat{c}_{\boldsymbol{k}}\hat{c}_{-\boldsymbol{k}})]. \tag{17.134}$$

Then, since $\hat{c}_{\boldsymbol{k}}|0\rangle = 0$, we have

$$\hat{U}^{\dagger}_{\mathrm{BCS}}\hat{\beta}_{\boldsymbol{k}}\hat{U}_{\mathrm{BCS}}|0\rangle = 0 \tag{17.135}$$

showing that we may identify

$$|\mathrm{ground}\rangle_{\mathrm{BCS}} = \hat{U}_{\mathrm{BCS}}|0\rangle \tag{17.136}$$

via the condition (17.116). When the exponential in $\hat{U}_{\mathrm{BCS}}$ is expanded out, and applied to the vacuum state $|0\rangle$, great simplifications occur. Consider the operator

$$\hat{s}_{\boldsymbol{k}} = \hat{c}^{\dagger}_{\boldsymbol{k}}\hat{c}^{\dagger}_{-\boldsymbol{k}} - \hat{c}_{\boldsymbol{k}}\hat{c}_{-\boldsymbol{k}}. \tag{17.137}$$

We have

$$\hat{s}^2_{\boldsymbol{k}} = -\hat{c}^{\dagger}_{\boldsymbol{k}}\hat{c}^{\dagger}_{-\boldsymbol{k}}\hat{c}_{\boldsymbol{k}}\hat{c}_{-\boldsymbol{k}} - \hat{c}_{\boldsymbol{k}}\hat{c}_{-\boldsymbol{k}}\hat{c}^{\dagger}_{\boldsymbol{k}}\hat{c}^{\dagger}_{-\boldsymbol{k}} \tag{17.138}$$

so that $\hat{s}^2_{\boldsymbol{k}}|0\rangle = -|0\rangle$. It follows that

$$\begin{aligned}\exp(\theta_k \hat{s}_{\boldsymbol{k}})|0\rangle &= (1 + \theta_k \hat{s}_{\boldsymbol{k}} - \frac{\theta_k^2}{2} - \frac{\theta_k^3}{3}\hat{s}_{\boldsymbol{k}} \ldots)|0\rangle \\ &= (\cos\theta_k + \sin\theta_k\, \hat{s}_{\boldsymbol{k}})|0\rangle \\ &= (\cos\theta_k + \sin\theta_k\, \hat{c}^\dagger_{\boldsymbol{k}} \hat{c}^\dagger_{-\boldsymbol{k}})|0\rangle \qquad (17.139)\end{aligned}$$

and hence

$$|\text{ground}\rangle_{\text{BCS}} = \prod_{\boldsymbol{k}} (\cos\theta_k + \sin\theta_k\, \hat{c}^\dagger_{\boldsymbol{k}} \hat{c}^\dagger_{-\boldsymbol{k}})|0\rangle. \qquad (17.140)$$

As for the superfluid, (17.140) represents a coherent superposition of correlated pairs, with no restraint on the particle number.

We should emphasize that the above is only the barest outline of a simple version of BCS theory, with no electromagnetic interactions, from which many subtleties have been omitted. Consider, for example, the binding energy E_{b} of a pair, to calculate which one needs to evaluate the constant γ in (17.114). To a good approximation one finds (see for example Enz 1992) $E_{\text{b}} \approx 3\Delta^2/E_{\text{F}}$. One can also calculate the approximate spatial extension of a pair, which is denoted by the *coherence length* ξ and is of order $v_{\text{F}}/\pi\Delta$ where $k_{\text{F}} = mv_{\text{F}}$ is the Fermi momentum. If we compare E_{b} to the Coulomb repulsion at a distance ξ we find

$$E_{\text{b}}/(\alpha/\xi) \sim a_0/\xi \qquad (17.141)$$

where a_0 is the Bohr radius. Numerical values show that the right-hand side of (17.141), in conventional superconductors, is of order 10^{-3}. Hence the pairs are not really bound, only correlated, and as many as 10^6 pairs may have their centres of mass within one coherence length of each other. Nevertheless, the simple theory presented here contains the essential features which underlie all attempts to understand the dynamical occurrence of spontaneous symmetry breaking in fermionic systems.

We now proceed to an important application in particle physics.

Problems

17.1 Verify (17.29).

17.2 Verify (17.35).

17.3 Derive (17.43) and (17.44).

17.4 Let

$$\hat{U}_\lambda = \exp[\frac{1}{2}\lambda\theta(\hat{a}^2 - \hat{a}^{\dagger 2})]$$

where $[\hat{a}, \hat{a}^\dagger] = 1$ and λ, θ are real parameters.

(a) Show that $\hat{U}_\lambda$ is unitary.

(b) Let

$$\hat{I}_\lambda = \hat{U}_\lambda \hat{a} \hat{U}_\lambda^{-1}, \qquad \text{and} \qquad \hat{J}_\lambda = \hat{U}_\lambda \hat{a}^\dagger \hat{U}_\lambda^{-1}.$$

Show that

$$\frac{\mathrm{d}\hat{I}_\lambda}{\mathrm{d}\lambda} = \theta \hat{J}_\lambda$$

and that

$$\frac{\mathrm{d}^2\hat{I}_\lambda}{\mathrm{d}\lambda^2} = \theta^2 \hat{I}_\lambda.$$

(c) Hence show that

$$\hat{I}_\lambda = \cosh(\lambda\theta)\, \hat{a} + \sinh(\lambda\theta)\, \hat{a}^\dagger,$$

and thus finally (compare (17.38) and (17.48)) that

$$\hat{U}_1 \hat{a} \hat{U}_1^{-1} = \cosh\theta\, \hat{a} + \sinh\theta\, \hat{a}^\dagger \equiv \hat{\alpha}$$

and

$$\hat{U}_1 \hat{a}^\dagger \hat{U}_1^{-1} = \sinh\theta\, \hat{a} + \cosh\theta\, \hat{a}^\dagger \equiv \hat{\alpha}^\dagger,$$

where

$$\hat{U}_1 \equiv \hat{U}_{\lambda=1} = \exp[\frac{1}{2}\theta(\hat{a}^2 - \hat{a}^{\dagger 2})].$$

17.5 Insert the ansatz (17.84) for $\hat{\phi}$ into $\hat{\mathcal{L}}_\mathrm{G}$ of (17.69), with $\hat{V} = \hat{V}_\mathrm{SB}$ of (17.77), and show that the result for the constant term, and the quadratic terms in $\hat{h}$ and $\hat{\theta}$, is as given in (17.85).

17.6 Verify that when (17.107) is inserted in (17.97), the terms quadratic in the fields $\hat{H}$ and $\hat{\theta}$ reveal that $\hat{\theta}$ is a massless field, while the quanta of the $\hat{H}$ field have mass $\sqrt{2}\mu$.

17.7 Verify that the $\hat{\beta}$'s of (17.111) satisfy the required anticommutation relations if (17.112) holds.

17.8 Verify (17.115).

17.9 Derive (17.118) and (17.119).

18

Chiral Symmetry Breaking

In section 12.4.2 we arrived at a puzzle: there seemed good reason to think that a world consisting of u and d quarks and their antiparticles, interacting via the colour gauge fields of QCD, should exhibit signs of the non-Abelian *chiral symmetry* $\mathrm{SU}(2)_{\mathrm{f}\,5}$, which was exact in the massless limit $m_{\mathrm{u}}, m_{\mathrm{d}} \to 0$. But, as we showed, one of the simplest consequences of such a symmetry should be the existence of nucleon parity doublets, which are not observed. We can now resolve this puzzle by making the hypothesis (section 18.1) first articulated by Nambu (1960) and Nambu and Jona-Lasinio (1961a), that this chiral symmetry is *spontaneously broken* as a dynamical effect – presumably, from today's perspective, as a property of the QCD interactions, as discussed in section 18.1.1. If this is so, an immediate physical consequence should be the appearance of massless (Goldstone) bosons, one for every symmetry not respected by the vacuum. Indeed, returning to (12.168) which we repeat here for convenience,

$$\hat{T}^{(\frac{1}{2})}_{+5}|d\rangle = |\tilde{u}\rangle, \tag{18.1}$$

we now interpret the state $|\tilde{u}\rangle$ (which is degenerate with $|d\rangle$) as $|d + \text{`}\pi^{+}\text{'}\rangle$ where 'π^{+}' is a massless particle of positive charge, but a *pseudo*scalar (0^{-}) rather than a scalar (0^{+}) since, as we saw, $|\tilde{u}\rangle$ has opposite parity to $|u\rangle$. In the same way, 'π^{-}' and 'π^{0}' will be associated with $\hat{T}^{(\frac{1}{2})}_{-5}$ and $\hat{T}^{(\frac{1}{2})}_{3\,5}$. Of course, no such *massless* pseudoscalar particles are observed: but it is natural to hope that when the small up and down quark masses are included, the real pions $(\pi^{+}, \pi^{-}, \pi^{0})$ will emerge as 'anomalously light', rather than strictly massless. This is indeed how they do appear, particularly with respect to the octet of mesons, which differ only in $\mathrm{q}\bar{\mathrm{q}}$ spin alignment from the 0^{-} octet. As Nambu and Jona-Lasinio (1961a) said, 'it is perhaps not a coincidence that there exists such an entity [i.e. the Goldstone state(s)] in the form of the pion'.

If this was the only observable consequence of spontaneously breaking chiral symmetry, it would perhaps hardly be sufficient grounds for accepting the hypothesis. But there are two circumstances which greatly increase the phenomenological implications of the idea. First, the vector and axial vector symmetry currents $\hat{\boldsymbol{T}}^{(\frac{1}{2})\mu}$ and $\hat{\boldsymbol{T}}^{(\frac{1}{2})\mu}_{5}$ of the u-d strong interaction SU(2) symmetries (see (12.109) and (12.165)) happen to be the very same currents which enter into strangeness-conserving semileptonic weak interactions (such as $\mathrm{n} \to \mathrm{pe}^{-}\bar{\nu}_{\mathrm{e}}$ and $\pi^{-} \to \mu^{-}\bar{\nu}_{\mu}$), as we shall see in chapter 20. Thus some remarkable connections between weak- and strong-interaction parameters can be

established, such as the Goldberger–Treiman (1958) relation (see section 18.2) and the Adler–Weisberger (Adler 1965, Weisberger 1965) relation. Second, it turns out that the dynamics of the Goldstone modes, and their interactions with other hadrons such as nucleons, are strongly constrained by the underlying chiral symmetry of QCD; indeed, surprisingly detailed *effective theories* (see section 18.3) have been developed, which provide a very successful description of the low energy dynamics of the Goldstone degrees of freedom. Finally we shall introduce the subject of chiral anomalies in section 18.4.

It would take us too far from our main focus on gauge theories to pursue these interesting avenues in any detail. But we hope to convince the reader, in this chapter, that chiral symmetry breaking is an integral part of the Standard Model, being a fundamental property of QCD.

18.1 The Nambu analogy

We recall from section 12.4.2 that for 'almost massless' fermions it is natural to use the representation (3.40) for the Dirac matrices, in terms of which the Dirac equation reads

$$E\phi = \boldsymbol{\sigma}\cdot\boldsymbol{p}\phi + m\chi \tag{18.2}$$

$$E\chi = -\boldsymbol{\sigma}\cdot\boldsymbol{p}\chi + m\phi. \tag{18.3}$$

Nambu (1960) and Nambu and Jona-Lasinio (1961a) pointed out a remarkable analogy between (18.2) and (18.3) and equations (17.122) and (17.123) which describe the elementary excitations in a superconductor (in the case Δ is real), and which we repeat here for convenience:

$$\omega_l \cos\theta_l = \epsilon_l \cos\theta_l + \Delta \sin\theta_l \tag{18.4}$$

$$\omega_l \sin\theta_l = -\epsilon_l \sin\theta_l + \Delta \cos\theta_l. \tag{18.5}$$

In (18.4) and (18.5), $\cos\theta_l$ and $\sin\theta_l$ are respectively the components of the electron destruction operator $\hat{c}_{\boldsymbol{l}}$ and the electron creation operator $\hat{c}^\dagger_{-\boldsymbol{l}}$ in the quasiparticle operator $\hat{\beta}_{\boldsymbol{l}}$ (see (17.111)):

$$\hat{\beta}_{\boldsymbol{l}} = \cos\theta_l\, \hat{c}_{\boldsymbol{l}} - \sin\theta_l\, \hat{c}^\dagger_{-\boldsymbol{l}}. \tag{18.6}$$

The superposition in $\hat{\beta}_{\boldsymbol{l}}$ combines operators which transform differently under the U(1) (number) symmetry. The result of this spontaneous breaking of the U(1) symmetry is the creation of the gap Δ (or 2Δ for a number-conserving excitation), and the appearance of a massless mode. If Δ vanishes, (17.126) implies that $\theta_l = 0$, and we revert to the symmetry-respecting operators $\hat{c}_{\boldsymbol{l}}, \hat{c}^\dagger_{-\boldsymbol{l}}$. Consider now (18.2) and (18.3). Here ϕ and χ are the components of

FIGURE 18.1
The type of fermion–antifermion in the 'Nambu chiral condensate'.

definite chirality in the Dirac spinor ω (compare (12.149)), which is itself not a chirality eigenstate when $m \neq 0$. When m vanishes, the Dirac equation for ω decouples into two separate ones for the chirality eigenstates $\phi_R \equiv \begin{pmatrix} \phi \\ 0 \end{pmatrix}$ and $\phi_L \equiv \begin{pmatrix} 0 \\ \chi \end{pmatrix}$. Nambu therefore made the following analogy:

$$\begin{aligned} \text{Superconducting gap parameter } \Delta &\leftrightarrow \text{Dirac mass } m \\ \text{quasiparticle excitation} &\leftrightarrow \text{massive Dirac particle} \\ \text{U(1) number symmetry} &\leftrightarrow \text{U(1)}_5 \text{ chirality symmetry} \\ \text{Goldstone mode} &\leftrightarrow \text{massless boson.} \end{aligned}$$

In short, the mass of a Dirac particle arises from the (presumed) spontaneous breaking of a chiral (or γ_5) symmetry, and this will be accompanied by a massless boson.

Before proceeding we should note that there are features of the analogy, on both sides, which need qualification. First, the particle symmetry we want to interpret this way is $SU(2)_{f\,5}$ not $U(1)_5$, so the appropriate generalization (Nambu and Jona-Lasinio 1961b) has to be understood. Second, we must again note that the BCS electrons are charged, so that in the real superconducting case we are dealing with a spontaneously broken *local* U(1) symmetry, not a global one. By contrast, the $SU(2)_{f\,5}$ chiral symmetry is not gauged.

As usual, the quantum field theory vacuum is analogous to the many-body ground state. According to Nambu's analogy, therefore, the vacuum for a massive Dirac particle is to be pictured as a condensate of correlated pairs of massive fermions. Since the vacuum carries neither linear nor angular momentum, the members of a pair must have equal and opposite spin: they therefore have the same helicity. However, since the vacuum does *not* violate fermion number conservation, one has to be a fermion and the other an antifermion. This means (recalling the discussion after (12.147)) that they have opposite chirality. Thus a typical pair in the Nambu vacuum is as shown in figure 18.1. We may easily write down an expression for the Nambu vacuum, analogous to (17.140) for the BCS ground state. Consider solutions ϕ_+ and χ_+ of positive helicity in (18.2) and (18.3); then

$$E\phi_+ \quad = \quad |\boldsymbol{p}|\phi_+ + m\chi_+ \tag{18.7}$$

$$E\chi_+ \quad = \quad -|\boldsymbol{p}|\chi_+ + m\phi_+. \tag{18.8}$$

Comparing (18.7) and (18.8) with (18.4) and (18.5), we can read off the mixing coefficients $\cos\theta_p$ and $\sin\theta_p$ as (cf (17.127))

$$\cos\theta_p \quad = \quad \left[\frac{1}{2}\left(1+\frac{|\boldsymbol{p}|}{E}\right)\right]^{1/2} \tag{18.9}$$

$$\sin\theta_p \quad = \quad \left[\frac{1}{2}\left(1-\frac{|\boldsymbol{p}|}{E}\right]\right)^{1/2} \tag{18.10}$$

where $E=(m^2+\boldsymbol{p}^2)^{1/2}$. The Nambu vacuum is then given by[1]

$$|0\rangle_{\rm N} = \prod_{\boldsymbol{p},s}(\cos\theta_p - \sin\theta_p \hat{c}_s^\dagger(\boldsymbol{p})\hat{d}_s^\dagger(-\boldsymbol{p}))|0\rangle_{m=0}, \tag{18.11}$$

where $\hat{c}_s^\dagger$'s and $\hat{d}_s^\dagger$'s are the operators in *massless* Dirac fields. Depending on the sign of the helicity s, each pair in (18.11) carries ± 2 units of chirality. We may check this by noting that in the mode expansion of the Dirac field $\hat{\psi}$, $\hat{c}_s(\boldsymbol{p})$ operators go with u-spinors for which the γ_5 eigenvalue equals the helicity, while $\hat{d}_s^\dagger(-\boldsymbol{p})$ operators accompany v-spinors for which the γ_5 eigenvalue equals minus the helicity. Thus under a chiral transformation $\hat{\psi}' = {\rm e}^{-{\rm i}\beta\gamma_5}\hat{\psi}$, $\hat{c}_s \to {\rm e}^{-{\rm i}\beta s}\hat{c}_s$ and $\hat{d}_s^\dagger \to {\rm e}^{{\rm i}\beta s}\hat{d}_s^\dagger$, for a given s. Hence $\hat{c}_s^\dagger\hat{d}_s^\dagger$ acquires a factor ${\rm e}^{2{\rm i}\beta s}$. Thus the Nambu vacuum does not have a definite chirality, and operators carrying non-zero chirality can have non-vanishing vacuum expectation values. A mass term $\bar{\hat{\psi}}\hat{\psi}$ is of just this kind, since under $\hat{\psi} = {\rm e}^{-{\rm i}\beta\gamma_5}\hat{\psi}$ we find $\hat{\psi}^\dagger\gamma^0\hat{\psi} \to \hat{\psi}^\dagger{\rm e}^{{\rm i}\beta\gamma_5}\gamma^0{\rm e}^{-{\rm i}\beta\gamma_5}\hat{\psi} = \bar{\hat{\psi}}{\rm e}^{-2{\rm i}\beta\gamma_5}\hat{\psi}$. Thus, in analogy with (17.120), a Dirac mass is associated with a non-zero value for ${}_{\rm N}\langle 0|\bar{\hat{\psi}}\hat{\psi}|0\rangle_{\rm N}$.

In the original conception by Nambu and co-workers, the fermion under discussion was taken to be the nucleon, with 'm' the (spontaneously generated) nucleon mass. The fermion–fermion interaction – necessarily invariant under chiral transformations – was taken to be of the four-fermion type. As we have seen in volume 1, this is actually a non-renormalizable theory, but a physical cut-off was employed, somewhat analogous to the Fermi energy $E_{\rm F}$. Thus the nucleon mass could not be dynamically predicted, unlike the analogous gap parameter Δ in BCS theory. Nevertheless, a gap equation similar to (17.131) could be formulated, and it was possible to show that when it had a non-trivial solution, a massless bound state automatically appeared in the $\bar{\rm f}{\rm f}$ channel (Nambu and Jona-Lasinio 1961a). This work was generalized to the ${\rm SU(2)_{f\,5}}$ case by Nambu and Jona-Lasinio (1961b), who showed that if the chiral symmetry was broken explicitly by the introduction of a small nucleon mass (~ 5 MeV), then the Goldstone pions would have their observed non-zero (but small) mass. In addition, the Goldberger–Treiman (1958) relation was derived, and a number of other applications were suggested. Subsequently, Nambu with other collaborators (Nambu and Lurie 1962, Nambu

[1] A different phase convention is used for $\hat{d}_s^\dagger(-\boldsymbol{p})$ as compared to that for $\hat{c}^\dagger_{-\boldsymbol{k}}$ in (17.111).

and Schrauner 1962) showed how the amplitudes for the emission of a single 'soft' (nearly massless, low momentum) pion could be calculated, for various processes. These developments culminated in the Adler-Weisberger relation (Adler 1965, Weisberger 1965) which involves *two* soft pions.

This work was all done in the absence of an agreed theory of the strong interactions (the NJ-L theory was an illustrative working model of dynamically generated spontaneous symmetry breaking, but not a complete theory of strong interactions). QCD became widely accepted as that theory around 1973. In this case, of course, the 'fermions in question' are quarks, and the interactions between them are gluon exchanges, which conserve chirality as noted in section 12.4.2. The bulk of the masses of the qqq bound states which form baryons is then interpreted as being spontaneously generated, while a small explicit quark mass term in the Lagrangian is responsible for the non-zero pion mass. Let us therefore now turn to two-flavour QCD.

18.1.1 Two flavour QCD and $\mathbf{SU(2)}_{\mathrm{f\,L}}\times\mathbf{SU(2)}_{\mathrm{f\,R}}$

Let us begin with the massless case, for which the fermionic part of the Lagrangian is

$$\hat{\mathcal{L}}_q = \bar{\hat{u}}\,\mathrm{i}\hat{\not{D}}\hat{u} + \bar{\hat{d}}\,\mathrm{i}\hat{\not{D}}\hat{d} \tag{18.12}$$

where $\hat{u}$ and $\hat{d}$ now stand for the field operators,

$$\hat{D}^{\mu} = \partial^{\mu} + \mathrm{i}g_{\mathrm{s}}\boldsymbol{\lambda}/2 \cdot \hat{\boldsymbol{A}}^{\mu}, \tag{18.13}$$

and the $\boldsymbol{\lambda}$ matrices act on the colour (r,b,g) degree of freedom of the u and d quarks. This Lagrangian is invariant under

(i) $\mathrm{U(1)_f}$ 'quark number' transformations

$$\hat{q} \to \mathrm{e}^{-\mathrm{i}\alpha}\hat{q}; \tag{18.14}$$

(ii) $\mathrm{SU(2)_f}$ 'flavour isospin' transformations

$$\hat{q} \to \exp(-\mathrm{i}\boldsymbol{\alpha}\cdot\boldsymbol{\tau}/2)\,\hat{q}; \tag{18.15}$$

(iii) $\mathrm{U(1)_{f\,5}}$ 'axial quark number' transformations

$$\hat{q} \to \mathrm{e}^{-\mathrm{i}\beta\gamma_5}\hat{q}; \tag{18.16}$$

(iv) $\mathrm{SU(2)_{f\,5}}$ 'axial flavour isospin' transformations

$$\hat{q} \to \exp(-\mathrm{i}\boldsymbol{\beta}\cdot\boldsymbol{\tau}/2\gamma_5)\,\hat{q}, \tag{18.17}$$

where

$$\hat{q} = \begin{pmatrix} \hat{u} \\ \hat{d} \end{pmatrix}. \tag{18.18}$$

Symmetry (i) is unbroken, and its associated 'charge' operator (the quark number operator) commutes with all other symmetry operators, so it need not concern us further. Symmetry (ii) is the standard isospin symmetry of chapter 12, explicitly broken by the electromagnetic interactions (and by the difference in the masses m_{u} and m_{d}, when included). Symmetry (iii) does not correspond to any known conservation law; on the other hand, there are not any near-massless isoscalar 0^- mesons, either, such as must be present if the symmetry is spontaneously broken. The η meson is an isoscalar 0^- meson, but with a mass of 547 MeV it is considerably heavier than the pion. In fact, it can be understood as one of the Goldstone bosons associated with the spontaneous breaking of the larger group $\mathrm{SU}(3)_{\mathrm{f}\,5}$, which includes the s quark (see section 18.3.3). In that case, the symmetry (iii) becomes extended to

$$\hat{u} \to \mathrm{e}^{-\mathrm{i}\beta\gamma_5}\hat{u}, \quad \hat{d} \to \mathrm{e}^{-\mathrm{i}\beta\gamma_5}\hat{d}, \quad \hat{s} \to \mathrm{e}^{-\mathrm{i}\beta\gamma_5}\hat{s}, \tag{18.19}$$

but there is still a missing light isoscalar 0^- meson. It can be shown that its mass must be less than or equal to $\sqrt{3}\, m_\pi$ (Weinberg 1975), but no such particle exists. This is the famous 'U(1) problem': it was resolved by 't Hooft (1976a, 1986), by showing that the inclusion of instanton configurations (Belavin *et al.* 1975) in path integrals leads to violations of symmetry (iii) – see, for example, Weinberg (1996) section 23.5. Finally, symmetry (iv) is the one with which we are presently concerned.

The symmetry currents associated with (iv) are those already given in (12.165), but we give them again here in a slightly different notation which will be similar to the one used for weak interactions:

$$\hat{j}^{\mu}_{i,5} = \bar{\hat{q}}\gamma^{\mu}\gamma_5\frac{\tau_i}{2}\hat{q} \qquad i = 1, 2, 3. \tag{18.20}$$

Similarly the currents associated with (ii) are

$$\hat{j}^{\mu}_{i} = \bar{\hat{q}}\gamma^{\mu}\frac{\tau_i}{2}\hat{q} \qquad i = 1, 2, 3. \tag{18.21}$$

The corresponding 'charges' are (compare (12.166))

$$\hat{Q}_{i,5} \equiv \int \hat{j}^{0}_{i,5}\,\mathrm{d}^3\boldsymbol{x} = \int \hat{q}^{\dagger}\gamma_5\frac{\tau_i}{2}\hat{q}\,\mathrm{d}^3\boldsymbol{x}, \tag{18.22}$$

previously denoted by $\hat{T}^{(\frac{1}{2})}_{i,5}$, and (compare (12.101)),

$$\hat{Q}_i = \int \hat{q}^{\dagger}\frac{\tau_i}{2}\hat{q}\,\mathrm{d}^3\boldsymbol{x}, \tag{18.23}$$

previously denoted by $\hat{\boldsymbol{T}}^{(\frac{1}{2})}_{5}$. As with all symmetries, it is interesting to discover the *algebra* of the generators, which are the six charges $\hat{Q}_i, \hat{Q}_{i,5}$ in this case. Patient work with the anticommutation relations for the operators in

$\hat{q}(x)$ and $\hat{q}^\dagger(x)$ gives the results (problem 18.1)

$$[\hat{Q}_i, \hat{Q}_j] = \mathrm{i}\epsilon_{ijk}\hat{Q}_k \tag{18.24}$$
$$[\hat{Q}_i, \hat{Q}_{j,5}] = \mathrm{i}\epsilon_{ijk}\hat{Q}_{k,5} \tag{18.25}$$
$$[\hat{Q}_{i,5}, \hat{Q}_{j,5}] = \mathrm{i}\epsilon_{ijk}\hat{Q}_k. \tag{18.26}$$

Relation (18.24) has been seen before in (12.101), and simply says that the $\hat{Q}_i$'s obey a SU(2) algebra. A simple trick reduces the rather complicated algebra of (18.24)–(18.26) to something much simpler. Defining

$$\hat{Q}_{i,\mathrm{R}} = \frac{1}{2}(\hat{Q}_i + \hat{Q}_{i,5}) \qquad \hat{Q}_{i,\mathrm{L}} = \frac{1}{2}(\hat{Q}_i - \hat{Q}_{i,5}) \tag{18.27}$$

we find (problem 18.2)

$$[\hat{Q}_{i,\mathrm{R}}, \hat{Q}_{j,\mathrm{R}}] = \mathrm{i}\epsilon_{ijk}\hat{Q}_{k,\mathrm{R}} \tag{18.28}$$
$$[\hat{Q}_{i,\mathrm{L}}, \hat{Q}_{j,\mathrm{L}}] = \mathrm{i}\epsilon_{ijk}\hat{Q}_{k,\mathrm{L}} \tag{18.29}$$
$$[\hat{Q}_{i,\mathrm{R}}, \hat{Q}_{j,\mathrm{L}}] = 0. \tag{18.30}$$

The operators $\hat{Q}_{i,\mathrm{R}}, \hat{Q}_{i,\mathrm{L}}$ therefore behave like *two commuting (independent) angular momentum operators*, each obeying the algebra of SU(2). For this reason, the symmetry group of the combined symmetries (ii) and (iv) is called $\mathrm{SU}(2)_{\mathrm{f\,L}} \times \mathrm{SU}(2)_{\mathrm{f\,R}}$.

The decoupling effected by (18.27) has a simple interpretation. Referring to (18.22) and (18.23), we see that

$$\hat{Q}_{i,\mathrm{R}} = \int \hat{q}^\dagger \left(\frac{1+\gamma_5}{2}\right) \frac{\tau_i}{2} \hat{q}\, \mathrm{d}^3\boldsymbol{x} \tag{18.31}$$

and similarly for $\hat{Q}_{i,\mathrm{L}}$. But $((1 \pm \gamma_5)/2)$ are just the projection operators $P_{\mathrm{R,L}}$ introduced in section 12.3.2, which project out the chiral parts of any fermion field. Furthermore, it is easy to see that $P_{\mathrm{R}}^2 = P_{\mathrm{R}}$ and $P_{\mathrm{L}}^2 = P_{\mathrm{L}}$, so that $\hat{Q}_{i,\mathrm{R}}$ and $\hat{Q}_{i,\mathrm{L}}$ can also be written as

$$\hat{Q}_{i,\mathrm{R}} = \int \hat{q}_{\mathrm{R}}^\dagger \frac{\tau_i}{2} \hat{q}_{\mathrm{R}}\, \mathrm{d}^3\boldsymbol{x} \qquad \hat{Q}_{i,\mathrm{L}} = \int \hat{q}_{\mathrm{L}}^\dagger \frac{\tau_i}{2} \hat{q}_{\mathrm{L}}\, \mathrm{d}^3\boldsymbol{x}, \tag{18.32}$$

where $\hat{q}_{\mathrm{R}} = ((1+\gamma_5)/2)\hat{q}, \hat{q}_{\mathrm{L}} = ((1-\gamma_5)/2)\hat{q}$. In a similar way, the currents (18.20) and (18.21) can be written as

$$\hat{j}_i^\mu = \hat{j}_{i,\mathrm{R}}^\mu + \hat{j}_{i,\mathrm{L}}^\mu \qquad \hat{j}_{i,5}^\mu = \hat{j}_{i,\mathrm{R}}^\mu - \hat{j}_{i,\mathrm{L}}^\mu, \tag{18.33}$$

where

$$\hat{j}_{i,\mathrm{R}} = \bar{\hat{q}}_{\mathrm{R}} \gamma^\mu \frac{\tau_i}{2} \hat{q}_{\mathrm{R}} \qquad \hat{j}_{i,\mathrm{L}}^\mu = \bar{\hat{q}}_{\mathrm{L}} \gamma^\mu \frac{\tau_i}{2} \hat{q}_{\mathrm{L}}. \tag{18.34}$$

Thus the $\mathrm{SU}(2)_{\mathrm{L}}$ and $\mathrm{SU}(2)_{\mathrm{R}}$ refer to the two chiral components of the fermion fields, which is why it is called *chiral symmetry*.

Under infinitesimal SU(2) isospin and axial isospin transformations, $\hat{q}$ transforms by

$$\hat{q} \to \hat{q}' = (1 - \mathrm{i}\boldsymbol{\epsilon} \cdot \boldsymbol{\tau}/2 - \mathrm{i}\boldsymbol{\eta} \cdot \boldsymbol{\tau}/2\,\gamma_5)\hat{q}. \tag{18.35}$$

This can be rewritten in terms of $\hat{q}_\mathrm{R}$ and $\hat{q}_\mathrm{R}$, using

$$\hat{q} = \hat{q}_\mathrm{R} + \hat{q}_\mathrm{L}, \quad \gamma_5\hat{q}_\mathrm{R} = \hat{q}_\mathrm{R}, \quad \gamma_5\hat{q}_\mathrm{L} = -\hat{q}_\mathrm{L}. \tag{18.36}$$

We find that

$$\hat{q}'_\mathrm{R} = (1 - \mathrm{i}(\boldsymbol{\epsilon} + \boldsymbol{\eta}) \cdot \boldsymbol{\tau}/2)\hat{q}_\mathrm{R} \tag{18.37}$$

and similarly

$$\hat{q}'_\mathrm{L} = (1 - \mathrm{i}(\boldsymbol{\epsilon} - \boldsymbol{\eta}) \cdot \boldsymbol{\tau}/2)\hat{q}_\mathrm{L}. \tag{18.38}$$

Hence $\hat{q}_\mathrm{R}$ and $\hat{q}_\mathrm{L}$ transform quite independently[2], which is why $[\hat{Q}_{i,\mathrm{R}}, \hat{Q}_{j,\mathrm{L}}] = 0$.

This formalism allows us to see immediately why (18.12) is chirally invariant: problem 18.3 verifies that $\hat{\mathcal{L}}_q$ can be written as

$$\hat{\mathcal{L}}_q = \bar{\hat{q}}_\mathrm{R}\mathrm{i}\,\hat{\not{D}}q_\mathrm{R} + \bar{\hat{q}}_\mathrm{L}\mathrm{i}\,\hat{\not{D}}\hat{q}_\mathrm{L} \tag{18.39}$$

which is plainly invariant under (18.37) and (18.38), since $\hat{D}$ is flavour-blind.

There is as yet no formal proof that this $\mathrm{SU}(2)_\mathrm{L}{\times}\mathrm{SU}(2)_\mathrm{R}$ chiral symmetry is spontaneously broken in QCD, though it can be argued that the larger symmetry $\mathrm{SU}(3)_\mathrm{L}{\times}\mathrm{SU}(3)_\mathrm{R}$ – appropriate to three massless flavours – must be spontaneously broken (see Weinberg 1996, section 22.5). This is, of course, an issue that cannot be settled within perturbation theory (compare the comments after (17.132)). Numerical solutions of QCD on a lattice (see chapter 16) do provide strong evidence that baryons acquire large dynamical ($\mathrm{SU}(2)_{\mathrm{f}\,5}$-breaking) mass.

Even granted that chiral symmetry is spontaneously broken in massless two-flavour QCD, how do we know that it breaks in such a way as to leave the isospin ('R + L') symmetry unbroken? A plausible answer can be given if we restore the quark mass terms via

$$\hat{\mathcal{L}}_m = -m_\mathrm{u}\bar{\hat{u}}\hat{u} - m_\mathrm{d}\bar{\hat{d}}\hat{d} = -\frac{1}{2}(m_\mathrm{u} + m_\mathrm{d})\bar{\hat{q}}\hat{q} - \frac{1}{2}(m_\mathrm{u} - m_\mathrm{d})\bar{\hat{q}}\tau_3\hat{q}. \tag{18.40}$$

Now

$$\bar{\hat{q}}\hat{q} = \bar{\hat{q}}_\mathrm{L}\hat{q}_\mathrm{R} + \bar{\hat{q}}_\mathrm{R}\hat{q}_\mathrm{L} \tag{18.41}$$

and

$$\bar{\hat{q}}\tau_3\hat{q} = \bar{\hat{q}}_\mathrm{L}\tau_3 q_\mathrm{R} + \bar{\hat{q}}_\mathrm{R}\tau_3\hat{q}_\mathrm{L}. \tag{18.42}$$

Including these extra terms is somewhat analogous to switching on an external field in the ferromagnetic problem, which determines a preferred direction for the symmetry breaking. It is clear that neither of (18.41) and (18.42) preserves

[2] We may set $\boldsymbol{\gamma} = \boldsymbol{\epsilon} + \boldsymbol{\eta}$, and $\boldsymbol{\delta} = \boldsymbol{\epsilon} - \boldsymbol{\eta}$.

$\mathrm{SU}(2)_\mathrm{L}\times\mathrm{SU}(2)_\mathrm{R}$ since they treat the L and R parts differently. Indeed from (18.37) and (18.38) we find

$$\begin{aligned}\bar{\hat{q}}_\mathrm{L}\hat{q}_\mathrm{R}\to\bar{\hat{q}}'_\mathrm{L}q'_\mathrm{R} &= \bar{\hat{q}}_\mathrm{L}(1+\mathrm{i}(\boldsymbol{\epsilon}-\boldsymbol{\eta})\cdot\boldsymbol{\tau}/2)(1-\mathrm{i}(\boldsymbol{\epsilon}+\boldsymbol{\eta})\cdot\boldsymbol{\tau}/2)\hat{q}_\mathrm{R} &(18.43)\\ &= \bar{\hat{q}}_\mathrm{L}\hat{q}_\mathrm{R}-\mathrm{i}\boldsymbol{\eta}\cdot\bar{\hat{q}}_\mathrm{L}\boldsymbol{\tau}\hat{q}_\mathrm{R} &(18.44)\end{aligned}$$

and

$$\bar{\hat{q}}_\mathrm{R}\hat{q}_\mathrm{L}\to\bar{\hat{q}}_\mathrm{R}\hat{q}_\mathrm{L}+\mathrm{i}\boldsymbol{\eta}\cdot\bar{\hat{q}}_\mathrm{R}\boldsymbol{\tau}\hat{q}_\mathrm{L}. \tag{18.45}$$

Equations (18.44) and (18.45) confirm that the term $\bar{\hat{q}}\hat{q}$ in (18.40) is invariant under the isospin part of $\mathrm{SU}(2)_\mathrm{L}\times\mathrm{SU}(2)_\mathrm{R}$ (since $\boldsymbol{\epsilon}$ is not involved), but not invariant under the axial isospin transformations parametrized by $\boldsymbol{\eta}$. The $\bar{\hat{q}}\tau_3\hat{q}$ term explicitly breaks the third component of isospin (resembling an electromagnetic effect), but its magnitude may be expected to be smaller than that of the $\bar{\hat{q}}\hat{q}$ term, being proportional to the difference of the masses, rather than their sum. This suggests that the vacuum will 'align' in such a way as to preserve isospin, but break axial isospin.

18.2 Pion decay and the Goldberger–Treiman relation

We now discuss some of the rather surprising phenomenological implications of spontaneously broken chiral symmetry – specifically, the spontaneous breaking of the axial isospin symmetry. We start by ignoring any 'explicit' quark masses, so that the axial isospin current is conserved, $\partial_\mu\hat{j}^\mu_{i,5}=0$. From sections 17.4 and 17.5 (suitably generalized) we know that this current has non-zero matrix elements between the vacuum and a 'Goldstone' state, which in our case is the pion. We therefore set (cf (17.94))

$$\langle 0|\hat{j}^\mu_{i,5}(x)|\pi_j,p\rangle=-\mathrm{i}p^\mu f_\pi\mathrm{e}^{-\mathrm{i}p\cdot x}\delta_{ij} \tag{18.46}$$

where f_π is a constant with dimensions of mass, and which we expect to be related to a symmetry breaking vev. This is just what we shall find in section 18.3.1. Note that (18.46) is consistent with $\partial_\mu\hat{j}^\mu_{i,5}=0$ if $p^2=0$, i.e. if the pion is massless.

We treat f_π as a phenomenological parameter. Its value can be determined from the rate for the decay $\pi^+\to\mu^+\nu_\mu$ by the following reasoning. In chapter 20 we shall learn that the effective weak Hamiltonian density for this low energy *strangeness non-changing semileptonic transition* is

$$\begin{aligned}\hat{\mathcal{H}}_\mathrm{W}(x) &= \frac{G_\mathrm{F}}{\sqrt{2}}V_\mathrm{ud}\bar{\hat{\psi}}_\mathrm{d}(x)\gamma^\mu(1-\gamma_5)\hat{\psi}_\mathrm{u}(x)\\ &\quad\times[\hat{\psi}_{\nu_\mathrm{e}}(x)\gamma_\mu(1-\gamma_5)\hat{\psi}_\mathrm{e}(x)+\bar{\hat{\psi}}_{\nu_\mu}(x)\gamma_\mu(1-\gamma_5)\hat{\psi}_\mu(x)] &(18.47)\end{aligned}$$

where $G_{\rm F}$ is Fermi constant and $V_{\rm ud}$ is an element of the Cabibbo–Kobayashi–Maskawa (CKM) matrix (see section 20.7.3). Thus the lowest-order contribution to the S-matrix is

$$\begin{aligned}
&-\mathrm{i}\langle\mu^+, p_1; \nu_\mu, p_2| \int \mathrm{d}^4x \hat{\mathcal{H}}_{\rm W}(x)|\pi^+, p\rangle \\
&\quad = -\mathrm{i}\frac{G_{\rm F}}{\sqrt{2}} V_{\rm ud} \int \mathrm{d}^4x \langle\mu^+, p_1; \nu_\mu, p_2|\bar{\hat{\psi}}_{\nu_\mu}(x)\gamma_\mu(1-\gamma_5)\hat{\psi}_\mu(x)|0\rangle \\
&\qquad \times \langle 0|\bar{\hat{\psi}}_{\rm d}(x)\gamma^\mu(1-\gamma_5)\hat{\psi}_{\rm u}(x)|\pi^+, p\rangle. \qquad (18.48)
\end{aligned}$$

The leptonic matrix element gives $\bar{u}_\nu(p_2)\gamma_\mu(1-\gamma_5)v_\mu(p_1)\mathrm{e}^{\mathrm{i}(p_1+p_2)\cdot x}$. For the pionic one, we note that

$$\bar{\hat{\psi}}_{\rm d}(x)\gamma^\mu(1-\gamma_5)\hat{\psi}_{\rm u}(x) = \hat{j}_1^\mu(x) - \mathrm{i}\hat{j}_2^\mu(x) - \hat{j}_{1,5}^\mu(x) + \mathrm{i}\hat{j}_{2,5}^\mu(x) \qquad (18.49)$$

from (18.20) and (18.21). Further, the currents $\hat{j}_i^\mu$ can have no matrix elements between the vacuum (which is a 0^+ state) and the π (which is 0^-), by the following argument. From Lorentz invariance such a matrix element has to be a 4-vector. But since the initial and final parities are different, it would have to be an axial 4-vector[3]. However, the only 4-vector available is the pion's momentum p^μ which is an ordinary (not an axial) 4-vector. On the other hand, precisely for this reason the axial currents $\hat{j}_{i,5}^\mu$ do have a non-zero matrix element, as in (18.46). Noting that $|\pi^+\rangle = \frac{1}{\sqrt{2}}|\pi_1 + \mathrm{i}\pi_2\rangle$, we find that

$$\begin{aligned}
\langle 0|\bar{\hat{\psi}}_{\rm d}(x)\gamma^\mu(1-\gamma_5)\hat{\psi}_{\rm u}(x)|\pi^+, p\rangle &= -\frac{\mathrm{i}}{\sqrt{2}}\langle 0|\hat{j}_{1,5}^\mu - \mathrm{i}\hat{j}_{2,5}|\pi_1 + \mathrm{i}\pi_2\rangle \qquad (18.50) \\
&= -\sqrt{2}p^\mu f_\pi \mathrm{e}^{-\mathrm{i}p\cdot x} \qquad (18.51)
\end{aligned}$$

so that (18.48) becomes

$$\mathrm{i}(2\pi)^4\delta^4(p_1+p_2-p)[G_{\rm F}V_{\rm ud}\bar{u}_\nu(p_2)\gamma_\mu(1-\gamma_5)v(p_1)p^\mu f_\pi]. \qquad (18.52)$$

The quantity in brackets is, therefore, the invariant amplitude for the process, $\mathcal{M}$. Using $p = p_1 + p_2$, we may replace $\not{p}$ in (18.52) by m_μ, neglecting the neutrino mass.

Before proceeding, we comment on the physics of (18.52). The $(1-\gamma_5)$ factor acting on a v spinor selects out the $\gamma_5 = -1$ eigenvalue which, *if the muon was massless*, would correspond to positive helicity for the μ^+ (compare the discussion in section 12.4.2). Likewise, taking the $(1-\gamma_5)$ through the $\gamma^0\gamma^\mu$ factor to act on $u_\nu^\dagger$, it selects the negative helicity neutrino state. Hence the configuration is as shown in figure 18.2, so that the leptons carry off a net spin angular momentum. But this is forbidden, since the pion spin is zero. Hence the amplitude vanishes for massless muons and neutrinos. Now the muon, at least, is not massless, and some 'wrong' helicity is present in

[3]See chapter 4 of volume 1.

FIGURE 18.2
Helicities of *massless* leptons in $\pi^+ \to \mu^+\nu_\mu$ due to the 'V-A' interaction.

its wavefunction, in an amount proportional to m_μ. This is why, as we have just remarked after (18.52), the amplitude is proportional to m_μ. The rate is therefore proportional to m_μ^2. This is a very important conclusion, because it implies that the rate to muons is $\sim (m_\mu/m_\mathrm{e})^2 \sim (400)^2$ times greater than the rate to electrons – a result which agrees with experiment, while grossly contradicting the naive expectation that the rate with the larger energy release should dominate. This, in fact, is one of the main indications for the 'vector-axial vector', or 'V-A', structure of (18.47), as we shall see in more detail in section 20.2.

Problem 18.4 shows that the rate computed from (18.52) is

$$\Gamma_{\pi\to\mu\nu} = \frac{G_\mathrm{F}^2 m_\mu^2 f_\pi^2 (m_\pi^2 - m_\mu^2)^2}{4\pi m_\pi^3} |V_\mathrm{ud}|^2. \tag{18.53}$$

Including radiative corrections, the value

$$f_\pi \simeq 92 \text{ MeV} \tag{18.54}$$

can be extracted.

Consider now another matrix element of $\hat{j}^\mu_{i,5}$, this time between nucleon states. Following an analysis similar to that in section 8.8 for the matrix elements of the electromagnetic current operator between nucleon states, we write

$$\begin{aligned}&\langle \mathrm{N}, p'|\hat{j}^\mu_{i,5}(0)|\mathrm{N}, p\rangle \\ &= \bar{u}(p')\left[\gamma^\mu\gamma_5 F_1^5(q^2) + \frac{\mathrm{i}\sigma^{\mu\nu}}{2M} q_\nu \gamma_5 F_2^5(q^2) + q^\mu \gamma_5 F_3^5(q^2)\right]\frac{\tau_i}{2}u(p),\end{aligned} \tag{18.55}$$

where the F_i^5's are certain form factors, M is the nucleon mass, and $q = p - p'$. The spinors in (18.55) are understood to be written in flavour and Dirac space. Since (with massless quarks) $\hat{j}^\mu_{i,5}$ is conserved – that is $q_\mu \hat{j}^\mu_{i,5}(0) = 0$ – we find

$$\begin{aligned} 0 &= \bar{u}(p')[\not{q}\gamma_5 F_1^5(q^2) + q^2\gamma_5 F_3^5(q^2)]\frac{\tau_i}{2}u(p) \\ &= \bar{u}(p')[(\not{p} - \not{p}')\gamma_5 F_1^5(q^2) + q^2\gamma_5 F_3^5(q^2)]\frac{\tau_i}{2}u(p) \\ &= \bar{u}(p')[-2M\gamma_5 F_1^5(q^2) + q^2\gamma_5 F_3^5(q^2)]\frac{\tau_i}{2}u(p), \end{aligned} \tag{18.56}$$

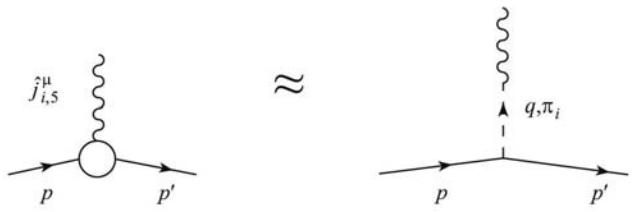


FIGURE 18.3
One pion intermediate state contribution to F_3^5.

using $\not{p}\gamma_5 = -\gamma_5 \not{p}$ and the Dirac equations for $u(p), \bar{u}(p')$. Hence the form factors F_1^5 and F_3^5 must satisfy

$$2MF_1^5(q^2) = q^2 F_3^5(q^2). \tag{18.57}$$

Now the matrix element (18.55) enters into neutron β-decay (as does the matrix element of $\hat{j}_i^{\mu}(0)$). Here, $q^2 \simeq 0$ and (18.57) appears to predict, therefore, that either $M = 0$ (which is certainly not so) or $F_1^5(0) = 0$. But $F_1^5(0)$ can be measured in β decay, and is found to be approximately equal to 1.26; it is conventionally called g_{A}. The only possible conclusion is that F_3^5 *must contain a part proportional to* $1/q^2$. Such a contribution can only arise from the propagator of a massless particle – which, of course, is the pion. This elegant physical argument, first given by Nambu (1960), sheds a revealing new light on the phenomenon of spontaneous symmetry breaking: the existence of the massless particle coupled to the symmetry current $\hat{j}^{\mu}_{i,5}$ 'saves' the conservation of the current.

We calculate the pion contribution to F_3^5 as follows. The process is pictured in figure 18.3. The pion-current matrix element is given by (18.46), and the (massless) propagator is i/q^2. For the $\pi - \mathrm{N}$ vertex, the conventional Lagrangian is

$$\mathrm{i}g_{\pi\mathrm{NN}}\hat{\pi}_i \bar{\hat{N}}\gamma_5\tau_i\hat{N}, \tag{18.58}$$

which is $\mathrm{SU}(2)_{\mathrm{f}}$-invariant and parity conserving since the pion field is a pseudoscalar, and so is $\bar{N}\gamma_5 N$. Putting these pieces together, the contribution of figure 18.3 to the current matrix element is

$$2g_{\pi\mathrm{NN}}\bar{u}(p')\gamma_5\frac{\tau_i}{2}u(p)\,\frac{\mathrm{i}}{q^2}\,(-\mathrm{i}q^{\mu}f_{\pi}), \tag{18.59}$$

and so

$$F_3^5(q^2) = \frac{1}{q^2}\,2g_{\pi\mathrm{NN}}f_{\pi} \tag{18.60}$$

from this contribution. Combining (18.57) with (18.60) we deduce

$$g_{\mathrm{A}} \equiv \lim_{q^2\to 0} F_1^5(q^2) = \frac{g_{\pi\mathrm{NN}}f_{\pi}}{M}, \tag{18.61}$$

the Goldberger–Treiman (1958) relation. Taking $M = 939$ MeV, $g_{\rm A} = 1.26$ and $f_\pi = 92$ MeV one obtains $g_{\pi{\rm NN}} \approx 12.9$, which is only 5% below the experimental value of this effective pion-nucleon coupling constant.

We can repeat the argument leading to the G-T relation but retaining $m_\pi^2 \neq 0$. Equation (18.46) tells us that $\partial_\mu \hat{j}^\mu_{i,5}/(m_\pi^2 f_\pi)$ behaves like a properly normalized pion field, at least when operating on a near mass-shell pion state. This means that the one-nucleon matrix element of $\partial_\mu \hat{j}^\mu_{i,5}$ is (cf (18.59))

$$2g_{\pi{\rm NN}}\bar{u}(p')\gamma_5 \frac{\tau_i}{2} u(p) \frac{\mathrm{i}}{q^2 - m_\pi^2} m_\pi^2 f_\pi, \tag{18.62}$$

while from (18.55) it is given by

$$\mathrm{i}\bar{u}(p')[-2M\gamma_5 F_1^5(q^2) + q^2\gamma_5 F_3^5(q^2)]\frac{\tau_i}{2}u(p). \tag{18.63}$$

Hence

$$-2MF_1^5(q^2) + q^2 F_3^5(q^2) = \frac{2g_{\pi{\rm NN}} m_\pi^2 f_\pi}{q^2 - m_\pi^2}. \tag{18.64}$$

Also, in place of (18.60) we now have

$$F_3^5(q^2) = \frac{1}{q^2 - m_\pi^2} 2g_{\pi{\rm NN}} f_\pi. \tag{18.65}$$

Equations (18.64) and (18.65) are consistent for $q^2 = m_\pi^2$ if

$$F_1^5(q^2 = m_\pi^2) = g_{\pi{\rm NN}} f_\pi / M. \tag{18.66}$$

$F_1^5(q^2)$ varies only slowly from $q^2 = 0$ to $q^2 = m_\pi^2$, since it contains no rapidly varying pion pole contribution, and so we recover the G-T relation again.

Amplitudes involving *two* Goldstone pions can be calculated by an extension of these techniques. However, a much more efficient method is available, through the use of *effective Lagrangians*, which capture the low energy dynamics of the Goldstone modes.

18.3 Effective Lagrangians

18.3.1 The linear and non-linear σ-models

We begin by considering the linear σ-model, which has the same Lagrangian as the one considered in section 17.6,

$$\hat{\mathcal{L}}_\Phi = (\partial_\mu \hat{\phi}^\dagger)(\partial^\mu \hat{\phi}) + \mu^2 \hat{\phi}^\dagger \hat{\phi} - \frac{\lambda}{4}(\hat{\phi}^\dagger \hat{\phi})^4, \tag{18.67}$$

but we shall interpret it differently here. The sign of the μ^2 term has been chosen to induce spontaneous symmetry breaking. In section 17.6, $\hat{\phi}$ was the SU(2) doublet

$$\hat{\phi} = \begin{pmatrix} \frac{1}{\sqrt{2}}(\hat{\phi}_1 + \mathrm{i}\hat{\phi}_2) \\ \frac{1}{\sqrt{2}}(\hat{\phi}_3 + \mathrm{i}\hat{\phi}_4) \end{pmatrix}, \tag{18.68}$$

in terms of which (18.67) becomes

$$\hat{\mathcal{L}}_\Phi = \frac{1}{2}\partial_\mu\hat{\phi}_a\partial^\mu\hat{\phi}_a + \frac{1}{2}\mu^2\hat{\phi}_a\hat{\phi}_a - \frac{\lambda}{16}(\hat{\phi}_a\hat{\phi}_a)^2, \tag{18.69}$$

where the sum on $a = 1$ to 4 is understood. Evidently (18.69) is invariant under transformations which preserve the 'dot product' $\hat{\phi}_a\hat{\phi}_a$, namely the transformations of SO(4). This group is discussed in appendix M, section M.4.3. We note there that the algebra of the generators of SO(4) is the same as that of SU(2) × SU(2), which is the algebra of the chiral charges in (18.28)–(18.30). This suggests that we should rewrite (18.69) in such a way as to reveal its $\mathrm{SU(2)_L}\times\mathrm{SU(2)_R}$ symmetry, rather than its O(4) symmetry. Three of the four fields will then be identified with the Goldstone bosons associated with the spontaneous breaking of the 'R – L' part; they will in turn be identified with the (massless) pions.

One way to bring out the chiral symmetry of (18.69) is to write

$$\hat{\phi} = \begin{pmatrix} (\hat{\pi}_2 + \mathrm{i}\hat{\pi}_1)/\sqrt{2} \\ (\hat{\sigma} - \mathrm{i}\hat{\pi}_3)/\sqrt{2} \end{pmatrix} = \frac{1}{\sqrt{2}}\hat{\Sigma}\begin{pmatrix} 0 \\ 1 \end{pmatrix}, \tag{18.70}$$

where

$$\hat{\Sigma} = \hat{\sigma} + \mathrm{i}\boldsymbol{\tau}\cdot\hat{\boldsymbol{\pi}}. \tag{18.71}$$

Then

$$\hat{\phi}^\dagger\hat{\phi} = \frac{1}{4}\mathrm{Tr}(\hat{\Sigma}^\dagger\hat{\Sigma}), \tag{18.72}$$

and (18.69) becomes

$$\hat{\mathcal{L}}_\Sigma = \frac{1}{4}\mathrm{Tr}(\partial_\mu\hat{\Sigma}^\dagger\partial^\mu\hat{\Sigma}) + \frac{\mu^2}{4}\mathrm{Tr}(\hat{\Sigma}^\dagger\hat{\Sigma}) - \frac{\lambda}{64}\mathrm{Tr}(\hat{\Sigma}^\dagger\hat{\Sigma})^2. \tag{18.73}$$

This Lagrangian is invariant under the $\mathrm{SU(2)_L}\times\ \mathrm{SU(2)_R}$ transformation

$$\hat{\Sigma} \to U_\mathrm{L}\hat{\Sigma}U_\mathrm{R}^\dagger \tag{18.74}$$

where

$$U_\mathrm{L} = \exp(-\mathrm{i}\boldsymbol{\alpha}_\mathrm{L}\cdot\boldsymbol{\tau}/2), \quad U_\mathrm{R} = \exp(-\mathrm{i}\boldsymbol{\alpha}_\mathrm{R}\cdot\boldsymbol{\tau}/2) \tag{18.75}$$

are two independent SU(2) transformations (remember that $\mathrm{Tr}AB = \mathrm{Tr}BA$). For the case of infinitesimal transformations, we find (problem 18.5)

$$\hat{\sigma} \quad\to\quad \hat{\sigma} - \boldsymbol{\eta}\cdot\hat{\boldsymbol{\pi}} \tag{18.76}$$

$$\hat{\boldsymbol{\pi}} \quad\to\quad \hat{\boldsymbol{\pi}} + \boldsymbol{\eta}\hat{\sigma} + \boldsymbol{\epsilon}\times\hat{\boldsymbol{\pi}}, \tag{18.77}$$

where

$$\boldsymbol{\eta} = (\boldsymbol{\epsilon}_{\rm R} - \boldsymbol{\epsilon}_{\rm L})/2, \quad \boldsymbol{\epsilon} = (\boldsymbol{\epsilon}_{\rm R} + \boldsymbol{\epsilon}_{\rm L})/2. \tag{18.78}$$

Evidently $\boldsymbol{\epsilon}_{\rm R} = \boldsymbol{\eta} + \boldsymbol{\epsilon}$ and $\boldsymbol{\epsilon}_{\rm L} = \boldsymbol{\epsilon} - \boldsymbol{\eta}$, which we may compare with the L and R transformation of the quark fields in (18.37), (18.38).

With the sign of μ^2 as in (18.73), the classical potential has a minimum at

$$\hat{\sigma}^2 + \hat{\boldsymbol{\pi}}^2 = 4\mu^2/\lambda \equiv v^2, \tag{18.79}$$

which we interpret as the symmetry breaking condition

$$\langle 0|\hat{\sigma}^2 + \hat{\boldsymbol{\pi}}^2|0\rangle = v^2. \tag{18.80}$$

Let us choose the particular ground state

$$\langle 0|\hat{\sigma}|0\rangle = v, \quad \langle 0|\hat{\boldsymbol{\pi}}|0\rangle = 0, \tag{18.81}$$

which is actually the same as (17.103). Referring back to (18.76) and (18.77) we see that this vacuum is invariant under 'L + R' transformations with parameters $\boldsymbol{\epsilon}$, but not under 'L − R' transformations with parameters $\boldsymbol{\eta}$. These correspond respectively to the $\mathrm{SU(2)_f}$ flavour isospin, and $\mathrm{SU(2)_{f5}}$ axial flavour isospin, transformations on the quark fields. So this vacuum spontaneously breaks the axial isospin symmetry. Fluctuations away from this minimum are described by fields $\hat{\boldsymbol{\pi}}$ and $\hat{s} = \hat{\sigma} - v$. Placing this shift into (18.73) we find that $\hat{\mathcal{L}}_\Sigma$ becomes $\hat{\mathcal{L}}_s$ where

$$\hat{\mathcal{L}}_s = \frac{1}{2}\partial_\mu \hat{s}\partial^\mu \hat{s} - \mu^2\hat{s}^2 + \frac{1}{2}\partial_\mu \hat{\boldsymbol{\pi}}\cdot\partial^\mu\hat{\boldsymbol{\pi}} - \frac{\lambda}{4}v\hat{s}(\hat{s}^2 + \hat{\boldsymbol{\pi}}^2) - \frac{\lambda}{16}(\hat{s}^2 + \hat{\boldsymbol{\pi}}^2)^2, \tag{18.82}$$

discarding an irrelevant constant. As expected, the field $\hat{s}$ is massive (with mass $\sqrt{2}\mu$), while the fields $\hat{\boldsymbol{\pi}}$ are massless, and may be identified with the Goldstone modes associated with the spontaneous breaking of the axial isospin symmetry.

The Lagrangian $\hat{\mathcal{L}}_s$ incorporates the correct symmetries, and can be used to calculate $\pi - \pi$ scattering, for example (in the massless limit). But it is not the most efficient Lagrangian to use, as we can see from the following considerations. Consider the amplitude for $\pi^+ - \pi^0$ scattering, in tree approximation (Donoghue *et al.* 1992). The contributing terms in $\hat{\mathcal{L}}_s$ are

$$\hat{\mathcal{L}}_{\pi-\pi} = -\frac{\lambda}{16}(\hat{\boldsymbol{\pi}}^2)^2 - \frac{\lambda}{4}v\hat{s}\hat{\boldsymbol{\pi}}^2, \tag{18.83}$$

which we can rewrite in terms of the charged and neutral fields as

$$\hat{\mathcal{L}}_{\pi-\pi} = -\frac{\lambda}{16}(2\hat{\pi}_+^\dagger\hat{\pi}_+ + \hat{\pi}^{0\ 2})^2 - \frac{\lambda}{4}v\hat{s}(2\hat{\pi}_+^\dagger\hat{\pi}_+ + \hat{\pi}^{0\ 2}). \tag{18.84}$$

Then the terms responsible for $\pi^+ - \pi^0$ scattering at tree level are

$$-\frac{\lambda}{4}\hat{\pi}_+^\dagger\hat{\pi}_+\hat{\pi}^{0\ 2} - \frac{\lambda}{2}v\hat{s}\left(\hat{\pi}_+^\dagger\hat{\pi}_+ + \frac{1}{2}\hat{\pi}^{0\ 2}\right). \tag{18.85}$$

The first of these represents a four-pion contact interaction with amplitude

$$-\mathrm{i}\lambda/2, \tag{18.86}$$

while the second contributes an s-exchange graph in the t-channel with amplitude

$$(-\mathrm{i}\lambda v/2)^2 \frac{\mathrm{i}}{q^2 - 2\mu^2}, \tag{18.87}$$

where q is the 4-momentum transfer $q = p'_+ - p_+ = p_0 - p'_0$. The sum of these is

$$-\mathrm{i}\lambda/2 \frac{q^2}{q^2 - 2\mu^2}, \tag{18.88}$$

which reduces to $\mathrm{i}q^2/v^2$ for $q \approx 0$. Thus, despite the apparent constant 4-boson piece (18.86), the total amplitude in fact *vanishes* as $q^2 \to 0$, due to a cancellation.

This cancellation is not an accident. It is generally true that Goldstone fields enter only via their derivatives, which bring factors of momenta into the amplitudes. We drew attention to this following equation (17.85), and the same is true of the $\hat{\boldsymbol{\theta}}$ fields in (17.107). This suggests that it is both possible, and more efficient, to recast $\hat{\mathcal{L}}_s$ into a form in which only the derivatives of the Goldstone fields enter. Equation (17.107) indicates how to do this: we define new pion fields (but call them the same) by

$$\hat{\Sigma} = (v + \hat{S})\hat{U}, \quad \hat{U} = \exp(\mathrm{i}\boldsymbol{\tau} \cdot \hat{\boldsymbol{\pi}}/v), \tag{18.89}$$

where $\hat{S}$ is invariant under $\mathrm{SU}(2)_\mathrm{L} \times \mathrm{SU}(2)_\mathrm{R}$, and where $\hat{U}$ transforms by

$$\hat{U} \to U_\mathrm{L} \hat{U} U_\mathrm{R}^\dagger. \tag{18.90}$$

Now $\hat{\Sigma}^\dagger \hat{\Sigma} = (v + \hat{S})^2$, and the Goldstone modes have been transformed away from the potential terms in $\hat{\mathcal{L}}_\Sigma$, reappearing in the derivative terms instead. We write the transformed Lagrangian as $\hat{\mathcal{L}}_S$ where

$$\hat{\mathcal{L}}_S = \frac{1}{2}\partial_\mu \hat{S} \partial^\mu \hat{S} - \mu^2 \hat{S}^2 + \frac{1}{4}(v + \hat{S})^2 \mathrm{Tr}(\partial_\mu \hat{U} \partial^\mu \hat{U}^\dagger) - \frac{1}{4}\lambda v \hat{S}^3 - \frac{\lambda}{16}\hat{S}^4, \tag{18.91}$$

where we have used

$$\hat{U}^\dagger \partial^\mu \hat{u} + \partial^\mu \hat{U}^\dagger \hat{U} = 0, \tag{18.92}$$

which follows from the unitary condition $\hat{U}^\dagger \hat{U} = 1$.

When $\partial_\mu \hat{U}$ is expanded in powers of $\hat{\boldsymbol{\pi}}$, we recover a kinetic energy piece

$$\frac{1}{2}\partial_\mu \hat{\boldsymbol{\pi}} \cdot \partial^\mu \hat{\boldsymbol{\pi}}, \tag{18.93}$$

and all other terms involve derivatives of $\hat{\boldsymbol{\pi}}$. In particular, the term with

the lowest number of derivatives which contributes to the $\pi - \pi$ scattering amplitude is

$$\frac{1}{6v^2}[(\hat{\boldsymbol{\pi}} \cdot \partial_\mu \hat{\boldsymbol{\pi}})(\hat{\boldsymbol{\pi}} \cdot \partial^\mu \hat{\boldsymbol{\pi}}) - \hat{\boldsymbol{\pi}}^2 \partial_\mu \hat{\boldsymbol{\pi}} \cdot \partial^\mu \hat{\boldsymbol{\pi}}], \tag{18.94}$$

since the $\hat{S} - \hat{\boldsymbol{\pi}} - \hat{\boldsymbol{\pi}}$ vertex already has two derivatives. The reader may check that the amplitude for $\pi^+\pi^0 \to \pi^+\pi^0$ calculated from (18.94) is $\mathrm{i}q^2/v^2$, exactly as before, but this time without having to go through the cancellation argument.

The fields in $\hat{\Sigma}$ on the one hand, and in $\hat{S}$ and $\hat{U}$ on the other, are related non-linearly, but a physical amplitude calculated with either representation has turned out to be the same, in this simple case. It is in fact generally true that such non-linear field redefinitions lead to the same physics (Haag 1958, Coleman, Wess and Zumino 1969, Callan, Coleman, Wess and Zumino 1969). It is clearly advantageous to work with $\hat{\mathcal{L}}_S$, which builds in the desired derivatives of the Goldstone modes.

Indeed, we can simplify matters even further. Since $\hat{S}$ is invariant under $\mathrm{SU}(2)_\mathrm{L} \times \mathrm{SU}(2)_\mathrm{R}$, the full symmetry of the Lagrangian is maintained with only the field $\hat{U}$, transforming by (18.90), and we may as well discard $\hat{S}$ altogether. The dynamics of the Goldstone sector are then described by the *non-linear σ-model*, with Lagrangian

$$\hat{\mathcal{L}}_2 = \frac{v^2}{4}\mathrm{Tr}(\partial_\mu \hat{U} \partial^\mu \hat{U}^\dagger). \tag{18.95}$$

This is the most general Lagrangian that involves the Goldstone fields, exhibits the desired symmetry, and contains only two derivatives.

Since $\hat{\mathcal{L}}_2$ is invariant under the $\mathrm{SU}(2)_\mathrm{L} \times \mathrm{SU}(2)_\mathrm{R}$ transformations (18.75), we can calculate the associated Noether currents (problem 18.6), obtaining

$$\hat{j}^\mu_{i,\mathrm{L}}(\hat{U}) = \frac{-\mathrm{i}v^2}{8}\mathrm{Tr}[\tau_i \hat{U} \partial^\mu \hat{U}^\dagger - \tau_i (\partial^\mu \hat{U})\hat{U}^\dagger] = \frac{-\mathrm{i}v^2}{4}\mathrm{Tr}(\tau_i \hat{U} \partial^\mu \hat{U}^\dagger), \tag{18.96}$$

and

$$\hat{j}^\mu_{i,\mathrm{R}}(\hat{U}) = \frac{\mathrm{i}v^2}{8}\mathrm{Tr}[\tau_i (\partial^\mu \hat{U}^\dagger)\hat{U} - \tau_i \hat{U}^\dagger \partial^\mu \hat{U}] = \frac{-\mathrm{i}v^2}{4}\mathrm{Tr}(\tau_i \hat{U}^\dagger \partial^\mu \hat{U}). \tag{18.97}$$

The axial 'R – L' current is then

$$\hat{j}^\mu_{i5}(\hat{U}) = \frac{\mathrm{i}v^2}{4}\mathrm{Tr}[\tau_i (\hat{U} \partial^\mu \hat{U}^\dagger - \hat{U}^\dagger \partial^\mu \hat{U})], \tag{18.98}$$

and the vector 'R + L' current is

$$\hat{j}^\mu_i(\hat{U}) = \frac{\mathrm{i}v^2}{4}\mathrm{Tr}[\tau_i (\hat{U} \partial^\mu \hat{U}^\dagger + \hat{U}^\dagger \partial^\mu \hat{U})]. \tag{18.99}$$

Expanding (18.98) in powers of the pion field, we find

$$\hat{j}^\mu_{i,5}(\hat{U}) = v \partial^\mu \hat{\pi}_i + \ldots, \tag{18.100}$$

which we may compare with (17.93). Just as in equation (17.94), (18.100) implies that this axial current has a matrix element between the vacuum and the one-Goldstone state:

$$\langle 0|\hat{j}^{\mu}_{i,5}(\hat{U})|\pi_j, p\rangle = -\mathrm{i}p^{\mu}v\mathrm{e}^{-\mathrm{i}p\cdot x}\delta_{ij}. \tag{18.101}$$

Now comes the pay-off: this is the *same* symmetry current which enters into weak interactions, for which we already defined the vacuum-to-one-particle matrix element in terms of the pion decay constant f_π, via equation (18.46). Comparing (18.101) with (18.46) we identify

$$v = f_\pi. \tag{18.102}$$

Thus, finally, the dynamics of our massless pions, to lowest order in an expansion in powers of momenta, is given by the Lagrangian

$$\hat{\mathcal{L}}_2 = \frac{1}{4}f_\pi^2 \mathrm{Tr}(\partial_\mu \hat{U}\partial^\mu \hat{U}^\dagger). \tag{18.103}$$

It is quite remarkable that the low energy dynamics of the (massless) Goldstone modes is completely determined in terms of one constant, measurable in π decay.

The Lagrangian of (18.103) is an example of an *effective Lagrangian*. By this is meant, broadly, any Lagrangian which involves the presumed relevant degrees of freedom (here the Goldstone modes), and respects desired symmetries of the theory. Evidently it is implied that there is some 'underlying theory', couched in terms of different degrees of freedom (here quarks and gluons), from which the symmetries have been abstracted. It is important to realize that an effective Lagrangian may or may not be renormalizable. Whereas our starting Lagrangian $\hat{\mathcal{L}}_\sigma$ is renormalizable, $\hat{\mathcal{L}}_2$ is not: clearly the latter contains terms with arbitrarily many pion fields, which are operators of arbitrarily high dimension, compensated by negative powers of f_π^2. As it stands, $\hat{\mathcal{L}}_2$ can only be used at tree level – as, for example, in the calculation of $\pi - \pi$ scattering using the interaction(18.94), for which the amplitude has an energy dependence of the form E^2/f_π^2, where E is the order of magnitude of the particles' energy or momentum. This interaction has mass dimension 6, and its coupling $1/f_\pi^2$ has dimension $(\text{mass})^{-2}$, like the 4-fermion interaction considered in section 11.8. It is therefore not renormalizable. However, the argument of section 11.8 suggests that a loop-by-loop renormalization programme is possible, and this was shown to be the case by Weinberg (1979a). Each loop built from the interaction (18.94) will carry an extra two powers of energy, to compensate the $1/f_\pi^2$ in the coupling. Thus f_π (or perhaps this multiplied by factors like 4 and π, if we are lucky) provides the energy scale characteristic of a non-renormalizable theory: as we go up in energy, we need more loops. But, at each loop order new divergences appear, which require additional counter terms for renormalization. Thus at any given order in

E^2/f_π^2, we must ensure that our effective Lagrangian contains all the appropriate counter terms which are allowed by the symmetry. For example, at one-loop order for $\hat{\mathcal{L}}_2$, we need to include the 4-derivative terms

$$\hat{\mathcal{L}}_4 = c_1 \mathrm{Tr}(\partial_\mu \hat{U} \partial^\mu \hat{U}^\dagger \partial_\nu \hat{U} \partial^\nu \hat{U}^\dagger) + c_2 \mathrm{Tr}(\partial_\mu \hat{U} \partial_\nu \hat{U}^\dagger \partial^\mu \hat{U} \partial^\nu \hat{U}^\dagger). \qquad (18.104)$$

To perform a one-loop calculation, one uses $\hat{\mathcal{L}}_2$ at tree-level and in one-loop diagrams, and $\hat{\mathcal{L}}_4$ at tree-level only.

Real pions, however, are not massless, nor are real quarks. We need to extend our effective Lagrangian to include *explicit* chiral symmetry breaking mass terms.

18.3.2 Inclusion of explicit symmetry breaking: masses for pions and quarks

Consider the term

$$\hat{\mathcal{L}}_{m_\pi} = \frac{m_\pi^2}{4} \mathrm{Tr}(\hat{U} + \hat{U}^\dagger). \qquad (18.105)$$

This is invariant only under the restricted set of transformations with $\boldsymbol{\alpha}_\mathrm{L} = \boldsymbol{\alpha}_\mathrm{R}$, that is transformations such that $U_\mathrm{R} = U_\mathrm{L}$, for then $\mathrm{Tr}\hat{U} \to \mathrm{Tr}(U_\mathrm{R} \hat{U} U_\mathrm{R}^\dagger) = \mathrm{Tr}\hat{U}$. Such transformations form the SU(2) flavour isospin group. The term (18.105) breaks the axial isospin group explicitly, which would correspond to transformations with $\boldsymbol{\alpha}_\mathrm{L} = -\boldsymbol{\alpha}_\mathrm{R}$, or equivalently $U_\mathrm{L} = U_\mathrm{R}^\dagger$, under which $\hat{U} \to U_\mathrm{L} \hat{U} U_\mathrm{L}$. Expanding (18.105) to second order in the pion fields, we find the term

$$\hat{\mathcal{L}}_{\mathrm{quad},m_\pi} = -\frac{1}{2} m_\pi^2 \boldsymbol{\pi}^2 \qquad (18.106)$$

which, together with (18.93), shows that the pion field now has mass m_π. Higher-order terms can be added, m_π^2 counting as equivalent to two derivatives. The low energy expansion is now an expansion in both the energy E and the pion mass m_π. This is called *chiral perturbation theory*, or ChPT for short.

For example, to calculate $\pi - \pi$ scattering to order E^2, we use $\hat{\mathcal{L}}_2 + \hat{\mathcal{L}}_{m_\pi}$ at tree-level, expanded up to fourth power in the pion fields. The result is to change the amplitude for $\pi^+\pi^0 \to \pi^+\pi^0$ from $\mathrm{i}(p'_+ - p_+)^2/f_\pi^2$ to $\mathrm{i}[(p'_+ - p_+)^2 - m_\pi^2]/f_\pi^2$. By considering the general $\pi - \pi$ amplitude, predictions for the scattering lengths can be made for low energy observables, for example the s-wave scattering lengths a_0 and a_2 in the isospin 0 and 2 channels. The results (first calculated by Weinberg 1966 using current algebra techniques) are

$$a_0 = \frac{7 m_\pi^2}{32\pi f_\pi^2} = 0.16\, m_\pi^{-1}, \quad a_2 = -\frac{m_\pi^2}{16\pi f_\pi^2} = -0.045\, m_\pi^{-1} \qquad (18.107)$$

The experimental values are $a_0 = 0.26 \pm 0.05\ m_\pi^{-1}$ and $a_2 = -0.028 \pm 0.012\ m_\pi^{-1}$, as given by Donoghue *et al.* (1992). The next order in ChPT improves upon these results.

A systematic exposition of ChPT at the one-loop level was given by Gasser and Leutwyler (1984). Bijnens *et al.* (1996) carried the $\pi - \pi$ calculation to two-loop order; see also Colangelo *et al.* (2001).

It is clear that there must be some relation between the masses of the u and d quarks (in the SU(2) flavour case) and the pion mass, since the latter must vanish in the limit $m_{\mathrm{u}} = m_{\mathrm{d}} = 0$. To see this connection, we consider the quark mass term in the 2-flavour QCD Lagrangian, which is

$$-\bar{\hat{q}}\,\boldsymbol{m}_2\hat{q}, \quad \boldsymbol{m}_2 = \begin{pmatrix} m_{\mathrm{u}} & 0 \\ 0 & m_{\mathrm{d}} \end{pmatrix}. \tag{18.108}$$

Let us now redefine the quark fields (compare (17.107) and (18.17)) by

$$\hat{q} = \exp[-\mathrm{i}\boldsymbol{\tau}\cdot\hat{\boldsymbol{\pi}}\gamma_5/(2f_\pi)]\,\hat{f}. \tag{18.109}$$

This transformation is a perfectly good parametrization of the Goldstone fields associated with the axial symmetry (18.17), and effectively removes them from the new fermion fields $\hat{f}$. The quark mass term now becomes

$$-\bar{\hat{f}}\exp[-\mathrm{i}\boldsymbol{\tau}\cdot\hat{\boldsymbol{\pi}}\gamma_5/(2f_\pi)]\,\boldsymbol{m}_2\exp[-\mathrm{i}\boldsymbol{\tau}\cdot\hat{\boldsymbol{\pi}}/(2f_\pi)]\hat{f}. \tag{18.110}$$

We now make the assumption that the axial SU(2) is spontaneously broken in QCD, by imposing a non-zero vev on the symmetry-breaking operator $\bar{\hat{f}}\hat{f}$:

$$\langle 0|\bar{\hat{f}}_i\hat{f}_j|0\rangle = -f_\pi^2 B\delta_{ij} \quad (i,j=1,2). \tag{18.111}$$

Expanding (18.110) up to second order in the pion fields, retaining just the expectation value of the fermion bilinear[4], we find a mass term

$$-\frac{1}{2}B(m_{\mathrm{u}}+m_{\mathrm{d}})\hat{\boldsymbol{\pi}}^2, \tag{18.112}$$

from which the relation (Gasser and Leutwyler 1982)

$$m_\pi^2 = -\frac{(m_{\mathrm{u}}+m_{\mathrm{d}})}{f_\pi^2}\langle 0|\bar{\hat{f}}\hat{f}|0\rangle \tag{18.113}$$

follows, where $\bar{\hat{f}}\hat{f}$ represents either $\bar{\hat{f}}_{\mathrm{u}}\hat{f}_{\mathrm{u}}$ or $\bar{\hat{f}}_{\mathrm{d}}\hat{f}_{\mathrm{d}}$. From (18.113) we can see that the *square* of the pion mass is proportional to the average u-d quark mass (provided of course that B does not accidentally vanish), and goes to zero as they do; $\langle 0|\bar{\hat{f}}\hat{f}|0\rangle$ is the 'chiral condensate' (cf figure 18.1).

Lattice QCD (see chapter 16) can be used to test equation (18.113), since simulations can be done for a range of quark masses, and the relation between m_π^2 and $m_{\mathrm{u,d}}$ can be checked. Conversely, ChPT can assist lattice QCD calculations by guiding the extrapolation of the calculated results to quark

[4] A formal justification of this step is provided by Weinberg (1996), section 19.6.

mass values lighter than can presently be simulated. For example, Noaki *et al.* (2008) have reported the results of such a calculation, using 2 light dynamical quark flavours, in the overlap fermion formalism (Neuberger 1998a, 1998b), which preserves chiral symmetry at finite lattice spacing. Their pion masses ranged from 290 MeV to 750 MeV, and they compared their results with the predictions of ChPT at one-loop (Gasser and Leutwyler 1984) and two-loop (Colangelo *et al.* 2001). They found good fits to the ChPT formulae, and extracted quark masses (in the $\overline{\rm MS}$ scheme at the scale 2 GeV) of about 4.5 MeV; they also found $|\langle 0|\bar{\hat{f}}\hat{f}|0\rangle \sim (235\,{\rm GeV})^3$, in the $\overline{\rm MS}$ scheme at 2 GeV scale. Studies by this and other groups are continuing, with 3 light flavours, lighter pion masses, and other lattice fermion formalisms.

18.3.3 Extension to $\mathbf{SU(3)_{f\,L}{\times}SU(3)_{f\,R}}$

To the extent that the strange quark is also 'light' on hadronic scales, the QCD Lagrangian has the larger symmetry of ${\rm SU(3)_{f\,L}{\times}SU(3)_{f\,R}}$, which breaks spontaneously so as to preserve the flavour symmetry ${\rm SU(3)_f}$, and produce an SU(3) octet of pseudoscalar Goldstone bosons: $\pi^\pm, \pi^0, {\rm K}^\pm, {\rm K}^0, \bar{\rm K}^0$ and η_8 (see figure 12.4). The effective Lagrangian approach to the dynamics of the Goldstone fields can be easily extended to chiral SU(3). One simply replaces $\hat{U} = \exp({\rm i}\boldsymbol{\tau}\cdot\hat{\boldsymbol{\pi}}/f_\pi)$ by $\hat{V} = \exp({\rm i}\boldsymbol{\lambda}\cdot\hat{\boldsymbol{\phi}}/f_\pi)$ where

$$\frac{1}{\sqrt{2}}\sum_{a=1}^{8}\lambda_a\hat{\phi}_a = \begin{pmatrix} \frac{1}{\sqrt{2}}\hat{\pi}^0 + \frac{1}{\sqrt{6}}\hat{\eta}_8 & \hat{\pi}^+ & \hat{K}^+ \\ \hat{\pi}^- & -\frac{1}{\sqrt{2}}\hat{\pi}^0 + \frac{1}{\sqrt{6}}\hat{\eta}_8 & \hat{K}^0 \\ \hat{K}^- & \bar{K}^0 & -\frac{2}{\sqrt{6}}\hat{\eta}_8 \end{pmatrix}. \quad (18.114)$$

One easily verifies that the kinetic terms in

$$\hat{\mathcal{L}}_2 = \frac{f_\pi^2}{4}{\rm Tr}\partial_\mu\hat{V}\partial^\mu\hat{V} \quad (18.115)$$

have the correct normalization, using ${\rm Tr}\lambda_a\lambda_b = 2\delta_{ab}$. The 3-flavour quark mass term is now

$$-\bar{\hat{f}}\exp[-{\rm i}\boldsymbol{\lambda}\cdot\hat{\boldsymbol{\phi}}\gamma_5/(2f_\pi)]\,\boldsymbol{m}_3\exp[-{\rm i}\boldsymbol{\lambda}\cdot\hat{\boldsymbol{\phi}}\gamma_5/(2f_\pi)]\,\hat{f} \quad (18.116)$$

where

$$\boldsymbol{m}_3 = \begin{pmatrix} m_{\rm u} & 0 & 0 \\ 0 & m_{\rm d} & 0 \\ 0 & 0 & m_{\rm s} \end{pmatrix}. \quad (18.117)$$

The axial SU(3) symmetry breaking vev is

$$\langle 0|\bar{\hat{f}}_i\hat{f}_j|0\rangle = -f_\pi^2 B\delta_{ij} \quad (i,j=1,2,3) \quad (18.118)$$

and the meson mass term is

$$-\frac{B}{2}{\rm Tr}\,\{(\boldsymbol{\lambda}\cdot\hat{\boldsymbol{\phi}})^2\boldsymbol{m}_3\}. \quad (18.119)$$

This yields (problem 18.7)

$$m_{\pi^+}^2 = m_{\pi^0}^2 = B(m_\mathrm{u} + m_\mathrm{d}), \tag{18.120}$$

$$m_{K^+}^2 = B(m_\mathrm{u} + m_\mathrm{s}), \tag{18.121}$$

$$m_{K^0}^2 = B(m_\mathrm{d} + m_\mathrm{s},) \tag{18.122}$$

$$m_{\eta_8}^2 = \frac{1}{3}B(m_\mathrm{u} + m_\mathrm{d} + 4m_\mathrm{s}), \tag{18.123}$$

and there is also a term which mixes π^0 and η_8:

$$m_{\pi\,\eta}^2 = \frac{B}{\sqrt{3}}(m_\mathrm{u} - m_\mathrm{d}). \tag{18.124}$$

It is interesting that the charged and neutral pions have the same mass, even though we have made no assumption about the ratio of m_u to m_d. The observed pion mass differences arise from electromagnetism.

If we ignore for the moment electromagnetic mass differences, we can deduce from (18.120)–(18.122) the relation

$$\frac{m_\pi^2}{2m_\mathrm{K}^2 - m_\pi^2} = \frac{m_\mathrm{u} + m_\mathrm{d}}{m_\mathrm{s}}. \tag{18.125}$$

The left-hand side is approximately equal to 0.04, so we learn that the non-strange quarks are about 1/25 times as heavy as the strange quark. We also obtain

$$m_{\eta_8}^2 = \frac{1}{3}(4m_\mathrm{K}^2 - m_\pi^2), \tag{18.126}$$

which is the Gell-Mann–Okubo formula for the (squared) masses of the pseudoscalar meson octet (Gell-Mann 1961, Okubo 1962). Using average values for the K and π masses, the relation (18.126) predicts $m_{\eta_8}^2 = 566$ MeV, quite close to the η (548 MeV).

Further progress requires the inclusion of electromagnetic effects, since m_u and m_d are themselves comparable to the electromagnetic mass differences. Including these effects, Weinberg (1996) estimates

$$\frac{m_\mathrm{d}}{m_\mathrm{s}} \approx 0.050, \quad \frac{m_\mathrm{u}}{m_\mathrm{s}} \approx 0.027; \tag{18.127}$$

see also Leutwyler (1996). Note that the d quark is almost twice as heavy as the u quark: according to QCD, the origin of SU(2) isospin symmetry is not that $m_\mathrm{u} \approx m_\mathrm{d}$, but that both are very small compared with, say, $\Lambda_{\overline{\mathrm{MS}}}$.

All the results we have given are subject to correction by the inclusion of higher-order effects in the ChPT expansion. In the case of chiral SU(3), the fourth-order Lagrangian $\hat{\mathcal{L}}_4$ contains 8 terms (Gasser and Leutwyler 1984, 1985). Donoghue *et al.* (1992) give a clear exposition of ChPT to one-loop order.

18.4 Chiral anomalies

In all our discussions of symmetries so far – unbroken, approximate, and spontaneously broken – there is one result on which we have relied, and never queried. We refer to Noether's theorem, as discussed in section 12.3.1. This states that for every continuous symmetry of a Lagrangian, there is a corresponding conserved current. We demonstrated this result in some special cases, but we have now to point out that while it is undoubtedly valid at the level of the *classical* Lagrangian and field equations, we did not investigate whether quantum corrections might violate the classical conservation law. This can, in fact, happen and when it does the afflicted current (or its divergence) is said to be 'anomalous', or to contain an 'anomaly'. General analysis shows that anomalies occur in renormalizable theories of fermions coupled to both vector and axial vector currents. We may therefore expect to find anomalies among the vector and axial vector flavour currents which we have been discussing.

One way of understanding how anomalies arise is through consideration of the renormalization process, which is in general necessary once we get beyond the classical ('tree level') approximation. As we saw in volume 1, this will invariably entail some *regularization* of divergent integrals. But the specific example of the $O(e^2)$ photon self-energy studied in section 11.3 showed that a simple cut-off form of regularization already violated the current conservation (or gauge invariance) condition (11.21). In that case, it was possible to find alternative regularizations which respected electromagnetic current conservation, and were satisfactory. Anomalies arise when *both* axial and vector symmetry currents are present, since it is not possible to find a regularization scheme which preserves both vector and axial vector current conservation (Adler 1970, Jackiw 1972, Adler and Bardeen 1969).

We shall not attempt an extended discussion of this technical subject. But we do want to alert the reader to the existence of these anomalies; to indicate how they arise in one simple model; and to explain why, in some cases, they are to be welcomed, while in others they must be eliminated.

We consider the classic case of $\pi^0 \to \gamma\gamma$, in the context of spontaneously broken chiral flavour symmetry, with massless quarks and pions. The axial isospin current $\hat{j}^\mu_{i,5}(x)$ should then be conserved, but we shall see that this implies that the amplitude for $\pi^0 \to \gamma\gamma$ must vanish, as first pointed out by Veltman (1967) and Sutherland (1967). We begin by writing the matrix element of $\hat{j}^\mu_{3,5}(x)$ between the vacuum and a 2γ state, in momentum space, as

$$\int \mathrm{d}^4x \mathrm{e}^{-\mathrm{i}q\cdot x}\langle\gamma, k_1, \epsilon_1; \gamma, k_2, \epsilon_2|\hat{j}^\mu_{3,5}(x)|0\rangle$$
$$= (2\pi)^4\delta^4(k_1 + k_2 - q)\epsilon^*_{1\nu}(k_1)\epsilon^*_{2\lambda}(k_2)\mathcal{M}^{\mu\nu\lambda}(k_1, k_2). \quad (18.128)$$

FIGURE 18.4
The amplitude considered in (18.128), and the one pion intermediate state contribution to it.

As in figure 18.3, one contribution to $\mathcal{M}^{\mu\nu\lambda}$ has the form (constant/q^2) due to the massless π^0 propagator, shown in figure 18.4. This is, once again, because when chiral symmetry is spontaneously broken, the axial current connects the pion state to the vacuum, as described by the matrix element (18.46). The contribution of the process shown in figure 18.4 to $\mathcal{M}^{\mu\nu\lambda}$ is then

$$\mathrm{i}q^\mu f_\pi \frac{\mathrm{i}}{q^2} \mathrm{i}A\epsilon^{\nu\lambda\alpha\beta}k_{1\alpha}k_{2\beta} \tag{18.129}$$

where the $\pi^0 \to \gamma\gamma$ amplitude is $A\epsilon^{\nu\lambda\alpha\beta}\epsilon^*_{1\nu}(k_1)\epsilon^*_{2\lambda}(k_2)k_{1\alpha}k_{2\beta}$. Note that this automatically incorporates electromagnetic gauge invariance (the amplitude vanishes when the polarization vector of either photon is replaced by its 4-momentum, due to the antisymmetry of the ϵ symbol), and it is symmetrical under interchange of the photon labels. Now consider replacing $\hat{j}^\mu_{3,5}(x)$ in (18.128) by $\partial_\mu \hat{j}^\mu_{3,5}(x)$, which should be zero. A partial integration then shows that this implies that

$$q_\mu \mathcal{M}^{\mu\nu\lambda} = 0 \tag{18.130}$$

which with (18.129) implies that $A = 0$, and hence that $\pi^0 \to \gamma\gamma$ is forbidden. It is important to realize that all other contributions to $\mathcal{M}^{\mu\nu\lambda}$, apart from the π^0 one shown in figure 18.4, will *not* have the $1/q^2$ factor in (18.129), and will therefore give a vanishing contribution to $q_\mu \mathcal{M}^{\mu\nu\lambda}$ at $q^2 = 0$ which is the on-shell point for the (massless) pion.

It is of course true that $m_\pi^2 \neq 0$. But estimates (Adler 1969) of the consequent corrections suggest that the predicted rate of $\pi^0 \to \gamma\gamma$ for real π^0's is far too small. Consequently, there is a problem for the hypothesis of spontaneously broken (approximate) chiral symmetry.

In such a situation it is helpful to consider a detailed calculation performed within a specific model. This is supplied by Itzykson and Zuber (1980), section 11.5.2; in essentials it is the same as the one originally considered by Steinberger (1949) in the first calculation of the $\pi^0 \to \gamma\gamma$ rate, and subsequently by Bell and Jackiw (1969) and by Adler (1969). It employs (scalar) σ and (pseudoscalar) π^0 meson fields, augmented by a fermion of mass m

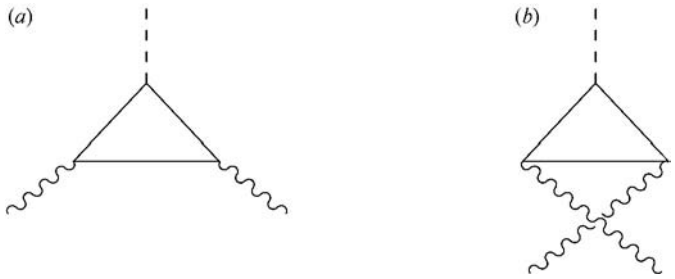


FIGURE 18.5
The two $O(\alpha)$ graphs contributing to $\pi^0 \to \gamma\gamma$ decay.

and charge $+e$, representing the proton. To order α, there are two graphs to consider, shown in figure 18.5(a) and (b). It turns out that the fermion loop integral is actually convergent. In the limit $q^2 \to 0$ the result is

$$A = \frac{e^2}{4\pi^2 f_\pi} \tag{18.131}$$

where A is the $\pi^0 \to \gamma\gamma$ amplitude introduced above. Problem 18.8 evaluates the $\pi^0 \to \gamma\gamma$ rate using (18.131), to give

$$\Gamma(\pi^0 \to 2\gamma) = \frac{\alpha^2}{64\pi^3}\frac{m_\pi^3}{f_\pi^2}. \tag{18.132}$$

(18.132) is in very good agreement with experiment.

In principle, various possibilities now exist. But a careful analysis of the 'triangle' graph contributions to the matrix element $\mathcal{M}^{\mu\nu\lambda}$ of (18.128), shown in figure 18.6, reveals that the fault lies in assuming that a regularization exists such that for these amplitudes the conservation equation $q_\mu \langle\gamma\gamma|\hat{j}^\mu_{3,5}(0)|0\rangle = 0$ can be maintained, at the same time as electromagnetic gauge variance. In fact, no such regularization can be found. When the amplitudes of figure 18.6 are calculated using an (electromagnetic) gauge invariant procedure, one finds a non-zero result for $q_\mu \langle\gamma\gamma|\hat{j}^\mu_{3,5}(0)|0\rangle$ (again the details are given in Itzykson and Zuber (1980)). This implies that $\partial_\mu \hat{j}^\mu_{3,5}(x)$ is not zero after all, the calculation producing the specific value

$$\partial_\mu \hat{j}^\mu_{3,5}(x) = -\frac{e^2}{32\pi^2}\epsilon^{\alpha\nu\beta\lambda}\hat{F}_{\alpha\nu}\hat{F}_{\beta\lambda} \tag{18.133}$$

where the F's are the usual electromagnetic field strengths.

Equation (18.133) means that (18.130) is no longer valid, so that A need no longer vanish: indeed, (18.133) predicts a definite value for A, so we need to see if it is consistent with (18.131). Taking the vacuum $\to 2\gamma$ matrix element

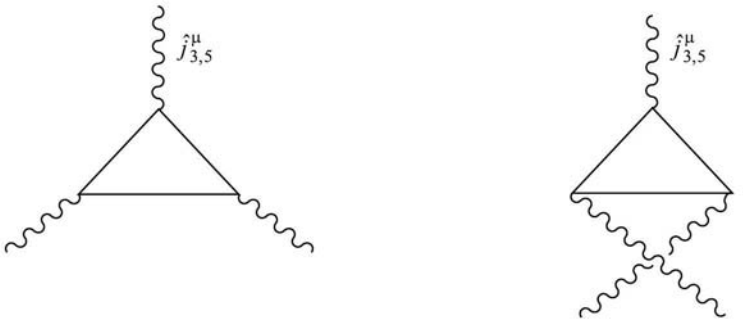


FIGURE 18.6
$O(\alpha)$ contributions to the matrix element in (18.128).

of (18.133) produces (problem 18.9)

$$\mathrm{i}q_\mu \mathcal{M}^{\mu\nu\lambda} = \frac{e^2}{4\pi^2}\epsilon^{\alpha\nu\beta\lambda}k_{1\alpha}k_{2\beta} \tag{18.134}$$

which is indeed consistent with (18.128) and (18.131), after suitably interchanging the labels on the ϵ symbol.

Equation (18.133) is therefore a typical example of 'an anomaly' – the violation, at the quantum level, of a symmetry of the classical Lagrangian. It might be thought that the result (18.133) is only valid to order α (though the $O(\alpha^2)$ correction would presumably be very small). But Adler and Bardeen (1969) showed that such 'triangle' loops give the *only* anomalous contributions to the $\hat{j}^{\mu}_{i,5} - \gamma - \gamma$ vertex, so that (18.133) is true to all orders in α.

The triangles considered above actually used a fermion with integer charge (the proton). We clearly should use quarks, which carry fractional charge. In this case, the previous numerical value for A is multiplied by the factor $\tau_3 Q^2$ for each contributing quark. For the u and d quarks of chiral SU(2) $\times$ SU(2), this gives 1/3. Consequently agreement with experiment is lost unless there exist three replicas of each quark, identical in their electromagnetic and SU(2) $\times$ SU(2) properties. Colour supplies just this degeneracy, and thus the $\pi^0 \to \gamma\gamma$ rate is important evidence for such a degree of freedom, as we noted in chapter 14.

In the foregoing discussion, the axial isospin current was associated with a global symmetry; only the electromagnetic currents (in the case of $\pi^0 \to \gamma\gamma$) were associated with a local (gauged) symmetry, and they remained conserved (anomaly free). If, however, we have an anomaly in a current associated with a local symmetry, we will have a serious problem. The whole rather elaborate construction of a quantum gauge field theory relies on current conservation equations such as (11.21) or (13.130) to eliminate unwanted gauge degrees of freedom, and ensure unitarity of the S-matrix. So anomalies in currents

coupled to gauge fields cannot be tolerated. As we shall see in chapter 20, and is already evident from (18.48), axial currents are indeed present in weak interactions and they are coupled to the $\mathrm{W}^{\pm}, \mathrm{Z}^0$ gauge fields. Hence, if this theory is to be satisfactory at the quantum level, all anomalies must somehow cancel away. That this is possible rests essentially on the observation that the anomaly (18.133) is independent of the mass of the circulating fermion. Thus cancellations are in principle possible between quark and lepton 'triangles' in the weak interaction case. Bouchiat *et al.* (1972) were the first to point out that, for each generation of quarks and leptons, the anomalies will cancel between quarks and leptons if the fractionally charged quarks come in three colours. The condition that anomalies cancel in the gauged currents of the Standard Model is the remarkably simple one (Ryder 1996, p384):

$$N_{\mathrm{c}}(Q_{\mathrm{u}} + Q_{\mathrm{d}}) + Q_{\mathrm{e}} = 0 \tag{18.135}$$

where N_{c} is the number of colours and $Q_{\mathrm{u}}, Q_{\mathrm{d}}$ and Q_{e} are the charges (in units of e) of the 'u', 'd', and 'e' type fields in each generation. Clearly (18.135) is true for each generation of the Standard Model; the condition indicates a remarkable connection, at some deep level, between the facts that quarks come in three colours and have charges which are 1/3 fractions. The Standard Model provides no explanation for this connection. Anomaly cancellation is a powerful constraint on possible theories ('t Hooft 1980, Weinberg 1996 section 22.4).

Problems

18.1 Verify (18.24)–(18.26).

18.2 Verify (18.28)–(18.30).

18.3 Show that $\mathcal{L}_q$ of (18.12) can be written as (18.39).

18.4 Show that the rate for $\pi^+ \to \mu^+\nu_\mu$, calculated from the lowest-order matrix element (18.52), is given by (18.53).

18.5 Verify the transformation equations (18.76) and (18.77).

18.6

(a) Consider a Lagrangian $\hat{\mathcal{L}}(\hat{\phi}_r, \partial_\mu\hat{\phi}_r)$ where the $\hat{\phi}_r$ could be either bosonic or fermionic fields. Let the fields transform by an infinitesimal local (x-dependent) transformation

$$\hat{\phi}_r \to \hat{\phi}_r - \mathrm{i}\epsilon_\alpha(x)T^\alpha_{rs}\hat{\phi}_s \quad (\text{sum on } s).$$

Show that the change in $\hat{\mathcal{L}}$ may be written as

$$\delta\hat{\mathcal{L}} = \hat{j}^{\alpha\mu}(x)\partial_\mu\epsilon_\alpha(x) + \epsilon_\alpha(x)\partial_\mu\hat{j}^{\alpha\mu}(x)$$

where

$$\hat{j}^{\alpha\mu}(x) = -\mathrm{i}T^{\alpha}_{rs}\hat{\phi}_s \frac{\partial\hat{\mathcal{L}}}{\partial(\partial_\mu\hat{\phi}_r)} = \frac{\partial(\delta\hat{\mathcal{L}})}{\partial(\partial_\mu\epsilon_\alpha(x)))}$$

and

$$\partial_\mu\hat{j}^{\alpha\mu}(x) = \frac{\partial(\delta\hat{\mathcal{L}})}{\partial\epsilon_\alpha(x)}. \qquad (1)$$

Deduce that if $\hat{\mathcal{L}}$ is invariant under the global form of this transformation (i.e. constant ϵ_α), then the current defined by (1) is conserved. [This procedure for finding conserved currents for global symmetries is due to Gell-Mann and Levy (1960).]

(b) Apply the method of part (a) to verify the form of the currents (18.96) and (18.97).

18.7 Verify equations (18.120)–(18.124).

18.8 Verify (18.132), and calculate the π^0 lifetime in seconds.

18.9 Verify (18.134).

19

Spontaneously Broken Local Symmetry

In earlier parts of this book we have briefly indicated why we might want to search for a *gauge* theory of the weak interactions. The reasons include: (i) the goal of unification (e.g. with the U(1) gauge theory QED), as mentioned in section 1.3.5; and (ii) certain 'universality' phenomena (to be discussed more fully in chapter 20), which are reminiscent of a similar situation in QED (see comment (ii) in section 2.6, and also section 11.6), and which are particularly characteristic of a non-Abelian gauge theory, as pointed out in section 13.1 after equation (13.44). However, we also know from section 1.3.5 that weak interactions are short-ranged, so that their mediating quanta must be massive. At first sight, this seems to rule out the possibility of a gauge theory of weak interactions, since a simple gauge boson mass violates gauge invariance, as we pointed out for the photon in section 11.3 and for non-Abelian gauge quanta in section 13.3.1, and will review again in the following section. Nevertheless, there is a way of giving gauge field quanta a mass, which is by '*spontaneously breaking*' the gauge (i.e. local) symmetry. This is the topic of the present chapter. The detailed application to the electroweak theory will be made in chapter 21.

19.1 Massive and massless vector particles

Let us begin by noting an elementary (classical) argument for why a gauge field quantum cannot have mass. The electromagnetic potential satisfies the Maxwell equation (cf (2.22))

$$\Box A^\nu - \partial^\nu(\partial_\mu A^\mu) = j^\nu_{\rm em} \tag{19.1}$$

which, as discussed in section 2.3, is invariant under the gauge transformation

$$A^\mu \to A'^\mu = A^\mu - \partial^\mu \chi. \tag{19.2}$$

However, if A^μ were to represent a *massive* field, the relevant wave equation would be

$$(\Box + M^2)A^\nu - \partial^\nu(\partial_\mu A^\mu) = j^\nu_{\rm em}. \tag{19.3}$$

To get this, we have simply replaced the massless 'Klein–Gordon' operator $\Box$ by the corresponding massive one, $\Box + M^2$ (compare sections 3.1 and 5.3).

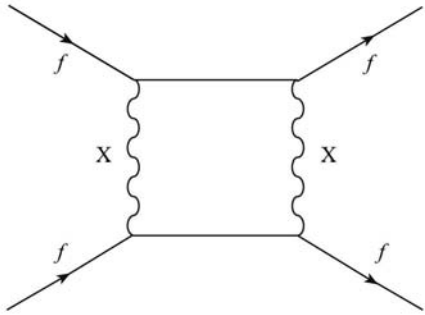


FIGURE 19.1
Fermion–fermion scattering via exchange of two X bosons.

Equation (19.3) is manifestly *not* invariant under (19.2), and it is precisely the mass term M^2A^ν that breaks the gauge invariance. The same conclusion follows in a Lagrangian treatment; to obtain (19.3) as the corresponding Euler-Lagrange equation, one adds a mass term $+\frac{1}{2}M^2A_\mu A^\mu$ to the Lagrangian of (7.66) (see also sections 11.4 and 13.3.1), and this clearly violates invariance under (19.2). Similar reasoning holds for the non-Abelian case too. Perhaps, then, we must settle for a theory involving massive charged vector bosons, $\mathrm{W}^\pm$ for example, without it being a gauge theory.

Such a theory is certainly possible, but it will not be *renormalizable*, as we now discuss. Consider figure 19.1, which shows some kind of fermion–fermion scattering (we need not be more specific), proceeding in fourth order of perturbation theory via the exchange of two massive vector bosons, which we will call X-particles. To calculate this amplitude, we need the propagator for the X-particle, which can be found by following the 'heuristic' route outlined in section 7.3.2 for photons. We consider the momentum-space version of (19.3) for the corresponding X^ν field, but without the current on the right-hand side (so as to describe a free field):

$$[(-k^2 + M^2)g^{\nu\mu} + k^\nu k^\mu]\tilde{X}_\mu(k) = 0, \tag{19.4}$$

which should be compared with (7.90). Apart from the 'iϵ', the propagator should be proportional to the inverse of the quantity in the square brackets in (19.4). Problem 19.1 shows that *unlike* the (massless) photon case, this inverse does exist, and is given by

$$\frac{-g^{\mu\nu} + k^\mu k^\nu/M^2}{k^2 - M^2}. \tag{19.5}$$

A proper field-theoretic derivation would yield this result multiplied by an overall factor 'i' as usual, and would also include the 'iϵ' via $k^2 - M^2 \to k^2 - M^2 + \mathrm{i}\epsilon$. We remark immediately that (19.5) gives nonsense in the limit

$M \to 0$, thus indicating already that a massless vector particle seems to be a very different kind of thing from a massive one (we can't just take the massless limit of the latter).

Now consider the loop integral in figure 19.1. At each vertex we will have a coupling constant g, associated with an interaction Lagrangian having the general form $g\bar{\hat{\psi}}\gamma_\mu\hat{\psi}\hat{X}^\mu$ (a $\gamma_\mu\gamma_5$ coupling could also be present but will not affect the argument). Just as in QED, this 'g' is dimensionless but, as we warned the reader in section 11.8, this may not guarantee renormalizability, and indeed this is a case where it does not. To get an idea of why this might be so, consider the leading divergent behaviour of figure 19.1. This will be associated with the $k^\mu k^\nu$ terms in the numerator of (19.5), so that the leading divergence is effectively

$$\sim \int \mathrm{d}^4k \left(\frac{k^\mu k^\nu}{k^2}\right)\left(\frac{k^\rho k^\sigma}{k^2}\right)\frac{1}{\not{k}}\frac{1}{\not{k}} \tag{19.6}$$

for high k-values (we are not troubling to get all the indices right, we are omitting the spinors altogether, and we are looking only at the large k part of the propagators). Now the first two bracketed terms in (19.6) behave like a constant at large k, so that the divergence is effectively

$$\sim \int \mathrm{d}^4k \frac{1}{\not{k}}\frac{1}{\not{k}} \tag{19.7}$$

which is quadratically divergent, and indeed exactly what we would get in a 'four-fermion' theory – see (11.98) for example. This strongly suggests that the theory is non-renormalizable.

Where have these dangerous powers of k in the numerator of (19.6) come from? The answer is simple and important. They come from the *longitudinal* polarization state of the massive X-particle, as we shall now explain. The free-particle wave equation is

$$(\Box + M^2)X^\nu - \partial^\nu(\partial_\mu X^\mu) = 0 \tag{19.8}$$

and plane wave solutions have the form

$$X^\nu = \epsilon^\nu \mathrm{e}^{-\mathrm{i}k\cdot x}. \tag{19.9}$$

Hence the polarization vectors ϵ^ν satisfy the condition

$$(-k^2 + M^2)\epsilon^\nu + k^\nu k_\mu \epsilon^\mu = 0. \tag{19.10}$$

Taking the 'dot' product of (19.10) with k_ν leads to

$$M^2 k\cdot\epsilon = 0, \tag{19.11}$$

which implies (for $M^2 \neq 0$!)

$$k\cdot\epsilon = 0. \tag{19.12}$$

Equation (19.12) is a covariant condition, which has the effect of ensuring that there are just three independent polarization vectors, as we expect for a spin-1 particle. Let us take $k^\mu = (k^0, 0, 0, |\boldsymbol{k}|)$; then the x- and y-directions are 'transverse' while the z-direction is 'longitudinal'. Now, in the rest frame of the X, such that $k_{\rm rest} = (M, 0, 0, 0)$, (19.12) reduces to $\epsilon^0 = 0$, and we may choose three independent ϵ's as

$$\epsilon^\mu(k_{\rm rest}, \lambda) = (0, \boldsymbol{\epsilon}(\lambda)) \tag{19.13}$$

with

$$\boldsymbol{\epsilon}(\lambda = \pm 1) = \mp 2^{-1/2}(1, \pm \mathrm{i}, 0) \tag{19.14}$$

$$\boldsymbol{\epsilon}(\lambda = 0) = (0, 0, 1). \tag{19.15}$$

The $\boldsymbol{\epsilon}$'s are 'orthonormalized' so that (cf (7.86))

$$\boldsymbol{\epsilon}(\lambda)^* \cdot \boldsymbol{\epsilon}(\lambda') = \delta_{\lambda\lambda'}. \tag{19.16}$$

These states have definite spin projection ($\lambda = \pm 1, 0$) along the z-axis. For the result in a general frame, we can Lorentz transform $\epsilon^\mu(k_{\rm rest}, \lambda)$ as required. For example, in a frame such that $k^\mu = (k^0, 0, 0, |\boldsymbol{k}|)$, we find

$$\epsilon^\mu(k, \lambda = \pm 1) = \epsilon^\mu(k_{\rm rest}, \lambda = \pm 1) \tag{19.17}$$

as before, but the longitudinal polarization vector becomes (problem 19.2)

$$\epsilon^\mu(k, \lambda = 0) = M^{-1}(|\boldsymbol{k}|, 0, 0, k^0). \tag{19.18}$$

Note that $k \cdot \epsilon^\mu(k, \lambda = 0) = 0$ as required.

From (19.17) and (19.18) it is straightforward to verify the result (problem 19.3)

$$\sum_{\lambda = 0, \pm 1} \epsilon^\mu(k, \lambda)\epsilon^{\nu *}(k, \lambda) = -g^{\mu\nu} + k^\mu k^\nu / M^2. \tag{19.19}$$

Consider now the propagator for a spin-1/2 particle, given in (7.63):

$$\frac{\mathrm{i}(\not{k} + m)}{k^2 - m^2 + \mathrm{i}\epsilon}. \tag{19.20}$$

Equation (7.64) shows that the factor in the numerator of (19.20) arises from the spin sum

$$\sum_s u_\alpha(k, s)\bar{u}_\beta(k, s) = (\not{k} + m)_{\alpha\beta}. \tag{19.21}$$

In just the same way, the massive spin-1 propagator is given by

$$\frac{\mathrm{i}[-g^{\mu\nu} + k^\mu k^\nu / M^2]}{k^2 - M^2 + \mathrm{i}\epsilon}, \tag{19.22}$$

the numerator in (19.22) arising from the spin sum (19.19). Thus the dangerous factor $k^\mu k^\nu/M^2$ can be traced to the spin sum (19.19): in particular, at large values of k the longitudinal state $\epsilon^\mu(k, \lambda = 0)$ is proportional to k^μ, and this is the origin of the numerator factors $k^\mu k^\nu/M^2$ in (19.22).

We shall not give further details here (see also section 22.1.2), but merely state that theories with massive charged vector bosons are indeed non-renormalizable. Does this matter? In section 11.8 we explained why it is thought that the relevant theories at presently accessible energy scales should be renormalizable theories. And, apart from anything else, they are much more predictive. Is there, then, any way of getting rid of the offending '$k^\mu k^\nu$' terms in the X-propagator, so as (perhaps) to render the theory renormalizable? Consider the photon propagator of chapter 7 repeated here:

$$\frac{\mathrm{i}[-g^{\mu\nu} + (1-\xi)k^\mu k^\nu/k^2]}{k^2 + \mathrm{i}\epsilon}. \tag{19.23}$$

This contains somewhat similar factors of $k^\mu k^\nu$ (admittedly divided by k^2 rather than M^2), but they are gauge-dependent, and can in fact be 'gauged away' entirely, by choice of the gauge parameter ξ (namely by taking $\xi = 1$). But, as we have seen, such 'gauging' – essentially the freedom to make gauge transformations – seems to be possible only in a massless vector theory.

A closely related point is that, as section 7.3.1 showed, free photons exist in only two polarization states (electromagnetic waves are purely transverse), instead of the three we might have expected for a vector (spin-1) particle – and as do indeed exist for massive vector particles. This gives another way of seeing in what way a massless vector particle is really very different from a massive one: the former has only two (spin) degrees of freedom, while the latter has three, and it is not at all clear how to 'lose' the offending longitudinal state smoothly (certainly not, as we have seen, by letting $M \to 0$ in (19.5)).

These considerations therefore suggest the following line of thought: is it possible somehow to create a theory involving massive vector bosons, in such a way that the dangerous $k^\mu k^\nu$ term can be 'gauged away', making the theory renormalizable? The answer is yes, via the idea of *spontaneous breaking* of gauge symmetry. This is the natural generalization of the spontaneous global symmetry breaking considered in chapter 17. By way of advance notice, the crucial formula is (19.74) for the propagator in such a theory, which is to be compared with (19.22).

The first serious challenge to the then widely held view that electromagnetic gauge invariance requires the photon to be massless was made by Schwinger (1962), as we pointed out in section 11.4. Soon afterwards, Anderson (1963) argued that several situations in solid state physics could be interpreted in terms of an effectively massive electromagnetic field. He outlined a general framework for treating the phenomenon of the acquisition of mass by a gauge boson, and discussed its possible relevance to contemporary attempts (Sakurai 1960) to interpret the recently discovered vector mesons ($\rho, \omega, \phi \ldots$) as the gauge quanta associated with a local extension of hadronic flavour sym-

metry. From his discussion, it is clear that Anderson had his doubts about the hadronic application, precisely because, as he remarked, gauge bosons can only acquire a mass if the symmetry is spontaneously broken. This has the consequence, as we saw in chapter 17, that the multiplet structure ordinarily associated with a non-Abelian symmetry would be lost. But we know that flavour symmetry, even if admittedly not exact, certainly leads to identifiable multiplets, which are at least approximately degenerate in mass. It was Weinberg (1967) and Salam (1968) who made the correct application of these ideas, to the generation of mass for the gauge quanta associated with the weak force. There is, however, nothing specifically relativistic about the basic mechanism involved, nor need we start with the non-Abelian case. In fact, the physics is well illustrated by the non-relativistic Abelian (i.e. electromagnetic) case – which is nothing but the physics of superconductivity. Our presentation is influenced by that of Anderson (1963).

19.2 The generation of 'photon mass' in a superconductor: Ginzburg–Landau theory and the Meissner effect

In chapter 17, section 17.7, we gave a brief introduction to some aspects of the BCS theory of superconductivity. We were concerned mainly with the nature of the BCS ground state, and with the non-perturbative origin of the energy gap for elementary excitations. In particular, as noted after (17.128), we omitted completely all electromagnetic couplings of the electrons in the 'microscopic' Hamiltonian. It is certainly possible to complete the BCS theory in this way, so as to include within the same formalism a treatment of electromagnetic effects (e.g. the Meissner effect) in a superconductor. We refer interested readers to the book by Schrieffer (1964), chapter 8. Instead, we shall follow a less 'microscopic' and somewhat more 'phenomenological' approach, which has a long history in theoretical studies of superconductivity, and is in some ways actually closer (at least formally) to our eventual application in particle physics.

In section 17.3.1 we introduced the concept of an 'order parameter', a quantity which was a measure of the 'degree of ordering' of a system below some transition temperature. In the case of superconductivity, the order parameter (in this sense) is taken to be a complex scalar field ψ, as originally proposed by Ginzburg and Landau (1950), well before the appearance of BCS theory. Subsequently, Gorkov (1959) and others showed how the Ginzburg–Landau description could be derived from BCS theory, in certain domains of temperature and magnetic field. This work all relates to static phenomena. More recently, an analogous 'effective theory' for time-dependent phenomena

(at zero temperature) has been derived from a BCS-type model (Aitchison *et al.* 1995). For the moment, we shall follow a more qualitative approach.

The Ginzburg–Landau field ψ is commonly referred to as the 'macroscopic wave function'. This terminology originates from the recognition that in the BCS ground state a macroscopic number of Cooper pairs have 'condensed' into the state of lowest energy, a situation similar to that in the Bogoliubov superfluid. Further, this state is highly *coherent*, all pairs having the same total momentum (namely zero, in the case of (17.140)). These considerations suggest that a successful phenomenology can be built by invoking the idea of a macroscopic wavefunction ψ, describing the condensate. Note that ψ is a 'bosonic' quantity, referring essentially to *paired* electrons. Perhaps the single most important property of ψ is that it is assumed to be normalized to the total density of Cooper pairs $n_{\rm c}$ via the relation

$$|\psi|^2 = n_{\rm c} = n_{\rm s}/2 \tag{19.24}$$

where $n_{\rm s}$ is the density of superconducting electrons. The quantities $n_{\rm c}$ and $n_{\rm s}$ will depend on temperature T, tending to zero as T approaches the superconducting transition temperature $T_{\rm c}$ from below. The precise connection between ψ and the microscopic theory is indirect; in particular, ψ has no knowledge of the coordinates of individual electron pairs. Nevertheless, as an 'empirical' order parameter, it may be thought of as in some way related to the ground state 'pair' expectation value introduced in (17.121): in particular, the charge associated with ψ is taken to be $-2e$, and the mass is $2m_{\rm e}$.

The Ginzburg–Landau description proceeds by considering the quantum-mechanical electromagnetic current associated with ψ, in the presence of a static external electromagnetic field described by a vector potential $\boldsymbol{A}$. This current was considered in section 2.4, and is given by the gauge-invariant form of (A.7), namely

$$\boldsymbol{j}_{\rm em} = \frac{-2e}{4m_{\rm e}\mathrm{i}}[\psi^*(\boldsymbol{\nabla} + 2\mathrm{i}e\boldsymbol{A})\psi - \{(\boldsymbol{\nabla} + 2\mathrm{i}e\boldsymbol{A})\psi\}^*\psi]. \tag{19.25}$$

Note that we have supplied an overall factor of $-2e$ to turn the Schrödinger 'number density' current into the appropriate electromagnetic current. Assuming now that, consistently with (19.24), ψ is varying primarily through its *phase* degree of freedom ϕ, rather than its modulus $|\psi|$, we can rewrite (19.25) as

$$\boldsymbol{j}_{\rm em} = -\frac{2e^2}{m_{\rm e}}\left(\boldsymbol{A} + \frac{1}{2e}\boldsymbol{\nabla}\phi\right)|\psi|^2 \tag{19.26}$$

where $\psi = \mathrm{e}^{\mathrm{i}\phi}|\psi|$. We easily verify that (19.26) is invariant under the gauge transformation (2.41), which can be written in this case as

$$\boldsymbol{A} \rightarrow \boldsymbol{A} + \boldsymbol{\nabla}\chi \tag{19.27}$$

$$\phi \rightarrow \phi - 2e\chi. \tag{19.28}$$

We now replace $|\psi|^2$ in (19.26) by $n_{\rm s}/2$ in accordance with (19.24), and take the curl of the resulting equation to obtain

$$\boldsymbol{\nabla} \times \boldsymbol{j}_{\rm em} = -\left(\frac{e^2 n_{\rm s}}{m_{\rm e}}\right)\boldsymbol{B}. \tag{19.29}$$

Equation (19.29) is known as the London equation (London 1950), and is one of the fundamental phenomenological relations in superconductivity.

The significance of (19.29) emerges when we combine it with the (static) Maxwell equation

$$\boldsymbol{\nabla} \times \boldsymbol{B} = \boldsymbol{j}_{\rm em}. \tag{19.30}$$

Taking the curl of (19.30), and using $\boldsymbol{\nabla} \times (\boldsymbol{\nabla} \times \boldsymbol{B}) = \boldsymbol{\nabla}(\boldsymbol{\nabla} \cdot \boldsymbol{B}) - \boldsymbol{\nabla}^2 \boldsymbol{B}$ and $\boldsymbol{\nabla} \cdot \boldsymbol{B} = 0$, we find

$$\boldsymbol{\nabla}^2 \boldsymbol{B} = \left(\frac{e^2 n_{\rm s}}{m_{\rm e}}\right)\boldsymbol{B}. \tag{19.31}$$

The variation of magnetic field described by (19.31) is a very characteristic one encountered in a number of contexts in condensed matter physics. First, we note that the quantity $(e^2 n_{\rm s}/m_{\rm e})$ must – in our units – have the dimensions of (length)$^{-2}$, by comparison with the left-hand side of (19.31). Let us write

$$\left(\frac{e^2 n_{\rm s}}{m_{\rm e}}\right) = \frac{1}{\lambda^2}. \tag{19.32}$$

Next, consider for simplicity one-dimensional variation

$$\frac{{\rm d}^2 \boldsymbol{B}}{{\rm d}x^2} = \frac{1}{\lambda^2}\boldsymbol{B} \tag{19.33}$$

in the half-plane $x \geq 0$, say. Then the solutions of (19.33) have the form

$$\boldsymbol{B}(x) = \boldsymbol{B}_0 \exp -(x/\lambda); \tag{19.34}$$

the exponentially growing solution is rejected as unphysical. The field therefore penetrates only a distance of order λ into the region $x \geq 0$. The range parameter λ is called the *screening length*. This expresses the fact that, in a medium such that (19.29) holds, the magnetic field will be 'screened out' from penetrating further into the medium.

The physical origin of the screening is provided by Lenz's law: when a magnetic field is applied to a system of charged particles, induced EMFs are set up which accelerate the particles, and the magnetic effect of the resulting currents tends to cancel (or screen) the applied field. On the atomic scale this is the cause of atomic diamagnetism. Here the effect is occurring on a macroscopic scale (as mediated by the 'macroscopic wavefunction' ψ), and leads to the Meissner effect – the exclusion of flux from the interior of a superconductor. In this case, screening currents are set up within the superconductor, over distances of order λ from the exterior boundary of the material. These

exactly cancel – perfectly screen – the applied flux density in the interior. With $n_s \sim 4 \times 10^{28}$ m^{-3} (roughly one conduction electron per atom) we find

$$\lambda = \left(\frac{m_e}{n_s e^2}\right)^{1/2} \approx 10^{-8} \text{ m}, \tag{19.35}$$

which is the correct order of magnitude for the thickness of the surface layer within which screening currents flow, and over which the applied field falls to zero. As $T \to T_c$, $n_s \to 0$ and λ becomes arbitrarily large, so that flux is no longer screened.

It is quite simple to interpret equation (19.31) in terms of an 'effective non-zero photon mass'. Consider the equation (19.8) for a free massive vector field. Taking the divergence via ∂_ν leads to

$$M^2 \partial_\nu X^\nu = 0 \tag{19.36}$$

(cf (19.11)), and so (19.8) can be written as

$$(\Box + M^2) X^\nu = 0, \tag{19.37}$$

which simply expresses the fact that each component of X^ν has mass M. Now consider the static version of (19.37), in the rest frame of the X-particle in which (see equation (19.13)) the $\nu = 0$ component vanishes. Equation (19.37) reduces to

$$\boldsymbol{\nabla}^2 \boldsymbol{X} = M^2 \boldsymbol{X} \tag{19.38}$$

which is exactly the same in form as (19.31) (if $\boldsymbol{X}$ were the electromagnetic field $\boldsymbol{A}$, we could take the curl of (19.38) to obtain (19.31) via $\boldsymbol{B} = \boldsymbol{\nabla} \times \boldsymbol{A}$). The connection is made precise by making the association

$$M^2 = \left(\frac{e^2 n_s}{m_e}\right) = \frac{1}{\lambda^2}. \tag{19.39}$$

Equation (19.39) shows very directly another way of understanding the 'screening length $\leftrightarrow$ photon mass' connection: in our units $\hbar = c = 1$, a mass has the dimension of an inverse length, and so we naturally expect to be able to interpret λ^{-1} as an equivalent mass (for the photon, in this case).

The above treatment conveys much of the essential physics behind the phenomenon of 'photon mass generation' in a superconductor. In particular, it suggests rather strongly that a *second* field, in addition to the electromagnetic one, is an essential element in the story (here, it is the ψ field). This provides a partial answer to the puzzle about the discontinuous change in the number of spin degrees of freedom in going from a massless to a massive gauge field: actually, some other field has to be supplied. Nevertheless, many questions remain unanswered so far. For example, how is all the foregoing related to what we learned in chapter 17 about spontaneous symmetry breaking? Where is the Goldstone mode? Is it really all gauge invariant? And

what about Lorentz invariance? Can we provide a Lagrangian description of the phenomenon? The answers to these questions are mostly contained in the model to which we now turn, which is due to Higgs (1964) and is essentially the *local* version of the U(1) Goldstone model of section 17.5.

19.3 Spontaneously broken local U(1) symmetry: the Abelian Higgs model

This model is just $\hat{\mathcal{L}}_{\rm G}$ of (17.69) and (17.77), extended so as to be locally, rather than merely globally, U(1) invariant. Due originally to Higgs (1964), it provides a deservedly famous and beautifully simple model for investigating what happens when a *gauge* symmetry is spontaneously broken.

To make (17.69) locally U(1) invariant, we need only replace the ∂'s by $\hat{D}$'s according to the rule (7.123), and add the Maxwell piece. This produces

$$\hat{\mathcal{L}}_{\rm H} = [(\partial^\mu + {\rm i}q\hat{A}^\mu)\hat{\phi}]^\dagger[(\partial_\mu + {\rm i}q\hat{A}_\mu)\hat{\phi}] - \frac{1}{4}\hat{F}_{\mu\nu}\hat{F}^{\mu\nu} - \frac{1}{4}\lambda(\hat{\phi}^\dagger\hat{\phi})^2 + \mu^2(\hat{\phi}^\dagger\hat{\phi}). \quad (19.40)$$

(19.40) is invariant under the local version of (17.72), namely

$$\hat{\phi}(x) \to \hat{\phi}'(x) = {\rm e}^{-{\rm i}\hat{\alpha}(x)}\hat{\phi}(x) \quad (19.41)$$

when accompanied by the gauge transformation on the potentials

$$\hat{A}^\mu(x) \to \hat{A}'^\mu(x) = \hat{A}^\mu(x) + \frac{1}{q}\partial^\mu\hat{\alpha}(x). \quad (19.42)$$

Before proceeding any further, we note at once that this model contains four field degrees of freedom – two in the complex scalar Higgs field $\hat{\phi}$, and two in the massless gauge field $\hat{A}^\mu$.

We learned in section 17.5 that the form of the potential terms in (19.40) (specifically the μ^2 one) does not lend itself to a natural particle interpretation, which only appears after making a 'shift to the classical minimum', as in (17.84). But there is a remarkable difference between the global and local cases. In the present (local) case, the phase of $\hat{\phi}$ is completely arbitrary, since any change in $\hat{\alpha}$ of (19.41) can be compensated by an appropriate transformation (19.42) on $\hat{A}^\mu$, leaving $\hat{\mathcal{L}}_{\rm H}$ the same as before. Thus the field $\hat{\theta}$ in (17.84) can be 'gauged away' altogether, if we choose! But $\hat{\theta}$ was precisely the Goldstone field, in the global case. This must mean that there is somehow no longer any physical manifestation of the massless mode. This is the first unexpected result in the local case. We may also be reminded of our desire to 'gauge away' the longitudinal polarization states for a 'massive gauge' boson: we shall return to this later.

However, a degree of freedom (the Goldstone mode) cannot simply disappear. Somehow the system must keep track of the fact that we started with four degrees of freedom. To see what is going on, let us study the field equation for $\hat{A}^\nu$, namely

$$\Box\hat{A}^\nu - \partial^\nu(\partial_\mu\hat{A}^\mu) = \hat{j}^\nu_{\rm em}, \tag{19.43}$$

where $\hat{j}^\nu_{\rm em}$ is the electromagnetic current contained in (19.40). This current can be obtained just as in (7.141), and is given by

$$\hat{j}^\nu_{\rm em} = {\rm i}q(\hat{\phi}^\dagger\partial^\nu\hat{\phi} - (\partial^\nu\hat{\phi}^\dagger)\hat{\phi}) - 2q^2\hat{A}^\nu\hat{\phi}^\dagger\hat{\phi}. \tag{19.44}$$

We now insert the field parametrization (cf (17.84))

$$\hat{\phi}(x) = \frac{1}{\sqrt{2}}(v + \hat{h}(x))\exp(-{\rm i}\hat{\theta}(x)/v) \tag{19.45}$$

into (19.44) where $v/\sqrt{2} = 2^{1/2}|\mu|/\lambda^{\frac{1}{2}}$ is the position of the minimum of the classical potential as a function of $|\phi|$, as in (17.81). We obtain (problem 19.4)

$$\hat{j}^\nu_{\rm em} = -v^2q^2\left(\hat{A}^\nu - \frac{\partial^\nu\hat{\theta}}{vq}\right) + \text{terms quadratic and cubic in the fields.} \tag{19.46}$$

The linear part of the right-hand side of (19.46) is directly analogous to the non-relativistic current (19.26), interpreting $\hat{\theta}$ as essentially playing the role of ϕ, and $|\psi|^2$ the role of v^2. Retaining just the linear terms in (19.46) (the others would appear on the right-hand side of equation (19.47) following, where they would represent interactions), and placing this $\hat{j}^\nu_{\rm em}$ in (19.43), we obtain

$$\Box\hat{A}^\nu - \partial^\nu\partial_\mu\hat{A}^\mu = -v^2q^2\left(\hat{A}^\nu - \frac{\partial^\nu\hat{\theta}}{vq}\right). \tag{19.47}$$

Now a gauge transformation on $\hat{A}^\nu$ has the form shown in (19.42), for arbitrary $\hat{\alpha}$. So we can certainly regard the whole expression $(\hat{A}^\nu - \partial^\nu\hat{\theta}/vq)$ as a perfectly acceptable gauge field. Let us define

$$\hat{A}'^\nu = \hat{A}^\nu - \frac{\partial^\nu\hat{\theta}}{vq}. \tag{19.48}$$

Then, since we know that the left-hand side of (19.47) is invariant under (19.42), the resulting equation for $\hat{A}'^\nu$ is

$$\Box\hat{A}'^\nu - \partial^\nu\partial_\mu\hat{A}'^\mu = -v^2q^2\hat{A}'^\nu, \tag{19.49}$$

or

$$(\Box + v^2q^2)\hat{A}'^\nu - \partial^\nu\partial_\mu\hat{A}'^\mu = 0. \tag{19.50}$$

But (19.50) is nothing but the equation (19.8) for a free massive vector field, with mass $M = vq$! This fundamental observation was first made, in the relativistic context, by Englert and Brout (1964), Higgs (1964), and Guralnik *et al.* (1964); for a full account, see Higgs (1966).

The foregoing analysis shows us two things. First, the current (19.46) is indeed a relativistic analogue of (19.26), in that it provides a 'screening' (mass generation) effect on the gauge field. Second, equation (19.48) shows how the *phase* degree of freedom of the Higgs field $\hat{\phi}$ has been incorporated into a new gauge field $\hat{A}'^{\nu}$, which is massive, and therefore has '*three*' spin degrees of freedom. In fact, we can go further. If we imagine plane wave solutions for $\hat{A}'^{\nu}$, $\hat{A}^{\nu}$ and $\hat{\theta}$, we see that the $\partial^{\nu}\hat{\theta}/vq$ part of (19.48) will contribute something proportional to k^{ν}/M to the polarization vector of A'^{ν} (recall $M = vq$). But this is exactly the (large k) behaviour of the longitudinal polarization vector of a massive vector particle. We may therefore say that the massless gauge field $\hat{A}^{\nu}$ has 'swallowed' the Goldstone field $\hat{\theta}$ via (19.48) to make the massive vector field $\hat{A}'^{\nu}$. The Goldstone field disappears as a massless degree of freedom, and reappears, via its gradient, as the longitudinal part of the massive vector field. In this way the four degrees of freedom are all now safely accounted for: three are in the massive vector field $\hat{A}'^{\nu}$, and one is in the real scalar field $\hat{h}$ (to which we shall return shortly).

In this (relativistic) case, we know from Lorentz covariance that all the components (transverse and longitudinal) of the vector field must have the same mass, and this has of course emerged automatically from our covariant treatment. But the transverse and longitudinal degrees of freedom respond differently in the non-relativistic (superconductor) case. There, the longitudinal part of $\boldsymbol{A}$ couples strongly to longitudinal excitations of the electrons: primarily, as Bardeen (1957) first recognized, to the collective density fluctuation mode of the electron system – that is, to plasma oscillations. This is a high frequency mode, and is essentially the one discussed in section 17.3.2, after equation (17.46). When this aspect of the dynamics of the electrons is included, a fully gauge invariant description of the electromagnetic properties of superconductors, within the BCS theory, is obtained (Schreiffer 1964, chapter 8).

We return to equations (19.48)–(19.50). Taking the divergence of (19.50) leads, as we have seen, to the condition

$$\partial_{\mu}\hat{A}'^{\mu} = 0 \tag{19.51}$$

on $\hat{A}'^{\mu}$. It follows that in order to interpret the relation (19.48) as a gauge transformation on $\hat{A}^{\nu}$ we must, to be consistent with (19.51), regard $\hat{A}^{\mu}$ as being in a gauge specified by

$$\partial_{\mu}\hat{A}^{\mu} = \frac{1}{vq}\Box\hat{\theta} = \frac{1}{M}\Box\hat{\theta}. \tag{19.52}$$

In going from the situation described by $\hat{A}^{\mu}$ and $\hat{\theta}$ to one described by $\hat{A}'^{\mu}$

alone via (19.48), we have evidently chosen a gauge function (cf (19.42))

$$\hat{\alpha}(x) = -\hat{\theta}(x)/v. \tag{19.53}$$

Recalling then the form of the associated local phase change on $\hat{\phi}(x)$

$$\hat{\phi}(x) \to \hat{\phi}'(x) = \mathrm{e}^{-\mathrm{i}\hat{\alpha}(x)}\hat{\phi}(x) \tag{19.54}$$

we see that the phase of $\hat{\phi}$ in (19.45) has been reduced to zero, in this choice of gauge. Thus it is indeed possible to 'gauge $\hat{\theta}$ away' in (19.45), but then the vector field we must use is $\hat{A}'^{\mu}$, satisfying the massive equation (19.50) (ignoring other interactions). In superconductivity, the choice of gauge which takes the macroscopic wavefunction to be real (i.e. $\phi = 0$ in (19.26)) is called the 'London gauge'. In the next section we shall discuss a subtlety in the argument which applies in the case of real superconductors, and which leads to the phenomenon of flux quantization.

The fact that this 'Higgs mechanism' leads to a massive vector field can be seen very economically by working in the particular gauge for which $\hat{\phi}$ is real, and inserting the parametrization (cf (19.45))

$$\hat{\phi} = \frac{1}{\sqrt{2}}(v + \hat{h}) \tag{19.55}$$

into the Lagrangian $\hat{\mathcal{L}}_{\mathrm{H}}$. Retaining only the terms quadratic in the fields one finds (problem 19.5)

$$\begin{aligned}\hat{\mathcal{L}}_{\mathrm{H}}^{\mathrm{quad}} &= -\frac{1}{4}(\partial_\mu \hat{A}_\nu - \partial_\nu \hat{A}_\mu)(\partial^\mu \hat{A}^\nu - \partial^\nu \hat{A}^\mu) + \frac{1}{2}q^2 v^2 \hat{A}_\mu \hat{A}^\mu \\ &\quad + \frac{1}{2}\partial_\mu \hat{h}\partial^\mu \hat{h} - \mu^2 \hat{h}^2.\end{aligned} \tag{19.56}$$

The first line of (19.56) is exactly the Lagrangian for a spin-1 field of mass vq – i.e. the Maxwell part with the addition of a mass term (note that the sign of the mass term is correct for the spatial (physical) degrees of freedom); the second line is the Lagrangian of a scalar particle of mass $\sqrt{2}\mu$. The latter is the mass of excitations of the Higgs field $\hat{h}$ away from its vacuum value (compare the global U(1) case discussed in section 17.5). The necessity for the existence of one or more massive *scalar* particles ('Higgs bosons'), when a gauge symmetry is spontaneously broken in this way, was pointed out by Higgs (1964).

We may now ask: what happens if we start with a certain phase $\hat{\theta}$ for $\hat{\phi}$ but do *not* make use of the gauge freedom in $\hat{A}^\nu$ to reduce $\hat{\theta}$ to zero? We shall see in section 19.5 that the equation of motion, and hence the propagator, for the vector particle *depends on the choice of gauge*; furthermore, Feynman graphs involving quanta corresponding to the degree of freedom associated with the phase field $\hat{\theta}$ will have to be included for a consistent theory, even though this must be an unphysical degree of freedom, as follows from the fact

that a gauge can be chosen in which this field vanishes. That the propagator is gauge dependent should, on reflection, come as a relief. After all, if the massive vector boson generated in this way were *simply* described by the wave equation (19.50), all the troubles with massive vector particles outlined in section 19.1 would be completely unresolved. As we shall see, a different choice of gauge from that which renders $\hat{\phi}$ real has precisely the effect of ameliorating the bad high-energy behaviour associated with (19.50). This is ultimately the reason for the following wonderful fact: *massive vector theories, in which the vector particles acquire mass through the spontaneous symmetry breaking mechanism, are renormalizable* ('t Hooft 1971b).

However, before discussing other gauges than the one in which $\hat{\phi}$ is given by (19.55), we first explore another interesting aspect of superconductivity.

19.4 Flux quantization in a superconductor

Though a slight diversion, it is convenient to include a discussion of flux quantization at this point, while we have a number of relevant results assembled. Apart from its intrinsic interest, the phenomenon may also provide a useful physical model for the 'confining' property of QCD, as already discussed in sections 1.3.6 and 16.5.3.

Our discussion of superconductivity so far has dealt, in fact, with only one class of superconductors, called type I; these remain superconducting throughout the bulk of the material (exhibiting a complete Meissner effect), when an external magnetic field of less than a certain critical value is applied. There is a quite separate class – type II superconductors – which allow partial entry of the external field, in the form of thin filaments of flux. Within each filament the field is high, and the material is not superconducting. Outside the core of the filaments, the material is superconducting and the field dies off over the characteristic penetration length λ. Around each filament of magnetic flux there circulates a vortex of screening current; the filaments are often called vortex lines. It is as if numerous thin cylinders, each enclosing flux, had been drilled in a block of type I material, thereby producing a non-simply connected geometry.

In real superconductors, screening currents are associated with the macroscopic pair wavefunction (field) ψ. For type II behaviour to be possible, $|\psi|$ must vanish at the centre of a flux filament, and rise to the constant value appropriate to the superconducting state over a distance $\xi < \lambda$, where ξ is the 'coherence length' of section 17.7. According to the Ginzburg–Landau (GL) theory, a more precise criterion is that type II behaviour holds if $\xi < 2^{1/2}\lambda$; both ξ and λ are, of course, temperature-dependent. The behaviour of $|\psi|$ and B in the vicinity of a flux filament is shown in figure 19.2. Thus, whereas for simple type I superconductivity, $|\psi|$ is simply set equal to a constant, in the

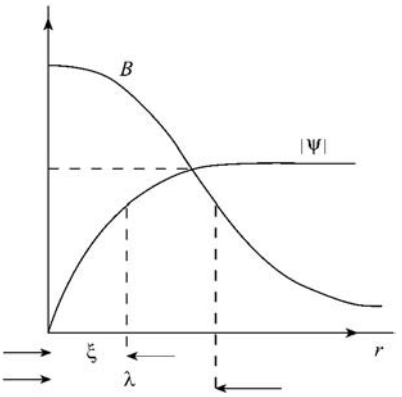


FIGURE 19.2
Magnetic field B and modulus of the macroscopic (pair) wavefunction $|\psi|$ in the neighbourhood of a flux filament.

type II case $|\psi|$ has the variation shown in this figure. Solutions of the coupled GL equations for $\boldsymbol{A}$ and ψ can be obtained which exhibit this behaviour.

An important result is that the flux through a vortex line is quantized. To see this, we write

$$\psi = \mathrm{e}^{\mathrm{i}\phi}|\psi| \tag{19.57}$$

as before. The expression for the electromagnetic current is

$$\boldsymbol{j}_{\mathrm{em}} = -\frac{q^2}{m}\left(\boldsymbol{A} - \frac{\boldsymbol{\nabla}\phi}{q}\right)|\psi|^2 \tag{19.58}$$

as in (19.26), but in (19.58) we are leaving the charge parameter q undetermined for the moment; the mass parameter m will be unimportant. Rearranging, we have

$$\boldsymbol{A} = -\frac{m}{q^2|\psi|^2}\boldsymbol{j}_{\mathrm{em}} + \frac{\boldsymbol{\nabla}\phi}{q}. \tag{19.59}$$

Let us integrate equation (19.59) around any closed loop $\mathcal{C}$ in the type II superconductor, which encloses a flux (or vortex) line. Far enough away from the vortex, the screening currents $\boldsymbol{j}_{\mathrm{em}}$ will have dropped to zero, and hence

$$\oint_{\mathcal{C}} \boldsymbol{A}\cdot \mathrm{d}\boldsymbol{s} = \frac{1}{q}\oint_{\mathcal{C}} \boldsymbol{\nabla}\phi \cdot \mathrm{d}\boldsymbol{s} = \frac{1}{q}[\phi]_{\mathcal{C}} \tag{19.60}$$

where $[\phi]_{\mathcal{C}}$ is the change in phase around $\mathcal{C}$. If the wavefunction ψ is single-valued, the change in phase $[\phi]_{\mathcal{C}}$ for any closed path can only be zero or an integer multiple of 2π. Transforming the left-hand side of (19.60) by Stokes' Theorem, we obtain the result that the flux Φ through any surface spanning

$\mathcal{C}$ is quantized:

$$\Phi = \int \boldsymbol{B} \cdot \mathrm{d}\boldsymbol{S} = \frac{2\pi n}{q} = n\Phi_0 \tag{19.61}$$

where $\Phi_0 = 2\pi/q$ is the flux equation (or $2\pi\hbar/q$ in ordinary units). It is not entirely self-evident why ψ should be single-valued, but experiments do indeed demonstrate the phenomenon of flux quantization, in units of Φ_0 with $|q| = 2e$ (which may be interpreted as the charge on a Cooper pair, as usual). The phenomenon is seen in non-simply connected specimens of type I superconductors (i.e. ones with holes in them, such as a ring), and in the flux filaments of type II materials; in the latter case each filament carries a single flux quantum Φ_0.

It is interesting to consider now a situation – so far entirely hypothetical – in which a magnetic monopole is placed in a superconductor. Dirac showed (1931) that for consistency with quantum mechanics the monopole strength g_m had to satisfy the 'Dirac quantization conduction'

$$qg_\mathrm{m} = n/2 \tag{19.62}$$

where q is any electronic charge, and n is an integer. It follows from (19.62) that the flux $4\pi g_\mathrm{m}$ out of any closed surface surrounding the monopole is quantized in units of Φ_0. Hence a flux filament in a superconductor can originate from, or be terminated by, a Dirac monopole (with the appropriate sign of g_m), as was first pointed out by Nambu (1974).

This is the basic model which, in one way or another, underlies many theoretical attempts to understand confinement. The monopole–antimonopole pair in a type II superconducting vacuum, joined by a quantized magnetic flux filament, provides a model of a meson. As the distance between the pair – the length of the filament – increases, so does the energy of the filament, at a rate proportional to its length, since the flux cannot spread out in directions transverse to the filament. This is exactly the kind of linearly rising potential energy required by hadron spectroscopy (see equations (1.33) and (16.145)). The configuration is stable, because there is no way for the flux to leak away; it is a conserved quantized quantity.

For the eventual application to QCD, one will want (presumably) particles carrying non-zero values of the colour quantum numbers to be confined. These quantum numbers are the analogues of electric charge in the U(1) case, rather than of magnetic charge. We imagine, therefore, interchanging the roles of magnetism and electricity in all of the foregoing. Indeed, the Maxwell equations have such a symmetry when monopoles are present, as well as charges. The essential feature of the superconducting ground state was that it involved the coherent state formed by condensation of electrically charged bosonic fermion pairs. A vacuum which confined filaments of $\boldsymbol{E}$ rather than $\boldsymbol{B}$ may be formed as a coherent state of condensed magnetic monopoles (Mandelstam 1976, 't Hooft 1976). These $\boldsymbol{E}$ filaments would then terminate on electric charges. Now magnetic monopoles do not occur naturally as solutions of QED: they would have to be introduced by hand. Remarkably enough,

however, solutions of the magnetic monopole type do occur in the case of non-Abelian gauge field theories, whose symmetry is spontaneously broken to an electromagnetic $\mathrm{U}(1)_{\mathrm{em}}$ gauge group. Just this circumstance can arise in a grand unified theory which contains $\mathrm{SU}(3)_{\mathrm{c}}$ and a residual $\mathrm{U}(1)_{\mathrm{em}}$. Incidentally, these monopole solutions provide an illuminating way of thinking about charge quantization: as Dirac (1931) pointed out, the existence of just one monopole implies, from his quantization condition (19.62), that charge is quantized.

When these ideas are applied to QCD, $\boldsymbol{E}$ and $\boldsymbol{B}$ must be understood as the appropriate colour fields (i.e. they carry an $\mathrm{SU}(3)_{\mathrm{c}}$ index). The group structure of SU(3) is also quite different from that of U(1) models, and we do not want to be restricted just to static solutions (as in the GL theory, here used as an analogue). Whether in fact the real QCD vacuum (ground state) is formed as some such coherent plasma of monopoles, with confinement of electric charges and flux, is a subject of continuing research; other schemes are also possible. As so often stressed, the difficulty lies in the non-perturbative nature of the confinement problem.

19.5 't Hooft's gauges

We must now at last grasp the nettle and consider what happens if, in the parametrization

$$\hat{\phi} = |\hat{\phi}| \exp(\mathrm{i}\hat{\theta}(x)/v) \tag{19.63}$$

we do not choose the gauge (cf (19.52))

$$\partial_\mu \hat{A}^\mu = \Box\hat{\theta}/M. \tag{19.64}$$

This was the gauge that enabled us to transform away the phase degree of freedom and reduce the equation of motion for the electromagnetic field to that of a massive vector boson. Instead of using the modulus and phase as the two independent degrees of freedom for the complex Higgs field $\hat{\phi}$, we now choose to parametrize $\hat{\phi}$, quite generally, by the decomposition

$$\hat{\phi} = 2^{-1/2}[v + \hat{\chi}_1(x) + \mathrm{i}\hat{\chi}_2(x)], \tag{19.65}$$

where the vacuum values of $\hat{\chi}_1$ and $\hat{\chi}_2$ are zero. Substituting this form for $\hat{\phi}$ into the master equation for $\hat{A}^\nu$ (obtained from (19.43) and (19.44))

$$\Box\hat{A}^\nu - \partial^\nu(\partial_\mu\hat{A}^\mu) = \mathrm{i}q[\hat{\phi}^\dagger\partial^\nu\hat{\phi} - (\partial^\nu\hat{\phi})^\dagger\hat{\phi}] - 2q^2\hat{A}^\nu\hat{\phi}^\dagger\hat{\phi}, \tag{19.66}$$

leads to the equation of motion

$$\begin{aligned}(\Box + M^2)\hat{A}^\nu - \partial^\nu(\partial_\mu\hat{A}^\mu) \;&=\; -M\partial^\nu\hat{\chi}_2 + q(\hat{\chi}_2\partial^\nu\hat{\chi}_1 - \hat{\chi}_1\partial^\nu\hat{\chi}_2) \\ &\quad - q^2\hat{A}^\nu(\hat{\chi}_1^2 + 2v\hat{\chi}_1 + \hat{\chi}_2^2)\end{aligned} \tag{19.67}$$

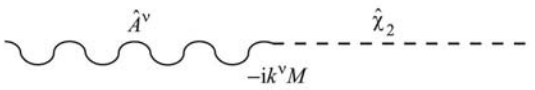


FIGURE 19.3
$\hat{A}^\nu - \hat{\chi}_2$ coupling.

with $M = qv$. At first sight this just looks like the equation of motion of an ordinary massive vector field $\hat{A}^\nu$ coupled to a rather complicated current. However, this certainly cannot be right, as we can see by a count of the degrees of freedom. In the previous gauge we had four degrees of freedom, counted either as two for the original massless $\hat{A}^\nu$ plus one each for $\hat{\theta}$ and $\hat{h}$, or as three for the massive $\hat{A}'^\nu$ and one for $\hat{h}$. If we take this new equation at face value, there seem to be three degrees of freedom for the massive field $\hat{A}^\nu$, and one for each of $\hat{\chi}_1$ and $\hat{\chi}_2$, making *five* in all. Actually, we know perfectly well that we can make use of the freedom gauge choice to set $\hat{\chi}_2$ to zero, say, reducing $\hat{\phi}$ to a real quantity and eliminating a spurious degree of freedom: we have then returned to the form (19.55). In terms of (19.67), the consequence of the unwanted degree of freedom is quite subtle, but it is basic to all gauge theories and already appeared in the photon case, in section 7.3.2. The difficulty arises when we try to calculate the propagator for $\hat{A}^\nu$ from equation (19.67).

The operator on the left-hand side can be simply inverted, as was done in section 19.1, to yield (apparently) the standard massive vector boson propagator

$$\mathrm{i}(-g^{\mu\nu} + k^\mu k^\nu/M^2)/(k^2 - M^2). \tag{19.68}$$

However, the current on the right-hand side of (19.67) is rather peculiar: instead of having only terms corresponding to $\hat{A}^\nu$ coupling to two or three particles, there is also a term involving only one field. This is the term $-M\partial^\nu\hat{\chi}_2$, which tells us that $\hat{A}^\nu$ actually couples directly to the scalar field χ_2 via the gradient coupling $(-M\partial^\nu)$. In momentum space this corresponds to a coupling strength $-\mathrm{i}k^\nu M$ and an associated vertex as shown in figure 19.3. Clearly, for a scalar particle, the momentum 4-vector is the only quantity that can couple to the vector index of the vector boson. The existence of this coupling shows that the propagators of $\hat{A}^\nu$ and $\hat{\chi}_2$ are necessarily mixed: the complete vector propagator must be calculated by summing the infinite series shown diagrammatically in figure 19.4. This complication is, of course, completely eliminated by the gauge choice $\hat{\chi}_2 = 0$. However, we are interested in pursuing the case $\hat{\chi}_2 \neq 0$.

In figure 19.4 the only unknown factor is the propagator for $\hat{\chi}_2$. This can be easily found by substituting (19.65) into $\hat{\mathcal{L}}_\mathrm{H}$ and examining the part which is quadratic in the $\hat{\chi}$'s; we find (problem 19.6)

$$\hat{\mathcal{L}}_\mathrm{H} = \frac{1}{2}\partial_\mu\hat{\chi}_1\partial^\mu\hat{\chi}_1 + \frac{1}{2}\partial_\mu\hat{\chi}_2\partial^\mu\hat{\chi} - \mu^2\hat{\chi}_1^2 + \text{ cubic and quartic terms.} \tag{19.69}$$

FIGURE 19.4
Series for the full $\hat{A}^\nu$ propagator.

FIGURE 19.5
Formal summation of the series in figure 19.4.

Equation (19.69) confirms that $\hat{\chi}_1$ is a massive field with mass $\sqrt{2}\mu$ (like the $\hat{h}$ in (19.56)), while $\hat{\chi}_2$ is massless. The $\hat{\chi}_2$ propagator is therefore i/k^2. Now that all the elements of the diagrams are known, we can formally sum the series by generalizing the well known result ((cf 10.12)and (11.27))

$$(1-x)^{-1} = 1 + x + x^2 + x^3 + \ldots \tag{19.70}$$

Diagrammatically, we rewrite the propagator of figure 19.4 as in figure 19.5 and perform the sum. Inserting the expressions for the propagators and vector-scalar coupling, and keeping track of the indices, we finally arrive at the result (problem 19.7)

$$\mathrm{i}\left(\frac{-g^{\mu\lambda} + k^\mu k^\lambda/M^2}{k^2 - M^2}\right)(g^\nu_\lambda - k^\nu k_\lambda/k^2)^{-1} \tag{19.71}$$

for the full propagator. But the inverse required in (19.71) is precisely (with a lowered index) the one we needed for the photon propagator in (7.91) – and, as we saw there, it does not exist. At last the fact that we are dealing with a gauge theory has caught up with us!

As we saw in section 7.3.2, to obtain a well-defined gauge field propagator we need to *fix the gauge*. A clever way to do this in the present (spontaneously broken) case was suggested by 't Hooft (1971b). His proposal was to set

$$\partial_\mu \hat{A}^\mu = M\xi\hat{\chi}_2 \tag{19.72}$$

where ξ is an arbitrary gauge parameter[1] (not to be confused with the superconducting coherence length). This condition is manifestly covariant, and moreover it effectively reduces the degrees of freedom by one. Inserting (19.72)

[1]We shall not enter here into the full details of quantization in such a gauge: we shall effectively treat (19.72) as a classical field relation.

into (19.67) we obtain

$$(\Box + M^2)\hat{A}^\nu - \partial^\nu(\partial_\mu \hat{A}^\mu)(1 - 1/\xi) = q(\hat{\chi}_2 \partial^\nu \hat{\chi}_1 - \hat{\chi}_1 \partial^\nu \hat{\chi}_2) - q^2 \hat{A}^\nu(\hat{\chi}_1^2 + 2v\hat{\chi}_1 + \hat{\chi}_2^2). \quad (19.73)$$

The operator appearing on the left-hand side now *does* have an inverse (see problem 19.8) and yields the general form for the gauge boson propagator

$$\mathrm{i}\left[-g^{\mu\nu} + \frac{(1-\xi)k^\mu k^\nu}{k^2 - \xi M^2}\right](k^2 - M^2)^{-1}. \quad (19.74)$$

This propagator is very remarkable[2]. The standard massive vector boson propagator

$$\mathrm{i}(-g^{\mu\nu} + k^\mu k^\nu / M^2)(k^2 - M^2)^{-1} \quad (19.75)$$

is seen to correspond to the limit $\xi \to \infty$, and in this gauge the high-energy disease outlined in section 19.1 appears to threaten renormalizability (in fact, it can be shown that there is a consistent set of Feynman rules for this gauge, and the theory is renormalizable thanks to many cancellations of divergences). For any finite ξ, however, the high-energy behaviour of the gauge boson propagator is actually $\sim 1/k^2$, which is as good as the *renormalizable* theory of QED (in Lorentz gauge). Note, however, that there seems to be another pole in the propagator (19.74) at $k^2 = \xi M^2$: this is surely unphysical since it depends on the arbitrary parameter ξ. A full treatment ('t Hooft 1971b) shows that this pole is always cancelled by an exactly similar pole in the propagator for the $\hat{\chi}_2$ field itself. These finite-ξ gauges are called *R gauges* (since they are 'manifestly renormalizable') and typically involve unphysical Higgs fields such as $\hat{\chi}_2$. The infinite-ξ gauge is known as the *U gauge* (U for unitary) since only physical particles appear in this gauge. For tree diagram calculations, of course, it is easiest to use the U gauge Feynman rules: the technical difficulties with this gauge choice only enter in loop calculations, for which the R gauge choice is easier.

Notice that in our master formula (19.74) for the gauge boson propagator the limit $M \to 0$ may be safely taken (compare the remarks about this limit for the 'naive' massive vector boson propagator in section 19.1). This yields the massless vector boson (photon) propagator in a general ξ-gauge, exactly as in equation (7.122) or (19.23).

We now proceed with the generalization of these ideas to the non-Abelian SU(2) case, which is the one relevant to the electroweak theory. The general non-Abelian case was treated by Kibble (1967).

[2] A vector boson propagator of similar form was first introduced by Lee and Yang (1962), but their discussion was not within the framework of a spontaneously broken theory, so that Higgs particles were not present, and the physical limit was obtained *only* as $\xi \to 0$.

19.6 Spontaneously broken local SU(2)×U(1) symmetry

We shall limit our discussion of the spontaneous breaking of a local non-Abelian symmetry to the particular case needed for the electroweak part of the Standard Model. This is, in fact, just the local version of the model studied in section 17.6. As noted there, the Lagrangian $\hat{\mathcal{L}}_\Phi$ of (17.97) is invariant under global SU(2) transformations of the form (17.100), and also global U(1) transformations (17.101). Thus in the local version we shall have to introduce three SU(2) gauge fields (as in section 13.1), which we call $\hat{W}_i^\mu(x)(i = 1, 2, 3)$, and one U(1) gauge field $\hat{B}^\mu(x)$. We recall that the scalar field $\hat{\phi}$ is an SU(2)-doublet

$$\hat{\phi} = \begin{pmatrix} \hat{\phi}^+ \\ \hat{\phi}^0 \end{pmatrix}, \tag{19.76}$$

so that the SU(2) covariant derivative acting on $\hat{\phi}$ is as given in (13.10), namely

$$\hat{D}^\mu = \partial^\mu + \mathrm{i}g\boldsymbol{\tau} \cdot \hat{\boldsymbol{W}}^\mu/2. \tag{19.77}$$

To this must be added the U(1) piece, which we write as $\mathrm{i}g'\hat{B}^\mu/2$, the $\frac{1}{2}$ being for later convenience. The Lagrangian (without gauge-fixing and ghost terms) is therefore

$$\hat{\mathcal{L}}_{\mathrm{G}\Phi} = (\hat{D}_\mu\hat{\phi})^\dagger(\hat{D}^\mu\hat{\phi}) + \mu^2\hat{\phi}^\dagger\hat{\phi} - \frac{\lambda}{4}(\hat{\phi}^\dagger\hat{\phi})^2 - \frac{1}{4}\hat{\boldsymbol{F}}_{\mu\nu} \cdot \hat{\boldsymbol{F}}^{\mu\nu} - \frac{1}{4}\hat{G}_{\mu\nu}\hat{G}^{\mu\nu} \tag{19.78}$$

where

$$\hat{D}^\mu\hat{\phi} = (\partial^\mu + \mathrm{i}g\boldsymbol{\tau} \cdot \hat{\boldsymbol{W}}^\mu/2 + \mathrm{i}g'\hat{B}^\mu/2)\hat{\phi}, \tag{19.79}$$

$$\hat{\boldsymbol{F}}^{\mu\nu} = \partial^\mu\hat{\boldsymbol{W}}^\nu - \partial^\nu\hat{\boldsymbol{W}}^\mu - g\hat{\boldsymbol{W}}^\mu \times \hat{\boldsymbol{W}}^\nu, \tag{19.80}$$

and

$$\hat{G}^{\mu\nu} = \partial^\mu\hat{B}^\nu - \partial^\nu\hat{B}^\mu. \tag{19.81}$$

We must now decide how to choose the non-zero vacuum expectation value that breaks this symmetry. The essential point for the electroweak application is that, after symmetry breaking, we should be left with three massive boson gauge bosons (which will be the $\mathrm{W}^\pm$ and Z^0) and one massless gauge boson, the photon. We may reasonably guess that the massless boson will be associated with a symmetry that is *un*broken by the vacuum expectation value. Put differently, we certainly do not want a 'superconducting' massive photon to emerge from the theory in this case, as the physical vacuum is not an electromagnetic superconductor. This means that we do not want to give a vacuum value to a charged field (as is done in the BCS ground state). On the other hand, we do want it to behave as a 'weak' superconductor, generating mass for $\mathrm{W}^\pm$ and Z^0. The choice suggested by Weinberg (1967) was

$$\langle 0|\hat{\phi}|0\rangle = \begin{pmatrix} 0 \\ v/\sqrt{2} \end{pmatrix} \tag{19.82}$$

where $v/\sqrt{2} = \sqrt{2}\mu/\lambda^{1/2}$, which we already considered in the global case in section 17.6. As pointed out there, (19.82) implies that the vacuum remains invariant under the combined transformation of 'U(1) + third component of SU(2) isospin' – that is, (19.82) implies

$$\left(\frac{1}{2} + t_3^{(\frac{1}{2})}\right) \langle 0|\hat{\phi}|0\rangle = 0 \tag{19.83}$$

and hence

$$\langle 0|\hat{\phi}|0\rangle \to (\langle 0|\hat{\phi}|0\rangle)' = \exp\left[\mathrm{i}\alpha\left(\frac{1}{2} + t_3^{(1/2)}\right)\right] \langle 0|\hat{\phi}|0\rangle = \langle 0|\hat{\phi}|0\rangle, \tag{19.84}$$

where as usual $t_3^{(1/2)} = \tau_3/2$ (we are using lowercase t for isospin now, anticipating that it is the *weak*, rather than hadronic, isospin – see chapter 21).

We now need to consider oscillations about (19.82) in order to see the physical particle spectrum. As in (17.107) we parametrize these conveniently as

$$\hat{\phi} = \exp(-\mathrm{i}\hat{\boldsymbol{\theta}}(x)\cdot\boldsymbol{\tau}/2v)\begin{pmatrix} 0 \\ \frac{1}{\sqrt{2}}(v + \hat{H}(x)) \end{pmatrix} \tag{19.85}$$

(compare (19.45)). However this time, in contrast to (17.107) but just as in (19.55), we can reduce the phase fields $\hat{\boldsymbol{\theta}}$ to zero by an appropriate gauge transformation, and it is simplest to examine the particle spectrum in this (*unitary*) gauge. Substituting

$$\hat{\phi} = \begin{pmatrix} 0 \\ \frac{1}{\sqrt{2}}(v + \hat{H}(x)) \end{pmatrix} \tag{19.86}$$

into (19.78) and retaining only terms which are second order in the fields (i.e. kinetic energies or mass terms) we find (problem 19.9)

$$\begin{aligned}
\hat{\mathcal{L}}^{\text{free}}_{\text{G}\Phi} \;=\; & \frac{1}{2}\partial_\mu\hat{H}\partial^\mu\hat{H} - \mu^2\hat{H}^2 \\
& - \frac{1}{4}(\partial_\mu\hat{W}_{1\nu} - \partial_\nu\hat{W}_{1\mu})(\partial^\mu\hat{W}_1^\nu - \partial^\nu\hat{W}_1^\mu) + \frac{1}{8}g^2v^2\hat{W}_{1\mu}\hat{W}_1^\mu \\
& - \frac{1}{4}(\partial_\mu\hat{W}_{2\nu} - \partial_\nu\hat{W}_{2\mu})(\partial^\mu\hat{W}_2^\nu - \partial^\nu\hat{W}_1^\mu) + \frac{1}{8}g^2v^2\hat{W}_{2\mu}\hat{W}_2^\mu \\
& - \frac{1}{4}(\partial_\mu\hat{W}_{3\nu} - \partial_\nu\hat{W}_{3\mu})(\partial^\mu\hat{W}_3^\nu - \partial^\nu\hat{W}_3^\mu) - \frac{1}{4}\hat{G}_{\mu\nu}\hat{G}^{\mu\nu} \\
& + \frac{1}{8}v^2(g\hat{W}_{3\mu} - g'\hat{B}_\mu)(g\hat{W}_3^\mu - g'\hat{B}^\mu).
\end{aligned} \tag{19.87}$$

The first line of (19.87) tells us that we have a scalar field of mass $\sqrt{2}\mu$ (the Higgs boson, again). The next two lines tell us that the components $\hat{W}_1$ and $\hat{W}_2$ of the triplet $(\hat{W}_1, \hat{W}_2, \hat{W}_3)$ acquire a mass (cf (19.56) in the U(1) case)

$$M_1 = M_2 = gv/2 \equiv M_{\text{W}}. \tag{19.88}$$

The last two lines show us that the fields $\hat{W}_3$ and $\hat{B}$ are mixed. But they can easily be unmixed by noting that the last term in (19.87) involves only the combination $g\hat{W}_3^\mu - g'\hat{B}^\mu$, which evidently acquires a mass. This suggests introducing the normalized linear combination

$$\hat{Z}^\mu = \cos\theta_{\rm W}\hat{W}_3^\mu - \sin\theta_{\rm W}\hat{B}^\mu \tag{19.89}$$

where

$$\cos\theta_{\rm W} = g/(g^2+g'^2)^{1/2} \qquad \sin\theta_{\rm W} = g'/(g^2+g'^2)^{1/2}, \tag{19.90}$$

together with the orthogonal combination

$$\hat{A}^\mu = \sin\theta_{\rm W}\hat{W}_3^\mu + \cos\theta_{\rm W}\hat{B}^\mu. \tag{19.91}$$

We then find that the last two lines of (19.87) become

$$-\frac{1}{4}(\partial_\mu\hat{Z}_\nu - \partial_\nu\hat{Z}_\mu)(\partial_\mu\hat{Z}^\nu - \partial^\nu\hat{Z}^\mu) + \frac{1}{8}v^2(g^2+g'^2)\hat{Z}_\mu\hat{Z}^\mu - \frac{1}{4}\hat{F}_{\mu\nu}\hat{F}^{\mu\nu}, \tag{19.92}$$

where

$$\hat{F}_{\mu\nu} = \partial_\mu\hat{A}_\nu - \partial_\nu\hat{A}_\mu. \tag{19.93}$$

Thus

$$M_{\rm Z} = \frac{1}{2}v(g^2+g'^2)^{1/2} = M_{\rm W}/\cos\theta_{\rm W} \tag{19.94}$$

and

$$M_{\rm A} = 0. \tag{19.95}$$

Counting degrees of freedom as in the local U(1) case, we originally had 12 in (19.78) – three massless $\hat{W}$'s and one massless $\hat{B}$, which is 8 degrees of freedom in all, together with 4 $\hat{\phi}$-fields, all with the same mass. After symmetry breaking, we have three massive vector fields $\hat{W}_1$, $\hat{W}_2$ and $\hat{Z}$ with 9 degrees of freedom, one massless vector field $\hat{A}$ with 2, and one massive scalar $\hat{H}$. Of course, the physical application will be to identify the $\hat{W}$ and $\hat{Z}$ fields with those physical particles, the $\hat{A}$ field with the massless photon, and the $\hat{H}$ field with the Higgs boson. In the gauge (19.86), the W and Z particles have propagators of the form (19.22).

The identification of $\hat{A}^\mu$ with the photon field is made clearer if we look at the form of $D_\mu\hat{\phi}$ written in terms of $\hat{A}_\mu$ and $\hat{Z}_\mu$, discarding the $\hat{W}_1$, $\hat{W}_2$ pieces:

$$\begin{aligned} D_\mu\hat{\phi} \;=\; & \left\{\partial_\mu + {\rm i}g\sin\theta_{\rm W}\left(\frac{1+\tau_3}{2}\right)\hat{A}_\mu \right. \\ & \left. + \frac{{\rm i}g}{\cos\theta_{\rm W}}\left[\frac{\tau_3}{2} - \sin^2\theta_{\rm W}\left(\frac{1+\tau_3}{2}\right)\right]\hat{Z}_\mu\right\}\hat{\phi}. \end{aligned} \tag{19.96}$$

Now the operator $(1+\tau_3)$ acting on $\langle 0|\hat{\phi}|0\rangle$ gives zero, as observed in (19.83),

and this is why $\hat{A}_\mu$ does not acquire a mass when $\langle 0|\hat{\phi}|0\rangle \neq 0$ (gauge fields coupled to *unbroken* symmetries of $\langle 0|\hat{\phi}|0\rangle$ do *not* become massive). Although certainly not unique, this choice of $\hat{\phi}$ and $\langle 0|\hat{\phi}|0\rangle$ is undoubtedly very economical and natural. We are interpreting the zero eigenvalue of $(1+\tau_3)$ as the electromagnetic charge of the vacuum, which we do not wish to be non-zero. We then make the identification

$$e = g \sin\theta_{\mathrm{W}} \tag{19.97}$$

in order to get the right 'electromagnetic D_μ' in (19.96).

We emphasize once more that the particular form of (19.87) corresponds to a *choice of gauge*, namely the unitary one (cf the discussions in sections 19.3 and 19.5). There is always the possibility of using other gauges, as in the Abelian case, and this will in general be advantageous when doing loop calculations involving renormalization. We would then return to a general parametrization such as (cf (19.65) and (17.95))

$$\hat{\phi} = \begin{pmatrix} 0 \\ v/\sqrt{2} \end{pmatrix} + \frac{1}{\sqrt{2}} \begin{pmatrix} \hat{\phi}_2 - \mathrm{i}\hat{\phi}_1 \\ \hat{\sigma} - \mathrm{i}\hat{\phi}_3 \end{pmatrix}, \tag{19.98}$$

and add 't Hooft gauge-fixing terms

$$-\frac{1}{2\xi}\Big\{ \sum_{i=1,2} (\partial_\mu \hat{W}_i^\mu + \xi M_{\mathrm{W}} \hat{\phi}_i)^2 + (\partial_\mu \hat{Z}^\mu + \xi M_{\mathrm{Z}} \hat{\phi}_3)^2 + (\partial_\mu \hat{A}^\mu)^2 \Big\}. \tag{19.99}$$

In this case the gauge boson propagators are all of the form (19.74), and ξ-dependent. In such gauges, the Feynman rules will have to involve graphs corresponding to exchange of quanta of the 'unphysical' fields $\hat{\phi}_i$, as well as those of the physical Higgs scalar $\hat{\sigma}$. These will also have to be suitable ghost interactions in the non-Abelian sector as discussed in section 13.3.3. The complete Feynman rules of the electroweak theory are given in Appendix B of Cheng and Li (1984), for example.

The model introduced here is actually the 'Higgs sector' of the Standard Model, but without any couplings to fermions. We have seen how, by supposing that the potential in (19.78) has the symmetry-breaking sign of the parameter μ^2, the $\mathrm{W}^\pm$ and Z^0 gauge bosons can be given masses. This seems to be an ingenious and even elegant 'mechanism' for arriving at a renormalizable theory of massive vector bosons. One may of course wonder whether this 'mechanism' is after all purely phenomenological, somewhat akin to the GL theory of a superconductor. In the latter case, we know that it can be derived from 'microscopic' BCS theory, and this naturally leads to the question whether there could be a similar underlying 'dynamical' theory, behind the Higgs sector. It is, in fact, quite simple to construct a theory in which the Higgs fields $\hat{\phi}$ appear as bound, or composite, states of heavy fermions.

But generating masses for the gauge bosons is not the only job that the Higgs sector does, in the Standard Model: it also generates masses for all

the fermions. As we will see in chapter 22, the gauge symmetry of the weak interactions is a *chiral* one, which requires that there should be no explicit fermion masses in the Lagrangian. We saw in chapter 18 how there is good evidence that the strong QCD interactions break chiral symmetry spontaneously, but that there is also a need for small Lagrangian masses for the quarks, which break chiral symmetry explicitly (so as to give mass to the pions, for example). The leptons are of course not coupled to QCD, and we have to assume Lagrangian masses for them too. Thus for both quarks and leptons chiral-symmetry-breaking mass terms seem to be required. The only way to preserve the weak chiral gauge symmetry is to assume that these fermion masses must, in their turn, be interpreted as arising 'spontaneously' also; that is, *not* via an explicit mass term in the Lagrangian. The dynamical generation of quark and lepton masses would, in fact, be closely analogous to the generation of the energy gap in the BCS theory, as we saw in section 18.1. So we may ask: is it possible to find a dynamical theory which generates masses for *both* the vector bosons, *and* the fermions? Such theories are generically known as 'technicolour models' (Weinberg 1979b, Susskind 1979), and they have been intensively studied (see, for example, Peskin 1997). One problem is that such theories are already tightly constrained by the precision electroweak experiments (see chapter 22), and meeting these constraints seems to require rather elaborate kinds of models. However, technicolour theories do offer the prospect of a new strongly interacting sector, which could possibly be probed at the LHC. But such ideas take us beyond the scope of the present volume. Within the Standard Model, one proceeds along what seems a more phenomenological route, attributing the masses of fermions to their couplings with the Higgs field, in a way that will be explained in chapter 22: briefly, the couplings have the Yukawa form $g_{\rm f}\bar{\hat{f}}\hat{f}\hat{\phi}$, so that when $\hat{\phi}$ develops a vev v, the fermions gain a mass $m_{\rm f} = g_{\rm f}v$.

We now turn, in the last part of the book, to weak interactions and the electroweak theory.

Problems

19.1 Show that

$$[(-k^2 + M^2)\, g^{\nu\mu} + k^\nu k^\mu] \left(\frac{-g_{\mu\rho} + k_\mu k_\rho / M^2}{k^2 - M^2} \right) = g^\nu_\rho .$$

19.2 Verify (19.18).

19.3 Verify (19.19).

19.4 Verify (19.46).

19.5 Insert (19.55) into $\hat{\mathcal{L}}_{\rm H}$ of (19.40) and derive (19.56) for the quadratic terms.

19.6 Insert (19.65) into $\hat{\mathcal{L}}_{\rm H}$ of (19.40) and derive the quadratic terms of (19.69).

19.7 Derive (19.71).

19.8 Write the left-hand side of (19.73) in momentum space (as in (19.4)), and show that the inverse of the factor multiplying $\tilde{\hat{A}}^\mu$ is (19.74) without the 'i' (cf problem 19.1).

19.9 Verify (19.87).

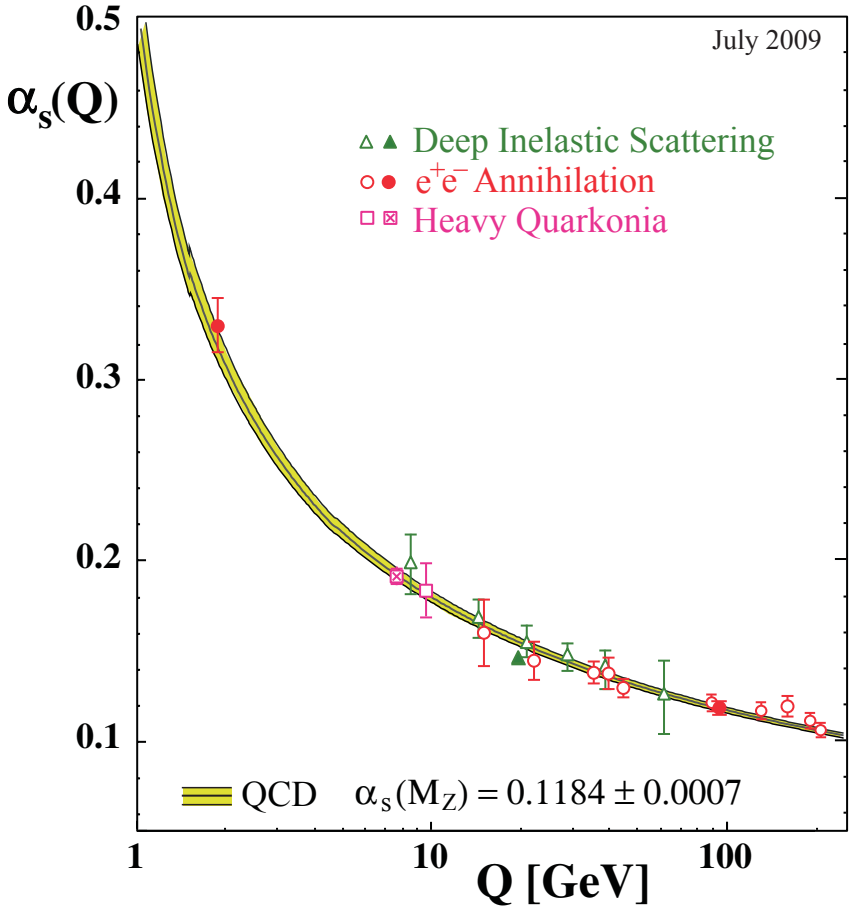


Plate I
Comparison between measurements of α_s and the theoretical prediction, as a function of the energy scale Q (Bethke 2009). (See figure 15.5 on page 129.)

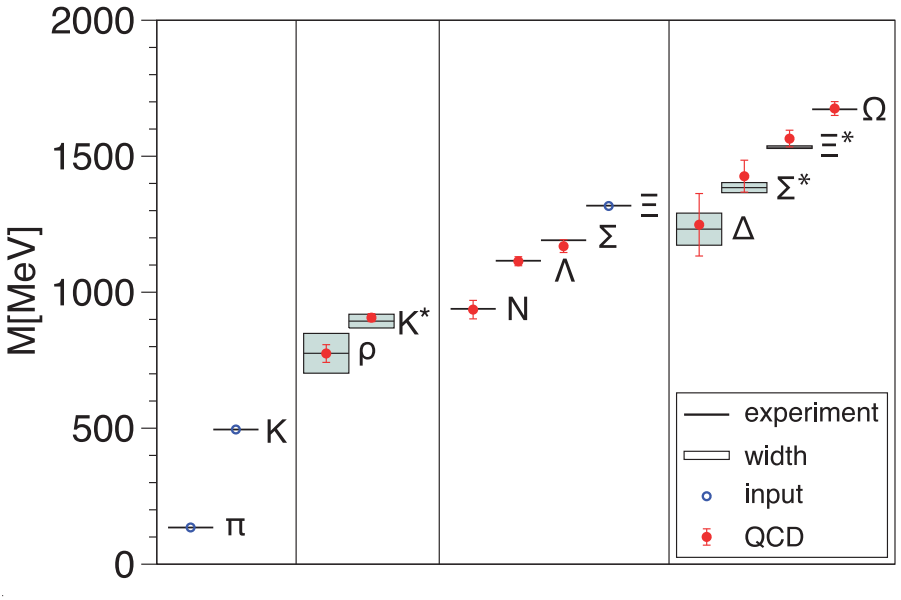


Plate II
The light hadron spectrum of QCD, from Dürr *et al.* (2008). (See figure 16.12 on page 190.)

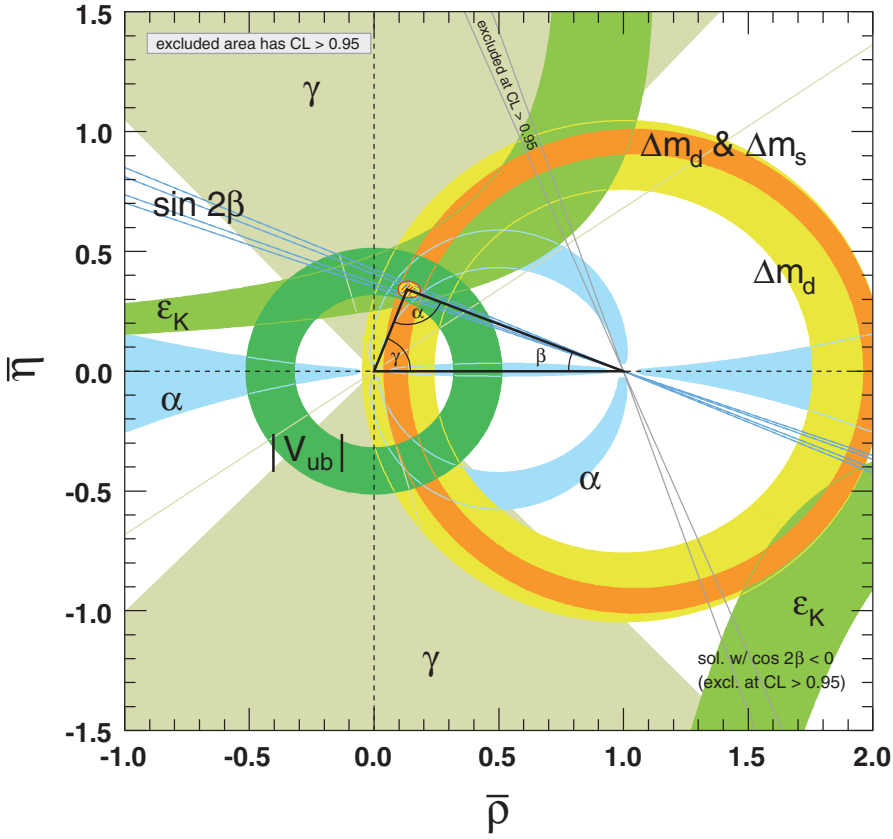


Plate III

Constraints in the $\bar{\rho}, \bar{\eta}$ plane. The shaded areas have 95% CL. [Figure reproduced, courtesy Michael Barnett for the Particle Data Group, from the review of the CKM Quark-Mixing Matrix by A Ceccucci, Z Ligeti and Y Sakai, section 11 in the *Review of Particle Physics*, K Nakamura *et al.* (Partcle Data Group) *Journal of Physics* G **37** (2010) 075021, IOP Publishing Limited.] (See figure 20.11 on page 323.)

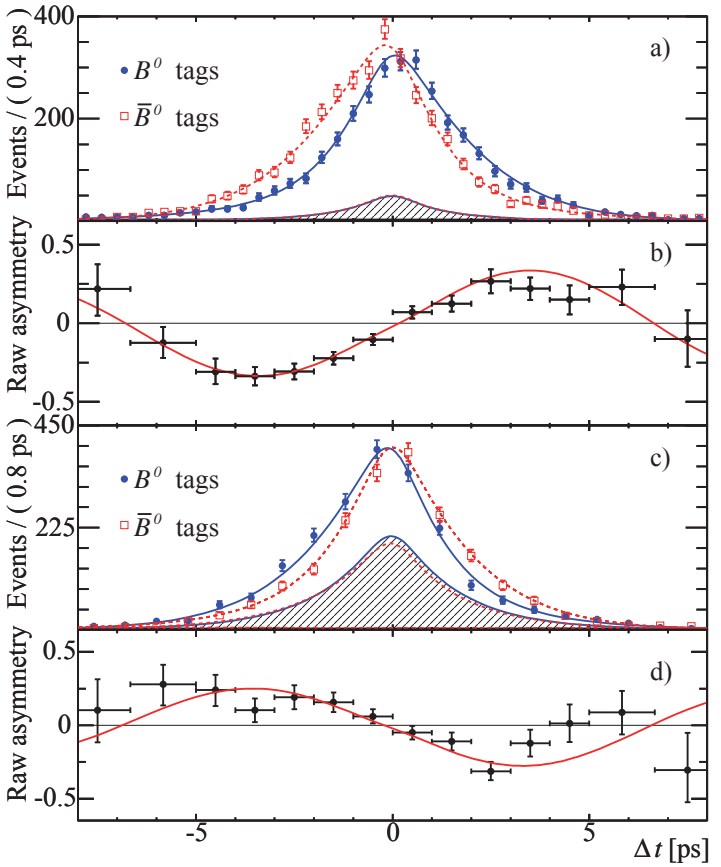


Plate IV
(a) Number of $\eta_f = -1$ candidates in the signal region with a B^0 tag (N_{B^0}) and with a $\bar{\mathrm{B}}^0$ tag ($N_{\bar{\mathrm{B}}^0}$), and (b) the measured asymmetry $(N_{\mathrm{B}^0} - N_{\bar{\mathrm{B}}^0})/(N_{\mathrm{B}^0} + N_{\bar{\mathrm{B}}^0})$, as functions of t; (c) and (d) are the corresponding distributions for the $\eta_f = +1$ candidates. Figure reprinted with permission from Aubert *et al.* (BaBar Collaboration) *Phys. Rev. Lett.* **99** 171803 (2007). Copyright 2007 by the American Physical Society. (See figure 21.7 on page 341.)

Part VIII

Weak Interactions and the Electroweak Theory


Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

20

Introduction to the Phenomenology of Weak Interactions

Public letter to the group of the Radioactives at the district society meeting in Tübingen:

Physikalisches Institut
der Eidg. Technischen Hochschule
Gloriastr.
Zürich

Zürich, 4. Dec. 1930

Dear Radioactive Ladies and Gentlemen,

As the bearer of these lines, to whom I graciously ask you to listen, will explain to you in more detail, how because of the 'wrong' statistics of the N and ^{6}Li nuclei and the continuous β-spectrum, I have hit upon a desperate remedy to save the 'exchange theorem' of statistics and the law of conservation of energy. Namely, the possiblity that there could exist in the nuclei electrically neutral particles, that I wish to call neutrons, which have the spin $\frac{1}{2}$ and obey the exclusion principle and which further differ from light quanta in that they do not travel with the velocity of light. The mass of the neutrons should be of the same order of magnitude as the electron mass and in any event not larger than 0.01 proton masses. – The continuous β-spectrum would then become understandable by the assumption that in β-decay, a neutron is emitted in addition to the electron such that the sum of the energies of the neutron and electron is constant.
...

I admit that on a first look my way out might seem to be quite unlikely, since one would certainly have seen the neutrons by now if they existed. But nothing ventured nothing gained, and the seriousness of the matter with the continuous β-spectrum is illustrated by a quotation of my honoured predecessor in office, Mr. Debye, who recently told me in Brussels: 'Oh, it is best not to think about it, like the new taxes.' Therefore one should earnestly discuss each way of salvation. – So, dear Radioactives, examine and judge it. – Unfortunately I cannot appear in Tübingen personally, since I am indispensable here in Zürich because of a ball on the

night of 6/7 December. – With my best regards to you, and also Mr. Back, your humble servant,

W. Pauli

Quoted from Winter (2000), pages 4–5.

At the end of the previous chapter we arrived at an important part of the Lagrangian of the Standard Model, namely the terms involving just the gauge and Higgs fields. The full electroweak Lagrangian also includes, of course, the couplings of these fields to the quarks and leptons. We could at this point simply write these couplings down, with little motivation, and proceed at once to discuss the empirical consequences. But such an approach, though economical, would assume considerable knowledge of weak interaction phenomenology on the reader's part. We prefer to keep this book as self-contained as possible, and so in the present chapter we shall provide an introduction to this phenomenology, following a 'semi-historical' route (for fuller historical treatments we refer the reader to Marshak *et al.* 1969, or to Winter 2000, for example).

Much of what we shall discuss is still, for many purposes, a very useful approximation to the full theory at energies well below the masses of the $\mathrm{W}^{\pm}$ ($\sim$80 GeV) and Z^0 ($\sim$90 GeV). The reason for this is that in the electroweak theory (chapter 22), tree-level amplitudes have a structure very similar to that in the purely electromagnetic case, namely (see equation (8.101))

$$j^{\mu}_{\mathrm{wk}} \, \frac{(-g_{\mu\nu} + q_\mu q_\nu / M^2_{\mathrm{W,Z}})}{q^2 - M^2_{\mathrm{W,Z}}} \, j^{\nu}_{\mathrm{wk}} \tag{20.1}$$

where j^{μ}_{wk} is a weak current, and we are using (19.75) for the propagator of the exchanged W or Z bosons. For $q^2 \ll M^2_{\mathrm{W,Z}}$, (20.1) becomes proportional to the product of two currents; this 'current–current' form was for many years the basis of weak interaction phenomenology, as we now describe.

20.1 Fermi's 'current–current' theory of nuclear β-decay, and its generalizations

The first quantum field theory of a weak interaction process was proposed by Fermi (1934a,b) for nuclear β-decay, building on the 'neutrino hypothesis' of Pauli. In 1930, Pauli (in his 'Dear Radioactive Ladies and Gentlemen' letter) had suggested that the continuous e^- spectrum in β-decay could be

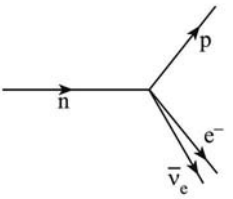


FIGURE 20.1
Four-fermion interaction for neutron β-decay.

understood by supposing that, in addition to the e^-, the decaying nucleus also emitted a light, spin-$\frac{1}{2}$, electrically neutral particle, which he called the 'neutron'. In this first version of the proposal, Pauli regarded his hypothetical particle as a constituent of the nucleus. This had the attraction of solving not only the problem with the continuous e^- spectrum, but a second problem as well – what he called the 'wrong' statistics of the ^{14}N and ^{6}Li nuclei. Taking ^{14}N for definiteness, the problem was as follows. Assuming that the nucleus was somehow composed of the only particles (other than the photon) known in 1930, namely electrons and protons, one requires 14 protons and 7 electrons for the known charge of 7. This implies a half-odd integer value for the total nuclear spin. But data from molecular spectra indicated that the nitrogen nuclei obeyed Bose–Einstein, not Fermi–Dirac statistics, so that – if the usual 'spin-statistics' connection were to hold – the spin of the nitrogen nucleus should be an integer, not a half-odd integer. This second part of Pauli's hypothesis was quite soon overtaken by the discovery of the (real) neutron by Chadwick (1932), after which it was rapidly accepted that nuclei consisted of protons and (Chadwick's) neutrons.

However, the β-spectrum problem remained, and at the Solvay Conference in 1933 Pauli restated his hypothesis (Pauli 1934), using now the name 'neutrino' which had meanwhile been suggested by Fermi. Stimulated by the discussions at the Solvay meeting, Fermi then developed his theory of β-decay. In the new picture of the nucleus, neither the electron nor the neutrino were to be thought of as nuclear constituents. Instead, the electron-neutrino pair had somehow to be created and emitted in the transition process of the nuclear decay, much as a photon is created and emitted in nuclear γ-decay. Indeed, Fermi relied heavily on the analogy with electromagnetism. The basic process was assumed to be the transition neutron→proton, with the emission of an $e^-\nu$ pair, as shown in figure 20.1. The n and p were then regarded as 'elementary' and without structure (point-like); the whole process took place at a single space-time point, like the emission of a photon in QED. Further, Fermi conjectured that the nucleons participated via a weak interaction analogue of the electromagnetic transition *currents* frequently encountered in volume 1 for QED. In this case, however, rather than having the 'charge conserving' form

of $\bar{u}_{\rm p}\gamma^{\mu}u_{\rm p}$ for instance, the 'weak current' had the form $\bar{u}_{\rm p}\gamma^{\mu}u_{\rm n}$, in which the charge of the nucleon changed. The lepton pair was also charged, obviously. The whole interaction then had to be Lorentz invariant, implying that the $\mathrm{e}^{-}\nu$ pair had also to appear in a similar (4-vector) 'current' form. Thus a 'current–current' amplitude was proposed, of the form

$$A\bar{u}_{\rm p}\gamma^{\mu}u_{\rm n}\bar{u}_{\mathrm{e}^-}\gamma_{\mu}u_{\nu}, \tag{20.2}$$

where A was a constant. Correspondingly, the process was described field theoretically in terms of the local interaction density

$$A\bar{\hat{\psi}}_{\rm p}(x)\gamma^{\mu}\hat{\psi}_{\rm n}(x)\bar{\hat{\psi}}_{\rm e}(x)\gamma_{\mu}\hat{\psi}_{\nu}(x). \tag{20.3}$$

The discovery of positron β-decay soon followed, and then of electron capture; these processes were easily accommodated by adding to (20.3) its Hermitian conjugate

$$A\bar{\hat{\psi}}_{\rm n}(x)\gamma^{\mu}\hat{\psi}_{\rm p}(x)\bar{\hat{\psi}}_{\nu}(x)\gamma_{\mu}\hat{\psi}_{\rm e}(x), \tag{20.4}$$

taking A to be real. The sum of (20.3) and (20.4) gave a good account of many observed characteristics of β-decay, when used to calculate transition probabilities in first-order perturbation theory.

Soon after Fermi's theory was presented, however, it became clear that the observed selection rules in some nuclear transitions could not be accounted for by the forms (20.3) and (20.4). Specifically, in 'allowed' transitions (where the orbital angular momentum carried by the leptons is zero) it was found that, while for many transitions the nuclear spin did not change ($\Delta J = 0$), for others – of comparable strength – a change of nuclear spin by one unit ($\Delta J = 1$) occurred. Now, in nuclear decays the energy release is very small ($\sim$ few MeV) compared to the mass of a nucleon, and so the non-relativistic limit is an excellent approximation for the nucleon spinors. It is then easy to see (problem 20.1) that, in this limit, the interactions (20.3) and (20.4) imply that the nucleon spins cannot 'flip'. Hence some other interaction(s) must be present. Gamow and Teller (1936) introduced the general four-fermion interaction, constructed from bilinear combinations of the nucleon pair and of the lepton pair, but not their derivatives. For example, the combination

$$\bar{\hat{\psi}}_{\rm p}(x)\hat{\psi}_{\rm n}(x)\bar{\hat{\psi}}_{\rm e}(x)\hat{\psi}_{\nu}(x) \tag{20.5}$$

could occur, and also

$$\bar{\hat{\psi}}_{\rm p}(x)\sigma_{\mu\nu}\hat{\psi}_{\rm n}(x)\bar{\hat{\psi}}_{\rm e}\sigma^{\mu\nu}\hat{\psi}_{\nu}(x) \tag{20.6}$$

where

$$\sigma_{\mu\nu} = \frac{\mathrm{i}}{2}(\gamma_{\mu}\gamma_{\nu} - \gamma_{\nu}\gamma_{\mu}). \tag{20.7}$$

The non-relativistic limit of (20.5) gives $\Delta J = 0$, but (20.6) allows $\Delta J = 1$.

Other combinations are also possible, as we shall discuss shortly. Note that the interaction must always be Lorentz invariant.

Thus began a long period of difficult experimentation to establish the correct form of the β-decay interaction. With the discovery of the muon (in 1937) and the pion (ten years later) more weak decays became experimentally accessible, for example μ decay

$$\mu^- \to \mathrm{e}^- + \nu + \nu \tag{20.8}$$

and π decay

$$\pi^- \to \mathrm{e}^- + \nu. \tag{20.9}$$

Note that we have deliberately called all the neutrinos just 'ν', without any particle/antiparticle indication, or lepton flavour label; we shall have more to say on these matters in section 20.3. There were hopes that the couplings of the pairs (p,n), (ν, e^-) and (ν, μ^-) might have the same form ('universality') but the data was incomplete, and in part apparently contradictory.

The breakthrough came in 1956, when Lee and Yang (1956) suggested that parity was not conserved in all weak decays. Hitherto, it had always been assumed that any physical interaction had to be such that parity was conserved, and this assumption had been built into the structure of the proposed β-decay interactions, such as (20.3), (20.5) or (20.6). Once it was looked for properly, following the analysis of Lee and Yang, parity violation was indeed found to be a strikingly evident feature of weak interactions.

20.2 Parity violation in weak interactions, and V-A theory

20.2.1 Parity violation

In 1957, the experiment of Wu *et al.* (1957) established for the first time that parity was violated in a weak interaction, specifically nuclear β-decay. The experiment involved a sample of ^{60}Co ($J = 5$) cooled to 0.01 K in a solenoid. At this temperature most of the nuclear spins are aligned by the magnetic field, and so there is a net polarization $\langle \boldsymbol{J} \rangle$, which is in the direction opposite to the applied magnetic field. ^{60}Co decays to ^{60}Ni ($J = 4$), a $\Delta J = 1$ transition. The degree of ^{60}Co alignment was measured from observations of the angular distribution of γ-rays from ^{60}Ni. The relative intensities of electrons emitted along and against the magnetic field direction were measured, and the results were consistent with a distribution of the form

$$I(\theta) = 1 - \langle \boldsymbol{J} \rangle \cdot \boldsymbol{p}/E \tag{20.10}$$

$$= 1 - Pv\cos\theta \tag{20.11}$$

where $v, \boldsymbol{p}$ and E are respectively the electron speed, momentum and energy, P is the magnitude of the polarization, and θ is the angle of emission of the electron with respect to $\langle \boldsymbol{J} \rangle$.

Why does this indicate parity violation? To see this, we recall from the discussion of the parity operation $\mathbf{P}$ in section 4.2.1 that the angular momentum $\boldsymbol{J}$ is an axial vector such that $\langle \boldsymbol{J} \rangle \to \langle \boldsymbol{J} \rangle$ under $\mathbf{P}$, while $\boldsymbol{p}$ is a polar vector transforming by $\boldsymbol{p} \to -\boldsymbol{p}$. Hence, in the parity-transformed system, the distribution (20.11) would have the form

$$I_{\mathbf{P}}(\theta) = 1 + Pv\cos\theta \tag{20.12}$$

The difference between (20.12) and (20.11) implies that, by performing the measurement, we can *determine* which of the two coordinate systems we must in fact be using. The two are inequivalent, in contrast to all the other coordinate system equivalences which we have previously studied (e.g. under three-dimensional rotations, and Lorentz transformations). This is an operational consequence of 'parity violation'. The crucial point in this example, evidently, is the appearance of the *pseudoscalar* quantity $\langle \boldsymbol{J} \rangle \cdot \boldsymbol{p}$ in (20.10), alongside the obviously scalar quantity '1'.

The Fermi theory, employing only vector currents, needs a modification to accommodate this result. We saw in section 4.2.1 that a combination of vector ('V') and axial vector ('A') currents would be parity-violating. Indeed, after many years of careful experiments, and many false trails, it was eventually established (always, of course, to within some experimental error) that the currents participating in Fermi's current–current interaction are, in fact, certain combinations of V-type and A-type currents, for both nucleons and leptons.

20.2.2 V-A theory: chirality and helicity

Quite soon after the discovery of parity violation, Sudarshan and Marshak (1958), and then Feynman and Gell-Mann (1958) and Sakurai (1958), proposed a specific form for the current–current interaction, namely the V-A ('V minus A') structure. For example, in place of the leptonic combination $\bar{u}_{\mathrm{e}^-}\gamma_\mu u_\nu$, these authors proposed the form $\bar{u}_{\mathrm{e}^-}\gamma_\mu(1-\gamma_5)u_\nu$, being the difference (with equal weight) of a V-type and an A-type current. For the part involving the nucleons the proposal was slightly more complicated, having the form $\bar{u}_{\mathrm{p}}\gamma_\mu(1 - r\gamma_5)u_{\mathrm{n}}$ where r had the empirical value $r \approx 1.2$. From our present perspective, of course, the hadronic transition is actually occurring at the quark level, so that rather than a transition $\mathrm{n} \to \mathrm{p}$ we now think in terms of a $\mathrm{d} \to \mathrm{u}$ one. In this case, the remarkable fact is that the appropriate current to use is, once again, essentially the simple 'V-A' one, $\bar{u}_{\mathrm{u}}\gamma_\mu(1-\gamma_5)u_{\mathrm{d}}$[1]. *This V-A structure for quarks and leptons is fundamental to the Standard Model.*

[1] We shall see in section 20.7 that a slight modification is necessary.

We must now at once draw the reader's attention to a rather remarkable feature of this V-A structure, which is that the $(1-\gamma_5)$ factor can be thought of as acting either on the u spinor or on the $\bar{u}$ spinor. Consider, for example, a term $\bar{u}_{\mathrm{e}^-}\gamma_\mu(1-\gamma_5)u_\nu$. We have

$$\begin{aligned}\bar{u}_{\mathrm{e}^-}\gamma_\mu(1-\gamma_5)u_\nu &= u^\dagger_{\mathrm{e}^-}\beta\gamma_\mu(1-\gamma_5)u_\nu \\ &= u^\dagger_{\mathrm{e}^-}(1-\gamma_5)\beta\gamma_\mu u_\nu \\ &= [(1-\gamma_5)u_{\mathrm{e}^-}]^\dagger\beta\gamma_\mu u_\nu \\ &= \overline{[(1-\gamma_5)u_{\mathrm{e}^-}]}\,\gamma_\mu u_\nu. \end{aligned} \tag{20.13}$$

To understand the significance of this, it is advantageous to work in the representation (3.40) of the Dirac matrices, in which γ_5 is diagonal, namely

$$\gamma_5 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \quad \boldsymbol{\alpha} = \begin{pmatrix} \boldsymbol{\sigma} & 0 \\ 0 & -\boldsymbol{\sigma} \end{pmatrix} \quad \beta = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \quad \boldsymbol{\gamma} = \begin{pmatrix} 0 & -\boldsymbol{\sigma} \\ \boldsymbol{\sigma} & 0 \end{pmatrix}. \tag{20.14}$$

Readers who have not worked through problem 9.4 might like to do so now; we may also suggest a backward glance at section 12.4.2 and chapter 17.

First of all it is clear that any combination '$(1-\gamma_5)u$' is an eigenstate of γ_5 with eigenvalue -1:

$$\gamma_5(1-\gamma_5)u = (\gamma_5 - 1)u = -(1-\gamma_5)u \tag{20.15}$$

using $\gamma_5^2 = 1$. In the terminology of section 12.4.2, '$(1-\gamma_5)u$' has definite *chirality*, namely L ('left-handed'), meaning that it belongs to the eigenvalue -1 of γ_5. We may introduce the projection operators P_{R}, P_{L} of section 12.4.2,

$$P_{\mathrm{L}} \equiv \left(\frac{1-\gamma_5}{2}\right) \qquad P_{\mathrm{R}} \equiv \left(\frac{1+\gamma_5}{2}\right) \tag{20.16}$$

satisfying

$$P_{\mathrm{R}}^2 = P_{\mathrm{R}} \qquad P_{\mathrm{L}}^2 = P_{\mathrm{L}} \qquad P_{\mathrm{R}}P_{\mathrm{L}} = P_{\mathrm{L}}P_{\mathrm{R}} = 0 \qquad P_{\mathrm{R}} + P_{\mathrm{L}} = 1, \tag{20.17}$$

and define

$$u_{\mathrm{L}} \equiv P_{\mathrm{L}}u, \qquad u_{\mathrm{R}} \equiv P_{\mathrm{R}}u \tag{20.18}$$

for any u. Then

$$\begin{aligned}\bar{u}_1\gamma_\mu\left(\frac{1-\gamma_5}{2}\right)u_2 &= \bar{u}_1\gamma_\mu P_{\mathrm{L}}u_2 = \bar{u}_1\gamma_\mu P_{\mathrm{L}}^2 u_2 \\ &= \bar{u}_1\gamma_\mu P_{\mathrm{L}}u_{2\mathrm{L}} = \bar{u}_1 P_{\mathrm{R}}\gamma_\mu u_{2\mathrm{L}} \\ &= u_1^\dagger P_{\mathrm{L}}\beta\gamma_\mu u_{2\mathrm{L}} = \bar{u}_{1\mathrm{L}}\gamma_\mu u_{2\mathrm{L}} \end{aligned} \tag{20.19}$$

which formalizes (20.13) and emphasizes the fact that *only the chiral L components of the u spinors enter into weak interactions*, a remarkably simple statement.

To see the physical consequences of this, we need the forms of the Dirac spinors in this new representation, which we shall now derive explicitly, for convenience. As usual, positive energy spinors are defined as solutions of $(\not{p} - m)u = 0$, so that writing

$$u = \begin{pmatrix} \phi \\ \chi \end{pmatrix} \tag{20.20}$$

we obtain

$$\begin{aligned} (E - \boldsymbol{\sigma} \cdot \boldsymbol{p})\phi &= m\chi \\ (E + \boldsymbol{\sigma} \cdot \boldsymbol{p})\chi &= m\phi. \end{aligned} \tag{20.21}$$

A convenient choice of 2-component spinors ϕ, χ is to take them to be *helicity eigenstates* (see section 3.3). For example, the eigenstate ϕ_+ with positive helicity $\lambda = +1$ satisfies

$$\boldsymbol{\sigma} \cdot \boldsymbol{p}\phi_+ = |\boldsymbol{p}|\phi_+ \tag{20.22}$$

while the eigenstate ϕ_- with $\lambda = -1$ satisfies (20.22) with a minus on the right-hand side. Thus the spinor $u(p, \lambda = +1)$ can be written as

$$u(p, \lambda = +1) = N \begin{pmatrix} \phi_+ \\ \frac{(E - |\boldsymbol{p}|)}{m}\phi_+ \end{pmatrix}. \tag{20.23}$$

The normalization N is fixed as usual by requiring $\bar{u}u = 2m$, from which it follows (problem 20.2) that $N = (E + |\boldsymbol{p}|)^{1/2}$. Thus finally we have

$$u(p, \lambda = +1) = \begin{pmatrix} \sqrt{E + |\boldsymbol{p}|}\phi_+ \\ \sqrt{E - |\boldsymbol{p}|}\phi_+ \end{pmatrix}. \tag{20.24}$$

Similarly

$$u(p, \lambda = -1) = \begin{pmatrix} \sqrt{E - |\boldsymbol{p}|}\phi_- \\ \sqrt{E + |\boldsymbol{p}|}\phi_- \end{pmatrix}. \tag{20.25}$$

Now we have agreed that only the chiral 'L' components of all u-spinors enter into weak interactions, in the Standard Model. But from the explicit form of γ_5 given in (20.14), we see that when acting on any spinor u, the projector P_{L} 'kills' the top two components:

$$P_{\mathrm{L}} \begin{pmatrix} \phi \\ \chi \end{pmatrix} = \begin{pmatrix} 0 \\ \chi \end{pmatrix}. \tag{20.26}$$

In particular

$$P_{\mathrm{L}} u(p, \lambda = +1) = \begin{pmatrix} 0 \\ \sqrt{E - |\boldsymbol{p}|}\phi_+ \end{pmatrix} \tag{20.27}$$

and

$$P_{\mathrm{L}} u(p, \lambda = -1) = \begin{pmatrix} 0 \\ \sqrt{E + |\boldsymbol{p}|}\phi_- \end{pmatrix}. \tag{20.28}$$

Equations (20.27) and (20.28) are very important. In particular, equation (20.27) implies that in the limit of zero mass m (and hence $E \to |\boldsymbol{p}|$), only the *negative helicity* u-spinor will enter. More quantitatively, using

$$\frac{\sqrt{E-|\boldsymbol{p}|}}{\sqrt{E+|\boldsymbol{p}|}} = \frac{\sqrt{E^2-\boldsymbol{p}^2}}{(E+|\boldsymbol{p}|)} \approx \frac{m}{2E} \qquad \text{for } m \ll E, \tag{20.29}$$

we can say that *positive helicity components of all fermions are suppressed in V-A matrix elements, relative to the negative helicity components, by factors of order* (m/E). Bearing in mind that the helicity operator $\boldsymbol{\sigma}\cdot\boldsymbol{p}/|\boldsymbol{p}|$ is a pseudoscalar, this 'unequal' treatment for $\lambda=+1$ and $\lambda=-1$ components is, of course, precisely related to the parity violation built in to the V-A structure.

A similar analysis may be done for the v-spinors. They satisfy $(\not{p}+m)v=0$ and the normalization $\bar{v}v=-2m$. We must however remember the 'small subtlety' to do with the labelling of v-spinors, discussed in section 3.4.3: the 2-component spinors χ_- in $v(p,\lambda=+1)$ actually satisfy $\boldsymbol{\sigma}\cdot\boldsymbol{p}\chi_- = -|\boldsymbol{p}|\chi_-$, and similarly the χ_+'s in $v(p,\lambda=-1)$ satisfy $\boldsymbol{\sigma}\cdot\boldsymbol{p}\chi_+ = |\boldsymbol{p}|\chi_+$. We then find (problem 20.3) the results

$$v(p,\lambda=+1) = \begin{pmatrix} -\sqrt{E-|\boldsymbol{p}|}\chi_- \\ \sqrt{E+|\boldsymbol{p}|}\chi_- \end{pmatrix} \tag{20.30}$$

and

$$v(\lambda=-1) = \begin{pmatrix} \sqrt{E+|\boldsymbol{p}|}\chi_+ \\ -\sqrt{E-|\boldsymbol{p}|}\chi_+ \end{pmatrix}. \tag{20.31}$$

Once again, the action of P_{L} removes the top two components, leaving the result that, in the massless limit, only the $\lambda=+1$ state survives. Recalling the 'hole theory' interpretation of section 3.4.3, this would mean that *the positive helicity components of all antifermions dominate in V-A interactions*, negative helicity components being suppressed by factors of order m/E. The proportionality of the negative helicity amplitude to the mass of the antifermion is of course exactly as noted for $\pi^+ \to \mu^+\nu_\mu$ decay in section 18.2.

We should emphasize that although the above results, stated in italics, were derived in the convenient representation (20.14) for the Dirac matrices, they actually hold independently of any choice of representation. This can be shown by using general helicity projection operators.

In Pauli's original letter, he suggested that the mass of the neutrino might be of the same order as the electron mass. Immediately after the discovery of parity violation, it was realized that the result could be elegantly explained by the assumption that the neutrinos were strictly massless particles (Landau 1957, Lee and Yang 1957 and Salam 1957). In this case, u and v spinors satisfy the same equation $\not{p}(u \text{ or } v) = 0$, which reduces via (20.21) (in the

$m = 0$ limit) to the two independent two-component 'Weyl' equations.

$$E\phi_0 = \boldsymbol{\sigma} \cdot \boldsymbol{p}\, \phi_0 \qquad E\chi_0 = -\boldsymbol{\sigma} \cdot \boldsymbol{p}\, \chi_0. \tag{20.32}$$

Remembering that $E = |\boldsymbol{p}|$ for a massless particle, we see that ϕ_0 has positive helicity and χ_0 negative helicity. In this strictly massless case, helicity is Lorentz invariant, since the direction of $\boldsymbol{p}$ cannot be reversed by a velocity transformation with $v < c$. Furthermore, each of the equations in (20.32) violates parity, since E is clearly a scalar while $\boldsymbol{\sigma} \cdot \boldsymbol{p}$ is a pseudoscalar (note that when $m \neq 0$ we can infer from (20.21) that, in this representation, $\phi \leftrightarrow \chi$ under $\mathbf{P}$, which is consistent with (20.32) and with the form of β in (20.14)). Thus the (massless) neutrino could be 'blamed' for the parity violation. In this model, neutrinos have one definite helicity, either positive or negative. As we have seen, the massless limit of the (four-component) V-A theory leads to the same conclusion.

Which helicity is actually chosen by Nature was determined in a classic experiment by Goldhaber *et al.* (1958), involving the K-capture reaction

$$\mathrm{e}^- + {}^{152}\mathrm{Eu} \rightarrow \nu + {}^{152}\mathrm{Sm}^*, \tag{20.33}$$

as described by Bettini (2008), for example. They found that the helicity of the emitted neutrino was (within errors) 100% *negative*, a result taken as confirming the '2-component' neutrino theory, and the V-A theory.

We now know that neutrinos are not massless. This information does not come from studies of nuclear decays, but rather from a completely different phenomenon – that of *neutrino oscillations*, which we shall mention again in the following section, and treat more fully in section 21.4. Neutrino masses are so small that the existence of the 'wrong helicity' component cannot be detected experimentally in processes such as (20.33), or indeed in any of the reactions we shall discuss, apart from neutrino oscillations.

In section 4.2.2 we introduced the charge conjugation operation $\mathbf{C}$ (see also section 7.5.2). As we noted there, $\mathbf{C}$ is not a good symmetry in weak interactions. The V-A interaction treats a negative helicity fermion very differently from a negative helicity antifermion, while one is precisely transformed into the other under $\mathbf{C}$. However, it is clear that the helicity operator itself is odd under $\mathbf{P}$. Thus the $\mathbf{CP}$ conjugate of a negative helicity fermion is positive helicity antifermion, which is what the V-A interaction selects. It may easily be verified (problem 20.4) that the '2-component' theory of (20.32) automatically incorporates $\mathbf{CP}$ invariance. Elegance notwithstanding, however, there are $\mathbf{CP}$-violating weak interactions, as mentioned in section 4.2.3. How this is accommodated within the Standard Model we shall discuss in section 20.7.3.

For charged fermions the distinction between particle and antiparticle is clear; but is there a conserved quantum number which we can use instead of charge to distinguish a neutrino from an antineutrino? That is the question to which we now turn.

20.3 Lepton number and lepton flavours

In section 1.2.1 of volume 1 we gave a brief discussion of leptonic quantum numbers ('lepton flavours'), adopting a traditional approach in which the data is interpreted in terms of conserved quantum numbers carried by neutrinos, which serve to distinguish neutrinos from antineutrinos. We must now examine the matter more closely, in the light of what we have learned about the helicity properties of the V-A interaction.

In 1995, Davis (1955) – following a suggestion made by Pontecorvo (1946) – argued as follows. Consider the e^- capture reaction $e^- + p \to \nu + n$, which was of course well established. Then in principle the inverse reaction $\nu + n \to e^- + p$ should also exist. Of course, the cross section is extremely small, but by using a large enough target volume this might perhaps be compensated. Specifically, the reaction $\nu + {}^{37}_{17}\mathrm{Cl} \to e^- + {}^{37}_{18}\mathrm{Ar}$ was proposed, the argon being detected through its radioactive decay. Suppose, however, that the 'neutrinos' actually used are those which accompany electrons in β^--decay. If (as was supposed in section 1.2.1) these are to be regarded as antineutrinos, '$\bar{\nu}$', carrying a conserved lepton number, then the reaction

$$\text{`}\bar{\nu}\text{'} + {}^{37}_{17}\mathrm{Cl} \to e^- + {}^{37}_{18}\mathrm{Ar} \tag{20.34}$$

should *not* be observed. If, on the other hand, the 'ν' in the capture process and the '$\bar{\nu}$' in β-decay are not distinguished by the weak interaction, the reaction (20.34) should be observed. Davis found no evidence for reaction (20.34), at the expected level of cross section, a result which could clearly be interpreted as confirming the 'conserved electron number hypothesis'.

However, another interpretation is possible. The e^- in β-decay has predominately negative helicity, and its accompanying '$\bar{\nu}$' has predominately positive helicity. The fraction of the other helicity present is of the order m/E, where $E \sim$ few Mev, and the neutrino mass is less than 1eV; this is, therefore, an almost undetectable 'contamination' of negative helicity component in the '$\bar{\nu}$'. Now the property of the V-A interaction is that it conserves helicity in the zero mass limit (in which chirality is the same as helicity). Hence the positive helicity '$\bar{\nu}$' from β^--decay will (predominately) produce a positive helicity lepton, which must be the e^+ not the e^-. Thus the property of the V-A interaction, together with the very small value of the neutrino mass, conspire effectively to forbid (20.34), independently of any considerations about 'lepton number'.

Indeed, the 'helicity-allowed' reaction

$$\text{`}\bar{\nu}\text{'} + p \to e^+ + n \tag{20.35}$$

was observed by Reines and Cowan (1956) (see also Cowan *et al.* 1956). Reaction (20.35) too, of course, can be interpreted in terms of '$\bar{\nu}$' carrying a lepton number of -1, equal to that of the e^+. It was also established that only 'ν'

produced e^- via (20.34), where 'ν' is the helicity -1 state (or, on the other interpretation, the carrier of lepton number $+1$).

The situation may therefore be summarized as follows. In the case of e^- and e^+, all four 'modes' – $e^-(\lambda = +1), e^-(\lambda = -1), e^+(\lambda = +1), e^+(\lambda = -1)$ – are experimentally accessible via electromagnetic interactions, even though only two generally dominate in weak interactions ($e^-(\lambda = -1)$ and $e^+(\lambda = +1)$). Neutrinos, on the other hand, seem to interact only weakly. In their case, we may if we wish say that the participating states are (in association with e^- or e^+) $\bar{\nu}_e$ $(\lambda = +1)$ and $\nu_e(\lambda = -1)$, to a very good approximation. But we may also regard these two states as simply two different helicity states of one particle, rather than of a particle and its antiparticle. As we have seen, the helicity rules do the job required just as well as the lepton number rules. In short, the question is: are these 'neutrinos' distinguished only by their helicity, or is there an additional distinguishing characteristic ('electron number')? In the latter case we should expect the 'other' two states $\bar{\nu}_e(\lambda = -1)$ and $\nu_e(\lambda = +1)$ to exist as well as the ones known from weak interactions.

If, in fact, no quantum number – other than the helicity – exists which distinguishes the neutrino states, then we would have to say that the **C**-conjugate of a neutrino state is a neutrino, not an antineutrino – that is, 'neutrinos are their own antiparticles'. A neutrino would be a fermionic state somewhat like a photon, which is of course also its own antiparticle. Such '**C**-self-conjugate' fermions are called *Majorana fermions* (Majorana 1937), in contrast to the *Dirac* variety, which have all four possible modes present (2 helicities, 2 particle/antiparticle). We discussed Majorana fermions in sections 4.2.2 and 7.5.2.

The distinction between the 'Dirac' and 'Majorana' neutrino possibilities becomes an essentially 'metaphysical' one in the limit of strictly massless neutrinos, since then (as we have seen) a given helicity state cannot be flipped by going to a suitably moving Lorentz frame, nor by any weak (or electromagnetic) interaction, since they both conserve chirality which is the same as helicity in the massless limit. We would have just the two states $\nu_e(\lambda = -1)$ and $\bar{\nu}_e(\lambda = +1)$, and no way of creating $\nu_e(\lambda = +1)$ or $\bar{\nu}_e(\lambda = -1)$. The '$-$' label then becomes superfluous. Unfortunately, the massless limit is approached smoothly, and neutrino masses are, in fact, so small that the 'wrong helicity' supression factors will make it very difficult to see the presence of the possible states $\nu_e(\lambda = +1), \bar{\nu}_e(\lambda = -1)$.

One much-discussed experimental test case (see, for example, the review by Vogel and Piepke in Nakamura *et al.* 2010) concerns 'neutrinoless double β-decay', which is the process $A \to A' + e^- + e^-$, where A, A′ are nuclei. If the neutrino emitted in the first β-decay carries no electron-type conserved quantum number, then in principle it can initiate a second weak interaction, exactly as in Davis' original argument, via the diagram shown in figure 20.2. Note that this is a second-order weak process, so that the amplitude contains the very small factor G_F^2. Furthermore, the ν emitted along with the e^- at the first vertex will be predominately $\lambda = +1$, but in the second vertex

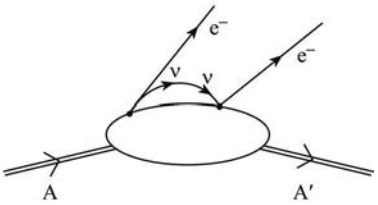


FIGURE 20.2
Double β-decay without emission of a neutrino, a test for Majorana-type neutrinos.

the V-A interaction will 'want' it to have $\lambda = -1$, like the outgoing e^-. Thus there is bound to be one 'm/E' suppression factor, whichever vertex we choose to make 'easy'. (In the case of 3-state neutrino mixing – see section 21.4 – the quantity 'm' will be an appropriately averaged mass.) There is also a complicated nuclear physics overlap factor. The expected half-lives of neutrinoless double β decays depend on the decaying nucleus, but are typically longer than $10^{24} - 10^{25}$ years. Evidently, the observation of this rare process is a formidable experimental challenge; as yet, no confirmed observation exists (see also section 21.4.5).

In the same way, '$\bar{\nu}$' particles accompanying the μ^-'s in π^- decay

$$\pi^- \to \mu^- + \text{`}\bar{\nu}\text{'} \tag{20.36}$$

are observed to produce only μ^+'s when they interact with matter, not μ^-'s. Again this can be interpreted either in terms of helicity conservation or in terms of conservation of a leptonic quantum number L_μ. We shall assume the analogous properties are true for the '$\bar{\nu}$''s accompanying τ leptons.

On the other hand, helicity arguments alone would allow the reaction

$$\text{`}\bar{\nu}\text{'} + \mathrm{p} \to \mathrm{e}^+ + \mathrm{n} \tag{20.37}$$

to proceed, but as we saw in section 1.2.1 the experiment of Danby *et al.* (1962) found no evidence for it. Thus there is evidence, in this type reaction, for a flavour quantum number distinguishing neutrinos which interact in association with one kind of charged lepton from those which interact in association with a different charged lepton. The electroweak sector of the Standard Model was originally formulated on the assumption that the three lepton flavours L_e, L_μ and L_τ are conserved, and that the neutrinos are massless. It turns out that these two assumptions are related, in the sense that if neutrinos have mass, then (barring degeneracies) 'neutrino oscillations' can occur, in which a state of one lepton flavour can acquire a component of another, as it propagates. Compelling evidence accumulated during the 2000s for oscillations of neutrinos caused by non-zero masses and neutrino mixing.

Strictly speaking, neutrino masses and oscillations lie outside the framework of the original Standard Model, and they are sometimes so regarded. Apart from anything else, the phenomenology of massive neutrinos has to allow for the possibility that they are Majorana, rather than Dirac, fermions. For the moment, we shall continue with a semi-historical path, and proceed with weak interaction phenomenology on the basis of the original Standard Model, with massless neutrinos. We return to the question of neutrino mass when we discuss neutrino oscillations (along with analogous oscillations in meson systems) in chapter 21.

20.4 The universal current × current theory for weak interactions of leptons

After the breakthroughs of parity violation and V-A theory, the earlier hopes (Pontecorvo 1947, Klein 1948, Puppi 1948, Lee, Rosenbluth and Yang 1949, Tiomno and Wheeler 1949) were revived of a universal weak interaction among the pairs of particles (p,n), $(\nu_{\rm e}, {\rm e}^-), (\nu_\mu, \mu^-)$, using the V-A modification to Fermi's theory. From our modern standpoint, this list has to be changed by the replacement of (p,n) by the corresponding quarks (u,d), and by the inclusion of the third lepton pair (ν_τ, τ^-) as well as two other quark pairs (c,s) and (t,b). It is to these pairs that the 'V-A' structure applies, as already indicated in section 20.2.2, and a certain form of 'universality' does hold, as we now describe.

Because of certain complications which arise, we shall postpone the discussion of the quark currents until section 20.7, concentrating here on the leptonic currents[2]. In this case, Fermi's original vector-like current $\bar{\hat{\psi}}_{\rm e}\gamma^\mu\hat{\psi}_\nu$ becomes modified to a *total leptonic charged current*

$$\hat{j}^\mu_{\rm CC}({\rm leptons}) = \hat{j}^\mu_{\rm wk}({\rm e}) + \hat{j}^\mu_{\rm wk}(\mu) + \hat{j}^\mu_{\rm wk}(\tau) \tag{20.38}$$

where, for example,

$$\hat{j}^\mu_{\rm wk}({\rm e}) = \bar{\hat{\nu}}_{\rm e}\gamma^\mu(1-\gamma_5)\hat{e}. \tag{20.39}$$

In (20.39) we are now adopting, for the first time, a useful shorthand whereby the field operator for the electron field, say, is denoted by $\hat{e}(x)$ rather than $\hat{\psi}_{\rm e}(x)$, and the 'x' argument is suppressed. The 'charged' current terminology refers to the fact that these weak current operators $\hat{j}^\mu_{\rm wk}$ carry net charge, in contrast to an electromagnetic current operator such as $\bar{\hat{e}}\gamma^\mu\hat{e}$ which is electrically neutral. We shall see in section 20.6 that there are also electrically neutral weak currents.

[2]Very much the same complications arise for the leptonic currents too, in the case of massive neutrinos, as we shall see in section 21.4.

The interaction Hamiltonian density accounting for all leptonic weak interactions is then taken to be

$$\hat{\mathcal{H}}^{\rm lep}_{\rm CC} = \frac{G_{\rm F}}{\sqrt{2}} \hat{j}^{\mu}_{\rm CC}({\rm leptons}) \hat{j}^{\dagger}_{{\rm CC}\mu}({\rm leptons}). \tag{20.40}$$

Note that

$$(\bar{\hat{\nu}}_{\rm e}\gamma^{\mu}(1-\gamma_5)\hat{e})^{\dagger} = \bar{\hat{e}}\gamma^{\mu}(1-\gamma_5)\hat{\nu}_{\rm e} \tag{20.41}$$

and similarly for the other bilinears. The currents can also be written in terms of the chiral components of the fields (recall section 20.2.2) using

$$2\bar{\hat{\nu}}_{\rm eL}\gamma^{\mu}\hat{e}_{\rm L} = \bar{\hat{\nu}}_{\rm e}\gamma^{\mu}(1-\gamma_5)\hat{e}, \tag{20.42}$$

for example. 'Universality' is manifest in the fact that all the lepton pairs have the same form of the V-A coupling, and the same 'strength parameter' $G_{\rm F}/\sqrt{2}$ multiplies all of the products in (20.40).

The terms in (20.40), when it is multiplied out, describe many physical processes. For example, the term

$$\frac{G_{\rm F}}{\sqrt{2}} \bar{\hat{\nu}}_{\mu}\gamma^{\mu}(1-\gamma_5)\hat{\mu}\, \bar{\hat{e}}\gamma_{\mu}(1-\gamma_5)\hat{\nu}_{\rm e} \tag{20.43}$$

describes μ^- decay:

$$\mu^- \to \nu_{\mu} + {\rm e}^- + \bar{\nu}_{\rm e}, \tag{20.44}$$

as well as all the reactions related by 'crossing' particles from one side to the other, for example

$$\nu_{\mu} + {\rm e}^- \to \mu^- + \nu_{\rm e}. \tag{20.45}$$

The value of $G_{\rm F}$ can be determined from the rate for process (20.44) (see for example Renton 1990, section 6.1.2), and it is found to be

$$G_{\rm F} \simeq 1.166 \times 10^{-5} {\rm GeV}^{-2}. \tag{20.46}$$

This is a convenient moment to notice that the theory is *not renormalizable* according to the criteria discussed in section 11.8 at the end of the previous volume: $G_{\rm F}$ has dimensions $({\rm mass})^{-2}$. We shall return to this aspect of Fermi-type V-A theory in section 22.1.

There are also what we might call 'diagonal' terms in which the same lepton pair is taken from $\hat{j}^{\mu}_{\rm wk}$ and $\hat{j}^{\dagger}_{{\rm wk}\mu}$, for example

$$\frac{G_{\rm F}}{\sqrt{2}} \bar{\hat{\nu}}_{\rm e}\gamma^{\mu}(1-\gamma_5)\hat{e}\, \bar{\hat{e}}\gamma_{\mu}(1-\gamma_5)\hat{\nu}_{\rm e} \tag{20.47}$$

which describes reactions such as

$$\bar{\nu}_{\rm e} + {\rm e}^- \to \bar{\nu}_{\rm e} + {\rm e}^-. \tag{20.48}$$

The cross section for (20.48) was measured by Reines, Gurr and Sobel (1976)

after many years of effort; the value obtained was consistent with the Glashow–Salam–Weinberg theory (see section 22.3), with the parameter $\sin^2\theta_{\rm W} = 0.29 \pm 0.05$.

It is interesting that some seemingly rather similar processes are forbidden to occur, to first order in $\hat{\mathcal{H}}^{\rm lep}_{\rm wk}$, for example

$$\bar{\nu}_\mu + {\rm e}^- \to \bar{\nu}_\mu + {\rm e}^-. \tag{20.49}$$

For reasons which will become clearer in section 20.6, (20.49) is called a 'neutral current' process, in contrast to all the others (such as β-decay or μ-decay) we have discussed so far, which are called 'charged current' processes. If the lepton pairs are arranged so as to have no net lepton number (for example ${\rm e}^-\bar{\nu}_{\rm e}, \mu^+\nu_\mu, \nu_\mu\bar{\nu}_\mu$ etc.) then pairs with non-zero charge occur in charged current processes, while those with zero charge participate in neutral current processes. In the case of (20.48), the leptons can be grouped either as $(\bar{\nu}_{\rm e}{\rm e}^-)$ which is charged, or as $(\bar{\nu}_{\rm e}\nu_{\rm e})$ or $({\rm e}^+{\rm e}^-)$ which are neutral. On the other hand, there is no way of pairing the leptons in (20.49) so as to cancel the lepton number and have non-zero charge. So (20.49) is a *purely* 'neutral current' process, while *some* 'neutral current' contribution could be present in (20.48), in principle. In 1973 such neutral current processes were discovered (Hasert *et al.* 1973), generating a whole new wave of experimental activity. Their existence had, in fact, been *predicted* in the first version of the Standard Model, due to Glashow (1961). Today we know that charged current processes are mediated by the $\rm W^\pm$ bosons, and the neutral current ones by the $\rm Z^0$. We shall discuss the neutral current couplings in section 20.6.

20.5 Calculation of the cross section for $\nu_\mu + {\rm e}^- \to \mu^- + \nu_{\rm e}$

After so much qualitative discussion it is time to calculate something. We choose the process (20.45), sometimes called inverse muon decay, which is a pure 'charged current' process. The amplitude, in the Fermi-like V-A current theory, is

$$\mathcal{M} = -{\rm i}(G_{\rm F}/\sqrt{2})\bar{u}(\mu,k')\gamma_\mu(1-\gamma_5)u(\nu_\mu,k)\bar{u}(\nu_{\rm e},p')\gamma^\mu(1-\gamma_5)u({\rm e},p). \tag{20.50}$$

We shall be interested in energies much greater than any of the leptons, and so we shall work in the *massless limit*; this is mainly for ease of calculation – the full expressions for non-zero masses can be obtained with more effort.

From the general formula (6.129) for $2 \to 2$ scattering in the CM system, we have, neglecting all masses,

$$\frac{{\rm d}\sigma}{{\rm d}\Omega} = \frac{1}{64\pi^2 s}|\overline{\mathcal{M}}|^2 \tag{20.51}$$

where $|\overline{\mathcal{M}}|^2$ is the appropriate spin-averaged matrix element squared, as in

(8.183) for example. In the case of neutrino-electron scattering, we must average over initial electron states for unpolarized electrons and sum over the final muon polarization states. For the neutrinos there is no averaging over initial neutrino helicities, since only left-handed (massless) neutrinos participate in the weak interaction. Similarly, there is no sum over final neutrino helicities. However, for convenience of calculation, we can in fact sum over both helicity states of both neutrinos since the $(1-\gamma_5)$ factors guarantee that right-handed neutrinos contribute nothing to the cross section. As for the eμ scattering example in section 8.7, the calculation then reduces to a product of traces:

$$|\overline{\mathcal{M}}|^2 = \left(\frac{G_F^2}{2}\right) \mathrm{Tr}[\not{k}'\gamma_\mu(1-\gamma_5)\ \not{k}\gamma_\nu(1-\gamma_5)]\frac{1}{2}\mathrm{Tr}[\not{p}'\gamma^\mu(1-\gamma_5)\ \not{p}\gamma^\nu(1-\gamma_5)], \tag{20.52}$$

all lepton masses being neglected. We define

$$|\overline{\mathcal{M}}|^2 = \left(\frac{G_F^2}{2}\right) N_{\mu\nu}E^{\mu\nu} \tag{20.53}$$

where the $\nu_\mu \to \mu^-$ tensor $N_{\mu\nu}$ is given by

$$N_{\mu\nu} = \mathrm{Tr}[\not{k}'\gamma_\mu(1-\gamma_5)\ \not{k}\gamma_\nu(1-\gamma_5)] \tag{20.54}$$

without a $1/(2s+1)$ factor, and the $e^- \to \nu_e$ tensor is

$$E^{\mu\nu} = \frac{1}{2}\mathrm{Tr}[\not{p}'\gamma^\mu(1-\gamma_5)\ \not{p}\gamma^\nu(1-\gamma_5)] \tag{20.55}$$

including a factor of $\frac{1}{2}$ for spin averaging.

Since this calculation involves a couple of new features, let us look at it in some detail. By commuting the $(1-\gamma_5)$ factor through two γ matrices ($\not{p}\gamma^\nu$) and using the result that

$$(1-\gamma_5)^2 = 2(1-\gamma_5) \tag{20.56}$$

the tensor $N_{\mu\nu}$ may be written as

$$\begin{aligned} N_{\mu\nu} &= 2\mathrm{Tr}[\not{k}'\gamma_\mu(1-\gamma_5)\ \not{k}\gamma_\nu] \\ &= 2\mathrm{Tr}(\not{k}'\gamma_\mu\ \not{k}\gamma_\nu) - 2\mathrm{Tr}(\gamma_5\ \not{k}\gamma_\nu\ \not{k}'\gamma_\mu). \end{aligned} \tag{20.57}$$

The first trace is the same as in our calculation of eμ scattering (cf (8.186)):

$$\mathrm{Tr}(\not{k}'\gamma_\mu\ \not{k}\gamma_\nu) = 4[k'_\mu k_\nu + k'_\nu k_\mu + (q^2/2)g_{\mu\nu}]. \tag{20.58}$$

The second trace must be evaluated using the result

$$\mathrm{Tr}(\gamma_5\ \not{a}\ \not{b}\ \not{c}\ \not{d}) = 4\mathrm{i}\epsilon_{\alpha\beta\gamma\delta}a^\alpha b^\beta c^\gamma d^\delta \tag{20.59}$$

(see equation (J.37) in appendix J of volume 1). The totally antisymmetric

tensor $\epsilon_{\alpha\beta\gamma\delta}$ is just the generalization of ϵ_{ijk} to four dimensions, and is defined by

$$\epsilon_{\alpha\beta\gamma\delta} = \begin{cases} +1 & \text{for } \epsilon_{0123} \text{ and all even permutations of } 0,1,2,3 \\ -1 & \text{for } \epsilon_{1023} \text{ and all odd permutations of } 0,1,2,3 \\ 0 & \text{otherwise.} \end{cases} \tag{20.60}$$

Its appearance here is a direct consequence of parity violation. Notice that this definition has the consequence that

$$\epsilon_{0123} = +1 \tag{20.61}$$

but

$$\epsilon^{0123} = -1. \tag{20.62}$$

We will also need to contract two ϵ tensors. By looking at the possible combinations, it should be easy to convince yourself of the result

$$\epsilon_{ijk}\epsilon_{ilm} = \begin{vmatrix} \delta_{jl} & \delta_{jm} \\ \delta_{kl} & \delta_{km} \end{vmatrix} \tag{20.63}$$

i.e.

$$\epsilon_{ijk}\epsilon_{ilm} = \delta_{jl}\delta_{km} - \delta_{kl}\delta_{jm}. \tag{20.64}$$

For the four-dimensional ϵ tensor one can show (see problem 20.6)

$$\epsilon_{\mu\nu\alpha\beta}\epsilon^{\mu\nu\gamma\delta} = -2! \begin{vmatrix} \delta^\gamma_\alpha & \delta^\gamma_\beta \\ \delta^\delta_\alpha & \delta^\delta_\beta \end{vmatrix} \tag{20.65}$$

where the minus sign arises from (20.62) and the 2! from the fact that the two indices are contracted.

We can now evaluate $N_{\mu\nu}$. We obtain, after some rearrangement of indices, the result for the $\nu_\mu \to \mu^-$ tensor:

$$N_{\mu\nu} = 8[(k'_\mu k_\nu + k'_\nu k_\mu + (q^2/2)g_{\mu\nu}) - \mathrm{i}\epsilon_{\mu\nu\alpha\beta}k^\alpha k'^\beta]. \tag{20.66}$$

For the electron tensor $E^{\mu\nu}$ we have a similar result (divided by 2):

$$E^{\mu\nu} = 4[(p'^\mu p^\nu + p'^\nu p^\mu + (q^2/2)g^{\mu\nu}) - \mathrm{i}\epsilon^{\mu\nu\gamma\delta}p_\gamma p'_\delta]. \tag{20.67}$$

Next, we have to perform the contraction $N_{\mu\nu}E^{\mu\nu}$ in (20.53). In the case of elastic $\mathrm{e}^-\mu^-$ scattering considered in section 8.7, the analogous contraction between the tensors $L_{\mu\nu}$ and $M^{\mu\nu}$ was simplified by using the conditions $q^\mu L_{\mu\nu} = q^\nu L_{\mu\nu} = 0$ (see (8.189)), which followed from electromagnetic current conservation at the electron vertex (see (8.188)): $q^\mu \bar{u}(k')\gamma_\mu u(k) = 0$. Here, the analogous vertex is $\bar{u}(\mu, k')\gamma_\mu(1-\gamma_5)u(\nu_\mu, k)$. In this case, when we contract this with $q^\mu = (k-k')^\mu$ we find a non-zero result:

$$(m_{\nu_\mu} - m_\mu)\bar{u}(\mu, k')u(\nu_\mu, k) + (m_\mu + m_{\nu_\mu})\bar{u}(\mu, k')\gamma_5 u(\nu_\mu, k), \tag{20.68}$$

using the on-shell conditions for the spinors. (In the electromagnetic case, there was no γ_5 term, and the intial and final masses were the same.) The quantity (20.68) vanishes only when the lepton masses vanish, and that is the approximation we shall make: i.e. we shall neglect all lepton masses. Then

$$q^\mu N_{\mu\nu} = q^\nu N_{\mu\nu} = 0, \tag{20.69}$$

and we may write

$$p' = p + q \tag{20.70}$$

and drop all terms involving q in the contraction with $N_{\mu\nu}$. In the antisymmetric term, however, we have

$$\epsilon^{\mu\nu\gamma\delta} p_\gamma (p_\delta + q_\delta) = \epsilon^{\mu\nu\gamma\delta} p_\gamma q_\delta \tag{20.71}$$

since the term with p_δ vanishes because of the antisymmetry of $\epsilon_{\mu\nu\gamma\delta}$. Thus we arrive at

$$E^{\mu\nu}_{\mathrm{eff}} = 8p^\mu p^\nu + 2q^2 g^{\mu\nu} - 4\mathrm{i}\epsilon^{\mu\nu\gamma\delta} p_\gamma q_\delta. \tag{20.72}$$

We must now evaluate the '$N \cdot E$' contraction in (20.53). Since we are neglecting all masses, it is easiest to perform the calculation in invariant form before specializing to the 'laboratory' frame. The usual Mandelstam variables are (neglecting all masses)

$$s = 2k \cdot p \tag{20.73}$$

$$u = -2k' \cdot p \tag{20.74}$$

$$t = -2k \cdot k' = q^2 \tag{20.75}$$

satisfying

$$s + t + u = 0. \tag{20.76}$$

The result of performing the contraction

$$N_{\mu\nu} E^{\mu\nu} = N_{\mu\nu} E^{\mu\nu}_{\mathrm{eff}} \tag{20.77}$$

may be found using the result (20.65) for the contraction of two ϵ tensors (see problem 20.6): the answer for $\nu_\mu \mathrm{e}^- \to \mu^- \nu_\mathrm{e}$ is

$$N_{\mu\nu} E^{\mu\nu} = 16(s^2 + u^2) + 16(s^2 - u^2) \tag{20.78}$$

where the first term arises from the symmetric part of $N_{\mu\nu}$ similar to $L_{\mu\nu}$, and the second term from the antisymmetric part involving $\epsilon_{\mu\nu\alpha\beta}$. We have also used

$$t = q^2 = -(s + u) \tag{20.79}$$

valid in the approximation in which we are working. Thus for $\nu_\mu \mathrm{e}^- \to \mu^- \nu_\mathrm{e}$ we have

$$N_{\mu\nu} E^{\mu\nu} = +32s^2 \tag{20.80}$$

and with

$$\frac{\mathrm{d}\sigma}{\mathrm{d}\Omega} = \frac{1}{64\pi^2 s}\left(\frac{G_\mathrm{F}^2}{2}\right) N_{\mu\nu} E^{\mu\nu} \tag{20.81}$$

we finally obtain the result

$$\frac{\mathrm{d}\sigma}{\mathrm{d}\Omega} = \frac{G_\mathrm{F}^2 s}{4\pi^2}. \tag{20.82}$$

The total cross section is then

$$\sigma = \frac{G_\mathrm{F}^2 s}{\pi}. \tag{20.83}$$

Since $t = -2p^2(1-\cos\theta)$, where p is the CM momentum and θ the CM scattering angle, (20.82) can alternatively be written in invariant form as (problem 20.7)

$$\frac{\mathrm{d}\sigma}{\mathrm{d}t} = \frac{G_\mathrm{F}^2}{\pi}. \tag{20.84}$$

All other purely leptonic processes may be calculated in an analogous fashion (see Bailin 1982 and Renton 1990 for further examples).

When we discuss deep inelastic neutrino scattering in section 20.7.2, we shall be interested in neutrino 'laboratory' cross sections, as in the electron scattering case of chapter 9. A simple calculation gives $s \simeq 2m_e E$ (neglecting squares of lepton masses by comparison with $m_e E$), where E is the 'laboratory' energy of a neutrino incident, in this example, on a stationary electron. It follows that *the total 'laboratory' cross section in this Fermi-like current–current model rises linearly with E*. We shall return to the implications of this in section 20.7.2.

The process (20.45) was measured by Bergsma *et al.* (1983) using the CERN wide band beam ($E_\nu \sim 20$ GeV). The ratio of the observed number of events to that expected for pure V-A was quoted as 0.98 ± 0.12.

20.6 Leptonic weak neutral currents

The first observations of the weak neutral current process $\bar{\nu}_\mu \mathrm{e}^- \to \bar{\nu}_\mu \mathrm{e}^-$ were reported by Hasert *et al.* (1973), in a pioneer experiment using the heavy-liquid bubble chamber Gargamelle at CERN, irradiated with a $\bar{\nu}_\mu$ beam. As in the case of the charged currents, much detailed experimental work was necessary to determine the precise form of the neutral current couplings. They are, of course, *predicted* by the Glashow–Salam–Weinberg theory, as we shall explain in chapter 22. For the moment, we continue with the current–current approach, parametrizing the currents in a convenient way.

There are two types of 'neutral current' couplings, those involving neutrinos of the form $\bar{\hat{\nu}}_l \ldots \hat{\nu}_l$, and those involving the charged leptons of the form

$\bar{\hat{l}}\ldots\hat{l}$. We shall assume the following form for these currents (with one eye on the GSW theory to come):

(1) neutrino neutral current

$$g_{\mathrm{N}} c^{\nu_l} \bar{\hat{\nu}}_l \gamma^\mu \left(\frac{1-\gamma_5}{2}\right) \hat{\nu}_l \qquad l = \mathrm{e}, \mu, \tau; \tag{20.85}$$

(2) charged lepton neutral current

$$g_{\mathrm{N}} \bar{\hat{l}} \gamma^\mu \left[c^l_{\mathrm{L}} \frac{(1-\gamma_5)}{2} + c^l_{\mathrm{R}} \frac{(1+\gamma_5)}{2} \right] \hat{l} \qquad l = \mathrm{e}, \mu, \tau. \tag{20.86}$$

This is, of course, by no means the most general possible parametrization. The neutrino coupling is retained as pure 'V-A', while the coupling in the charged lepton sector is now a combination of 'V-A' and 'V+A' with certain coefficients c^l_{L} and c^l_{R}. We may also write the coupling in terms of 'V' and 'A' coefficients defined by $c^l_{\mathrm{V}} = c^l_{\mathrm{L}} + c^l_{\mathrm{R}}, c^l_{\mathrm{A}} = c^l_{\mathrm{L}} - c^l_{\mathrm{R}}$. An overall factor g_{N} determines the strength of the neutral currents as compared to the charged ones; the c's determine the relative amplitudes of the various neutral current processes.

As we shall see, an essential feature of the GSW theory is its prediction of weak neutral current processes, with couplings determined in terms of one parameter of the theory called 'θ_{W}', the 'weak mixing angle' (Glashow 1961, Weinberg 1967). The GSW predictions for the parameter g_{N} and the c's are (see equations (22.59)–(22.62))

$$g_{\mathrm{N}} = g/\cos\theta_{\mathrm{W}}, \quad c^{\nu_l} = \frac{1}{2}, \quad c^l_{\mathrm{L}} = -\frac{1}{2} + a, \quad c^l_{\mathrm{R}} = a \tag{20.87}$$

for $l = \mathrm{e}, \mu, \tau$, where $a = \sin^2\theta_{\mathrm{W}}$ and g is the SU(2) gauge coupling. Note that a strong form of 'universality' is involved here too: the coefficients are independent of the 'flavour' e, μ or τ, for both neutrinos and charged leptons.

The following reactions are available for experimental measurement (in addition to the charged current process (20.45) already discussed):

$$\nu_\mu \mathrm{e}^- \rightarrow \nu_\mu \mathrm{e}^-, \quad \bar{\nu}_\mu \mathrm{e}^- \rightarrow \bar{\nu}_\mu \mathrm{e}^- \quad (\mathrm{NC}) \tag{20.88}$$

$$\nu_{\mathrm{e}} \mathrm{e}^- \rightarrow \nu_{\mathrm{e}} \mathrm{e}^-, \quad \bar{\nu}_{\mathrm{e}} \mathrm{e}^- \rightarrow \bar{\nu}_{\mathrm{e}} \mathrm{e}^- \quad (\mathrm{NC} + \mathrm{CC}) \tag{20.89}$$

where 'NC' means neutral current and 'CC' charged current. Formulae for these cross sections are given in section 22.3. The experiments are discussed and reviewed in Commins and Bucksbaum (1983), Renton (1990), and by Winter (2000). All observations are in excellent agreement with the GSW predictions, with θ_{W} determined as $\sin^2\theta_{\mathrm{W}} \simeq 0.23$. The reader must note, however, that modern precision measurements are sensitive to higher-order (loop) corrections, which must be included in comparing the full GSW theory

with experiment (see section 22.6). The simultaneous fit of data from all four reactions in terms of the single parameter $\theta_{\rm W}$ provides already strong confirmation of the theory – and indeed such confirmation was already emerging in the late 1970's and early 1980's, before the actual discovery of the $\rm W^{\pm}$ and $\rm Z^0$ bosons. It is also interesting to note that the presence of vector (V) interactions in the neutral current processes may suggest the possibility of some kind of link with electromagnetic interactions, which are of course also 'neutral' (in this sense) and vector-like. In the GSW theory, this linkage is provided essentially through the parameter $\theta_{\rm W}$, as we shall see.

20.7 Quark weak currents

We now turn our attention to the weak interactions of quarks. We shall begin by considering an earlier world, when only two generations (four flavours) were known.

20.7.1 Two generations

The original version of V-A theory was framed in terms of a nucleonic current of the form $\bar{\hat{\psi}}_{\rm p}\gamma^\mu(1-r\gamma_5)\hat{\psi}_{\rm n}$. With the acceptance of quark substructure it was natural to re-interpret such a hadronic transition by a charged current of the form $\bar{\hat{u}}\gamma^\mu(1-\gamma_5)\hat{d}$, very similar to the charged lepton currents; indeed, here was a further example of 'universality', this time between quarks and leptons. Detailed comparison with experiment showed, however, that such $\rm d \to u$ transitions were very slightly weaker than the analogous leptonic ones; this could be established by comparing the rates for $\rm n \to p e^- \bar{\nu}_e$ and $\bar{\mu} \to \nu_\mu \rm e^- \bar{\nu}_e$.

But for quarks (or their hadronic composites) there is a further complication, which is the very familiar phenomenon of flavour change in weak hadronic processes (recall the discussion in section 1.2.2). The first step towards the modern theory of quark currents was taken by Cabibbo (1963); in a sense, it restored universality. Cabibbo postulated that the strength of the hadronic weak interaction was *shared* between the $\Delta S = 0$ and $\Delta S = 1$ transitions (where S is the strangeness quantum number), the latter being relatively suppressed as compared to the former. According to Cabibbo's hypothesis, phrased in terms of quarks, the total weak charged current for u, d and s quarks is

$$\hat{j}^\mu_{\rm Cab}({\rm u,d,s}) = \cos\theta_{\rm C}\,\bar{\hat{u}}\gamma^\mu\frac{(1-\gamma_5)}{2}\hat{d} + \sin\theta_{\rm C}\,\bar{\hat{u}}\gamma^\mu\frac{(1-\gamma_5)}{2}\hat{s}, \qquad (20.90)$$

where $\theta_{\rm C}$ is the 'Cabibbo angle' (not to be confused with $\theta_{\rm W}$). We can now postulate a total weak charged current

$$\hat{j}^\mu_{\rm CC}({\rm total}) = \hat{j}^\mu_{\rm CC}({\rm leptons}) + \hat{j}^\mu_{\rm Cab}({\rm u,d,s}), \qquad (20.91)$$

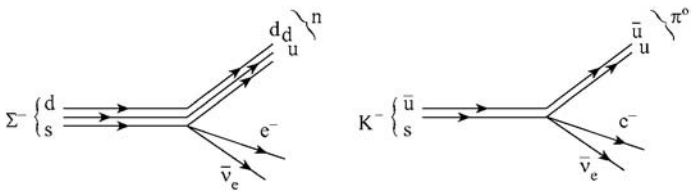


FIGURE 20.3
Strangeness-changing semi-leptonic weak decays.

where $\hat{j}^{\mu}_{\rm CC}$(leptons) is given by (20.38), and then generalize (20.40) to

$$\hat{\mathcal{H}}^{\rm tot}_{\rm CC} = \frac{G_{\rm F}}{\sqrt{2}} \hat{j}^{\mu}_{\rm CC}({\rm total}) \hat{j}^{\dagger}_{{\rm CC}\mu}({\rm total}). \tag{20.92}$$

The effective interaction (20.92) describes a great many processes. The purely leptonic ones discussed previously are, of course, present in the term $\hat{j}^{\mu}_{\rm CC}({\rm leptons})\hat{j}^{\dagger}_{{\rm CC}\mu}({\rm leptons})$. But there are also now all the *semi-leptonic* processes such as the $\Delta S = 0$ (strangeness conserving) one

$$\mathrm{d} \to \mathrm{u} + \mathrm{e}^- + \bar{\nu}_{\rm e}, \tag{20.93}$$

and the $\Delta S = 1$ (strangeness changing) one

$$\mathrm{s} \to \mathrm{u} + \mathrm{e}^- + \bar{\nu}_{\rm e}. \tag{20.94}$$

The notion that the 'total current' should be the sum of a hadronic and a leptonic part is already familiar from electromagnetism – see, for example, equation (8.91).

The transition (20.94), for example, is the underlying process in semi-leptonic decays such as

$$\Sigma^- \to \mathrm{n} + \mathrm{e}^- + \bar{\nu}_{\rm e} \tag{20.95}$$

and

$$\mathrm{K}^- \to \pi^0 + \mathrm{e}^- + \bar{\nu}_{\rm e} \tag{20.96}$$

as indicated in figure 20.3.

The 's' quark is assigned $S = -1$ and charge $-\frac{1}{3}e$. The s $\to$ u transition is then referred to as one with '$\Delta S = \Delta Q$', meaning that the change in the quark (or hadronic) strangeness is equal to the change in the quark (or hadronic) charge: both the strangeness and the charge increase by 1 unit. Prior to the advent of the quark model, and the Cabibbo hypothesis, it had been established empirically that all known strangeness-changing semileptonic decays satisfied the rules $|\Delta S| = 1$ and $\Delta S = \Delta Q$. The u-s current in (20.90) satisfies these rules automatically. Note, for example, that the process apparently similar to (20.95), $\Sigma^+ \to \mathrm{n} + \mathrm{e}^+ + \nu_{\rm e}$, is forbidden in the lowest order (it

requires a double quark transition from suu to udd). All known data on such decays can be fit with a value $\sin\theta_{\rm C} \simeq 0.22$ for the Cabibbo angle $\theta_{\rm C}$. This relatively small angle is therefore a measure of the suppression of $|\Delta S| = 1$ processes relative to $\Delta S = 0$ ones.

The Cabibbo current can be written in a more compact form by introducing the 'mixed' field

$$\hat{d}' \equiv \cos\theta_{\rm C}\hat{d} + \sin\theta_{\rm C}\hat{s}. \tag{20.97}$$

Then

$$\hat{j}^{\mu}_{\rm Cab}({\rm u,d,s}) = \bar{\hat{u}}\gamma^{\mu}\frac{(1-\gamma_5)}{2}\hat{d}'. \tag{20.98}$$

In 1970 Glashow, Iliopuolos and Maiani (GIM) (1970) drew attention to a theoretical problem with the interaction (20.92) if used in *second* order. Now it is, of course, the case that this interaction is not renormalizable, as noted previously for the purely leptonic one (20.40), since $\rm G_F$ has dimensions of an inverse mass squared. As we saw in section 11.7, this means that one-loop diagrams will typically diverge quadratically, so that the contribution of such a second-order process will be of order $(G_{\rm F}.G_{\rm F}\Lambda^2)$ where Λ is a cut-off, compared to the first-order amplitude $G_{\rm F}$. Recalling from (20.46) that $G_{\rm F} \sim 10^{-5}$ GeV^{-2}, we see that for $\Lambda \sim 10$ GeV such a correction could be significant if accurate enough data exists. GIM pointed out, in particular, that some second-order processes could be found which violated the (hitherto) well-established phenomenological selection rules, such as the $|\Delta S| = 1$ and $\Delta S = \Delta Q$ rules already discussed. For example, there could be $\Delta S = 2$ amplitudes contributing to the $\rm K_L - K_S$ mass difference (see Renton 1990, section 9.1.6, for example), as well as contributions to unobserved decay modes such as

$$\rm K^+ \to \pi^+ + \nu + \bar{\nu} \tag{20.99}$$

which has a *neutral* lepton pair in association with a *strangeness change* for the hadron. In fact, experiment placed very tight limits on the rate for (20.99) – and still does: the branching fraction is $(1.7 \pm 1.1) \times 10^{-10}$ (Nakamura *et al.* 2010). This seemed to imply a surprisingly low value of the cut-off, say ~ 3 GeV (Mohapatra *et al.* 1968).

Partly in order to address this problem, and partly as a revival of an earlier lepton-quark symmetry proposal (Bjorken and Glashow 1964), GIM introduced a fourth quark, now called c (the charm quark) with charge $\frac{2}{3}e$. Note that in 1970 the τ-lepton had not been discovered, so only two lepton family pairs $(\nu_{\rm e}, {\rm e})$, (ν_μ, μ) were known; this fourth quark therefore did restore the balance, via the two quark family pairs (u,d), (c,s). In particular, a second quark current could now be hypothesized, involving the (c,s) pair. GIM postulated that the c-quark was coupled to the 'orthogonal' d-s combination (cf (20.97))

$$\hat{s}' = -\sin\theta_{\rm C}\hat{d} + \cos\theta_{\rm C}\hat{s}. \tag{20.100}$$

The complete four-quark charged current is then

$$\hat{j}^{\mu}_{\rm GIM}({\rm u,d,c,s}) = \bar{\hat{u}}\gamma^{\mu}\frac{(1-\gamma_5)}{2}\hat{d}' + \bar{\hat{c}}\gamma^{\mu}\frac{(1-\gamma_5)}{2}\hat{s}'. \tag{20.101}$$

The form (20.101) had already been suggested by Bjorken and Glashow (1964). The new feature of GIM was the observation that, assuming an exact $\mathrm{SU}(4)_{\rm f}$ symmetry for the four quarks (in particular, equal masses), all second-order contributions which could have violated the $|\Delta S| = 1, \Delta S = \Delta Q$ selection rules now vanished. Further, to the extent that the (unknown) mass of the charm quark functioned as an effective cut-off Λ, due to breaking of the $\mathrm{SU}(4)_{\rm f}$ symmetry, they estimated $m_{\rm c}$ to lie in the range 3-4 GeV, from the observed $\mathrm{K_L} - \mathrm{K_S}$ mass difference.

GIM went on to speculate that the non-renormalizability could be overcome if the weak interactions were described by an SU(2) Yang-Mills gauge theory, involving a triplet ($\mathrm{W^+}$, $\mathrm{W^-}$, $\mathrm{W^0}$) of gauge bosons. In this case, it is natural to introduce the idea of (weak) 'isospin', in terms of which the pairs ($\nu_{\rm e}$, e), (ν_μ, μ), (u,d′), (c, s′) are all $t = \frac{1}{2}$ doublets with $t_3 = \pm\frac{1}{2}$. Charge-changing currents then involve the 'raising' matrix

$$\tau_+ \equiv \frac{1}{2}(\tau_1 + {\rm i}\tau_2) = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \tag{20.102}$$

and charge-lowering ones the matrix $\tau_- = (\tau_1 - {\rm i}\tau_2)/2$. The full symmetry must also involve the matrix τ_3, given by the commutator $[\tau_+, \tau_-] = \tau_3$. Whereas τ_+ and τ_- would (in this model) be associated with transitions mediated by $\mathrm{W^\pm}$, transitions involving τ_3 would be mediated by $\mathrm{W^0}$, and would correspond to 'neutral current' transitions for quarks. We now know that things are slightly more complicated than this: the correct symmetry is the $\mathrm{SU}(2) \times \mathrm{U}(1)$ of Glashow (1961), also invoked by GIM. Skipping therefore some historical steps, we parametrize the *weak quark neutral current* as (cf (20.86) for the leptonic analogue)

$$g_{\rm N} \sum_{q={\rm u,c,d',s'}} \bar{\hat{q}}\gamma^{\mu}[c^q_{\rm L}\frac{(1-\gamma_5)}{2} + c^q_{\rm R}\frac{(1+\gamma_5)}{2}]\hat{q} \tag{20.103}$$

for the four flavours so far in play. In the GSW theory, the $c^q_{\rm L}$'s are predicted to be

$$c^{\rm u,c}_{\rm L} = \frac{1}{2} - \frac{2}{3}a \qquad c^{\rm u,c}_{\rm R} = -\frac{2}{3}a \tag{20.104}$$

$$c^{\rm d,s}_{\rm L} = -\frac{1}{2} + \frac{1}{3}a \qquad c^{\rm d,s}_{\rm R} = \frac{1}{3}a \tag{20.105}$$

where $a = \sin^2\theta_{\rm W}$ as before, and $g_{\rm N} = g/\cos\theta_{\rm W}$.

One feature of (20.103) is very important. Consider the terms

$$\bar{\hat{d}}'\{\ldots\}\hat{d}' + \bar{\hat{s}}'\{\ldots\}\hat{s}'. \tag{20.106}$$

It is simple to verify that, whereas either part of (20.106) alone contains a *strangeness changing neutral* combination such as $\bar{\hat{d}}\{\ldots\}\hat{s}$ or $\bar{\hat{s}}\{\ldots\}\hat{d}$, such combinations vanish in the sum, leaving the result *diagonal in quark flavour*. Thus there are no first-order neutral flavour-changing currents in this model, a result which will be extended to three flavours in section 20.7.3.

In 1974, Gaillard and Lee (1974) performed a full one-loop calculation of the $\mathrm{K_L} - \mathrm{K_S}$ mass difference in the GSW model as extended by GIM to quarks and using the renormalization techniques recently developed by 't Hooft (1971b). They were able to predict $m_\mathrm{c} \sim 1.5$ GeV for the charm quark mass, a result spectacularly confirmed by the subsequent discovery of the $\mathrm{c\bar{c}}$ states in charmonium, and of charmed mesons and baryons of the appropriate mass.

In summary, then, the essential feature of the quark weak currents in the two-generation model is that they have the universal V-A form, but the participating fields are $(\hat{u}, \hat{d}'), (\hat{c}, \hat{s}')$ where $\hat{d}'$ and $\hat{s}'$ are not the fields $\hat{d}, \hat{s}$ with definite mass, but rather are related to them by an orthogonal transformation:

$$\begin{pmatrix} \hat{d}' \\ \hat{s}' \end{pmatrix} = \begin{pmatrix} \cos\theta_\mathrm{C} & \sin\theta_\mathrm{C} \\ -\sin\theta_\mathrm{C} & \cos\theta_\mathrm{C} \end{pmatrix} \begin{pmatrix} \hat{d} \\ \hat{s} \end{pmatrix}. \tag{20.107}$$

In section 20.8 we shall enlarge this picture to three generations, where significant new features occur, specifically **CP** violation. In chapter 22 we shall see how this transformation from the 'mass' basis to the 'weak interaction' basis arises via the gauge-invariant interactions of the Standard Model.

20.7.2 Deep inelastic neutrino scattering

We now have enough theory to present another illustrative calculation within the framework of the 'current–current' model, this time involving neutrinos and quarks. We shall calculate cross sections for deep inelastic neutrino scattering from nucleons, using the parton model introduced (for electromagnetic interactions) in chapter 9. In particular, we shall consider the processes

$$\nu_\mu + \mathrm{N} \quad \rightarrow \quad \mu^- + \mathrm{X} \tag{20.108}$$

$$\bar{\nu}_\mu + \mathrm{N} \quad \rightarrow \quad \mu^+ + \mathrm{X} \tag{20.109}$$

which of course involve the charged currents, for both leptons and quarks. Studies of these reactions at Fermilab and CERN in the 1970s and 1980s played a crucial part in establishing the quark structure of the nucleon, in particular the quark distribution functions.

The general process is illustrated in figure 20.4. By now we are becoming accustomed to the idea that such processes are in fact mediated by the W^+, but we shall assume that the momentum transfers are such that the W-propagator is effectively constant. The effective lepton-quark interaction will then take the form

$$\hat{\mathcal{H}}^{\mathrm{eff}}_{\nu\mathrm{q}} = \frac{G_\mathrm{F}}{\sqrt{2}} \bar{\hat{\mu}} \gamma_\mu (1-\gamma_5) \hat{\nu}_\mu [\bar{\hat{u}} \gamma^\mu (1-\gamma_5) \hat{d} + \bar{\hat{c}} \gamma^\mu (1-\gamma_5) \hat{s}], \tag{20.110}$$

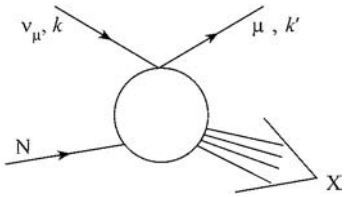


FIGURE 20.4
Inelastic neutrino scattering from a nucleon.

leading to expressions for the parton-level subprocess amplitudes which are exactly similar to that in (20.50) for $\nu_\mu + e^- \to \mu^- + \nu_e$. Note that we are considering only the four flavours u, d, c, s to be 'active', and we have set $\theta_C \approx 0$.

As in (20.53), the ν_μ cross section will have the general form

$$d\sigma^{(\nu)} \propto N_{\mu\nu} W^{\mu\nu}_{(\nu)}(q,p) \tag{20.111}$$

where $N_{\mu\nu}$ is the neutrino tensor of (20.67). The form of the weak hadron tensor $W^{\mu\nu}_{(\nu)}$ is deduced from Lorentz invariance. In the approximation of neglecting lepton masses, we can ignore any dependence on the 4-vector q since

$$q^\mu N_{\mu\nu} = q^\nu N_{\mu\nu} = 0. \tag{20.112}$$

Just as $N_{\mu\nu}$ contains the pseudotensor $\epsilon_{\mu\nu\alpha\beta}$ so too will $W^{\mu\nu}_{(\nu)}$ since parity is not conserved. In a manner similar to equation (9.10) for the case of electron scattering, and following the steps that led from (20.67) to (20.72), we define effective neutrino structure functions by

$$W^{\mu\nu}_{(\nu)} = (-g^{\mu\nu})W_1^{(\nu)} + \frac{1}{M^2} p^\mu p^\nu W_2^{(\nu)} - \frac{\mathrm{i}}{2M^2} \epsilon^{\mu\nu\gamma\delta} p_\gamma q_\delta W_3^{(\nu)}. \tag{20.113}$$

In general, the structure functions depend on two variables, say Q^2 and ν, where $Q^2 = -(k-k')^2$ and $\nu = p\cdot q/M$; but in the Bjorken limit approximate scaling is observed, as in the electron case:

$$\left.\begin{array}{c} Q^2 \to \infty \\ \nu \to \infty \end{array}\right\} \qquad x = Q^2/2M\nu \ \text{ fixed} \tag{20.114}$$

$$\nu W_2^{(\nu)}(Q^2,\nu) \ \to \ F_2^{(\nu)}(x) \tag{20.115}$$

$$M W_1^{(\nu)}(Q^2,\nu) \ \to \ F_1^{(\nu)}(x) \tag{20.116}$$

$$\nu W_3^{(\nu)}(Q^2,\nu) \ \to \ F_3^{(\nu)}(x) \tag{20.117}$$

where, as with (9.21) and (9.22), the physics lies in the assertion that the F's

are finite. This scaling can again be interpreted in terms of pointlike scattering from partons – which we shall take to have quark quantum numbers.

In the 'laboratory' frame (in which the nucleon is at rest) the cross section in terms of W_1, W_2 and W_3 may be derived in the usual way from (cf equation (9.11))

$$\mathrm{d}\sigma^{(\nu)} = \left(\frac{G_\mathrm{F}}{\sqrt{2}}\right)^2 \frac{1}{4k\cdot p} 4\pi M N_{\mu\nu} W^{\mu\nu}_{(\nu)} \frac{\mathrm{d}^3\boldsymbol{k}'}{2k'(2\pi)^3}. \tag{20.118}$$

In terms of 'laboratory' variables, one obtains (problem 20.9)

$$\frac{\mathrm{d}^2\sigma^{(\nu)}}{\mathrm{d}Q^2\mathrm{d}\nu} = \frac{G_\mathrm{F}^2}{2\pi}\frac{k'}{k}\left(W_2^{(\nu)}\cos^2(\theta/2) + W_1^{(\nu)} 2\sin^2(\theta/2) + \frac{k+k'}{M}\sin^2(\theta/2) W_3^{(\nu)}\right). \tag{20.119}$$

For an incoming antineutrino beam, the W_3 term changes sign.

In neutrino scattering it is common to use the variables x, ν and the 'inelasticity' y where

$$y = p\cdot q / p\cdot k. \tag{20.120}$$

In the 'laboratory' frame, $\nu = E - E'$ (the energy transfer to the nucleon) and $y = \nu/E$. The cross section can be written in the form (see problem 20.9)

$$\frac{\mathrm{d}^2\sigma^{(\nu)}}{\mathrm{d}x\mathrm{d}y} = \frac{G_\mathrm{F}^2}{2\pi} s\left(F_2^{(\nu)}\frac{1+(1-y)^2}{2} + xF_3^{(\nu)}\frac{1-(1-y)^2}{2}\right) \tag{20.121}$$

in terms of the Bjorken scaling functions, and we have assumed the relation

$$2xF_1^{(\nu)} = F_2^{(\nu)} \tag{20.122}$$

appropriate for spin$-\frac{1}{2}$ constituents.

We now turn to the parton-level subprocesses. Their cross sections can be straightforwardly calculated in the same way as for $\nu_\mu \mathrm{e}^-$ scattering in section 20.5. We obtain (problem 20.10)

$$\nu\mathrm{q}, \bar\nu\bar{\mathrm{q}} : \frac{\mathrm{d}^2\sigma}{\mathrm{d}x\mathrm{d}y} = \frac{G_\mathrm{F}^2}{\pi} sx\delta\left(x - \frac{Q^2}{2M\nu}\right) \tag{20.123}$$

$$\nu\bar{\mathrm{q}}, \bar\nu\mathrm{q} : \frac{\mathrm{d}^2\sigma}{\mathrm{d}x\mathrm{d}y} = \frac{G_\mathrm{F}^2}{\pi} sx(1-y)^2\delta\left(x - \frac{Q^2}{2M\nu}\right). \tag{20.124}$$

The factor $(1-y)^2$ in the $\nu\bar{\mathrm{q}}, \bar\nu\mathrm{q}$ cases means that the reaction is forbidden at $y = 1$ (backwards in the CM frame). This follows from the V-A nature of the current, and angular momentum conservation, as a simple helicity argument shows. Consider for example the case $\nu\bar{\mathrm{q}}$ shown in figure 20.5, with the helicities marked as shown. In our current–current interaction there are no gradient coupling terms and therefore no momenta in the momentum-space matrix element. This means that no orbital angular momentum is available to account for the reversal of net helicity in the initial and final states in figure 20.5. The lack of orbital angular momentum can also be inferred physically

FIGURE 20.5
Suppression of $\nu_\mu \bar{\mathrm{q}} \to \mu^- \bar{\mathrm{q}}$ for $y = 1$: (a) initial state helicities; (b) final state helicities at $y = 1$.

from the 'pointlike' nature of the current–current coupling. For the νq or $\bar{\nu}\bar{\mathrm{q}}$ cases, the initial and final helicities add to zero, and backward scattering is allowed.

The contributing processes are

$$\nu \mathrm{d} \to l^- \mathrm{u}, \qquad \bar{\nu}\bar{\mathrm{d}} \to l^+ \bar{\mathrm{u}} \tag{20.125}$$

$$\nu \bar{\mathrm{u}} \to l^- \bar{\mathrm{d}}, \qquad \bar{\nu}\mathrm{u} \to l^+ \mathrm{d}, \tag{20.126}$$

the first pair having the cross section (20.123), the second (20.124). Following the same steps as in the electron scattering case (sections 9.2 and 9.3) we obtain

$$F_2^{\nu \mathrm{p}} = F_2^{\bar{\nu}\mathrm{n}} = 2x[d(x) + \bar{u}(x)] \tag{20.127}$$

$$F_3^{\nu \mathrm{p}} = F_3^{\bar{\nu}\mathrm{n}} = 2[d(x) - \bar{u}(x)] \tag{20.128}$$

$$F_2^{\nu \mathrm{n}} = F_2^{\bar{\nu}\mathrm{p}} = 2x[u(x) + \bar{d}(x)] \tag{20.129}$$

$$F_3^{\nu \mathrm{n}} = F_3^{\bar{\nu}\mathrm{p}} = 2[u(x) - \bar{d}(x)]. \tag{20.130}$$

Inserting (20.127) and (20.128) into (20.121), for example, we find

$$\frac{\mathrm{d}^2\sigma^{(\nu \mathrm{p})}}{\mathrm{d}x\mathrm{d}y} = 2\sigma_0 x[d(x) + (1-y)^2 \bar{u}(x)] \tag{20.131}$$

where

$$\sigma_0 = \frac{G_\mathrm{F}^2 s}{2\pi} = \frac{G_\mathrm{F}^2 ME}{\pi} \simeq 1.5 \times 10^{-42} (E/\mathrm{GeV}) \mathrm{m}^2 \tag{20.132}$$

is the basic 'pointlike' total cross section (compare (20.83)). Note the small magnitude of this cross section, as compared with the electromagnetic one of equation (B.18) in volume 1, which was $\sigma \approx \frac{86.8}{(s/\mathrm{GeV}^2)} \times 10^{-37} \mathrm{m}^2$. Similarly, one finds

$$\frac{\mathrm{d}^2\sigma^{(\bar{\nu}\mathrm{p})}}{\mathrm{d}x\mathrm{d}y} = 2\sigma_0 x[(1-y)^2 u(x) + \bar{d}(x)]. \tag{20.133}$$

The corresponding results for νn and $\bar{\nu}$n are given by interchanging $u(x)$ and $d(x)$, and $\bar{u}(x)$ and $\bar{d}(x)$.

The target nuclei usually have high mass number (in order to increase the cross section), with approximately equal numbers of protons and neutrons; it

is then appropriate to average the 'n' and 'p' results to obtain an 'isoscalar' cross section $\sigma^{(\nu\mathrm{N})}$ or $\sigma^{(\bar{\nu}\mathrm{N})}$:

$$\frac{\mathrm{d}^2\sigma^{(\nu\mathrm{N})}}{\mathrm{d}x\mathrm{d}y} = \sigma_0 x[q(x) + (1-y)^2\bar{q}(x)] \tag{20.134}$$

$$\frac{\mathrm{d}^2\sigma^{(\bar{\nu}\mathrm{N})}}{\mathrm{d}x\mathrm{d}y} = \sigma_0 x[(1-y)^2 q(x) + \bar{q}(x)] \tag{20.135}$$

where $q(x) = u(x) + d(x)$ and $\bar{q}(x) = \bar{u}(x) + \bar{d}(x)$.

Many simple and striking predictions now follow from these quark parton results. For example, by integrating (20.134) and (20.135) over x we can write

$$\frac{\mathrm{d}\sigma^{(\nu\mathrm{N})}}{\mathrm{d}y} = \sigma_0[Q + (1-y)^2\bar{Q}] \tag{20.136}$$

$$\frac{\mathrm{d}\sigma^{(\bar{\nu}\mathrm{N})}}{\mathrm{d}y} = \sigma_0[(1-y)^2 Q + \bar{Q}] \tag{20.137}$$

where $Q = \int xq(x)\mathrm{d}x$ is the fraction of the nucleon's momentum carried by quarks, and similarly for $\bar{Q}$. These two distributions in y ('inelasticity distributions') therefore give a direct measure of the quark and antiquark composition of the nucleon. Figure 20.6 shows the inelasticity distributions as reported by the CDHS collaboration (de Groot *et al.* 1979), from which the authors extracted the ratio

$$\bar{Q}/(Q + \bar{Q}) = 0.15 \pm 0.03 \tag{20.138}$$

after applying radiative corrections. An even more precise value can be obtained by looking at the region near $y = 1$ for $\bar{\nu}$N which is dominated by $\bar{Q}$, the small Q contribution ($\propto (1-y)^2$) being subtracted out using νN data at the same y. This method yields

$$\bar{Q}/(Q + \bar{Q}) = 0.15 \pm 0.01. \tag{20.139}$$

Integrating (20.136) and (20.137) over y gives

$$\sigma^{(\nu\mathrm{N})} = \sigma_0(Q + \frac{1}{3}\bar{Q}) \tag{20.140}$$

$$\sigma^{(\bar{\nu}\mathrm{N})} = \sigma_0(\frac{1}{3}Q + \bar{Q}) \tag{20.141}$$

and hence

$$Q + \bar{Q} = 3(\sigma^{(\nu\mathrm{N})} + \sigma^{(\bar{\nu}\mathrm{N})})/4\sigma_0 \tag{20.142}$$

while

$$\bar{Q}/(Q + \bar{Q}) = \frac{1}{2}\left(\frac{3r-1}{1+r}\right) \tag{20.143}$$

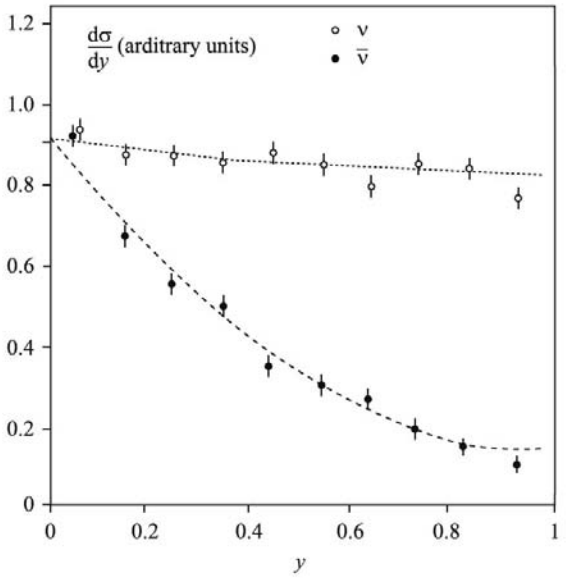


FIGURE 20.6
Charged-current inelasticity (y) distribution as measured by CDHS; figure from K Winter (2000) *Neutrino Physics* 2nd edn, courtesy Cambridge University Press.

where $r = \sigma^{(\nu \mathrm{N})}/\sigma^{(\bar{\nu} \mathrm{N})}$. From total cross section measurements, and including c and s contributions, the CHARM collaboration (Allaby *et al.* 1988) reported

$$Q + \bar{Q} = 0.492 \pm 0.006(\mathrm{stat}) \pm 0.019(\mathrm{syst}) \quad (20.144)$$

$$\bar{Q}/Q + \bar{Q} = 0.154 \pm 0.005(\mathrm{stat}) \pm 0.011(\mathrm{syst}). \quad (20.145)$$

The second figure is in good agreement with (20.139), and the first shows that only about 50% of the nucleon momentum is carried by charged partons, the rest being carried by the gluons, which do not have weak or electromagnetic interactions.

Equations (20.140) and (20.141), together with (20.132), predict that the total cross sections $\sigma^{\nu \mathrm{N}}$ and $\sigma^{\bar{\nu} \mathrm{N}}$ rise linearly with energy E. This (parton model) prediction was confirmed as early as 1975 (Perkins 1975), soon after the model's success in deep inelastic electron scattering; later data is included in figure 20.7. In fact, both $\sigma^{\nu \mathrm{N}}/E$ and $\sigma^{\bar{\nu} \mathrm{N}}/E$ are found to be independent of E up to $E \sim 350$ GeV (Nakamura *et al.* 2010).

Detailed comparison between the data at high energies and the earlier data of figure 20.7 at E_ν up to 15 GeV reveals that the $\bar{Q}$ fraction is increasing with energy. This is in accordance with the expectation of QCD corrections to the parton model (section 15.6): the $\bar{Q}$ distribution is large at small x,

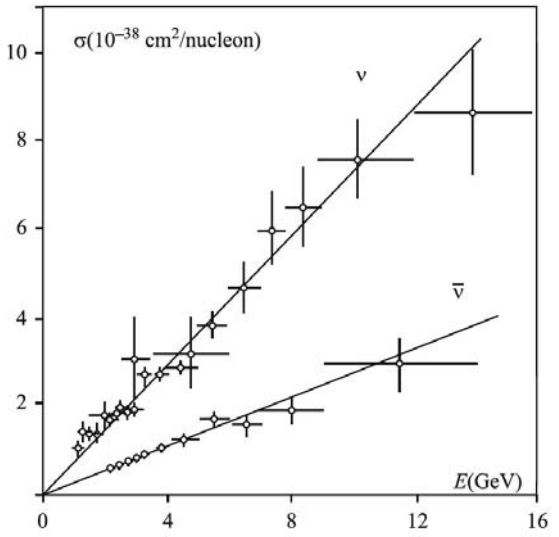


FIGURE 20.7
Low-energy ν and $\bar{\nu}$ cross-sections; figure from K Winter (2000) *Neutrino Physics* 2nd edn, courtesy Cambridge University Press.

and scaling violations embodied in the evolution of the parton distributions predict a rise at small x as the energy scale increases.

Returning now to (20.127)–(20.130), the two sum rules of (9.65) and (9.66) can be combined to give

$$3 = \int_0^1 \mathrm{d}x[u(x) + d(x) - \bar{u}(x) - \bar{d}(x)] \tag{20.146}$$

$$= \frac{1}{2}\int_0^1 \mathrm{d}x(F_3^{\nu\mathrm{p}} + F_3^{\nu\mathrm{n}}) \tag{20.147}$$

$$\equiv \int_0^1 \mathrm{d}x F_3^{\nu\mathrm{N}} \tag{20.148}$$

which is the Gross–Llewellyn Smith sum rule (1969), expressing the fact that the number of valence quarks per nucleon is three. The CDHS collaboration (de Groot *et al.* 1979), quoted

$$I_{\mathrm{GLLS}} \equiv \int_0^1 \mathrm{d}x F_3^{\nu\mathrm{N}} = 3.2 \pm 0.5. \tag{20.149}$$

In perturbative QCD there are corrections expressible as a power series in α_s, so that the parton model result is only reached as $Q^2 \to \infty$:

$$I_{\mathrm{GLLS}}(Q^2) = 3[1 + d_1\alpha_\mathrm{s}/\pi + d_2\alpha_\mathrm{s}^2/\pi^2 + \ldots] \tag{20.150}$$

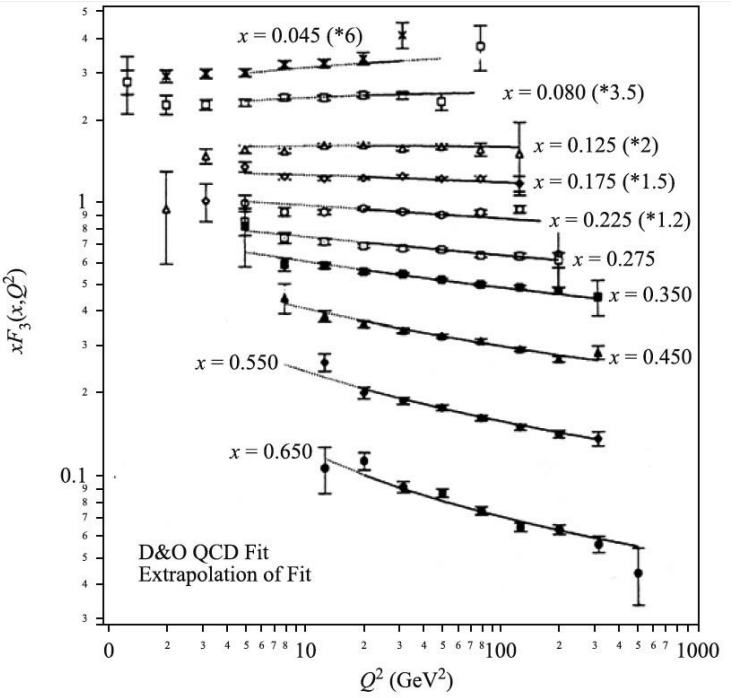


FIGURE 20.8
CCFR neutrino-iron structure functions $xF_3^{(\nu)}$ (Shaevitz *et al.* 1995). The solid line is the next-to-leading order (one-loop) QCD prediction, and the dotted line is an extrapolation to regions outside the kinematic cuts for the fit.

where $d_1 = -1$ (Altarelli *et al.* 1978a, 1978b), $d_2 = -55/12 + N_\mathrm{f}/3$ (Gorishnii and Larin 1986) where N_f is the number of active flavours. The CCFR collaboration (Shaevitz *et al.* 1995) measured I_GLLS in antineutrino-nucleon scattering at $\langle Q^2 \rangle \sim 3\mathrm{GeV}^2$. They obtained

$$I_\mathrm{GLLS}(\langle Q^2 \rangle = 3\ \mathrm{GeV}^2) = 2.50 \pm 0.02 \pm 0.08 \tag{20.151}$$

in agreement with the $\mathrm{O}(\alpha_\mathrm{s}^3)$ calculation of Larin and Vermaseren (1991) using $\Lambda_{\overline{MS}} = 250 \pm 50\mathrm{MeV}$.

The predicted Q^2 evolution of xF_3 is particularly simple since it is not coupled to the gluon distribution. To leading order, the xF_3 evolution is given by (cf (15.109))

$$\frac{\mathrm{d}}{\mathrm{d}\ln Q^2}(xF_3(x,Q^2)) = \frac{\alpha_\mathrm{s}(Q^2)}{2\pi}\int_x^1 P_\mathrm{qq}(z) xF_3\left(\frac{x}{z}, Q^2\right)\frac{\mathrm{d}z}{z}. \tag{20.152}$$

Figure 20.8, taken from Shaevitz *et al.* (1995) shows a comparison of the

CCFR data with the next-to-leading order calculation of Duke and Owens (1984). This fit yields a value of α_s at $Q^2 = M_Z^2$ given by

$$\alpha_s(M_Z^2) = 0.111 \pm 0.002 \pm 0.003. \tag{20.153}$$

The Adler sum rule (Adler 1963) involves the functions $F_2^{\bar{\nu}p}$ and $F_2^{\nu p}$:

$$I_A = \int_0^1 \frac{dx}{x}(F_2^{\bar{\nu}p} - F_2^{\nu p}). \tag{20.154}$$

In the simple model of (20.127)–(20.130), the right-hand side of I_A is just

$$2\int_0^1 dx(u(x) + \bar{d}(x) - d(x) - \bar{u}(x)) \tag{20.155}$$

which represents four times the average of I_3 (isospin) of the target, which is $\frac{1}{2}$ for the proton. This sum rule follows from the conservation of the charged weak current (as will be true in the Standard Model, since this is a gauge symmetry current, as we shall see in the following chapter). Its measurement, however, depends precisely on separating the non-isoscalar contribution (I_A vanishes for the isoscalar average 'N'). The BEBC collaboration (Allasia *et al.* 1984, 1985) reported:

$$I_A = 2.02 \pm 0.40; \tag{20.156}$$

in agreement with the expected value 2.

Relations (20.127)–(20.130) allow the F_2 functions for electron (muon) and neutrino scattering to be simply related. From (9.58) and (9.61) we have

$$F_2^{eN} = \frac{1}{2}(F_2^{ep} + F_2^{en}) = \frac{5}{18}x(u + \bar{u} + d + \bar{d}) + \frac{1}{9}x(s + \bar{s}) + \cdots \tag{20.157}$$

while (20.127) and (20.129) give

$$F_2^{\nu N} \equiv \frac{1}{2}(F_2^{\nu p} + F_2^{\nu n}) = x(u + d + \bar{u} + \bar{d}). \tag{20.158}$$

Assuming that the non-strange contributions dominate, the neutrino and charged lepton structure functions should be approximately in the ratio 18/5, which is the reciprocal of the mean squared charged of the u and d quarks in the nucleon. Figure 20.9 shows the neutrino results on F_2 and xF_3 together with those from several μN experiments scaled by the factor 18/5. The agreement is satisfactory for a tree-level parton model calculation.

From (20.127)–(20.130) we see that $F_2^{\nu N} - xF_3^{\nu N} = 2x(\bar{u} + \bar{d})$, which is just the sea distribution; figure 20.9 shows that this is concentrated at small x, as we already inferred in section 9.3.

We have mentioned QCD corrections to the simple parton model at several points. Clearly the full machinery introduced in chapter 16, in the context of deep inelastic charged lepton scattering, can be employed for the case of neutrino scattering also. For further access to this area we refer to Ellis *et al.* (1996), chapter 4, and Winter (2000) chapter 5.

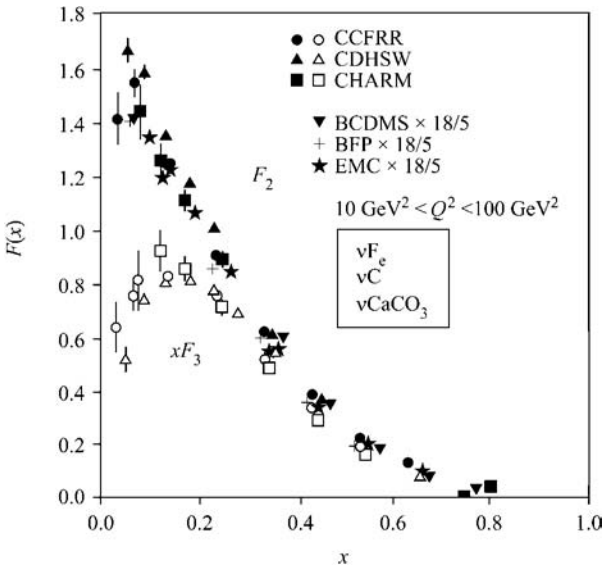


FIGURE 20.9
Comparison of neutrino results (experiments CCFRR, CDHSW and CHARM) on $F_2(x)$ and $xF_3(x)$ with those from muon production (experiments BCDMS, BFP and EMC) properly rescaled by the factor 18/5, for a Q^2 ranging between 10 and 1000 GeV^2; figure from K Winter (2000) *Neutrino Physics* 2nd edn, courtesy Cambridge University Press.

20.7.3 Three generations

We have seen in section 20.2.2 that the V-A interaction violates both **P** and **C**, and that it conserves **CP** in interactions with massless neutrinos. But we know (section 4.2.3) that **CP**-violating transitions occur, among states formed from quarks in the first two generations, albeit at a very slow rate. Is it possible, in fact, to incorporate **CP**-violation with only two generations of quarks?

To answer this question, we need to go back and examine the **CP**-transformation properties of the interactions in more detail. Rather than work with the current–current form, which is after all only an approximation valid for energies much less than $M_{\mathrm{W,Z}}$, we shall look at the actual gauge interactions of the electroweak theory. Given the form of those interactions, we want to know the condition for **CP**-violation to be present.

Consider then the particular interaction involved in u $\leftrightarrow$ d transitions:

$$V_{\mathrm{ud}}\bar{\hat{u}}\gamma^\mu\hat{d}\,\hat{W}_\mu + V_{\mathrm{ud}}^*\bar{\hat{d}}\gamma^\mu\hat{u}\,\hat{W}_\mu^\dagger - V_{\mathrm{ud}}\bar{\hat{u}}\gamma^\mu\gamma_5\hat{d}\,\hat{W}_\mu - V_{\mathrm{ud}}^*\bar{\hat{d}}\gamma^\mu\gamma_5\hat{u}\,\hat{W}_\mu^\dagger, \quad (20.159)$$

where $\hat{W}_\mu = (\hat{W}_\mu^1 - \mathrm{i}\hat{W}_\mu^2)/\sqrt{2}$ destroys the W^+ or creates the W^-. We have

written out the Hermitian conjugate terms explicitly, keeping the coupling $V_{\rm ud}$ complex for the sake of generality, and separating the vector from the axial vector parts. Problem 20.11 shows that the different parts of (20.159) transform under **C** as follows (normal ordering being understood in all cases):

$$\mathbf{C}: \bar{\hat{u}}\gamma^\mu\hat{d} \to -\bar{\hat{d}}\gamma^\mu\hat{u}, \quad \bar{\hat{u}}\gamma^\mu\gamma_5\hat{d} \to +\bar{\hat{d}}\gamma^\mu\gamma_5\hat{u}, \tag{20.160}$$

and we also know that under **C**, $\hat{W}_\mu \to -\hat{W}_\mu^\dagger$ (the dagger is as in the charged scalar field case, and the minus sign is as in the photon $\hat{A}_\mu$ case). Hence under **C**, (20.159) transforms into

$$V_{\rm ud}\bar{\hat{d}}\gamma^\mu\hat{u}\,\hat{W}_\mu^\dagger + V_{\rm ud}^*\bar{\hat{u}}\gamma^\mu\hat{d}\,\hat{W}_\mu + V_{\rm ud}\bar{\hat{d}}\gamma^\mu\gamma_5\hat{u}\,\hat{W}_\mu^\dagger + V_{\rm ud}^*\bar{\hat{u}}\gamma^\mu\gamma_5\hat{d}\,\hat{W}_\mu. \tag{20.161}$$

Under **P**, $\hat{W}_\mu$ behaves like an ordinary four-vector, so the 'vector·vector' products in (20.161) are even under **P**, while the 'axial vector·vector' products are odd under **P**. Thus finally, under the combined **CP** transformation (20.159) becomes

$$V_{\rm ud}\bar{\hat{d}}\gamma^\mu\hat{u}\,\hat{W}_\mu^\dagger + V_{\rm ud}^*\bar{\hat{u}}\gamma^\mu\hat{d}\,\hat{W}_\mu - V_{\rm ud}\bar{\hat{d}}\gamma^\mu\gamma_5\hat{u}\,\hat{W}_\mu^\dagger - V_{\rm ud}^*\bar{\hat{u}}\gamma^\mu\gamma_5\hat{d}\,\hat{W}_\mu. \tag{20.162}$$

Comparing (20.159) with (20.162) we deduce the essential result that this interaction conserves **CP** if and only if

$$V_{\rm ud} = V_{\rm ud}^*, \tag{20.163}$$

that is, if the coupling is real. The same is true for all the other couplings $V_{\rm ij}$.

The couplings we have introduced in this chapter so far only involve the real Fermi constant $G_{\rm F}$, and the elements of the Cabibbo-GIM matrix which enters into the relation between the weakly interacting fields $(\hat{d}', \hat{s}')$ and the fields with definite mass $(\hat{d}, \hat{s})$:

$$\begin{pmatrix} \hat{d}' \\ \hat{s}' \end{pmatrix} = \begin{pmatrix} \cos\theta_{\rm C} & \sin\theta_{\rm C} \\ -\sin\theta_{\rm C} & \cos\theta_{\rm C} \end{pmatrix} \begin{pmatrix} \hat{d} \\ \hat{s} \end{pmatrix} \equiv \mathbf{V}_{\rm CGIM} \begin{pmatrix} \hat{d} \\ \hat{s} \end{pmatrix}. \tag{20.164}$$

All these couplings are plainly real. But could we perhaps parametrize the $(\hat{d}', \hat{s}') \leftrightarrow (\hat{d}, \hat{s})$ differently, so as to smuggle in a complex, **CP**-violating, coupling?

This is the question that Kobayashi and Maskawa asked themselves in 1972 (Kobayashi 2009, Maskawa 2009). To answer it is not completely straightforward, because we can always change the phases of the quark fields by independent constant amounts. A rephasing of the quark fields in the transition $\mathrm{i} \leftrightarrow \mathrm{j}$ with coupling $V_{\rm ij}$ changes $V_{\rm ij}$ by the phase factor $\exp \mathrm{i}(\alpha_{\rm i} - \alpha_{\rm j})$. We need to know whether, after allowing for this rephasing of the quark fields, an 'irreducible' complex coupling can remain.

First of all, note that the matrix $\mathbf{V}_{\rm CGIM}$ appearing in (20.164) is orthogonal, and this property guaranteed the vanishing of tree-level neutral

strangeness-changing transitions, as we saw after (20.106). But this could just as well be achieved if the matrix was unitary. Now a general 2×2 matrix has 8 real parameters; unitarity gives 2 real conditions from the diagonal elements of $\mathbf{V}_{\rm CGIM}\mathbf{V}_{\rm CGIM}^{\dagger} = \mathbf{I}$, and one complex condition from the off-diagonal elements, leaving four real parameters. If all the elements are taken to be real from the beginning, the matrix becomes orthogonal, as in (20.164), and depends on only one real parameter, the 'rotation' in the 2-dimensional $\hat{d} - \hat{s}$ space. So in the general, unitary case, the matrix will have one real angle parameter, and three phase parameters. But we have four quark fields whose phases we can adjust. In fact, since only phase differences enter, we really only have three free phases at our disposal, but that is just enough to transform away the three phases in the unitary version of $\mathbf{V}_{\rm CGIM}$, leaving it in the real orthogonal form (20.164). Kobayashi and Maskawa therefore concluded that the 2-generation GIM-type theory could not accommodate **CP**-violation.

In a step which may seem natural now but was very bold in 1972, they decided to see if there was room for **CP**-violation in a 3-generation model. (Remember that there was no sign of any third generation particles at that time.) The matrix transforming from the mass basis to the weak basis is now a 3×3 unitary matrix $\mathbf{V}$, with 18 real parameters. There are three real diagonal conditions from unitarity, and three complex off-diagonal conditions, leaving 9 real parameters. If the elements of $\mathbf{V}$ are taken to be real, one has an orthogonal (rotation) matrix, which can be parametrized by three real Euler angles. That leaves 6 phase parameters in the general unitary $\mathbf{V}$. We also have 6 quark fields, with 5 phase differences which can be adjusted. Thus just one irreducible phase degree of freedom can remain in $\mathbf{V}$, after quark rephasing. Consequently, the three-generation model naturally accommodates **CP**-violation in the quark sector: this was the great discovery of Kobayashi and Maskawa (1973). It was another four years before the existence of the b quark was established, and more than twenty before the t quark was produced.

The 3-generation matrix $\mathbf{V}$, written out in full, is

$$\mathbf{V} = \begin{pmatrix} V_{\rm ud} & V_{\rm us} & V_{\rm ub} \\ V_{\rm cd} & V_{\rm cs} & V_{\rm cb} \\ V_{\rm td} & V_{\rm ts} & V_{\rm tb} \end{pmatrix}, \tag{20.165}$$

and is called the CKM matrix, after Cabibbo, Kobayashi, and Maskawa. Clearly, there is no unique parametrization of $\mathbf{V}$. One that has now become standard (Nakamura *et al.* 2010) is (Chau and Keung 1984)

$$\mathbf{V} = \begin{pmatrix} c_{12}c_{13} & s_{12}c_{13} & s_{13}\mathrm{e}^{-\mathrm{i}\delta} \\ -s_{12}c_{23} - c_{12}s_{23}s_{13}\mathrm{e}^{\mathrm{i}\delta} & c_{12}c_{23} - s_{12}s_{23}s_{13}\mathrm{e}^{\mathrm{i}\delta} & s_{23}c_{13} \\ s_{12}s_{23} - c_{12}c_{23}s_{13}\mathrm{e}^{\mathrm{i}\delta} & -c_{12}s_{23} - s_{12}c_{23}s_{13}\mathrm{e}^{\mathrm{i}\delta} & c_{23}c_{13} \end{pmatrix} \tag{20.166}$$

where $c_{ij} = \cos\theta_{ij}, s_{ij} = \sin\theta_{ij}$ with $i, j = 1, 2, 3$; the θ_{ij} may be thought of as the three Euler angles in an orthogonal $\mathbf{V}$, and δ is the remaining irreducible **CP**-violating phase. In the limit $\theta_{13} = \theta_{23} = 0$, this CKM matrix reduces to the Cabibbo-GIM matrix with $\theta_{12} \equiv \theta_{\rm C}$.

However, it would also be desirable to have a measure of **CP**-violation that was independent of quark rephasing. Consider one of the off-diagonal unitarity conditions,

$$V_{\rm ud}V_{\rm ub}^* + V_{\rm cd}V_{\rm cb}^* + V_{\rm td}V_{\rm tb}^* = 0. \tag{20.167}$$

(Note that the complex conjugate of this equation gives another, independent, condition.) The best-measured of these products is $V_{\rm cd}V_{\rm cb}^*$; dividing by this quantity, (20.167) can be written as

$$1 + z_1 + z_2 = 0, \tag{20.168}$$

where

$$z_1 = \frac{V_{\rm td}V_{\rm tb}^*}{V_{\rm cd}V_{\rm cb}^*}, \quad z_2 = \frac{V_{\rm ud}V_{\rm ub}^*}{V_{\rm cd}V_{\rm cb}^*}. \tag{20.169}$$

When viewed in the complex plane, relation (20.168) is the statement that the vectors (1,0), z_1 and z_2 close to form a triangle as shown in figure 20.10, one of 6 such *unitarity triangles* that can be formed. The area Δ of this triangle is

$$\Delta = \frac{1}{2}{\rm Im}(z_2 z_1^*) = \frac{1}{2}{\rm Im}\left(\frac{V_{\rm ud}V_{\rm ub}^*V_{\rm td}^*V_{\rm tb}}{|V_{\rm cd}|^2|V_{\rm cb}|^2}\right). \tag{20.170}$$

Recalling that a rephasing multiplies $V_{\rm ij}$ by ${\rm expi}(\alpha_{\rm i} - \alpha_{\rm j})$, we see that Δ is rephasing invariant; in particular, so is the numerator J where

$$J \equiv {\rm Im}(V_{\rm ud}V_{\rm tb}V_{\rm ub}^*V_{\rm td}^*) \tag{20.171}$$

is a *Jarlskog invariant* (Jarlskog 1985). J may be thought of as follows: (i) strike out the 'c' row and 's' column of $\mathbf{V}$; (ii) take the complex conjugate of the off-diagonal elements in the 2×2 matrix that remains; (iii) multiply the four elements and take the imaginary part. There is nothing special about this particular row and column: there are nine different ways of choosing to pair one row with one column, but all such Js are equal up to a sign, because of the unitarity of $\mathbf{V}$. In the parametrization (20.166), J takes the form

$$J = c_{12}s_{12}c_{23}s_{23}c_{13}^2 s_{13}\sin\delta, \tag{20.172}$$

which vanishes if any $\theta_{\rm ij} = 0,$ or $\pi/2$, or if $\delta = 0$ or π.

The CKM matrix is an integral part of the Standard Model, and testing its validity is an important experimental goal. Various tests are possible. Consider first the magnitudes of the CKM elements. These must satisfy six relations following from the unitarity of $\mathbf{V}$: namely, the sum of the squares of the absolute values of the elements of each row, and of each column, must add up to unity.

The magnitudes of the six elements of the first two rows have been determined from measurements of semileptonic decay rates: for example, the amplitude for the tree-level process ${\rm d} \to {\rm u} + {\rm e}^- + \bar{\nu}_{\rm e}$ is proportional to $V_{\rm ud}$.

But non-perturbative strong interaction effects enter into the amplitudes for corresponding measured hadronic transitions, such as $\mathrm{n} \to \mathrm{p} + \mathrm{e}^- + \bar{\nu}_\mathrm{e}$ or $\pi^- \to \pi^0 + \mathrm{e}^- + \bar{\nu}_\mathrm{e}$. In many cases these hadronic factors in the matrix elements can now be calculated by unquenched lattice QCD.

The status of the experimental determination of the moduli $|V_{\mathrm{ij}}|$ is regularly reviewed by the Particle Data Group. The current results for the unitarity checks are (Ceccucci *et al.* 2010)

$$|V_{\mathrm{ud}}|^2 + |V_{\mathrm{us}}|^2 + |V_{\mathrm{ub}}|^2 = 0.9999 \pm 0.0006 \quad (20.173)$$

$$|V_{\mathrm{cd}}|^2 + |V_{\mathrm{cs}}|^2 + |V_{\mathrm{cb}}|^2 = 1.101 \pm 0.074 \quad (20.174)$$

$$|V_{\mathrm{ud}}|^2 + |V_{\mathrm{cd}}|^2 + |V_{\mathrm{td}}|^2 = 1.002 \pm 0.005 \quad (20.175)$$

$$|V_{\mathrm{us}}|^2 + |V_{\mathrm{cs}}|^2 + |V_{\mathrm{ts}}|^2 = 1.098 \pm 0.074. \quad (20.176)$$

Evidently these results are fully consistent with the CKM prediction of unitarity.

The most accurate values of the nine magnitudes are obtained by a global fit to all the available measurements, imposing the constraints of 3-generation unitarity. The current result for the magnitudes, imposing these constraints, is (Ceccucci *et al.* 2010)

$$\mathbf{V} = \begin{pmatrix} 0.9428 \pm 0.00015 & 0.2253 \pm 0.0007 & 0.00347^{+0.00016}_{-0.00012} \\ 0.2252 \pm 0.0007 & 0.97345^{+0.00015}_{-0.00016} & 0.0410^{+0.0011}_{-0.0007} \\ 0.00862^{+0.00026}_{-0.00020} & 0.0403^{+0.0011}_{-0.0007} & 0.999152^{+0.000030}_{-0.000045} \end{pmatrix}, \quad (20.177)$$

and the Jarlskog invariant is $J = (2.91^{+0.19}_{-0.11}) \times 10^{-5}$.

From (20.177) it follows that the mixing angles are small, and moreover satisfy a definite hierarchy

$$1 \gg \theta_{12} \gg \theta_{23} \gg \theta_{13}. \quad (20.178)$$

In more physical terms, hadrons evidently prefer to decay semileptonically to the nearest generation. Also, because the elements $V_{\mathrm{ub}}, V_{\mathrm{cb}}, V_{\mathrm{td}}$ and V_{ts}, which connect the third generation to the first two, are all quite small, the physics of the first two generations is hardly influenced by the presence of the third. This reflects, in quantitative terms, the success of the Cabibbo-GIM description, and the fact that the **CP**-violation seen in the K-meson sector is so weak. **CP**-violation is much more visible in B physics, as Carter and Sanda (1980, 1981) were the first to suggest, and as we shall discuss in the following chapter.

Consider now the complex-valued off-diagonal unitarity conditions, in particular the condition (20.168). Following Wolfenstein (1983), we identify s_{12} as the small parameter λ, and write $V_{\mathrm{cb}} \simeq s_{23} = A\lambda^2$ and $V_{\mathrm{ub}} = s_{13}\exp(-\mathrm{i}\delta) = A\lambda^3(\rho - \mathrm{i}\eta)$ with $A \simeq 1$ and $|\rho - \mathrm{i}\eta| < 1$. This gives

$$\mathbf{V} = \begin{pmatrix} 1 - \lambda^2/2 & \lambda & A\lambda^3(\rho - \mathrm{i}\eta) \\ -\lambda & 1 - \lambda^2/2 & A\lambda^2 \\ A\lambda^3(1 - \rho - \mathrm{i}\eta) & -A\lambda^2 & 1 \end{pmatrix}, \quad (20.179)$$

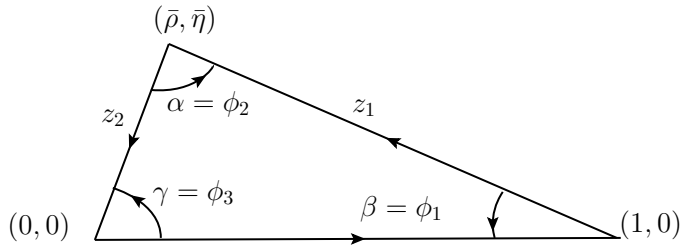


FIGURE 20.10
The unitarity triangle represented by (20.168).

neglecting terms of order λ^4 and higher. Then

$$z_1 = \frac{V_{\rm td}V_{\rm tb}^*}{V_{\rm cd}V_{\rm cb}^*} \simeq \rho + {\rm i}\eta - 1,\ z_2 = \frac{V_{\rm ud}V_{\rm ub}^*}{V_{\rm cd}V_{\rm cb}^*} \simeq -(\rho + {\rm i}\eta),\ J \simeq A^2\lambda^6\eta. \quad (20.180)$$

The unitarity triangle represented by the condition (20.168), or alternatively $-z_1 - z_2 = 1$, is therefore a triangle on the base (1,0), with sides $\rho + {\rm i}\eta$ and $1 - (\rho + {\rm i}\eta)$. Buras *et al.* (1994) showed that including terms up to order λ^5 changes (ρ, η) to $(\bar\rho, \bar\eta)$ where $\bar\rho = (1-\lambda^2/2)\rho, \bar\eta = (1-\lambda^2/2)\eta$. The top vertex of the triangle in figure 20.10 is therefore at the point $(\bar\rho, \bar\eta)$. The angles α, β and γ (also called ϕ_2, ϕ_1 and ϕ_3) are defined by

$$\alpha \equiv \phi_2 \equiv \arg\left(-\frac{V_{\rm td}V_{\rm tb}^*}{V_{\rm ud}V_{\rm ub}^*}\right) \approx \arg -\left(\frac{1-\bar\rho-{\rm i}\bar\eta}{\bar\rho+{\rm i}\bar\eta}\right) \quad (20.181)$$

$$\beta \equiv \phi_1 \equiv \arg\left(-\frac{V_{\rm cd}V_{\rm cb}^*}{V_{\rm td}V_{\rm tb}^*}\right) \approx \arg\left(\frac{1}{1-\bar\rho-{\rm i}\bar\eta}\right) \quad (20.182)$$

$$\gamma \equiv \phi_3 \equiv \arg\left(-\frac{V_{\rm ud}V_{\rm ub}^*}{V_{\rm cd}V_{\rm cb}^*}\right) \approx \arg(\bar\rho+{\rm i}\bar\eta) \quad (20.183)$$

The sides of this triangle are determined by the magnitudes of the CKM elements, and so another check is provided by the condition that the three sides should close to form a triangle. Further independent constraints are provided by measurements of the angles α, β, and γ which are directly related to **CP**-violation effects, as we shall discuss in the following chapter. Figure 20.11 shows a plot of all the constraints in the $\bar\rho, \bar\eta$ plane from many different measurements (combined following the approach of Charles *et al.* 2005 and Höcker *et al.* 2001), and the global fit, as presented by Ceccucci *et al.* (2010). The annular region labelled by $|V_{\rm ub}|$ represents, for example, the uncertainty in the determination of $|z_2| = |V_{\rm ud}V_{\rm ub}^*/V_{\rm cd}V_{\rm cb}^*|$, which is principally due to the uncertainty in $|V_{\rm ub}|$. The region labelled by $\Delta m_{\rm d}$ represents the constraint on

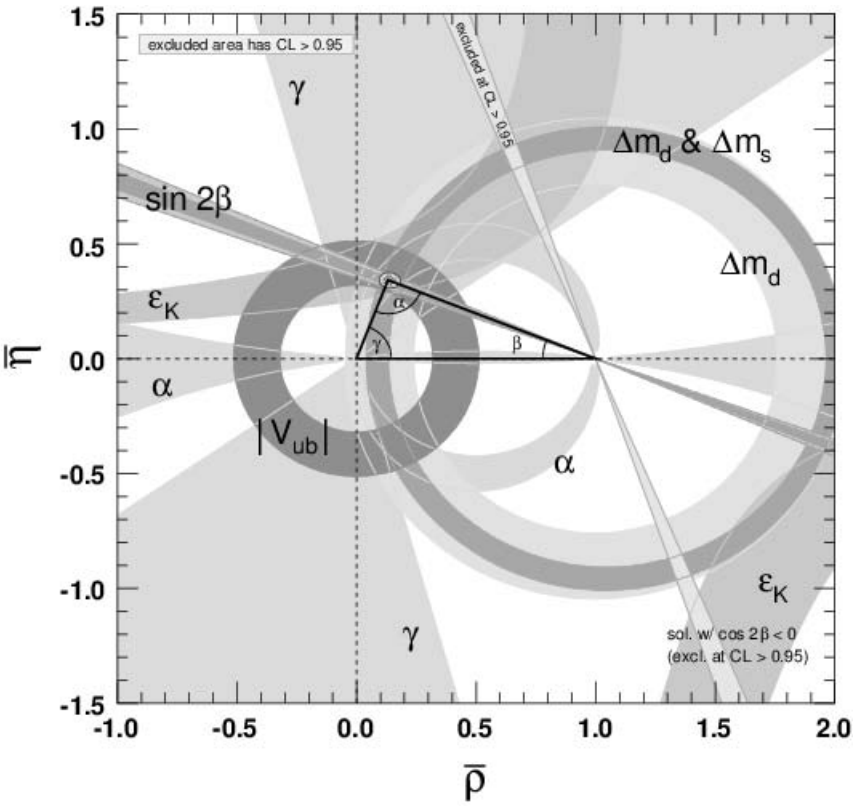


FIGURE 20.11
Constraints in the $\bar{\rho}, \bar{\eta}$ plane. The shaded areas have 95% CL. [Figure reproduced, courtesy Michael Barnett for the Particle Data Group, from the review of the CKM Quark-Mixing Matrix by A Ceccucci, Z Ligeti and Y Sakai, section 11 in the *Review of Particle Physics*, K Nakamura *et al.* (Partcle Data Group) *Journal of Physics* G **37** (2010) 075021, IOP Publishing Limited.] (See color plate III.)

$|z_1| = |V_{\rm td}V_{\rm tb}^*/V_{\rm cd}V_{\rm cb}^*|$, where $|V_{\rm td}|$ is deduced from the value of the $\rm B^0 - \bar{B}^0$ mass difference $\Delta m_{\rm d}$ measured in $\rm B^0 - \bar{B}^0$ oscillations mediated by top-quark dominated box diagrams (see section 21.2.1 in the following chapter); here the uncertainties are dominated by lattice QCD. Figure 20.11 represents an enormous experimental effort, especially in the decade 2000-2010. The 95% CL regions all overlap consistently. It is quite remarkable how the single **CP**- violating parameter, three-generation scheme of Kobayashi and Maskawa (1973) has withstood this searching test.

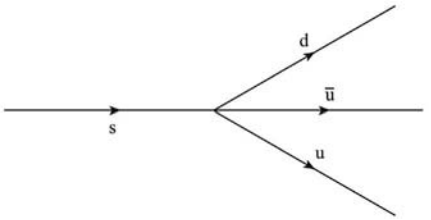


FIGURE 20.12
Effective four-fermion non-leptonic weak transition at the quark level.

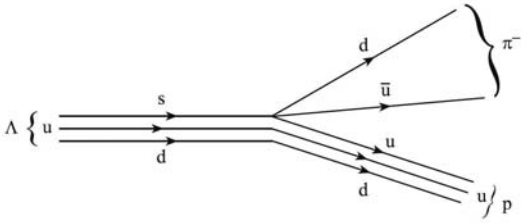


FIGURE 20.13
Non-leptonic weak decay of Λ^0 using the process of figure 20.12, with the addition of two 'spectator' quarks.

20.8 Non-leptonic weak interactions

The CKM 6-quark charged weak current, which replaces the GIM current (20.101), is

$$\hat{j}^{\mu}_{\rm CKM}({\rm u,d,s,c,t,b}) = \bar{\hat{u}}\gamma^{\mu}\frac{(1-\gamma_5)}{2}\hat{d}' + \bar{\hat{c}}\gamma^{\mu}\frac{(1-\gamma_5)}{2}\hat{s}' + \bar{\hat{t}}\gamma^{\mu}\frac{(1-\gamma_5)}{2}\hat{b}', \tag{20.184}$$

and the effective weak Hamiltonian of (20.92) (as modified by CKM) clearly contains the term

$$\hat{\mathcal{H}}^{q}_{\rm CC}(x) = \frac{G_{\rm F}}{\sqrt{2}}\hat{j}^{\mu}_{\rm CKM}(x)\hat{j}^{\dagger}_{\mu\rm CKM}(x) \tag{20.185}$$

in which no lepton fields are present (just as there are no quarks in (20.40)). This interaction is responsible, at the quark level, for transitions involving four quark (or antiquark) fields at a point. For example, the process shown in figure 20.12 can occur. By 'adding on' another two quark lines u and d, which undergo no weak interaction, we arrive at figure 20.13, which represents the non-leptonic decay $\Lambda^0 \to {\rm p}\pi^-$.

This figure is, of course, rather symbolic since there are strong QCD interactions (not shown) which are responsible for binding the three-quark systems into baryons, and the $q\bar{q}$ system into a meson. Unlike the case of deep inelastic lepton scattering, these QCD interactions can not be treated perturbatively, since the distance scales involved are typically those of the hadron sizes ($\sim$ 1 fm), where perturbation theory fails. This means that non-leptonic weak interactions among hadrons are difficult to analyze quantitatively, though progress can be made via lattice QCD. Similar difficulties also arise, evidently, in the case of semi-leptonic decays. In general, one has to begin in a phenomenological way, parametrizing the decay amplitudes in terms of appropriate form factors (which are analogous to the electromagnetic form factors introduced in chapter 8). In the case of transitions involving at least one heavy quark Q, Isgur and Wise (1989, 1990) noticed that a considerable simplification occurs in the limit $m_{\mathrm{Q}} \to \infty$. For example, one universal function (the 'Isgur–Wise form factor') is sufficient to describe a large number of hadronic form factors introduced for semi-leptonic transitions between two heavy pseudoscalar (0^-) or vector (1^-) mesons. For an introduction to the Isgur–Wise theory we refer to Donoghue *et al.* (1992).

The non-leptonic sector is, however, the scene of some very interesting physics, such as $\mathrm{K}^0 - \bar{\mathrm{K}}^0$ and $\mathrm{B}^0 - \bar{\mathrm{B}}^0$ oscillations, and **CP** violation in the $\mathrm{K}^0 - \bar{\mathrm{K}}^0, \mathrm{D}^0 - \bar{\mathrm{D}}^0$ and $\mathrm{B}^0 - \bar{\mathrm{B}}^0$ systems. We shall discuss these phenomena in the following chapter.

Problems

20.1 Show that in the non-relativistic limit ($|\boldsymbol{p}| \ll M$) the matrix element $\bar{u}_{\mathrm{p}}\gamma^{\mu}u_{\mathrm{n}}$ of (20.2) vanishes if p and n have different spin states.

20.2 Verify the normalization $N = (E + |\boldsymbol{p}|)^{1/2}$ in (20.23).

20.3 Verify (20.30) and (20.31).

20.4 Verify that equations (20.32) are invariant under **CP**.

20.5 The matrix γ_5 is defined by $\gamma_5 = \mathrm{i}\gamma^0\gamma^1\gamma^2\gamma^3$. Prove the following properties:

(a) $\gamma_5^2 = 1$ and hence that

$$(1 + \gamma_5)(1 - \gamma_5) = 0;$$

(b) from the anticommutation relations of the other γ matrices, show that

$$\{\gamma_5, \gamma_\mu\} = 0$$

and hence that

$$(1+\gamma_5)\gamma_0 = \gamma_0(1-\gamma_5)$$

and

$$(1+\gamma_5)\gamma_0\gamma_\mu = \gamma_0\gamma_\mu(1+\gamma_5).$$

20.6

(a) Consider the two-dimensional antisymmetric tensor ϵ_{ij} defined by

$$\epsilon_{12} = +1, \epsilon_{21} = -1, \qquad \epsilon_{11} = \epsilon_{22} = 0.$$

By explicitly enumerating all the possibilities (if necessary), convince yourself of the result

$$\epsilon_{ij}\epsilon_{kl} = +1(\delta_{ik}\delta_{jl} - \delta_{il}\delta_{jk}).$$

Hence prove that

$$\epsilon_{ij}\epsilon_{il} = \delta_{jl} \qquad \text{and} \qquad \epsilon_{ij}\epsilon_{ij} = 2$$

(remember, in two dimensions, $\sum_i \delta_{ii} = 2$).

(b) By similar reasoning to that in part (a) of this question, it can be shown that the product of two three-dimensional antisymmetric tensors has the form

$$\epsilon_{ijk}\epsilon_{lmn} = \begin{vmatrix} \delta_{il} & \delta_{im} & \delta_{in} \\ \delta_{jl} & \delta_{jm} & \delta_{jn} \\ \delta_{kl} & \delta_{km} & \delta_{kn} \end{vmatrix}.$$

Prove the results

$$\epsilon_{ijk}\epsilon_{imn} = \begin{vmatrix} \delta_{jm} & \delta_{jn} \\ \delta_{km} & \delta_{kn} \end{vmatrix} \qquad \epsilon_{ijk}\epsilon_{ijn} = 2\delta_{kn} \qquad \epsilon_{ijk}\epsilon_{ijk} = 3!$$

(c) Extend these results to the case of the four-dimensional (Lorentz) tensor $\epsilon_{\mu\nu\alpha\beta}$ (remember that a minus sign will appear as a result of $\epsilon_{0123} = +1$ but $\epsilon^{0123} = -1$).

20.7 Starting from the amplitude for the process

$$\nu_\mu + \mathrm{e}^- \to \mu^- + \nu_\mathrm{e}$$

given by the current–current theory of weak interactions,

$$\mathcal{M} = -\mathrm{i}(G_\mathrm{F}/\sqrt{2})\bar{u}(\mu)\gamma_\mu(1-\gamma_5)u(\nu_\mu)g^{\mu\nu}\bar{u}(\nu_\mathrm{e})\gamma_\nu(1-\gamma_5)u(\mathrm{e}),$$

verify the intermediate results given in section 20.5 leading to the result

$$\mathrm{d}\sigma/\mathrm{d}t = G_\mathrm{F}^2/\pi$$

(neglecting all lepton masses). Hence show that the local total cross section for this process rises linearly with s:

$$\sigma = G_{\rm F}^2 s/\pi.$$

20.8 The invariant amplitude for $\pi^+ \to e^+\nu$ decay may be written as (see (18.52))

$$\mathcal{M} = (G_{\rm F}V_{\rm ud}) f_\pi p^\mu \bar{u}(\nu)\gamma_\mu(1-\gamma_5)v({\rm e})$$

where p^μ is the 4-momentum of the pion, and the neutrino is taken to be massless. Evaluate the decay rate in the rest frame of the pion using the decay rate formula

$$\Gamma = (1/2m_\pi)|\overline{\mathcal{M}}|^2 {\rm dLips}(m_\pi^2; k_{\rm e}, k_\nu).$$

Show that the ratio of $\pi^+ \to {\rm e}^+\nu$ and $\pi^+ \to \mu^+\nu$ rates is given by

$$\frac{\Gamma(\pi^+ \to {\rm e}^+\nu)}{\Gamma(\pi^+ \to \mu^+\nu)} = \left(\frac{m_{\rm e}}{m_\mu}\right)^2 \left(\frac{m_\pi^2 - m_{\rm e}^2}{m_\pi^2 - m_\mu^2}\right)^2.$$

Repeat the calculation using the amplitude

$$\mathcal{M}' = (G_{\rm F}V_{\rm ud}) f_\pi p^\mu \bar{u}(\nu)\gamma_\mu(g_{\rm V} + g_{\rm A}\gamma_5)v({\rm e})$$

and retaining a finite neutrino mass. Discuss the $\rm e^+/\mu^+$ ratio in the light of your result.

20.9

(a) Verify that the inclusive inelastic neutrino-proton scattering differential cross section has the form

$$\frac{{\rm d}^2\sigma^{(\nu)}}{{\rm d}Q^2{\rm d}\nu} = \frac{G_{\rm F}^2 k'}{2\pi k}\left(W_2^{(\nu)}\cos^2(\theta/2) + W_1^{(\nu)} 2\sin^2(\theta/2) + \frac{(k+k')}{M}\sin^2(\theta/2) W_3^{(\nu)}\right)$$

in the notation of section 20.7.2.

(b) Using the Bjorken scaling behaviour

$$\nu W_2^{(\nu)} \to F_2^{(\nu)} \qquad M W_1^{(\nu)} \to F_1^{(\nu)} \qquad \nu W_3^{(\nu)} \to F_3^{(\nu)}$$

rewrite this expression in terms of the scaling functions. In terms of the variables x and y, neglect all masses and show that

$$\frac{{\rm d}^2\sigma^{(\nu)}}{{\rm d}x{\rm d}y} = \frac{G_{\rm F}^2}{2\pi} s[F_2^{(\nu)}(1-y) + F_1^{(\nu)} xy^2 + F_3^{(\nu)}(1-y/2)yx].$$

Remember that

$$\frac{k'\sin^2(\theta/2)}{M} = \frac{xy}{2}.$$

(c) Insert the Callan–Gross relation

$$2xF_1^{(\nu)} = F_2^{(\nu)}$$

to derive the result quoted in section 20.7.2:

$$\frac{\mathrm{d}^2\sigma^{(\nu)}}{\mathrm{d}x\mathrm{d}y} = \frac{G_\mathrm{F}^2}{2\pi} s F_2^{(\nu)} \left(\frac{1+(1-y)^2}{2} + \frac{xF_3^{(\nu)}}{F_2^{(\nu)}} \frac{1-(1-y)^2}{2} \right).$$

20.10 The differential cross section for $\nu_\mu \mathrm{q}$ scattering by charged currents has the same form (neglecting masses) as the $\nu_\mu \mathrm{e}^- \to \mu^- \nu_\mathrm{e}$ result of problem 20.7, namely

$$\frac{\mathrm{d}\sigma}{\mathrm{d}t}(\nu\mathrm{q}) = \frac{G_\mathrm{F}^2}{\pi}.$$

(a) Show that the cross section for scattering by antiquarks $\nu_\mu \bar{\mathrm{q}}$ has the form

$$\frac{\mathrm{d}\sigma}{\mathrm{d}t}(\nu\bar{\mathrm{q}}) = \frac{G_\mathrm{F}^2}{\pi}(1-y)^2.$$

(b) Hence prove the results quoted in section 20.7.2:

$$\frac{\mathrm{d}^2\sigma}{\mathrm{d}x\mathrm{d}y}(\nu\mathrm{q}) = \frac{G_\mathrm{F}^2}{\pi} sx\delta(x - Q^2/2M\nu)$$

and

$$\frac{\mathrm{d}^2\sigma}{\mathrm{d}x\mathrm{d}y}(\nu\bar{\mathrm{q}}) = \frac{G_\mathrm{F}^2}{\pi} sx(1-y^2)\delta(x - Q^2/2M\nu)$$

(where M is the nucleon mass).

(c) Use the parton model prediction

$$\frac{\mathrm{d}^2\sigma^{(\nu)}}{\mathrm{d}x\mathrm{d}y} = \frac{G_\mathrm{F}^2}{\pi} sx[q(x) + \bar{q}(x)(1-y)^2]$$

to show that

$$F_2^{(\nu)} = 2x[q(x) + \bar{q}(x)]$$

and

$$\frac{xF_3^{(\nu)}(x)}{F_2^{(\nu)}(x)} = \frac{q(x) - \bar{q}(x)}{q(x) + \bar{q}(x)}.$$

20.11 Verify the transformation laws (20.160).

21

CP Violation and Oscillation Phenomena

In this chapter we shall continue with the phenomenology of weak interactions, introducing two topics which have been the focus of intense experimental effort in the recent decade: **CP** violation in B meson decays, and oscillations in both neutral meson and neutrino systems. In the following chapter we take up again the gauge theory theme, with the Glasow-Salam-Weinberg electroweak theory.

CP violation was first discovered in the decays of neutral K mesons (Christenson *et al.* 1964), but we shall not follow a historical approach to this subject. Instead we shall concentrate on B-meson decays, where the effects are far larger, and much clearer to interpret theoretically than in the K-meson case. **CP** violation is reviewed in Branco *et al.* (1999), Bigi and Sanda (2000) and Harrison and Quinn (1998). We aim simply to illustrate the principles with some particular examples. In particular, we shall generally not discuss theoretical predictions; our main emphasis will be on describing selected experiments which have allowed determinations of the angles α, β and γ of the unitarity triangle, figure 20.10.

We saw in section 20.7.3 that, in the Standard Model, **CP** violation is attributable solely to one irreducible phase degree of freedom, δ, in the CKM matrix $\mathbf{V}$. Clearly, to measure this phase, it is necessary (as usual in quantum mechanics) to create situations where it enters into the *interference* between two complex amplitudes. Two situations may be distinguished (Carter and Sanda 1980):

(i) interference between two decay amplitudes $\mathrm{B}^0 \to \mathrm{X}$ and $\bar{\mathrm{B}}^0 \to \mathrm{X}$, where the B^0 and $\bar{\mathrm{B}}^0$ have been produced in a coherent state by mixing, and decay to a common hadronic final state X;

(ii) interference between two different amplitudes for a single B-meson to decay to a final state X.

Method (ii) ('direct **CP** violation') can be applied to charged as well as neutral mesons.

The mixing in method (i) is formally similar to that involved in neutrino oscillations, which we treat after the meson case. We shall therefore start in section 21.1 with an example illustrating method (ii). We set up the mixing formalism and apply it to **CP** violation in B decays in section 21.2; we discuss K decays in section 21.3. Neutrino oscillations are treated in section 21.4.

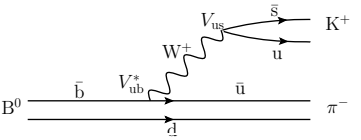


FIGURE 21.1
Tree diagram contribution to $B^0 \to K^+\pi^-$ via the quark transition $\bar{b} \to \bar{s}u\bar{u}$.

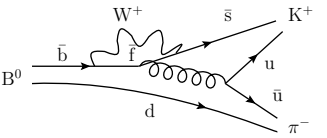


FIGURE 21.2
Penguin diagrams ($\bar{f} = \bar{u}, \bar{c}, \bar{t}$) contributing to $B^0 \to K^+\pi^-$ via the quark transition $\bar{b} \to \bar{s}\,u\,\bar{u}$.

21.1 Direct CP violation in B decays

Consider the decays

$$B^0 \to K^+\pi^- \quad \text{and} \quad \bar{B}^0 \to K^-\pi^-. \tag{21.1}$$

The first of these can proceed via the quark transitions shown in figure 21.1, which (in parton-like language) is a 'tree-diagram'. Of course, long-distance strong interaction effects will come into play in forming the hadronic states B^0, K^+ and π^-, and in final state interactions between the K^+ and π^-; we do not represent these strong interactions in figure 21.1, or in subsequent similar diagrams. We are specifically interested in the *weak phase* of figure 21.1, since it is this quantity which changes sign under the **CP** transformation ($V_{ij} \to V^*_{ij}$), and this phase change will lead to observable **CP** violation effects. By contrast, the strong interaction phases – which will play an important role – will be **CP** invariant, but we do not need to display them yet. So we write the amplitude for figure 21.1 as

$$A_T(B^0 \to K^-\pi^-) = V^*_{ub}V_{us}\ t_{\bar{u}}, \tag{21.2}$$

where the CKM couplings have been displayed.

There are, however, three order-α_s loop corrections to figure 21.1, shown in figure 21.2, where $\bar{f} = \bar{u}, \bar{c}$ and $\bar{t}$. We write the amplitude for the sum of

these three 'penguin' diagrams as

$$A_{\rm P}({\rm B}^0 \to {\rm K}^+\pi^-) = V_{\rm us}V_{\rm ub}^*\, p_{\bar{\rm u}} + V_{\rm cs}V_{\rm cb}^*\, p_{\bar{\rm c}} + V_{\rm ts}V_{\rm tb}^*\, p_{\bar{\rm t}}, \tag{21.3}$$

where $p_{\bar{\rm f}}$ is the penguin amplitude with $\bar{\rm f}$ in the loop. It is convenient to use a unitarity relation to rewrite $V_{\rm ts}V_{\rm tb}^*$ in terms of the other two related CKM products:

$$V_{\rm ts}V_{\rm tb}^* = -V_{\rm us}V_{\rm ub}^* - V_{\rm cs}V_{\rm cb}^*, \tag{21.4}$$

so that the total amplitude becomes

$$A({\rm B}^0 \to {\rm K}^+\pi^-) = V_{\rm ub}^*V_{\rm us}T_{{\rm K}\pi} + V_{\rm cs}V_{\rm cb}^*P_{{\rm K}\pi}, \tag{21.5}$$

where

$$T_{{\rm K}\pi} = t_{\bar{\rm u}} + p_{\bar{\rm u}} - p_{\bar{\rm t}}, \qquad P_{{\rm K}\pi} = p_{\bar{\rm c}} - p_{\bar{\rm t}}. \tag{21.6}$$

In terms of the parametrization (20.179), (21.5) becomes

$$A({\rm B}^0 \to {\rm K}^+\pi^-) = A\lambda^4(\rho + {\rm i}\eta)T_{{\rm K}\pi} + A\lambda^2(1 - \lambda^2/2)P_{{\rm K}\pi}. \tag{21.7}$$

Similarly, the amplitude for the charge-conjugate reaction is

$$A(\bar{\rm B}^0 \to {\rm K}^-\pi^+) = A\lambda^4(\rho - {\rm i}\eta)T_{{\rm K}\pi} + A\lambda^2(1 - \lambda^2/2)P_{{\rm K}\pi}. \tag{21.8}$$

We can now calculate the decay-rate asymmetry

$$\mathcal{A}_{{\rm K}\pi} = \frac{|A(\bar{\rm B}^0 \to {\rm K}^-\pi^+)|^2 - |A({\rm B}^0 \to {\rm K}^+\pi^-)|^2}{|A(\bar{\rm B}^0 \to {\rm K}^-\pi^+)|^2 + |A({\rm B}^0 \to {\rm K}^+\pi^-)|^2}. \tag{21.9}$$

To simplify things, let us take a common complex factor K out of the expressions (21.7) and (21.8) and write them as

$$A({\rm B}^0 \to {\rm K}^+\pi^-) = K({\rm e}^{{\rm i}\gamma} + R{\rm e}^{{\rm i}(\delta_{\rm P} - \delta_{\rm T})}) \tag{21.10}$$

$$A(\bar{\rm B}^0 \to {\rm K}^-\pi^+) = K({\rm e}^{-{\rm i}\gamma} + R{\rm e}^{{\rm i}(\delta_{\rm P} - \delta_{\rm T})}), \tag{21.11}$$

where (see equation (20.183)) γ is the phase of $\rho + {\rm i}\eta$, R is real, and $\delta_{\rm P} - \delta_{\rm T}$ is the difference in (strong) phases between $P_{{\rm K}\pi}$ and $T_{{\rm K}\pi}$. Then we easily find

$$\mathcal{A}_{{\rm K}\pi} = \frac{2R\sin\gamma\,\sin(\delta_{\rm T} - \delta_{\rm P})}{1 + R^2 + 2R\cos\gamma\,\cos(\delta_{\rm T} - \delta_{\rm P})}. \tag{21.12}$$

Thus we see that, for a **CP**-violating signal, there must be two interfering amplitudes leading to a common final state, and the amplitudes must have both different weak phases and different strong phases. An order of magnitude estimate of the effect can be made as follows. First, note that $P_{{\rm K}\pi}$ is not ultraviolet divergent, since it is the difference of two penguin contributions; its magnitude is expected to be of order $\alpha_{\rm s}/\pi \sim 0.05$. The tree contribution in (21.7) carries an extra factor of $\lambda^2 \sim 0.05$ as compared with the penguin contribution, so that R is of order 1. This indicates that the asymmetry should be significant.

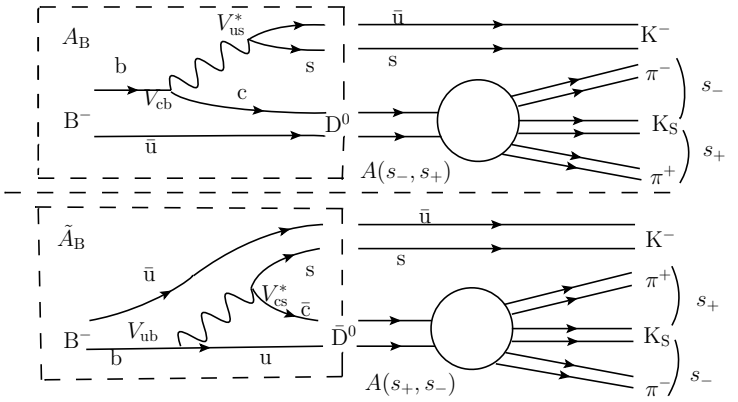


FIGURE 21.3
Left-hand part: tree diagram contributions to $B^- \to D^0 K^-$ (upper diagram, via quark transition $b \to c\bar{u}s$), and to $B^- \to \bar{D}^0 K^-$ (lower diagram, via quark transition $b \to u\bar{c}s$). Right-hand part: decays of D^0 and $\bar{D}^0$ to the common $\pi^+\pi^-\pi^0$ state.

Indeed non-zero values of $\mathcal{A}_{K\pi}$ have been reported by both the BaBar and Belle collaborations:

$$\text{BaBar (Aubert } et\ al.\text{ 2004)} : \mathcal{A}_{K\pi} = -0.133 \pm 0.030 \pm 0.009 \qquad (21.13)$$

$$\text{Belle (Chao } et\ al.\text{ 2005)} : \mathcal{A}_{K\pi} = -0.113 \pm 0.022 \pm 0.008 \qquad (21.14)$$

where the first error is statistical and the second is systematic.

Altough $\mathcal{A}_{K\pi}$ is sensitive to the **CP**-violating angle γ, it is not easy to extract γ cleanly from these measurements. Both the tree and the penguin amplitudes involve non-perturbative factors for producing a particular meson state from the corresponding $q\bar{q}$ state; the strong phases are also not calculable.

A decay with no penguin contributions, but still with two interfering channels, would have fewer uncertainties. (It is also less likely to be affected by new physics, which could provide short-distance corrections to penguin loops.) One such example is provided by the decays (i) $B^- \to D^0 K^-$ and (ii) $B^- \to \bar{D}^0 K^-$, which can interfere when the $(D^0 K^-)$ and $(\bar{D}^0 K^-)$ states decay to a common final state. Here the quark transition in (i) is $b \to c\bar{u}s$, and in (ii) is $b \to u\bar{c}s$; in neither case is a penguin contribution possible.

The tree-level diagrams which contribute are shown in the left-hand parts of figure 21.3 (we shall discuss the right-hand parts in a moment). We denote the amplitude for $B^- \to D^0 K^-$ by A_B, and note that $A_B \sim A\lambda^3$. The amplitude for $B^- \to \bar{D}^0 K^-$, $\tilde{A_B}$, differs in three ways from A_B: (i) it is colour-suppressed by a factor 1/3 since the $\bar{c}$ and u have to be colour matched; (ii)

it contains the factor $V_{\rm ub}V_{\rm cs}^* \sim A\lambda^3(\rho - {\rm i}\eta)$; (iii) it will have a different strong interaction phase. With these factors in mind, we write

$$\tilde{A_{\rm B}} = r_{\rm B} A_{\rm B} {\rm e}^{{\rm i}(\delta_{\rm B} - \gamma)} \tag{21.15}$$

where $\delta_{\rm B}$ is the difference in strong phases between $\tilde{A}_{\rm B}$ and $A_{\rm B}$, and $r_{\rm B}$ is the magnitude ratio of the amplitudes. Since $|\rho - {\rm i}\eta| \sim 0.38$, $r_{\rm B}$ is of order 0.1–0.2, allowing for the colour suppression.

Once again, the asymmetry is proportional to

$$|1 + r_{\rm B}{\rm exp}[{\rm i}(\delta_{\rm B} - \gamma)]\,|^2 - |1 + r_{\rm B}{\rm exp}[{\rm i}(\delta_{\rm B} + \gamma)]\,|^2 \approx 4 r_{\rm B} \sin\delta_{\rm B} \sin\gamma. \tag{21.16}$$

This involves γ, but the relative smallness of $r_{\rm B}$ tends to reduce the sensitivity to γ. An alternative determination of γ can be made (Attwood *et al.* 2001, Giri *et al.* 2003) by making use of three-body decays (to a common channel) of $\rm D^0$ and $\rm \bar{D}^0$, such as $\rm D^0$, $\rm \bar{D}^0 \to K_S \pi^+ \pi^-$. If we denote the amplitude for $\rm D^0 \to K_S \pi^+ \pi^-$ by $A(s_-, s_+)$ (see figure 21.3), where $s_- = (p_{\rm K} + p_{\pi^-})^2$ and $s_+ = (p_{\rm K} + p_{\pi^+})^2$ are the indicated invariant masses, then the amplitude for the $\rm B^-$ to decay to $\rm K^- K_S\, \pi^+ \pi^-$ via the $\rm D^0$ path is[1]

$$A_- = A_{\rm B} D[A(s_-, s_+) + r_{\rm B} {\rm e}^{{\rm i}(\delta_{\rm B} - \gamma)} A(s_+, s_-)], \tag{21.17}$$

and the amplitude for the charge conjugate reaction $\rm B^+ \to K_S\, \pi^- \pi^+$ is

$$A_+ = A_{\rm B} D[A(s_+, s_-) + r_{\rm B} {\rm e}^{{\rm i}(\delta_{\rm B} + \gamma)} A(s_+ - s_+)], \tag{21.18}$$

where D is the D meson propagator. The event rate for the $\rm B^-$ decay is then $\Gamma_-(s_-, s_+)$ where (Aubert *et al.* 2008)

$$\begin{aligned} \Gamma_-(s_-, s_+) \propto |A(s_-, s_+)|^2 + r_{\rm B}^2 |A(s_+, s_-)|^2 + \\ 2\,[x_- {\rm Re}\{A(s_-, s_+) A^*(s_+, s_-)\} + y_- {\rm Im}\{A(s_-, s_+) A^*(s_+, s_-)\}] \end{aligned} \tag{21.19}$$

and the rate for $\rm B^+$ decay is $\Gamma_+(s_-, s_+)$ where

$$\begin{aligned} \Gamma_+(s_-, s_+) \propto |A(s_+, s_-)|^2 + r_{\rm B}^2 |A(s_-, s_+)|^2 + \\ 2\,[x_+ {\rm Re}\{A(s_+, s_-) A^*(s_-, s_+)\} + y_+ {\rm Im}\{A(s_+, s_-) A^*(s_-, s_+)\}]\,. \end{aligned} \tag{21.20}$$

Here

$$x_- = r_{\rm B} \cos(\delta_{\rm B} - \gamma), \quad y_- = r_{\rm B} \sin(\delta_{\rm B} - \gamma) \tag{21.21}$$

$$x_+ = r_{\rm B} \cos(\delta_{\rm B} + \gamma), \quad y_+ = r_{\rm B} \sin(\delta_{\rm B} + \gamma). \tag{21.22}$$

The geometry of the **CP**-violating parameters is shown in figure 21.4. Note that the separation of the $\rm B^-$ and $\rm B^+$ positions in the (x, y) plane is equal to

[1] We are neglecting $\rm D^0$-$\rm \bar{D}^0$ mixing and **CP** asymmetries in D decays, which are at the 1% or less level (Grossman *et al.* 2005).

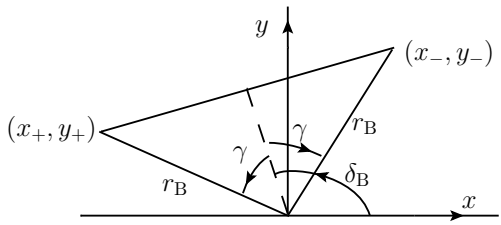


FIGURE 21.4
Geometry of the **CP**-violating parameters $x_\pm, y_\pm$.

$2r_B|\sin\gamma|$, and is a measure of direct **CP** violation. The angle between the lines connecting the B^- and B^+ centres to the origin (0,0) is equal to 2γ.

If the functional dependence of both the modulus and the phase of $A(s_-, s_+)$ were known, then the rates would depend on only three variables, r_B, δ_B, and γ (or equivalently on $x_\pm, y_\pm$). In fact, $A(s_-, s_+)$ can be determined from a Dalitz plot analysis of the decays of D^0 mesons coming from $D^{*+} \to D^0\pi^+$ decays produced in $e^+e^- \to c\bar{c}$ events; the charge of the low-momentum π^+ identifies the flavour of the D^0. Such an analysis is a well-established tool in the study of three-hadron final states, originating in the pioneer work of Dalitz (1953), in connection with the decay $K \to 3\pi$. The partial rate for $D^0(\bar{D}^0) \to K_S\pi^+\pi^-$ is (see the kinematics section of Nakamura *et al.* 2010)

$$d\Gamma \propto |A(s_-, s_+)|^2 ds_- ds_+. \tag{21.23}$$

The physical region in the s_-, s_+ plane is a bounded oval-like region, which would be uniformly populated if $A(s_-, s_+)$ were a constant. In reality, the decay is dominated by quasi two-body states, in particular

$$\begin{aligned} D^- &\to K^{*-}(s_-)\,\pi^+ \quad \text{(CA)} \\ &\to K^{*+}(s_+)\,\pi^- \quad \text{(DCS)} \\ &\to K_S\,\rho^0(s_0), \quad \text{(CP)} \end{aligned} \tag{21.24}$$

where (CA) means CKM-favoured, (DCS) means doubly CKM-suppressed, and (CP) means that it is a **CP** eigenstate. The Dalitz plot shows a dense band of events at $s_- = m^2_{K^{*-}}$ corresponding to the K^{*-} resonance, a band at $s_+ = m^2_{K^{*+}}$, and a band at $s_0 = m^2_\rho$, where $s_0 = (p_{\pi^+} + p_{\pi^-})^2$ and $s_+ + s_- + s_0 = m^2_D + m^2_K + 2m^2_\pi$.

The Dalitz-plot analysis proceeds by writing (Aubert *et al.* 2008) $A(s_-, s_+)$ as a coherent sum of terms representing the quasi two-body modes, together with a non-resonant background. Once $A(s_-, s_+)$ is known, it is inserted into $\Gamma_\mp(s_-, s_+)$ to obtain $(x_\pm, y_\pm)$ from the Dalitz plot distributions of the signal modes of the $B^\mp$ decays. From these, the quantities r_B, δ_B and finally δ can be inferred.

This method has been applied by both BaBar and Belle to determine γ. Their most recently published results are

$$\text{BaBar (Aubert } et\ al.\text{ 2010)}: \quad \gamma = 68 \pm 14 \pm 4 \pm 3^\circ \tag{21.25}$$

$$\text{Belle (Poluektov } et\ al.\text{ 2010)}: \quad \gamma = 78.4^{+10.8}_{-11.6} \pm 3 \pm 8.9^\circ \tag{21.26}$$

where the last uncertainty is due to the D-decay modelling (which ignores, for example, rescattering among the three final state particles). Both these experiments use decays $B^\pm \to DK^\pm, B^\pm \to D^*K^\pm$ with $D^* \to D\pi^0$ and $D^* \to D\gamma$; BaBar in addition uses the decays $D^0 \to K_S K^+ K^-$.

We now turn to the other main method of detecting **CP** violation, through the interference between decays of (for example) B^0 and $\bar{B}^0$ mesons that have been produced in a coherent state by mixing. For this we need to set up the formalism describing time-dependent mixing.

21.2 CP violation in B meson oscillations

B^0-$\bar{B}^0$ oscillations have been studied by the BaBar and Belle collaborations at the PEP2 and KEKB asymmetric e^+e^- colliders. These machines operate at a centre of mass energy equal to the mass of the $\Upsilon(4S)$ resonance state, which is some 20 MeV above the threshold for B^0 $\bar{B}^0$ production. If produced in a symmetric e^+e^- collider (with equal and opposite momenta for the e^+ and e^-), the produced B mesons would move very slowly, $v/c \sim 0.06$, covering a distance of only some 30 μm before decaying ($c\tau$ for B mesons is about 460 μm). This would make it impossible to resolve the decay vertices of the two Bs, as is required in order to observe B^0-$\bar{B}^0$ oscillations, since the accuracy of the decay vertex reconstruction is roughly 100 μm. Oddone (1989) suggested making e^+ e^- collisions with asymmetric energy colliding beams, so that the B mesons now move with the motion of the centre of mass, which can be considerable. For example, at PEP2 (e^- 9 GeV, e^+ 3.1 GeV) $\beta_{\rm cm} \sim 0.5$ and $\gamma_{\rm cm} \sim 1.15$, so that the distance travelled in the (asymmetric) lab frame during the lifetime of an average B meson is ~ 250 μm, which is measurable. At KEKB (e^- 8 GeV, e^- 3.5 GeV), $\beta_{\rm cm}\gamma_{\rm cm} \sim 0.425$.

Since the $\Upsilon(4S)$ state has $J = 1$, the decay $\Upsilon \to BB$ leaves the B mesons in a p wave state, which is forbidden for two identical spinless bosons; therefore one must be a B^0 and the other a $\bar{B}^0$, but we do not know which is which until one has been identified ('tagged') in some way. The flavour of the tagged B may be determined, for example, by the charge of the lepton emitted in the semi-leptonic decays $B^0 \to D^-\ell^+\nu_e, \bar{B}^0 \to D^+\ell^-\bar{\nu}_e$. We shall not describe the evolution of the BB coherent state following production; interested readers may consult Cohen *et al.* (2009) for an instructive discussion, which also covers neutrino oscillations. We shall be interested in the time dependence of the state of the meson which partners the tagged meson, once the correlated

state has been collapsed by the tagging at time $t = 0$ say; the partner meson will be reconstructed by its decay products. Note that the partner meson can decay earlier or later than the tagged one; its state vector has that time dependence which ensures that it becomes the correlate of the tagged particle at $t = 0$.

21.2.1 Time-dependent mixing formalism

We denote the neutral meson by B (which will usually be B^0, but could also be K^0 or D^0), and its **CP**-conjugate by $\bar{B}$. According to the theory of Weisskopf and Wigner (1930a, 1930b) (see also appendix A of Kabir 1968) a state that is initially in some superposition of $|B\rangle$ and $|\bar{B}\rangle$, say

$$|\psi(0)\rangle = a(0)|B\rangle + b(0)|\bar{B}\rangle, \tag{21.27}$$

will evolve in time to a general superposition

$$|\psi(t)\rangle = a(t)|B\rangle + b(t)|\bar{B}\rangle \tag{21.28}$$

governed by an effective Hamiltonian **H** with matrix elements, in the 2-state subspace,

$$\mathbf{H} = \mathbf{M} - \mathrm{i}\frac{\mathbf{\Gamma}}{2} = \begin{pmatrix} A & p^2 \\ q^2 & A \end{pmatrix} \tag{21.29}$$

where **M** and **Γ** are Hermitian, and the equality $M_{11} - i\Gamma_{11}/2 = M_{22} - i\Gamma_{22}/2 = A$ follows from **CPT** invariance, which we shall assume. If **CP** is a good symmetry, then

$$\begin{aligned} \langle\bar{B}|\mathbf{H}|B\rangle &= \langle\bar{B}|(\mathbf{CP})^{-1}(\mathbf{CP})\mathbf{H}(\mathbf{CP})^{-1}\mathbf{CP}|B\rangle \\ &= \langle B|\mathbf{H}|\bar{B}\rangle \end{aligned} \tag{21.30}$$

so that p would equal q. Since **M** and **Γ** are both Hermitian, this would imply that M_{12} and Γ_{12} are both real; in the **CP** non-invariant world, this is not the case.

The eigenvalues of **H** are

$$\omega_L \equiv m_L - i\Gamma_L/2 = A + p\,q, \quad \omega_H \equiv m_H - i\Gamma_H/2 = A - p\,q, \tag{21.31}$$

and the corresponding eigenstates are

$$\begin{aligned} |B_L\rangle &= (p\,|B\rangle + q|\bar{B}\rangle)/(|p|^2 + |q|^2)^{1/2} \\ |B_H\rangle &= (p\,|B\rangle - q|\bar{B}\rangle)/(|p|^2 + |q|^2)^{1/2}. \end{aligned} \tag{21.32}$$

The states $|B_L\rangle, |B_H\rangle$ have definite masses m_H, m_L and widths Γ_L and Γ_H. The widths Γ_L, Γ_H are equal to a very good approximation for B and D mesons, because the Q-values of both are large; in the case of K-mesons (see section 21.3), one state decays predominantly to 2π and the other to 3π, with different Q-values, and the lifetimes are very different.

Suppose now that at time $t = 0$ the 'tag' shows that a B^0 has decayed. Then the partner is a $\bar{\mathrm{B}}^0$ at $t = 0$, described by the superposition

$$|\bar{\mathrm{B}}^0\rangle = -\frac{\sqrt{|p|^2 + |q|^2}}{2q}(|\mathrm{B_H}\rangle - |\mathrm{B_L}\rangle). \tag{21.33}$$

At a later time t in the $\bar{\mathrm{B}}^0$ rest-frame, this state evolves to (problem 21.1)

$$|\bar{\mathrm{B}}^0_t\rangle = g_+(t)\,|\bar{\mathrm{B}}^0\rangle + (p/q)g_-(t)|\mathrm{B}^0\rangle \tag{21.34}$$

where

$$g_+(t) = \mathrm{e}^{-\mathrm{i}m_\mathrm{B}t}\mathrm{e}^{-\Gamma t/2}\cos(\Delta m_\mathrm{B}t/2) \tag{21.35}$$

$$g_-(t) = \mathrm{i}\mathrm{e}^{-\mathrm{i}m_\mathrm{B}t}\mathrm{e}^{-\Gamma t/2}\sin(\Delta m_\mathrm{B}t/2) \tag{21.36}$$

with $m_\mathrm{B} = \frac{1}{2}(m_\mathrm{H} + m_\mathrm{L})$ and $\Delta m_\mathrm{B} = m_\mathrm{H} - m_\mathrm{L}$. Note, from (21.34), that the state which started as a $\bar{\mathrm{B}}^0$ at $t = 0$ develops also a B^0 component at a later time. Similarly, if the tag shows that a $\bar{\mathrm{B}}^0$ has decayed, the partner meson at $t = 0$ is a B^0, and its state evolves to

$$|\mathrm{B}^0_t\rangle = (q/p)g_-(t)|\bar{\mathrm{B}}^0\rangle + g_+(t)|\mathrm{B}^0\rangle. \tag{21.37}$$

Consider first the semileptonic decays of B^0 and $\bar{\mathrm{B}}^0$, where the only transitions that can occur are

$$\mathrm{B}^0 \to \ell^+\nu_\ell X, \quad \bar{\mathrm{B}}^0 \to \ell^-\bar{\nu}_\ell X. \tag{21.38}$$

The state $|\bar{\mathrm{B}}^0_t\rangle$ of (21.34), however, which was pure $\bar{\mathrm{B}}^0$ at $t = 0$, will be able to decay to a positively charged lepton via the admixture of the $|\mathrm{B}^0\rangle$ component; similarly negatively charged leptons may appear in the decay of $|\mathrm{B}^0_t\rangle$. From (21.34) and (21.37) we obtain directly the amplitudes for these 'wrong sign' transitions:

$$\langle\ell^-\bar{\nu}_\ell X|\hat{\mathcal{H}}_\mathrm{sl}|\mathrm{B}^0_t\rangle = (q/p)g_-(t)\langle\ell^-\bar{\nu}_\ell X|\hat{\mathcal{H}}_\mathrm{sl}|\bar{\mathrm{B}}^0\rangle \tag{21.39}$$

and

$$\langle\ell^+\nu_\ell X|\hat{\mathcal{H}}_\mathrm{sl}|\bar{\mathrm{B}}^0_t\rangle = (p/q)g_-(t)\langle\ell^+\nu_\ell X|\hat{\mathcal{H}}_\mathrm{sl}|\mathrm{B}^0\rangle, \tag{21.40}$$

where $\hat{\mathcal{H}}_\mathrm{sl}$ is the relevant semileptonic part of the complete weak current–current Hamiltonian. Hence the semileptonic asymmetry is

$$\mathcal{A}_\mathrm{SL} = \frac{\Gamma(\bar{\mathrm{B}}^0_t \to \ell^+\nu_\ell X) - \Gamma(\mathrm{B}^0_t \to \ell^-\bar{\nu}_\ell X)}{\Gamma(\bar{\mathrm{B}}^0_t \to \ell^+\nu_\ell X) + \Gamma(\mathrm{B}^0_t \to \ell^-\bar{\nu}_\ell X)} = \frac{1 - |q/p|^4}{1 + |q/p|^4}, \tag{21.41}$$

independent of time. In (21.41) we have used the fact that $\langle\ell^-\bar{\nu}_\ell X|\hat{\mathcal{H}}_\mathrm{sl}|\bar{\mathrm{B}}^0\rangle = \langle\ell^+\nu_\ell X|\hat{\mathcal{H}}_\mathrm{sl}|\mathrm{B}^0\rangle^*$. The upper bound on $\mathcal{A}_\mathrm{SL}$ is of order 10^{-3} (Nakamura *et al.* 2010). At the present level of experimental precision, it is a very good approximation to take $|q/p| = 1$. Since $q/p = [(M_{12}^* - \mathrm{i}\Gamma_{12}^*/2)/(M_{12} - \mathrm{i}\Gamma_{12}/2)]^{1/2}$, it follows that in this approximation we can neglect Γ_{12}, and the phase of q/p is just minus the phase of M_{12}.

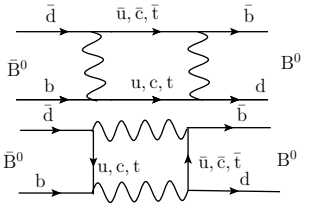


FIGURE 21.5
Box diagram contributions to B^0-$\bar{B}^0$ mixing.

In the Standard Model, the B^0-$\bar{B}^0$ mixing amplitude occurs via the box diagrams of figure 21.5. The box amplitude is approximately proportional to the product of the masses of the internal quarks, and in this case the t quark contribution dominates (the magnitudes of the CKM couplings are all comparable). The phase of M_{12} is then that of $(V_{\rm td}^* V_{\rm tb})^2$, which is the phase of $((1-\rho-{\rm i}\eta)^*)^2$ in the parametrization of (20.179), which in turn is equal to the angle 2β. Hence

$$(q/p) = {\rm e}^{-2{\rm i}\beta}, \tag{21.42}$$

neglecting terms of order λ^4. Equation (21.42) will be important in what follows.

From (21.34) we can now read off that the probability that the state $|\bar{\rm B}^0_t\rangle$ (which – we remind the reader – is the partner of the state tagged as a B^0 at $t=0$, and which is pure $\bar{B}^0$ at $t=0$) decays as a $\bar{B}^0$ at $t\neq 0$, is $|g_+(t)|^2 = \exp(-\Gamma t)\,\cos^2\Delta m_{\rm B}^2 t/2$. Similarly, the probability that this state decays as a B^0 at time t is $\exp(-\Gamma t)\,\sin^2\Delta m_{\rm B} t/2$, taking $|(p/q)|=1$. Hence the difference in these probabilities, normalized to their sum, is $\cos\Delta m_{\rm B} t$. Measurements of this *flavour asymmetry* yield the value of $\Delta m_{\rm B}$, currently (Nakamura *et al.* 2010)

$$\Delta m_{\rm B} = 3.3337 \pm 0.033 \times 10^{-10}\ {\rm MeV}. \tag{21.43}$$

More generally, we define decay amplitudes to final states $|f\rangle$ by

$$A_f = \langle f|\hat{\mathcal{H}}_{\rm wk}|{\rm B}^0\rangle \quad , \quad \bar{A}_f = \langle f|\hat{\mathcal{H}}_{\rm wk}|\bar{\rm B}^0\rangle \tag{21.44}$$

$$A_{\bar{f}} = \langle \bar{f}|\hat{\mathcal{H}}_{\rm wk}|{\rm B}^0\rangle \quad , \quad \bar{A}_{\bar{f}} = \langle \bar{f}|\hat{\mathcal{H}}_{\rm wk}|{\rm B}^0\rangle, \tag{21.45}$$

where ${\rm CP}|f\rangle = |\bar{f}\rangle$ and $\hat{\mathcal{H}}_{\rm wk}$ is the weak interaction Hamiltonian responsible for the transition. We can now calculate the rates for $|\bar{\rm B}^0_t\rangle$ to go to $|f\rangle$, and for $|{\rm B}^0_t\rangle$ to go to $|f\rangle$; up to a common normalization factor, which we omit, these rates are (problem 21.2)

$$\begin{aligned}\Gamma(\bar{\rm B}^0_t \to f) \;=\; & \frac{1}{2}{\rm e}^{-\Gamma t}\{|\bar{A}_f|^2 + |(p/q)A_f|^2 + (|\bar{A}_f|^2 - |(p/q)A_f|^2)\cos\Delta m_{\rm B} t \\ & +\, 2{\rm Im}(\bar{A}_f \frac{q}{p} A_f^*)\sin\Delta m_{\rm B} t\},\end{aligned} \tag{21.46}$$

and

$$\Gamma(\mathrm{B}^0_t \to f) = \frac{1}{2}\mathrm{e}^{-\Gamma t}\{|A_f|^2 + |(q/p)\bar{A}_f|^2 + (|A_f|^2 - |(q/p)\bar{A}_f|^2)\cos\Delta m_{\mathrm{B}} t - 2\mathrm{Im}(\bar{A}_f \frac{q}{p} A^*_f)\sin\Delta m_{\mathrm{B}} t\}. \quad (21.47)$$

The rates to $|\bar{f}\rangle$ are obtained by the substitutions $A_f \to A_{\bar{f}}, \bar{A}_f \to \bar{A}_{\bar{f}}$.

We can now derive the basic formulae for the time-dependent **CP** asymmetry of neutral B decays to a final state f common to B^0 and $\bar{\mathrm{B}}^0$ (problem 21.3):

$$\mathcal{A}_f = \frac{\Gamma(\bar{\mathrm{B}}^0_t \to f) - \Gamma(\mathrm{B}^0_t \to f)}{\Gamma(\bar{\mathrm{B}}^0_t \to f) + \Gamma(\mathrm{B}^0_t \to f)} = S_f \sin(\Delta m_{\mathrm{B}}\, t) - C_f \cos\Delta m_{\mathrm{B}}\, t) \quad (21.48)$$

where

$$S_f = \frac{2\mathrm{Im}\lambda_f}{1+|\lambda_f|^2}, \quad C_f = \frac{1-|\lambda_f|^2}{1+|\lambda_f|^2}, \quad \lambda_f = \frac{q}{p}\left(\frac{\bar{A}_f}{A_f}\right). \quad (21.49)$$

21.2.2 Determination of the angles $\alpha(\phi_2)$ and $\beta(\phi_1)$ of the unitarity triangle

A very large number of measurements have been made, constraining the parameters of the CKM matrix, or equivalently the unitarity triangle of figure 20.10. We shall limit our discussion to those measurements which determine the angles $\alpha(\phi_2)$ and $\beta(\phi_1)$ of the triangle.

(i) The angle β (ϕ_1)

One of the cleanest examples is the decay

$$\mathrm{B}^0 \to J/\psi + \mathrm{K}^0_{\mathrm{S,L}}. \quad (21.50)$$

The tree diagram is shown in figure 21.6(a), and the penguins in figure 21.6(b). The tree diagram contributes CKM factors $V^*_{\mathrm{cb}}V_{\mathrm{cs}} = A\lambda^2(1-\lambda^2/2)$. The $\bar{\mathrm{f}} = \bar{\mathrm{u}}$ penguin has factors $V^*_{\mathrm{ub}}V_{\mathrm{us}} = A\lambda^4(\rho - \mathrm{i}\eta)$ which is suppressed by two powers of λ; it also carries a loop factor $\sim \alpha_{\mathrm{s}}/\pi$, and it may therefore be safely neglected. The other two penguins have the same weak phase as the tree diagram. Hence to a good approximation we can write the amplitude as

$$A_{\psi\mathrm{K}} = (V^*_{\mathrm{cb}}V_{\mathrm{cs}})T_{\psi\mathrm{K}}. \quad (21.51)$$

There is one subtlety: to get the two final states from B^0 and $\bar{\mathrm{B}}^0$ to interfere, we need K^0-$\bar{\mathrm{K}}^0$ mixing to produce the (very nearly) **CP** eigenstates $\mathrm{K}^0_{\mathrm{S}}$ (**CP** $= +1$) and $\mathrm{K}^0_{\mathrm{L}}$ (**CP** $= -1$). (We shall discuss the $\mathrm{K}^0 - \bar{\mathrm{K}}^0$ system briefly in section 21.3.) This introduces a factor $(q/p)_{\mathrm{K}} = (V^*_{\mathrm{cd}}V_{\mathrm{cs}}/V_{\mathrm{cd}}V^*_{\mathrm{cs}})$, quite analogously to (21.42), but its effect on $\lambda_{\psi\mathrm{K}}$ is negligible. So, remembering

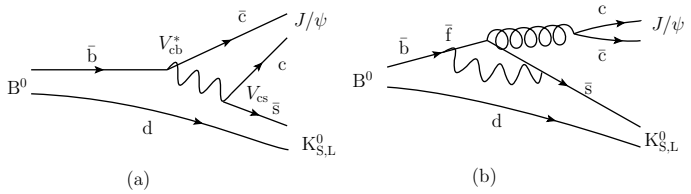


FIGURE 21.6
Tree (a) and penguin (b) contributions to $B^0 \to J/\psi + K^0_{S,L}$ via the quark transitions $\bar{b} \to \bar{c}c\bar{s}$.

that the relative orbital angular momentum of the two final state particles is $\ell = 1$, we have $\lambda_{\psi K_S} = -\exp(-2i\beta)$ and $S_f = \sin 2\beta$, while the $J/\psi K^0_L$ state has **CP**=+1 and $S_f = -\sin 2\beta$. Hence $S_{\psi K}$ measures $-\eta_f \sin 2\beta$, where η_f is the **CP** eigenvalue of the $J/\psi K^0_{S,L}$ state: the sinusoidal oscillations in the asymmetry $\mathcal{A}_{\psi K}$ for the two modes S, L will have the same amplitude and opposite phase.

Both BaBar and Belle have reported increasingly precise measurements of $\mathcal{A}_{\psi K}$ in these modes. The early results (Abashian *et al.* 2001, Aubert *et al.* 2001) were the first direct measurements of one of the angles of the unitarity triangle, offering a test of the consistency of the CKM mechanism for **CP** violation. Later measurements have achieved accuracies of about ±5%. The current world average for $\sin 2\beta$ is (see the review by Ceccucci *et al.* in Nakamura *et al.* 2010)

$$\sin 2\beta = 0.673 \pm 0.023. \tag{21.52}$$

Figure 21.7 shows the asymmetry (before corrections for experimental effects) for $\eta_f = -1$ and $\eta_f = +1$ candidates as measured by BaBar (Aubert *et al.* 2007a); Belle has reported similar results. We should note that a measurement of $\sin 2\beta$ still leaves ambiguities in β (for example, $\beta \to \pi/2 - \beta$), which can be resolved by other measurements (Ceccucci *et al.*, in Nakamura *et al.* 2010).

(ii) The angle $\alpha(\phi_2)$

The angle α is the phase between $V^*_{tb}V_{td}$ and $V^*_{ub}V_{ud}$. It can be measured in decays dominated by the quark transition $b \to u\,\bar{u}\,d$. Consider, for example, the decays $B^0 \to \pi^+\pi^-, \bar{B}^0 \to \pi^+\pi^-$. Figure 21.8 shows the tree graph (a) and penguin (b) contributions to $B^0 \to \pi^+\pi^-$. Exposing the weak phases as before, the amplitude is

$$\begin{aligned} A_{+-} &= V^*_{ub}V_{ud}(t + p_{\bar{u}} - p_{\bar{c}}) + V^*_{tb}V_{td}(p_{\bar{t}} - p_{\bar{c}}) \\ &\equiv V^*_{ub}V_{ud}T_{+-} + V^*_{tb}V_{td}P_{+-}. \end{aligned} \tag{21.53}$$

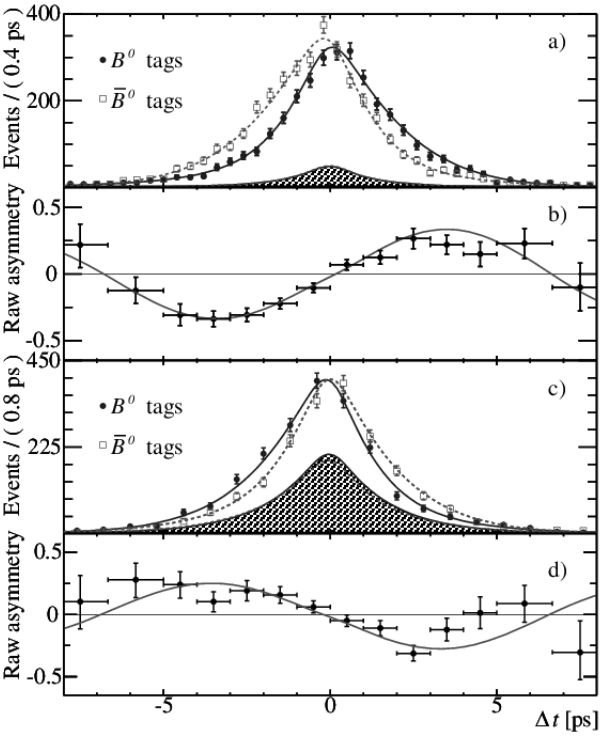


FIGURE 21.7
(a) Number of $\eta_f = -1$ candidates in the signal region with a B^0 tag (N_{B^0}) and with a $\bar{B}^0$ tag ($N_{\bar{B}^0}$), and (b) the measured asymmetry $(N_{B^0} - N_{\bar{B}^0})/(N_{B^0} + N_{\bar{B}^0})$, as functions of t; (c) and (d) are the corresponding distributions for the $\eta_f = +1$ candidates. Figure reprinted with permission from Aubert *et al.* (BaBar Collaboration) *Phys. Rev. Lett.* **99** 171803 (2007). Copyright 2007 by the American Physical Society. (See color plate IV.)

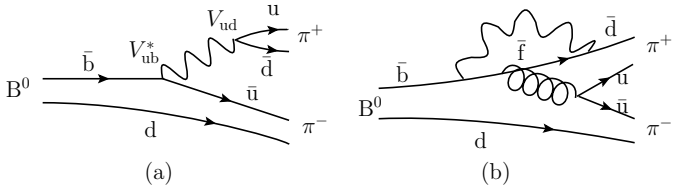


FIGURE 21.8
Tree graph (a) and penguin (b) contributions to $B^0 \to \pi^+\pi^-$, via quark transitions $\bar{b} \to \bar{d}u\bar{u}$.

Suppose first that the penguin contributions could be neglected. Then the asymmetry $\mathcal{A}_{\pi^+\pi^-}$ would measure

$$\begin{aligned} \mathrm{Im}\lambda_{\pi^+\pi^-} &= \mathrm{Im}\left(\mathrm{e}^{-2\mathrm{i}\beta}\frac{\bar{A}_{+-}}{A_{+-}}\right) = \mathrm{Im}\left(\mathrm{e}^{-2\mathrm{i}\beta}\frac{V_{\mathrm{ub}}V_{\mathrm{ud}}^*}{V_{\mathrm{ub}}^*V_{\mathrm{ud}}}\right) \\ &= \mathrm{Im}\,\mathrm{e}^{-2\mathrm{i}(\gamma+\beta)} = \sin 2\alpha \end{aligned} \tag{21.54}$$

where α is defined as $\pi-\beta-\gamma$. Unfortunately, this simple result is spoiled by the penguin contributions, which there is no good reason to ignore. However, Gronau and London (1990) showed how an isospin analysis could disentangle the tree and penguin parts. The method involves the three amplitudes $A_{+-}, A_{00}(\mathrm{B}^0\to\pi^0\pi^0)$, and $A_{+0}(\mathrm{B}^+\to\pi^+\pi^0)$.

First of all, note that Bose statistics for the 2π states requires them to have only the symmetric isospin states $I=0$ or 2, since the angular momentum is zero. Next, the effective non-leptonic weak Hamiltonian $\hat{\mathcal{H}}_{\mathrm{nl}}$ acting in the tree diagram transition contains both $\Delta I=1/2$ and $\Delta I=3/2$ pieces; combining with the initial $I=1/2$ of the B meson, the first piece will lead only to the $I=0$ final state, while the second contributes to both $I=0$ and $I=2$ final states. However, since the gluon in the penguin diagrams carries no isospin, these diagrams can only change the isospin by $\Delta I=1/2$, which connects only to the $I=0$ final state. The conclusion is that the $I=2$ final state is free of penguins, and carries the pure tree phase.

This information can be exploited as follows (Gronau and London 1990). First, the action of $\hat{\mathcal{H}}_{\mathrm{nl}}$ on the B^0 state can be written as

$$\hat{\mathcal{H}}_{\mathrm{nl}}|\tfrac{1}{2}\,-\tfrac{1}{2}\rangle = \frac{1}{\sqrt{2}}A_{3/2}|20\rangle + \frac{1}{\sqrt{2}}A_{1/2}|00\rangle \tag{21.55}$$

where as usual $|II_3\rangle$ is the state with isospin I and third component I_3. Expanding the states $\pi^+\pi^-, \pi^+\pi^0$ and $\pi^0\pi^0$ in terms of definite isospin states, one finds (problem 21.4)

$$\begin{aligned} A_{+-} &= \frac{1}{\sqrt{6}}A_{3/2} + \frac{1}{\sqrt{3}}A_{1/2} \\ A_{+0} &= \frac{\sqrt{3}}{2}A_{3/2} \\ A_{00} &= \frac{1}{\sqrt{3}}A_{3/2} - \frac{1}{\sqrt{6}}A_{1/2} \end{aligned} \tag{21.56}$$

where A_{ij} is the amplitude $\langle\pi^i\pi^j|\hat{\mathcal{H}}_{\mathrm{nl}}|B^{i+j}\rangle$. The $\pi^+\pi^0$ state can have only $I=2$, and arises solely from the tree diagram. Furthermore, the three complex amplitudes A_{+-}, A_{+0} and A_{00} are expressed in terms of only two reduced amplitudes $A_{3/2}$ and $A_{1/2}$, leading to one relation between them:

$$\frac{1}{\sqrt{2}}A_{+-} + A_{00} = A_{+0}, \tag{21.57}$$

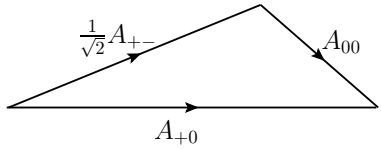


FIGURE 21.9
The triangle formed by the three amplitudes A_{ij} in equation (21.57).

which can be represented as a triangle in the complex plane, as shown in figure 21.9. There is a similar triangle for the charge conjugate processes

$$\frac{1}{\sqrt{2}}\bar{A}_{+-} + \bar{A}_{00} = \bar{A}_{+0}, \tag{21.58}$$

where the $\bar{A}$ amplitudes are obtained from the As by complex conjugating the CKM couplings, the strong phases remaining the same as usual.

Since A_{+0} is pure tree, its weak phase is well defined, namely that of $V_{\rm ub}^* V_{\rm ud}$, which is γ. It is convenient to define (Lipkin *et al.* 1991) $\tilde{A} = \exp(2{\rm i}\gamma)\bar{A}$, so that $\tilde{A}_{+0} = A_{+0}$. Then the two triangles have a common base, A_{+0}. The failure of the two triangles to overlap exactly is a measure of the penguin contribution. In principle, by measuring the asymmetry coefficients $S_{\pi^+\pi^-}, C_{\pi^+\pi^-}$, the branching fractions of all three modes, and C_{00}, one can construct the triangles. But unfortunately the relative orientation of the triangles is not known, which leads to various possible solutions to α in the range $0 < \alpha < 2\pi$. In addition, the data on $\pi^0\pi^0$ (with a branching ratio of order 10^{-6}) has sizeable experimental errors, and only a relatively loose constraint on α can be obtained.

A much better constraint can be found from the **CP** asymmetries in $\mathrm{B} \to \pi\rho$ decays (Snyder and Quinn 1993). The method is essentially a time-dependent version of the Dalitz plot analysis discussed in section 21.1. The available channels are

$$\begin{aligned} \mathrm{B}^0 &\to \{\rho^+\pi^-, \rho^-\pi^+, \rho^0\pi^0\} \to \pi^+\pi^-\pi^0 \\ \bar{\mathrm{B}}^0 &\to \{\rho^-\pi^+, \rho^+\pi^-, \rho^0\pi^0\} \to \pi^+\pi^-\pi^0 \end{aligned} \tag{21.59}$$

where all result in the final state $\pi^+\pi^-\pi^0$ after the decay of the ρ mesons, and interferences following B^0-$\bar{\mathrm{B}}^0$ mixing are possible.

Returning then to equations (21.46) and (21.47), the rate for the 3π decay following a B^0 tag at $t = 0$ is

$$\begin{aligned} \Gamma(\bar{\mathrm{B}}^0_t \to 3\pi) \;=\; & \frac{1}{4}\Gamma {\rm e}^{-\Gamma t}\Bigg[|A_{3\pi}|^2 + |\bar{A}_{3\pi}|^2 + (|\bar{A}_{3\pi}|^2 - |A_{3\pi}|^2)\cos\Delta m_{\rm B} t \\ & + 2{\rm Im}\left(\frac{q}{p}\bar{A}_{3\pi}A^*_{3\pi}\right)\sin\Delta m_{\rm B} t\Bigg], \end{aligned} \tag{21.60}$$

and there is a similar formula, with appropriate changes, for the case of a $\bar{\mathrm{B}}^0$ tag at $t = 0$. We now write

$$A_{3\pi} = f_+(s_+)F^+ + f_-(s_-)F^- + f_0(s_0)F^0 \tag{21.61}$$

and similarly

$$\bar{A}_{3\pi} = f_+(s_+)\bar{F}^+ + f_-(s_-)\bar{F}^- + f_0(s_0)\bar{F}^0, \tag{21.62}$$

where $s_+ = (p_{\pi^+} + p_{\pi^0})^2, s_- = (p_{\pi^-} + p_{\pi^0})^2, s_0 = (p_{\pi^+} + p_{\pi^-})^2$, satisfying $s_+ + s_- + s_0 = m_{\mathrm{B}}^2 + 2m_{\pi^+}^2 + m_{\pi^0}^2$. $f_\kappa(s_\kappa)$ is the sum of three relativistic Breit-Wigner resonance amplitudes, together with appropriate angular momentum and angle factors, corresponding to the $\rho(770), \rho(1450)$ and $\rho(1700)$ resonances. F^κ is the amplitude for the quasi two-body mode $\mathrm{B}^0 \to \rho^\kappa \pi^{\bar{\kappa}}$. Here κ takes the values $+, -$ and 0, and correspondingly $\bar{\kappa} = -, +, 0$. The amplitudes F^κ are complex and include the strong and weak transition phases, from tree and penguin diagrams; they are, however, independent of the Dalitz plot variables.

The $\rho\pi$ states have the same decomposition into tree and penguin parts as discussed previously for the $\pi\pi$ states, namely

$$F^\kappa = \mathrm{e}^{\mathrm{i}\gamma} T^\kappa + \mathrm{e}^{-\mathrm{i}\beta} P^\kappa, \tag{21.63}$$

where the magnitudes of the weak couplings have been absorbed into T^κ and P^κ. We can rewrite (21.63) as

$$\mathrm{e}^{\mathrm{i}\beta} F^\kappa = -\mathrm{e}^{-\mathrm{i}\alpha} T^\kappa + P^\kappa \equiv A^\kappa, \tag{21.64}$$

and similarly

$$\mathrm{e}^{\mathrm{i}\beta}(q/p)\bar{F}^\kappa = -\mathrm{e}^{\mathrm{i}\alpha} T^{\bar{\kappa}} + P^{\bar{\kappa}} \equiv \bar{A}^\kappa. \tag{21.65}$$

Then (21.61) and (21.62) become

$$A_{3\pi} = \sum_\kappa f_\kappa(s_\kappa) A^\kappa \tag{21.66}$$

$$(q/p)\bar{A}_{3\pi} = \sum_\kappa f_\kappa(s_\kappa) \bar{A}^\kappa, \tag{21.67}$$

disregarding a common overall phase $\mathrm{e}^{-\mathrm{i}\beta}$. When (21.66) and (21.67) are inserted into (21.60), it is clear that one obtains many terms, for example

$$\mathrm{Re}(f_+ f_-^*)\,\mathrm{Im}(\bar{A}^+ A^{-*} + \bar{A}^- A^{+*}), \quad \mathrm{Im}(f_+ f_-^*)\,\mathrm{Re}(\bar{A}^+ A^{-*} - \bar{A}^- A^{+*}), \tag{21.68}$$

and so on, in which different resonances interfere on the Dalitz plot. The strong, and known, rapid phase variation in these interference regions, via factors such as $f_+ f_-^*$, is a powerful tool for extracting the complex amplitudes $A^\kappa, \bar{A}^\kappa$, and hence via (21.64) and (21.65) the phase α. The quantities multiplying the interference terms $\mathrm{Re}(f_\kappa f_\sigma^*)$ and $\mathrm{Im}(f_\kappa f_\sigma^*)$ are the key degrees of

freedom which allow this analysis to determine the penguin contributions and the strong phases, and hence α. However, the resonance overlap regions cover a small fraction of the Dalitz plot, so that a substantial data sample (a few thousand events) is needed to constrain all the amplitude parameters.

An isospin analysis similar to that of the $\pi\pi$ states can be done for the $\rho\pi$ states, but now there is no reason to forbid the final state to have $I = 1$. Nevertheless, if charged B decays are also included, there are five physical amplitudes ($\rho^0 \to \pi^+\pi^-, \pi^-\pi^+, \pi^0\pi^0, \rho^+ \to \pi^+\pi^0, \rho^- \to \pi^-\pi^0$) which are expressible in terms of two pure tree ($\Delta I = 3/2$) transitions to $I = 1, 2$ final states. One of the pure tree amplitudes may be written (Gronau 1991) as the sum $A^+ + A^- + 2A^0$, and hence the ratio $(\bar{A}^+ + \bar{A}^- + 2\bar{A}^0)/(A^+ + A^- + 2A^0)$ has the phase 2α.

This approach has been followed by both BaBar and Belle, with the results

$$\text{BaBar (Aubert } \textit{et al.} \text{ 2007}b) \qquad \alpha = 87^{+45^\circ}_{-13^\circ} \tag{21.69}$$

$$\text{Belle (Kusaka } \textit{et al.} \text{ 2007)} \qquad 68 < \alpha < 95^\circ. \tag{21.70}$$

These results are consistent with the values of β and γ given in (21.52), (21.25) and (21.26), given the definition $\alpha = \pi - \beta - \gamma$.

Of course, this is only one (at present not very tight) consistency check. But there are now very many independent measurements of the magnitudes of the CKM matrix elements, as well as the angles. We shall not describe these here, referring the reader to the regular updates by the Particle Data Group (currently Nakamura *et al.* 2010). We showed in figure 20.11 the 2010 plot of the contraints in the $\bar{\rho}, \bar{\eta}$ plane, presented by Ceccucci *et al.*. They concluded that the 95% CL regions all overlapped consistently around the global fit region, though the consistency of $|V^{\rm ub}|$ and $\sin 2\beta$ was not very good. It would be premature to make too much of the minor reservation, though it may be noted that $\sin 2\beta$ could be sensitive to new physics via short-distance corrections to the box diagrams of figure 21.5, while $|V_{\rm ub}|$ is obtained from a tree-level process, and is thus unlikely to be affected by new physics. Overall, the consistency represented in figure 20.11 must be counted as a major triumph of the Standard Model, in particular of the original analysis by Kobayashi and Maskawa (1973). It must be remembered, though, that many extensions of the Standard Model allow considerable room for new **CP**-violating effects, which could be revealed by increasingly precise determinations of the CKM parameters.

21.3 CP violation in neutral K-meson decays

Although the formalism is similar, the phenomenology of **CP** violation in neutral K-meson decays is very different from that in neutral B-meson decays. In the K case, **CP** violation is a very small effect, typically at the level of

parts per thousand or smaller; its observation by Cristenson *et al.* (1964) was a historic achievement. But the neutral K system is most simply (and traditionally) approached by starting with the assumption that **CP** is conserved.

We will define $\mathbf{CP}|\mathrm{K}^0\rangle = -|\bar{\mathrm{K}}^0\rangle$; then the **CP** eigenstates are

$$|\mathrm{K}_\pm\rangle = \frac{1}{\sqrt{2}}(|\mathrm{K}^0\rangle \mp |\bar{\mathrm{K}}^0\rangle \tag{21.71}$$

The $\mathbf{CP} = 1$ state can decay to two pions in an s-state, but not to three pions if (as we are assuming to start with) **CP** is a good symmetry; the situation is the opposite for the $\mathbf{CP} = -1$ state. The Q-value for the three pion mode is very much smaller than for the two pion mode, with the result that the $|\mathrm{K}_+\rangle$ state, decaying to two pions, has a much shorter lifetime than the $|\mathrm{K}_-\rangle$ state: $\tau_{2\pi} \sim 0.9 \times 10^{-10} s$, $\tau_{3\pi} \sim 5 \times 10^{-8} s$. Due to **CP** violation, the actual eigenstates $|\mathrm{K_L}\rangle$ and $|\mathrm{K_S}\rangle$ of the effective Hamiltonian are slightly different from $|\mathrm{K}_\pm\rangle$ (see (21.75) and (21.76)), with masses m_S and m_L, and widths Γ_S and Γ_L. At this point, however, we shall associate m_S and Γ_S with $|\mathrm{K}_+\rangle$, and m_L and Γ_L with $|\mathrm{K}_-\rangle$.

A K^0 is produced in strangeness-conserving reactions such as $\mathrm{K}^+\mathrm{n} \to \mathrm{K}^0\mathrm{p}$, and a $\bar{\mathrm{K}}^0$ in $\mathrm{K}^- + \mathrm{p} \to \bar{\mathrm{K}}^0 + \mathrm{n}$, for example. However, the two states can mix following production, since (as usual) it is the Hamiltonian eigenstates which propagate in free space, and they are the superpositions $|\mathrm{K}_\pm\rangle$, assuming **CP** is conserved. Hence, as time proceeds following production, a state produced as a K^0 at time $t = 0$ will evolve into the state

$$|\mathrm{K}^0_t\rangle = \frac{1}{2}(\mathrm{e}^{-\Gamma_\mathrm{L} t/2 - \mathrm{i} m_\mathrm{L} t} + \mathrm{e}^{-\Gamma_\mathrm{S} t/2 - \mathrm{i} m_\mathrm{S} t})|\mathrm{K}^0\rangle + (\mathrm{e}^{-\Gamma_\mathrm{L} t/2 - \mathrm{i} m_\mathrm{L} t} - \mathrm{e}^{-\Gamma_\mathrm{S} t/2 - \mathrm{i} m_\mathrm{S} t})|\bar{\mathrm{K}}^0\rangle. \tag{21.72}$$

The probability that a $\mathrm{K}^0(\bar{\mathrm{K}}^0)$ will then be observed at time t following production (in the K-meson rest frame) is

$$P_{+(-)} = \frac{1}{4}[\mathrm{e}^{-\Gamma_\mathrm{L} t} + \mathrm{e}^{-\Gamma_\mathrm{S} t} + (-)2\mathrm{e}^{-(\Gamma_\mathrm{L} + \Gamma_\mathrm{S})t/2} \cos \Delta m t] \tag{21.73}$$

where $\Delta m = m_\mathrm{L} - m_\mathrm{S}$. This is the famous phenomenon of strangeness oscillations, predicted by Gell-Mann and Pais (1955). Experimentally, the strangeness of the state at time t is defined by the modes $\mathrm{K}^0 \to \pi^- \ell^+ \nu_\ell$ and $\bar{\mathrm{K}}^0 \to \pi^+ \ell^- \bar{\nu}_\ell$. The difference $P_+(t) - P_-(t)$ is measured, and although the oscillations are heavily damped by $\exp(-\Gamma_\mathrm{S} t)$, the mass difference can be determined:

$$\Delta m = (3.483 \pm 0.006) \times 10^{-12} \text{ meV}. \tag{21.74}$$

However, this is not the whole story. Christenson *et al.* (1964) found that, after many τ_S lifetimes, some 2π events were observed, indicating that the surviving state $\mathrm{K_L}$ was capable of decaying to 2π after all (albeit very rarely). The same conclusion follows from the fact that $P_+(t) - P_-(t)$ does not go to zero at long times, as it should from (21.73). Accordingly, the true

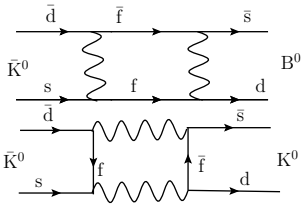


FIGURE 21.10
Box diagrams contributing to K^0-$\bar{\mathrm{K}}^0$ mixing.

Hamiltonian eigenstates are not quite the **CP** eigenstates, but rather

$$|\mathrm{K_L}\rangle = [(1+\bar\epsilon)|\mathrm{K}^0\rangle + (1-\bar\epsilon)|\bar{\mathrm{K}}^0\rangle]/\sqrt{2(1+|\bar\epsilon|^2)} \tag{21.75}$$

$$|\mathrm{K_S}\rangle = [(1+\bar\epsilon)|\mathrm{K}^0\rangle - (1-\bar\epsilon)|\bar{\mathrm{K}}^0\rangle]/\sqrt{2(1+|\bar\epsilon|^2)}. \tag{21.76}$$

This is a traditional parametrization in K-physics, similar to that in (21.54) with $q/p = (1-\bar\epsilon)/(1+\bar\epsilon)$ (this is why we chose **CP** $|\mathrm{K}^0\rangle = -|\mathrm{K}^0\rangle$). We now find that a state which starts at $t=0$ as a K^0 evolves to

$$|\mathrm{K}^0_t\rangle = g_+(t)|\mathrm{K}^0\rangle + \frac{1-\bar\epsilon}{1+\bar\epsilon}\, g_-(t)|\bar{\mathrm{K}}^0\rangle \tag{21.77}$$

where

$$g_\pm(t) = \mathrm{e}^{-\Gamma_\mathrm{L} t/2}\mathrm{e}^{-\mathrm{i}m_\mathrm{L} t}[1 \pm \mathrm{e}^{-\Delta\Gamma t/2}\mathrm{e}^{\mathrm{i}\Delta m t}], \tag{21.78}$$

with $\Delta\Gamma = \Gamma_\mathrm{S} - \Gamma_\mathrm{L}$, $\Delta m = m_\mathrm{L} - m_\mathrm{S}$, and we have omitted a normalization factor. Similarly, a state tagged as $\bar{\mathrm{K}}^0$ at $t=0$ evolves to

$$|\bar{\mathrm{K}}^0_t\rangle = \frac{1-\bar\epsilon}{1+\bar\epsilon} g_-(t)|\mathrm{K}^0\rangle + g_+(t)|\bar{\mathrm{K}}^0\rangle. \tag{21.79}$$

The K^0-$\bar{\mathrm{K}}^0$ mixing amplitude arises in the Standard Model from the box graphs shown in figure 21.9 (cf figure 21.5). These contain factors of m_f^2, but the magnitude of the four CKM couplings to the t quark are of order λ^{10}, compared with λ^6 for the c quark, so that the c quark diagram dominates, with a CKM factor of $(V_\mathrm{cs}V_\mathrm{cd}^*)^2$, which is real to a good approximation. This means that $\mathrm{Im}\bar\epsilon$ is very small. A comparison of the mass difference Δm predicted from figure 21.10 and the experimental value is complicated by uncertainties in the hadronic matrix element.

The traditional reactions in which **CP** violation is probed in K decays are the 2π modes, where one looks for the existence of $\mathrm{K_L} \to 2\pi$. There is also the semileptonic asymmetry. Three common observables are defined by

$$\eta_{00} = \frac{\langle\pi^0\pi^0|\hat{\mathcal{H}}_\mathrm{nl}|\mathrm{K_L}\rangle}{\langle\pi^0\pi^0|\hat{\mathcal{H}}_\mathrm{nl}|\mathrm{K_S}\rangle}, \quad \eta_{+-} = \frac{\langle\pi^+\pi^-|\hat{\mathcal{H}}_\mathrm{nl}|\mathrm{K_L}\rangle}{\langle\pi^+\pi^-|\hat{\mathcal{H}}_\mathrm{nl}|\mathrm{K_S}\rangle} \tag{21.80}$$

and

$$\delta_{\rm L} = \frac{\Gamma(\mathrm{K_L} \to \pi^-\ell^+\nu_\ell) - \Gamma(\mathrm{K_L} \to \pi^+\ell^-\bar{\nu_\ell})}{\Gamma(\mathrm{K_L} \to \pi^-\ell^+\nu_\ell) + \Gamma(\mathrm{K_L} \to \pi^+\ell^-\bar{\nu}_\ell)}. \tag{21.81}$$

The experimental numbers are (Nakamura *et al.* 2010)

$$|\eta_{00}| = (2.221 \pm 0.011) \times 10^{-3}, \quad |\eta_{+-}| = (2.232 \pm 0.011) \times 10^{-3} \tag{21.82}$$

$$\mathrm{Arg}\,\eta_{00} \approx 43.5^\circ, \quad \mathrm{Arg}\,\eta_{\pm} \approx 43.5^\circ \tag{21.83}$$

and

$$\delta_{\rm L} = (3.32 \pm 0.06) \times 10^{-3}. \tag{21.84}$$

It is useful to describe the final 2π states in terms of their isospin, which then have a definite strong interaction phase. As noted in connection with the B decays, the allowed isospin states are only $I = 0$ and $I = 2$, and one has

$$A_\pm \equiv A_{\mathrm{K}^0 \to \pi^+\pi^-} = \sqrt{\frac{2}{3}}\,|A_0|\mathrm{e}^{\mathrm{i}(\delta_0+\phi_0)} + \sqrt{\frac{1}{3}}\,|A_2|\mathrm{e}^{\mathrm{i}(\delta_2+\phi_2)} \tag{21.85}$$

$$\bar{A}_\pm \equiv A_{\bar{\mathrm{K}}^0 \to \pi^+\pi^-} = -\sqrt{\frac{2}{3}}\,|A_0|\mathrm{e}^{\mathrm{i}(\delta_0-\phi_0)} - \sqrt{\frac{1}{3}}\,|A_2|\mathrm{e}^{\mathrm{i}(\delta_2-\phi_2)} \tag{21.86}$$

where the minus sign arises from our choice $\mathbf{CP}|\mathrm{K}^0\rangle = -|\mathrm{K}^0\rangle$, and where δ_I and ϕ_I are the strong and weak phases, respectively, for the state with isospin I. Also,

$$A_{00} \equiv A_{\mathrm{K}^0 \to \pi^0\pi^0} = \sqrt{\frac{1}{3}}\,|A_0|\mathrm{e}^{\mathrm{i}(\delta_0+\phi_0)} - \sqrt{\frac{2}{3}}\,|A_2|\mathrm{e}^{\mathrm{i}(\delta_2+\phi_2)} \tag{21.87}$$

$$\bar{A}_{00} \equiv \bar{A}_{\bar{\mathrm{K}}^0 \to \pi^0\pi^0} = -\sqrt{\frac{1}{3}}\,|A_0|\mathrm{e}^{\mathrm{i}(\delta_0-\phi_0)} + \sqrt{\frac{2}{3}}\,|A_2|\mathrm{e}^{\mathrm{i}(\delta_2-\phi_2)}. \tag{21.88}$$

The significant fact experimentally is that $|A_2|/|A_0| \sim 1/22$, a manifestation of the '$\Delta I = 1/2$' rule in this case (i.e. $\Delta I = 3/2$ is suppressed; see, for example, Donoghue *et al.* 1992, section VIII-4). Inserting (21.85) and (21.86), (21.87) and (21.88), into (21.80) and retaining only first-order terms in $|A_2|/|A_0|$, and treating ϕ_0 and ϕ_2 as small, we find (problem 21.5)

$$\eta_{00} = \bar{\epsilon} + \mathrm{i}\phi_0 - \sqrt{2}\,\frac{|A_2|}{|A_0|}\mathrm{i}(\phi_2 - \phi_0)\mathrm{e}^{\mathrm{i}(\delta_2-\delta_0)} \tag{21.89}$$

$$\eta_{+-} = \bar{\epsilon} + \mathrm{i}\phi_0 + \frac{1}{\sqrt{2}}\,\frac{|A_2|}{|A_0|}\mathrm{i}(\phi_2 - \phi_0)\mathrm{e}^{\mathrm{i}(\delta_2-\delta_0)}. \tag{21.90}$$

These relations are usually written as

$$\eta_{00} = \epsilon - 2\epsilon', \quad \eta_{+-} = \epsilon + \epsilon', \tag{21.91}$$

where

$$\epsilon = \bar{\epsilon} + \mathrm{i}\phi_0, \quad \epsilon' = \mathrm{i}\frac{\mathrm{e}^{\mathrm{i}(\delta_2-\delta_0)}}{\sqrt{2}}\,\frac{|A_2|}{|A_0|}(\phi_2 - \phi_0). \tag{21.92}$$

The merit of this manoeuvering is that the parameter ϵ' involves only the **CP** violation in the transition amplitude ('direct **CP** violation'), while ϵ involves both a transition phase and the mixing parameter $\bar{\epsilon}$.

What can experiment tell us about ϵ and ϵ'? Consider first $\delta_{\rm L}$. Assuming that $|A(\mathrm{K}^0 \to \ell^+\nu_\ell\pi^-)| = |A(\bar{\mathrm{K}}^0 \to \ell^-\bar{\nu}_\ell\pi^+)|$ and that $A(\mathrm{K}^0 \to \ell^-\bar{\nu}_\ell\pi^+) = A(\bar{\mathrm{K}}^0 \to \ell^+\nu_\ell\pi^-) = 0$, we find

$$\delta_{\rm L} = 2\mathrm{Re}\,\bar{\epsilon}/(1+|\bar{\epsilon}|^2) \approx 2\mathrm{Re}\,\epsilon, \tag{21.93}$$

so that $\delta_{\rm L}$ is sensitive to the same parameter as appears in the $\mathrm{K_L} \to \pi\pi$ decays. An interesting observable is the ratio between the ratios of the decay rates to $\pi^+\pi^-$ and $\pi^0\pi^0$ of $\mathrm{K_S}$ and $\mathrm{K_L}$. One finds

$$\frac{1}{6}\left(1 - \frac{|\eta_{00}|^2}{|\eta_{+-}|^2}\right) \approx \mathrm{Re}\,(\epsilon'/\epsilon), \tag{21.94}$$

which from equation (21.82) is another small number, approximately equal to 1.64×10^{-3}. In the years before the B factories opened, ϵ' was the only window into **CP** violation in the transition amplitude. But all the branching ratios in (21.94) are of order 10^{-3}, and establishing a non-zero value of ϵ' was very difficult. The first claim for non-zero ϵ' was by the NA 31 experiment at CERN (Barr *et al.* 1993), a 3.5 standard deviation effect. But a contemporary experiment at Fermilab (Gibbons *et al.* 1993) found a result compatible with zero. The next generation of experiments produced agreement:

$$\mathrm{Re}\,(\epsilon'/\epsilon) = (2.07) \pm 0.28) \times 10^{-3} \text{ Alavi-Harati } et\ al.\ 2003 \text{ (KTeV)} \tag{21.95}$$

$$\mathrm{Re}\,(\epsilon'/\epsilon) = (1.47 \pm 0.22) \times 10^{-3} \text{ Batley } et\ al.\ 2002 \text{ (NA 48)}. \tag{21.96}$$

The current world average is $(1.65 \pm 0.26) \times 10^{-3}$. Fits to all the data also yield (Nakamura *et al.* 2010)

$$|\epsilon| = (2.228 \pm 0.011) \times 10^{-3}. \tag{21.97}$$

The experimental value of $\delta_{\rm L}$ gives us $\mathrm{Re}\,\epsilon \simeq 1.66 \times 10^{-3}$, and we can deduce that $\arg\epsilon \simeq \pi/4$. The phase of ϵ' is $\pi/2 + \delta_2 - \delta_0$ which happens also to be approximately $\pi/4$. It follows that ϵ'/ϵ is very nearly real.

Comparison of these small numbers with theoretical predictions is complicated by hadronic uncertainties, and it is beyond our scope to pursue that issue.

In closing this discussion of mesonic mixing and **CP** violation, we briefly discuss the charm sector. First, we note that D^0-$\bar{\mathrm{D}}^0$ mixing has been observed (Aubert *et al.* 2007c, Staric *et al.* 2007, Aaltonen *et al.* 2008). **CP**-violating effects in charm decays have been generally expected to be very small. A rough estimate of the direct **CP**-violating asymmetries in D decays can be made following the method of section 21.1. Consider, for example, the decays

$\mathrm{D}^0 \to \mathrm{K}^+\mathrm{K}^-$ and $\bar{\mathrm{D}}^0 \to \mathrm{K}^-\mathrm{K}^+$. As in (21.5) and (21.10) , the amplitude for the first decay is

$$A(\mathrm{D}^0 \to \mathrm{K}^+\mathrm{K}^-) = V_{\mathrm{cs}}^* V_{\mathrm{us}} T_{\mathrm{KK}} + V_{\mathrm{cb}}^* V_{\mathrm{ub}} P_{\mathrm{KK}} \quad (21.98)$$

$$= T(1 + r_{\mathrm{K}} \exp^{\mathrm{i}(\delta_{\mathrm{K}} - \gamma)}), \quad (21.99)$$

where r_{K} is the relative magnitude of the penguin contribution, and δ_{K} is the relative strong phase. The amplitude for the **CP**-conjugate process is the same, with γ replaced by $-\gamma$. The penguin contribution is CKM-suppressed by a factor $V_{\mathrm{cb}}^* V_{\mathrm{ub}} / V_{\mathrm{cs}}^* V_{\mathrm{us}} \sim \lambda^4$, and there is also a loop factor, so that r_{K} would seem to be of order 10^{-4}. The asymmetry is then

$$\mathcal{A}_{\mathrm{KK}}^{\mathrm{D}} = \frac{|A(\mathrm{D}^0 \to \mathrm{K}^+\mathrm{K}^-)|^2 - |A(\bar{\mathrm{D}}^0 \to \mathrm{K}^-\mathrm{K}^+)|^2}{|A(\mathrm{D}^0 \to \mathrm{K}^+\mathrm{K}^-)|^2 + |A(\bar{\mathrm{D}}^0 \to \mathrm{K}^-\mathrm{K}^+)|^2} \quad (21.100)$$

$$= 2 r_{\mathrm{K}} \sin\gamma \sin\delta_{\mathrm{K}}, \quad (21.101)$$

which is indeed very small. A similar argument predicts the asymmetry in the decays $\mathrm{D}^0 \to \pi^+\pi^-$ and $\bar{\mathrm{D}}^0 \to \pi^-\pi^+$ to be

$$\mathcal{A}_{\pi\pi}^{\mathrm{D}} = -2 r_{\mathrm{K}} \sin\gamma \sin\delta_{\mathrm{K}}. \quad (21.102)$$

Recently, however, the LHCb collaboration has published a measurement of the difference between the time-integrated **CP** asymmetries in the KK and $\pi\pi$ decays, which to a very good approximation can be identified with the difference between the direct asymmetries (21.101) and (21.102). The LHCb result is (Aaij *et al.* 2012)

$$\mathcal{A}_{\mathrm{KK}}^{\mathrm{D}} - \mathcal{A}_{\pi\pi}^{\mathrm{D}} = (-0.82 \pm 0.21 \pm 0.11)\%, \quad (21.103)$$

which is substantially larger than the estimates (21.101) and (21.102).

It is possible that this 3.5 σ effect (the first evidence for **CP**-violation in the charm sector) indicates the presence of some new physics. However, it must be noted that the mass scale of the charm quark, $m_{\mathrm{c}} \sim 1.3$ GeV, is not large enough to be safely in the perturbative QCD regime (as indicated by the parameter $\Lambda_{\overline{\mathrm{MS}}}/m_{\mathrm{c}}$), so that non-perturbative enhancements are possible. **CP**-violation in the charm sector promises to be an interesting area for experimental and theoretical exploration.

21.4 Neutrino mixing and oscillations

21.4.1 Neutrino mass and mixing

Experiments with solar, atmospheric, reactor and accelerator neutrinos have established the phenomenon of neutrino oscillations caused by non-zero neutrino masses, and mixing. We shall give an elementary introduction to this

topic, which is a highly active field of research in particle physics; there are analogies with the meson oscillations we have been considering.

It is fair to say that in the original Standard Model the neutrinos were taken to be massless, but there was no compelling theoretical reason for this, and the framework of the Standard Model can easily be extended to include massive neutrinos. However, one question immediately arises: are neutrinos Dirac or Majorana fermions? As explained in section 20.3, we do not yet know the answer, and it may be some time before we do. The way the mass terms enter the Lagrangian is, in fact, different in the two cases. We are familiar with the Dirac mass term

$$m\bar{\hat{\psi}}\hat{\psi} = m(\bar{\hat{\psi}}_{\mathrm{R}}\hat{\psi}_{\mathrm{L}} + \bar{\hat{\psi}}_{\mathrm{L}}\hat{\psi}_{\mathrm{R}}), \tag{21.104}$$

where $\hat{\psi}$ is a four-component Dirac field, and R and L refer to the chirality components. We learned in section 7.5.2 that a Majorana mass term can be written in the form

$$m\hat{\chi}_{\mathrm{L}}^{\mathrm{T}}\mathrm{i}\sigma_2\hat{\chi}_{\mathrm{L}} + \mathrm{h.c.} \tag{21.105}$$

where $\hat{\chi}_{\mathrm{L}}$ is a two-component field of L chirality. A similar expression could be written using a two-component R-chirality field. The difference in form between the Dirac and Majorana mass terms leads to a difference in the parametrization of neutrino mixing, as we shall see.

Suppose, first, that the neutrinos are Dirac particles, with both L and R chiralities (or equivalently either helicity) for a given mass. We remind the reader that this is not ruled out experimentally, since the non-observation of the 'wrong' helicity component may be accounted for by the appearance of a suppression factor (m/E), where m is a neutrino mass and E is an average neutrino energy (see section 20.2.2). We also assume that their interactions have the V-A structure indicated by the phenomenology of the previous chapter. Then only the L (R) chirality component of a neutrino (antineutrino) field feels the weak force; the R (L) component of a neutrino (antineutrino) field has no interactions of Standard Model type. But, just as in the quark case, it will in general be necessary to allow for the possibility that the L-components of the fields which have definite neutrino mass, call them $\hat{\nu}_{1\mathrm{L}}, \hat{\nu}_{2\mathrm{L}}, \hat{\nu}_{3\mathrm{L}}$, are not the same as the fields $\hat{\nu}_{\mathrm{eL}}, \hat{\nu}_{\mu\mathrm{L}}, \hat{\nu}_{\tau\mathrm{L}}$ which enter into the charged current V-A interaction. For Dirac neutrinos, we therefore write

$$\begin{pmatrix} \hat{\nu}_{\mathrm{e}} \\ \hat{\nu}_{\mu} \\ \hat{\nu}_{\tau} \end{pmatrix}_{\mathrm{L}} = \begin{pmatrix} U_{\mathrm{e}1} & U_{\mathrm{e}2} & U_{\mathrm{e}3} \\ U_{\mu 1} & U_{\mu 2} & U_{\mu 3} \\ U_{\tau 1} & U_{\tau 2} & U_{\tau 3} \end{pmatrix} \begin{pmatrix} \hat{\nu}_1 \\ \hat{\nu}_2 \\ \hat{\nu}_3 \end{pmatrix}_{\mathrm{L}} \equiv \mathbf{U} \begin{pmatrix} \hat{\nu}_1 \\ \hat{\nu}_2 \\ \hat{\nu}_3 \end{pmatrix}_{\mathrm{L}}, \tag{21.106}$$

where the unitary matrix $\mathbf{U}$ is the PMNS matrix, named after Pontecorvo (1957, 1958, 1967), and Maki, Nakagawa and Sakata (1962).

Now we showed in section 20.7.3 that the general 3×3 unitary matrix has three real (rotation angle) parameters, and 6 phase parameters, five of which we could get rid of by rephasing the quark fields by global U(1) transformations of the form $\hat{q}' = \exp(\mathrm{i}\theta)\hat{q}$. Such rephasing transformations are equally allowed

for the charged leptons, and also for Dirac neutrinos, since evidently the mass term (21.104) is invariant under a global U(1) transformation $\hat{\psi}' = \exp(\mathrm{i}\theta)\hat{\psi}$. Hence the matrix **U** will, in this Dirac case, have a parametrization of the CKM form, with one **CP**-violating phase.

The mixing described by (21.106) implies that the individual lepton flavour numbers $L_\mathrm{e}, L_\mu, L_\tau$ are no longer conserved. However, since we are here taking the neutrinos to be Dirac particles, there will be a quantum number carried by ν_e, ν_μ and ν_τ which is conserved by the interactions. This could, for example, be the total lepton number $L_\mathrm{e} + L_\mu + L_\tau$, assigning $L(\nu_\alpha) = 1$ for $\alpha = \mathrm{e}, \mu, \tau$, which would follow from invariance under the global U(1) transformation $\hat{\ell}'_\alpha = \exp(\mathrm{i}\delta)\hat{\ell}_\alpha, \hat{\nu}'_\alpha = \exp(\mathrm{i}\delta)\hat{\nu}_\alpha$, where δ is independent of the flavour α.

This 'Dirac' option, though simple, may be thought uneconomical, however. As noted, the R components of neutrino fields have no interactions of Standard Model type. The charged leptons do have electromagnetic interactions, of course, as do the quarks, which also have strong interactions. But the neutral neutrinos only have weak interactions, which involve only their L-components. Why, then, enlarge the field content to include hypothetical $\hat{\nu}_\mathrm{R}$ fields, which don't have any SM interactions? It seems more economical to make do with only the $\hat{\nu}_\mathrm{L}$ fields. In this case, the Dirac mass term (21.104) is not possible, but a Majorana mass term (21.105) can still exist. Clearly, such a mass term is *not* invariant under U(1) global phase transformations, and it breaks lepton number conservation explicitly. As in the Dirac case, the chiral L component will include a 'wrong' (i.e. positive) helicity component with an amplitude proportional to m/E.

The fact that global phase changes on the neutrino fields are now no longer freely available, because that symmetry is lost if they are Majorana fields, has implications for the mixing matrix, call it $\mathbf{U}_\mathrm{M}$, in this case. Since the three Majorana neutrino fields can no longer absorb phases, we have only the three phases from the charged leptons at our disposal, which leaves three phase parameters in $\mathbf{U}_\mathrm{M}$, after rephasing. The PMNS matrix in the Majorana case therefore has two more irreducible phase parameters than the CKM matrix, and is conventionally parametrized as

$$\mathbf{U}_\mathrm{M} = \mathbf{U}(\mathrm{CKM-type}) \times \mathrm{diag.}(1, \mathrm{e}^{\mathrm{i}\alpha_{21}/2}, \mathrm{e}^{\mathrm{i}\alpha_{31}/2}). \tag{21.107}$$

There are three **CP**-violating phases in the Majorana neutrino case.

The only information at present (2012) concerning the entries in **U** comes from neutrino oscillation experiments, which we shall discuss in the next sections. We shall see that the Majorana phases α_{21} and α_{31} cancel in the probabilities calculated for neutrino transitions, and no experiment so far is sensitive to **CP**-violating effects in the neutrino sector. We shall discuss how the values of the parameters θ_{12}, θ_{13} and θ_{23} can be inferred from the observed oscillations, and also the differences in the squared masses of the neutrinos. Anticipating these results, we state here that the two independent squared mass differences, $m_2^2 - m_1^2$ and $m_3^2 - m_1^2$, turn out to be very small indeed, and rather different from each other: namely approximately 7.6×10^{-5} eV^2 and

2.4×10^{-3} eV^2, respectively. The smaller value is associated with oscillations of solar or reactor neutrinos, and the larger with oscillations of atmospheric or accelerator neutrinos.

Data on the actual mass values are limited. There is a bound on the $\bar{\nu}_e$ mass from measurements of the electron spectrum near the end-point in tritium β-decay, which gives (Lobashev *et al.* 2003, Eitel *et al.* 2005)

$$m_{\bar{\nu}_e} < 2.3\text{eV} \quad 95\%\text{CL}. \tag{21.108}$$

A weaker limit on m_{ν_μ} comes from measurements of the muon spectrum in charged pion decay:

$$m_{\nu_\mu} < 0.19 \text{ MeV } 95\%\text{CL}. \tag{21.109}$$

The strongest upper bound comes from cosmology, assuming three neutrinos. The Cosmic Microwave Background data of the WMAP experiment, combined with supernovae data and data on galaxy clustering, can be used to obtain an upper limit on the sum of three neutrino masses (Spergel *et al.* 2007):

$$\sum_{i+1}^{3} m_{\nu_i} < 0.68 \text{ eV}, \quad 95\%\text{CL}. \tag{21.110}$$

Taking the squared mass differences as indicative of the actual mass scale, neutrino masses are evidently very much smaller than the masses of the other fermions in the Standard Model. We shall return to what this might tell us about the origin of neutrino mass in section 22.5, where we discuss how gauge-invariant masses are generated in the Standard Model.

Returning to the question of **CP** violation, we noted in section 4.2.3 that the **CP** violation present in the Standard Model was insufficient to account for the matter–antimatter asymmetry in the universe. However, we now see that it is possible to have **CP** violation in the lepton sector, in an extended Standard Model with massive neutrinos. Leptonic matter–antimatter asymmetries can be converted into baryon asymmetries in the very hot early universe by a non-perturbative process predicted by Standard Model dynamics – a process called leptogenesis (Fukugita and Yanagida 1986, Kuzmin, Rubakov and Shaposhnikov 1985). It has been argued that the Dirac and/or Majorana phases in the neutrino matrix **U** or $\mathbf{U}_{\mathrm{M}}$ can provide the **CP** violation necessary in leptogenesis models for the generation of the observed baryon asymmetry of the universe (Pascoli *et al.* 2007a, 2007b). If such a proposal should prove to be the case, the reach of Pauli's 'desperate remedy' will have been vast indeed.

21.4.2 Neutrino oscillations: formulae

The existence of neutrino oscillations means that if a neutrino of a given flavour $\nu_\alpha (\alpha = \mathrm{e}, \mu, \tau)$ with energy E is produced in a charged current weak interaction process, such as $\pi^+ \to \mu^+ \nu_\mu$, then at a sufficiently large distance L

from the ν_α source the probability $P(\nu_\alpha \to \nu_\beta; E, L)$ of detecting a neutrino of a different flavour ν_β is non-zero.[2] Such a flavour change will of course imply that the ν_α survival probability, $P(\nu_\alpha \to \nu_\alpha; E, L)$, is less than 1. We shall give a simplified version of the derivation of such probabilities, following the approach of the review by Nakamura and Petcov in section 13 of Nakamura *et al.* (2010). This review includes a large list of references to the time-dependent formalism; we mention here the contributions of Kayser (1981), Nauenberg (1999) and Cohen *et al.* (2009). We shall treat all the neutrinos as stable particles.

We consider the evolution of the state $|\nu_\alpha\rangle$ in the frame in which the detector which measures its flavour is at rest (the lab frame). As in the meson case, the states with simple space-time evolution in a vacuum are the mass eigenstates $|\nu_i\rangle$ $(i = 1, 2, 3)$, a superposition of which is equal to $|\nu_\alpha\rangle$:

$$|\nu_\alpha\rangle = \sum_j \mathrm{U}^*_{\alpha i} |\nu_i, p_i\rangle, \tag{21.111}$$

the complex conjugation arising from taking the dagger of the relation (21.106) for the field operators. Here **U** stands for either the Dirac or the Majorana matrix, and p_i is the 4-momentum of ν_i. Similarly,

$$|\bar{\nu}_\alpha\rangle = \sum_i \mathrm{U}_{\alpha i} |\bar{\nu}_i, p_i\rangle. \tag{21.112}$$

We will consider highly relativistic neutrinos, as is the case for the experiments under discussion. We will assume that there are no degeneracies among the masses m_j. The states in the superpositions (21.111) and (21.112) will all have, in general, different energies and momenta E_i, p_i. We shall also treat the evolution as occurring in one spatial dimension, taking all the momenta to lie in the direction from the source to the detector. Note that the fractional deviation of E_i and p_i from the massless case $E = p$ is of order m^2/E^2 which will be extremely small, of order one part in 10^{16}, say.

Suppose now that the neutrinos of flavour ν_α started in the state (21.111) at time $t = 0$ in the detector frame are detected at time T after production, having travelled a distance L. Then the amplitude for finding a neutrino of flavour ν_β at (L, T) is

$$\begin{aligned} A(\nu_\alpha \to \nu_\beta; L, E) &= \sum_i \mathrm{U}^*_{\alpha i}\, \mathrm{e}^{-\mathrm{i}E_i T + \mathrm{i}p_i L}\, \langle \nu_\beta | \nu_i, p_i \rangle \\ &= \sum_i \mathrm{U}^*_{\alpha i}\, \mathrm{U}_{\beta i} \mathrm{e}^{-\mathrm{i}E_i T + \mathrm{i}p_i L}. \end{aligned} \tag{21.113}$$

We make two immediate comments on (21.113). First, the Majorana phases in (21.54) cancel in $A(\nu_\alpha \to \nu_\beta; L, E) = \delta_{\alpha\beta}$, since the same phase appears

[2] We shall not indicate the chirality explicitly from now on, it being assumed that we are referring to the L (R) component for neutrinos (antineutrinos).

in $\mathrm{U}_{\alpha i}$ and $\mathrm{U}_{\beta i}$. We conclude that oscillation experiments cannot distinguish Majorana from Dirac neutrinos. Second, if the neutrinos were massless, the phase factors in (21.113) would all be unity, and then $A(\nu_\alpha \to \nu_\beta; L, E) = \delta_{\alpha\beta}$, from the unitarity of the matrix $\mathbf{U}$, so there would be no flavour change.

Flavour oscillations come about via the interference in $|A(\nu_\alpha \to \nu_\beta; L, T)|^2$ between phase factors that are slightly different from one another, because of the different masses. A typical interference phase is then $\phi_{ij} = (E_i - E_j)T - (p_i - p_j)L$. Following the review by Nakamura and Petcov in Nakamura *et al.* (2010), we note that

$$\frac{m_i^2 - m_j^2}{p_i + p_j} = \frac{(E_i^2 - p_i^2) - (E_j^2 - p_j^2)}{p_i + p_j} = (E_i - E_j)\frac{(E_i + E_j)}{(p_i + p_j)} - (p_i - p_j) \quad (21.114)$$

so that

$$\phi_{ij} = (E_i - E_j)\left[T - \frac{E_i + E_j}{p_i + p_j}L\right] + \frac{m_i^2 - m_j^2}{p_i + p_j}L. \quad (21.115)$$

Bearing in mind that the energies differ from the momenta by terms of order m^2/E^2, we see that the first term in (21.115) can be dropped, and the interference phase is, to a very good approximation,

$$\phi_{ij} = \frac{m_i^2 - m_j^2}{2E} \equiv \frac{\Delta m_{ij}^2}{2E} \quad (21.116)$$

where E is the average energy, or momentum, of the neutrinos. We therefore obtain the probability

$$\begin{aligned} P(\nu_\alpha \to \nu_\beta; L, E) &= \sum_i |\mathrm{U}_{\alpha i}|^2 |\mathrm{U}_{\beta i}|^2 \quad (21.117) \\ &\quad + 2\sum_{i>j} |\mathrm{U}_{\beta i} U^*_{\alpha i} \mathrm{U}_{\alpha j} \mathrm{U}^*_{\beta j}| \cos\left[\left(\frac{\Delta m_{ij}^2}{2E}\right)L - \phi_{\beta\alpha;ij}\right] \end{aligned}$$

where

$$\phi_{\alpha\beta;ij} = \mathrm{Arg}\,(\mathrm{U}_{\beta i}\mathrm{U}^*_{\alpha i}\mathrm{U}_{\alpha j}\mathrm{U}^*_{\beta j}). \quad (21.118)$$

A more useful expression can be obtained by using the unitarity of $\mathbf{U}$ (problem 21.6):

$$\begin{aligned} P(\nu_\alpha \to \nu_\beta; L, E) &= \delta_{\alpha\beta} - 4\sum_{i>j} \mathrm{Re}\,(\mathrm{U}_{\beta i}\mathrm{U}^*_{\alpha i}\mathrm{U}_{\alpha j}\mathrm{U}^*_{\beta j}) \sin^2 \frac{\Delta m_{ij}^2 L}{4E} \\ &\quad + 2\sum_{i>j} \mathrm{Im}\,(\mathrm{U}_{\beta i}\mathrm{U}^*_{\alpha i}\mathrm{U}_{\alpha j}\mathrm{U}^*_{\beta j}) \sin \frac{\Delta m_{ij}^2 L}{2E} \quad (21.119) \end{aligned}$$

The expression for $P(\bar\nu_\alpha \to \bar\nu_\beta; L, T)$ is the same, except for a change in sign

of the last term in (21.119):

$$P(\bar{\nu}_\alpha \to \bar{\nu}_\beta; L, E) = \delta_{\alpha\beta} - 4\sum_{i>j} \mathrm{Re}\,(\mathrm{U}_{\beta i}\mathrm{U}^*_{\alpha i}\mathrm{U}_{\alpha j}\mathrm{U}^*_{\beta j}) \sin^2 \frac{\Delta m^2_{ij} L}{4E} - 2\sum_{i>j} \mathrm{Im}\,(\mathrm{U}_{\beta i}\mathrm{U}^*_{\alpha i}\mathrm{U}_{\alpha j}\mathrm{U}^*_{\beta j}) \sin \frac{\Delta m^2_{ij} L}{2E}. \quad (21.120)$$

It follows from (21.119) and (21.120) that $P(\nu_\alpha \to \nu_\beta; L, E) = P(\bar{\nu}_\beta \to \bar{\nu}_\alpha; L, E)$, a consequence of **CPT** invariance. **CP** alone requires $P(\nu_\alpha \to \nu_\beta; L, E) = P(\bar{\nu}_\alpha \to \bar{\nu}_\beta; L, E)$. A measure of **CP** violation is provided by

$$\begin{aligned} \mathcal{A}^{(\beta\alpha)}_{\mathbf{CP}} &= P(\nu_\alpha \to \nu_\beta; L, E) - P(\bar{\nu}_\alpha \to \bar{\nu}_\beta; L, E) \\ &= 4\sum_{i>j} \mathrm{Im}\,(\mathrm{U}_{\beta i}\mathrm{U}^*_{\alpha i}\mathrm{U}_{\alpha j}\mathrm{U}^*_{\beta j}) \sin \frac{\Delta m^2_{ij}}{2E} L. \end{aligned} \quad (21.121)$$

The reader will recognize the Jarlskog (1985) invariants in (21.121). In this 3×3 mixing situation, which is exactly analogous to quark mixing, all these invariants are equal up to a sign, and (21.121) becomes (Krastev and Petcov 1988)

$$\begin{aligned} \mathcal{A}^{(\mu\mathrm{e})}_{\mathbf{CP}} &= -\mathcal{A}^{(\tau\mathrm{e})}_{\mathbf{CP}} = \mathcal{A}^{(\tau\mu)}_{\mathbf{CP}} \\ &= 4J_\nu \left[\sin\left(\frac{\Delta m^2_{32}}{2E} L\right) + \sin\left(\frac{\Delta m^2_{21}}{2E} L\right) + \sin\left(\frac{\Delta m^2_{13}}{2E} L\right) \right] \end{aligned} \quad (21.122)$$

where

$$J_\nu = \mathrm{Im}(\mathrm{U}_{\mu 3}\mathrm{U}^*_{\mathrm{e}3}\mathrm{U}_{\mathrm{e}2}\mathrm{U}^*_{\mu 2}). \quad (21.123)$$

If any one mass-squared difference is zero, say Δm^2_{21}, then $\Delta m^2_{32} = -\Delta m^2_{13}$, and the right-hand side of (21.122) vanishes: we need all three mass-squared differences to be non-zero, in order to get **CP** violation.

In proceeding to discuss the experimental situation, it will be useful to define an 'oscillation length' $\lambda_{ij}(E)$ given by

$$\lambda_{ij}(E) = 2E/\Delta m^2_{ij} \approx 0.4 \frac{(E/\mathrm{GeV})}{(\Delta m^2_{ij}/\mathrm{eV}^2)}\,\mathrm{km}. \quad (21.124)$$

In practice, the three-state mixing formalism can often be simplified, making use of what is now known about the neutrino mass spectrum. One squared mass difference is considerably smaller than the other:

$$|\Delta m^2_{21}| \sim 7.6 \times 10^{-5}\ \mathrm{eV}^2, \quad |\Delta m^2_{31}| \sim 2.4 \times 10^{-3}\mathrm{eV}^2. \quad (21.125)$$

Suppose now that $L/|\lambda_{31}(E)| \sim 1$, while $L/|\lambda_{21}(E)| \ll 1$. Then expression (21.119) reduces to (problem 21.7)

$$\begin{aligned} P(\nu_\alpha \to \nu_\beta; L, E) &\approx \delta_{\alpha\beta} - 4|\mathrm{U}_{\alpha 3}|^2[\delta_{\alpha\beta} - |\mathrm{U}_{\beta 3}|^2] \sin^2 \frac{\Delta m^2_{31}}{4E} L \\ &= P(\bar\nu_\alpha \to \bar\nu_\beta; L, E). \end{aligned} \tag{21.126}$$

In particular,

$$P(\bar\nu_\mathrm{e} \to \bar\nu_\mathrm{e}; L, E) = 1 - 4|\mathrm{U}_{\mathrm{e}3}|^2(1 - |\mathrm{U}_{\mathrm{e}3}|^2) \sin^2 \frac{\Delta m^2_{31}}{4E} L, \tag{21.127}$$

which can describe the survival probability of reactor $\bar\nu_\mathrm{e}$s, for example.

Adopting a parametrization of the form (20.166), with rows labelled by e, μ and τ, and columns by 1, 2, and 3, $|\mathrm{U}_{\mathrm{e}3}|^2$ is $\sin^2\theta_{\mathrm{e}3}$, which is found experimentally to be small (see the following section). It is often a good approximation to set $|U_{\mathrm{e}3}|$ to zero, in which case $|U_{\mu 3}|^2 = \sin^2\theta_{\mu 3}$. Then (21.126) gives the ν_μ survival probability

$$P(\nu_\mu \to \nu_\mu; L, E) = P(\bar\nu_\mu \to \bar\nu_\mu; L, E) \approx 1 - \sin^2 2\theta_{\mu 3} \sin^2(L/2\lambda_{31}(E)) \tag{21.128}$$

and the flavour-change probability

$$P(\nu_\mu \to \nu_\tau; L, E) = P(\bar\nu_\mu \to \bar\nu_\tau; L, E) \approx \sin^2 2\theta_{\mu 3} \sin^2(L/2\lambda_{31}(E)). \tag{21.129}$$

In this approximation, $P(\nu_\mu \to \nu_\mathrm{e}) = P(\bar\nu_\mu \to \bar\nu_\mathrm{e}) = 0$. Formulae (21.128) and (21.129) can be used to describe the dominant atmospheric ν_μ and $\bar\nu_\mu$ oscillations (see the following section), and the parameters $\theta_{\mu 3}$ and Δm^2_{31} (or Δm^2_{32}) are referred to as the atmospheric mixing angle and mass squared difference. The smaller mass squared difference Δm^2_{21}, and the angle θ_{e2}, are associated with solar ν_e oscillations.

The formulae (21.128) and (21.129) are, in fact, exactly what a simple 2-state mixing model would give. Suppose that the effective mixing matrix for the 2-state system has the form (see problem 1.6)

$$\begin{pmatrix} -a\cos 2\theta & a\sin 2\theta \\ a\sin 2\theta & a\cos 2\theta \end{pmatrix}, \tag{21.130}$$

where rows are labelled by e, μ and columns by 1, 3; then the survival probability is just

$$1 - \sin^2 2\theta \sin^2(La), \tag{21.131}$$

where we have taken $L \approx T$ as before. We can therefore identify the mixing parameter as

$$a = [2\lambda_{31}(E)]^{-1} = \frac{\Delta m^2_{31}}{4E}. \tag{21.132}$$

Note that the energies are here measured relative to a common average energy; if $|\ell\rangle$ is the lighter eigenstate and $|h\rangle$ the heavier, then

$$\begin{aligned} |\nu_\mathrm{e}\rangle &= \cos\theta|\ell\rangle + \sin\theta|h\rangle \\ |\nu_\mu\rangle &= -\sin\theta|\ell\rangle + \cos\theta|h\rangle. \end{aligned} \tag{21.133}$$

21.4.3 Neutrino oscillations: experimental results

Historically, the search for neutrino oscillations began when experiments by Davis *et al.* (1968) detected solar neutrinos (from ^{8}B decays) at a rate approximately one third of that predicted by the solar model calculations of Bahcall *et al.* (1968). Pontecorvo (1946) had proposed the experiment, in which the neutrinos are detected by the inverse β-decay process $\nu_e + {}^{37}\mathrm{Cl} \to e^- + {}^{37}\mathrm{Ar}$. The Davis experiment used 520 metric tons of liquid tetrachloroethylene (C_2Cl_4), buried 4850 feet underground in the Homestake gold mine, in South Dakota. Davis' findings provided the impetus to study solar neutrinos using Kamiokande, a 3000 ton imaging water Cerenkov detector situated about one kilometre underground in the Kamioka mine in Japan. Indeed, ^{8}B solar neutrinos were observed, and at a rate consistent with that of the Davis experiment (Hirata *et al.* 1989). Later results from the Homestake mine (Cleveland *et al.* 1998) reported a solar neutrino detection rate almost exactly one third of the updated calculations of Bahcall *et al.* (2001).

In a separate development, Kamiokande also reported (Hirata *et al.* 1988) an anomaly in the atmospheric neutrino flux. Atmospheric neutrinos are produced as decay products in hadronic showers which result from collisions of cosmic rays with nuclei in the upper atmosphere of the Earth. Production of electron and muon neutrinos is dominated by the decay chain $\pi^+ \to \mu^+ + \nu_\mu$, $\mu^+ \to e^+ + \bar{\nu}_\mu + \nu_e$ (and its charge-conjugate), which gives an expected value of about 2 for the ratio of $(\nu_\mu + \bar{\nu}_\mu)$ flux to $(\nu_e + \bar{\nu}_e)$ flux.[3] While the number of electron-like events was in good agreement with the Monte Carlo calculations based on atmospheric neutrino interactions in the detector, the number of muon-like events was about one half of the expected number, at the 4σ level.

This muon-like defect (and the lack of an electron-like defect) was later confirmed at the 9σ level by Super-Kamiokande (Fukuda *et al.* 1998). In this experiment, a marked dependence was observed on the zenith angle of the muon neutrinos. This angle is simply related to the distance travelled by the neutrinos from their point of production, which varies from about 20 km (from above the detector) to over 10,000 km (from below the detector). The Super-Kamiokande data was the first compelling evidence for neutrino oscillations. Interpreting their data in terms of a simple 2-state $\nu_\mu \leftrightarrow \nu_\tau$ model, as in (21.129), Fukuda *et al.* (1998) reported the values $\sin^2 2\theta_{\mu 3} > 0.82$, and $5 \times 10^{-4} < \Delta m^2_{31} < 6 \times 10^{-3}$ eV^2 at 90% CL.

We will postpone further discussion of the solar neutrino deficit for the moment, since it is complicated by interactions of the neutrinos with the Sun's matter (see the following subsection). We proceed to describe some of the main results which have come from the analysis of data from neutrinos produced in terrestrial accelerators and reactors.

[3] The detector could not measure the charge of the final state leptons, and therefore ν and $\bar{\nu}$ events could not be discriminated.

We begin with the CHOOZ experiment, which was the first experiment to limit the value of θ_{e3} (Apollonio *et al.* 1999, 2003). CHOOZ is the name of a nuclear power station situated near the French village of the same name. The experiment was designed to detect reactor $\bar{\nu}_e$s via the inverse β-decay reaction $\bar{\nu}_e + p \to e^+ + n$. The signature was a delayed coincidence between the prompt e^+ signal, and the signal from the neutron capture. The detector was located in an underground laboratory about 1 km from the neutrino source. It consisted of a central 5-ton target filled with 0.09 % Gd-doped liquid scintillator; Gd-doping was chosen to maximize the capture of the neutrons. The neutrino energy E was a few MeV, and L was 1 km. For these values $2\lambda_{21}(E)$ is greater than about 10 km, while $2\lambda_{31}(E)$ is about 0.3 km. The neglect of $\sin^2 L/2\lambda_{21}(E)$ is justified, and formula (21.127) can be used for the $\bar{\nu}_e$ survival probability. The experiment found no evidence for $\bar{\nu}_e$ disappearance, and reported the 90% CL upper limit of $\sin^2 2\theta_{e3} < 0.19$, for $|\Delta m^2_{31}| = 2 \times 10^{-3}$ eV^2. We shall for the moment set θ_{e3} to zero, and return to discuss its value at the end of the chapter.

The mass squared range $\Delta m^2 > 2\times10^{-3}$ eV^2 can be explored by accelerator-based long-baseline experiments, with typically $E \sim 1$ GeV and $L \sim$ several hundred kilometres. The K2K (KEK-to-Kamioka) experiment was the first accelerator-based experiment with a neutrino path length extending hundreds of kilometres. A horn-focused wide-band ν_μ beam with mean energy 1.3 GeV and path length 250 km was produced by 12 GeV protons from the KEK-PS and directed to the Super-Kamiokande detector. In this case, $L/2\lambda_{21}(E) \sim 10^{-2}$, which may be neglected. Then formulae (21.128) and (21.129) may be used, in the approximation $U_{e3} \approx 0$. The K2K data showed (Ahn *et al.* 2006) that $\sin^2 2\theta_{\mu3} \approx 1 (\theta_{\mu3} \approx \pi/4)$, and that $|\Delta m^2_{31}|$ had a value consistent with (21.125).

The first evidence for the appearance of ν_e in a ν_μ beam was obtained by the T2K collaboration (Abe *et al.* 2011). The ν_μ beam is produced using the high intensity proton accelerator at J-PARC, located in Tokai, Japan. The beam was directed 2.5° off-axis to the Super-Kamiokande detector at Kamioka, 295 km away. This configuration produces a narrow-band ν_μ beam, tuned at the first oscillation maximum $E_\nu = |\Delta m^2_{31}|L/2\pi \approx 0.6$ MeV, so as to reduce background from higher energy neutrino interactions. In the vacuum, the probability of the appearance of a ν_e in a ν_μ beam is given (in our customary effective 2-state mixing approximation) by (21.126) as

$$P(\nu_\mu \to \nu_e; L, E) = \sin^2\theta_{\mu3} \sin^2 2\theta_{e3} \sin^2 \frac{\Delta m^2_{31}}{4E} L; \qquad (21.134)$$

$P(\bar{\nu}_\mu \to \bar{\nu}_e; L, E)$ is given by the same expression. Taking $|\Delta m^2_{31}| = 2.4 \times 10^{-3}$ eV^2 and $\sin^2 2\theta_{\mu3} = 1$, the number of expected ν_e events was 1.5 ± 0.3(syst.) for $\sin^2 2\theta_{e3} = 0$, and 5.5 ± 1.0 events if $\sin^2\theta_{e3} = 0.1$. Six events were observed which passed all the ν_e selection criteria. As we will see in the following section, the value of $\sin^2 2\theta_{e3} = 0.1$ is entirely consistent with direct measurements of this quantity reported in 2012.

Another long baseline accelerator experiment is MINOS at Fermilab. Neutrinos are produced by the Neutrinos at the Main Injector facility (NuMI), using 120 GeV protons from the Fermilab main injector. The detector is a 5.4 kton iron-scintillator tracking calorimeter with a toroidal magnetic field, situated underground in the Soudan mine, 735 km from Fermilab. The neutrino energy spectrum from a wide-band beam is horn-focused to be enhanced in the 1-5 GeV range. The current MINOS results yield $|\Delta m^2_{31}| = (2.32^{+0.12}_{-0.08}) \times 10^{-3}$ eV^2, and $\sin^2 2\theta_{\mu 3} > 0.90$ at 90 % CL (Adamson *et al.* 2011).

A second reactor experiment, KamLAND at Kamioka, was designed to be sensitive to the smaller squared mass difference Δm^2_{21}, and thus to $\theta_{\mathrm{e}2}$. The Kamioka Liquid scintillator AntiNeutrino Detector is at the site of the former Kamioka experiment. The detector is essentially one kiloton of highly purified liquid scintillator surrounded by photomultiplier tubes. $\bar{\nu}_{\mathrm{e}}$s are detected as usual via the inverse β-decay reaction $\bar{\nu}_{\mathrm{e}} + \mathrm{p} \to \mathrm{e}^+ + \mathrm{n}$. KamLAND is surrounded by 55 nuclear power units, each an isotropic $\bar{\nu}_{\mathrm{e}}$ source. The flux-weighted average path length is $L \sim 180$ km, and the energy E ranges from about 2 MeV to about 8 MeV. For $E = 3$ MeV, $2\lambda_{21}(E) \sim 30$ km, which allows for more than one oscillation. In this case (21.119) reduces to

$$P(\bar{\nu}_{\mathrm{e}} \to \bar{\nu}_{\mathrm{e}}; L, E) = 1 - 4|\mathrm{U}_{\mathrm{e}1}|^2\,|\mathrm{U}_{\mathrm{e}2}|^2\,\sin^2(L/2\lambda_{21}(E)) \tag{21.135}$$

assuming $|\mathrm{U}_{\mathrm{e}3}| \approx 0$. In a parametrization of the form (20.166), this becomes

$$P(\bar{\nu}_{\mathrm{e}} \to \bar{\nu}_{\mathrm{e}}; L, E) = 1 - \sin^2 2\theta_{\mathrm{e}2}\,\sin^2(L/2\lambda_{21}(E)), \tag{21.136}$$

again a simple 2-state mixing result. Data shown in figure 21.11 (Abe *et al.* 2008) gives

$$|\Delta m^2_{21}| = 7.58^{+0.14+0.15}_{-0.13-0.15} \times 10^{-5}\ \mathrm{eV}^2 \tag{21.137}$$

$$\tan^2\theta_{\mathrm{e}2} = 0.56^{+0.10+0.10}_{-0.70-0.06}. \tag{21.138}$$

The KamLAND data showed for the first time the periodic behaviour of the $\bar{\nu}_{\mathrm{e}}$ survival probability.

We now return to the solar neutrino problem, taking up the story after Davis' results. Some doubts remained as to whether the solar calculations could be absolutely relied upon, for example because of the extreme sensitivity to the core temperature ($\propto T^{18}$). One particular class of ν_{e} could, however, be reliably calculated, namely those associated with the initial reaction $\mathrm{pp} \to {}^2\mathrm{H} + \mathrm{e}^+ + \nu_{\mathrm{e}}$ of the pp cycle. Whereas the Davis experiments allowed detection of the higher energy ν_{e}s (threshold 814 keV) from the B and Be stages of the cycle, the energy of the ν_{e}s from the pp stage cuts off at around 400 keV. Detectors using the reaction $\nu_{\mathrm{e}} + {}^{71}\mathrm{Ga} \to \mathrm{e}^- + {}^{71}\mathrm{Ge}$, which has a 233 keV energy threshold, were built (GALLEX, GNO and SAGE); their results (Altman *et al.* 2005, Abdurashitov *et al.* 2009) are in agreement, and again much smaller than the (updated) Bahcall *et al.* (2005) prediction.

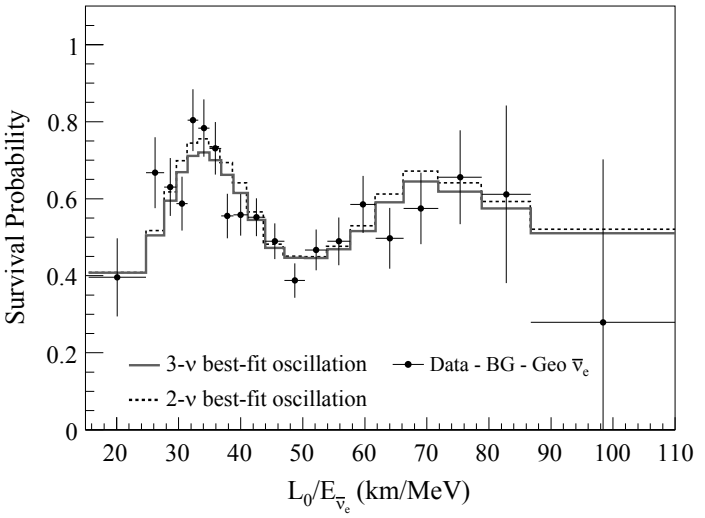


FIGURE 21.11
Ratio of the background and geo-neutrino subtracted $\bar{\nu}_e$ spectrum to the expectation for no-oscillation, as a function of L_0/E, where $L_0 = 180$ km. Figure reprinted with permission from S Abe *et al.* (KamLAND Collaboration) *Phys. Rev. Lett.* **100** 221803 (2008). Copyright 2008 by the American Physical Society.

In 1999, the Sudbury Neutrino Observatory (SNO) in Canada began observation. This experiment used 1 kiloton of ultra-pure heavy water (D_2O). It measured ^{8}B solar ν_es via both the CC reaction $\nu_e + d \rightarrow e^- + p + p$, and the NC reaction $\nu + d \rightarrow \nu + p + n$, as well as elastic $\nu\, e^-$ scattering. The CC reaction is sensitive only to ν_e, while the NC reaction is sensitive to all active neutrinos, as is $\nu\, e^-$ scattering. If the solar neutrino deficit were caused by neutrino oscillations, the solar neutrino fluxes measured by the CC and NC reactions would be significantly different. SNO found that, while the total neutrino flux was consistent with solar model expectations, the ratio of the ν_e flux to the total neutrino flux was about 1/3 (Ahmad *et al.* 2001, 2002). This number can be understood in terms of the effect of dense matter on the propagation of the ν_es, as we now discuss.

21.4.4 Matter effects in neutrino oscillations

We have assumed in the foregoing that neutrinos propagate in vacuum between the source and the detector. Since neutrinos interact only weakly, it might seem that this is always an excellent approximation. But in the same way that light travelling through a transparent medium can have its refractive index changed, so can a neutrino. In particular, the refractive index can be

different for ν_e and ν_μ. The difference in refractive indices is determined by the difference in the real parts of the forward $\nu_e e^-$ and $\nu_\mu e^-$ elastic scattering amplitudes (Wolfenstein 1978). The essential point is that the scattering can be coherent, with the spins and momenta of the particles remaining unchanged. This means that the effect is going to be proportional to the density of electrons in the matter traversed, N_e. The scattering amplitude, in turn, is proportional to G_F, so that a figure of merit for the effect is given by the product $G_F N_e$. This has the dimensions of an energy, and can be interpreted as an addition to the effective 2-state mixing matrix of (21.130). Detailed analysis, which we omit, shows that the correct addition is actually $+\sqrt{2}G_F N_e$, so that (21.130) is modified to

$$\begin{pmatrix} -\frac{\Delta m^2}{4E}\cos 2\theta + \sqrt{2}G_F N_e & \frac{\Delta m^2}{4E}\sin 2\theta \\ \frac{\Delta m^2}{4E}\sin 2\theta & \frac{\Delta m^2}{4E}\cos 2\theta \end{pmatrix}, \tag{21.139}$$

where now $\Delta m^2 = m_2^2 - m_1^2$, and $\theta = \theta_{e2}$. Two-state mixing now gives (problem 21.8) a new mixing angle θ_m such that

$$\tan 2\theta_m = \frac{\tan 2\theta}{1 - N_e/N_{res}}, \quad N_{res} = \frac{\Delta m^2 \cos 2\theta}{2\sqrt{2}G_F E}, \tag{21.140}$$

and the mass eigenstates $|1\rangle_m, |2\rangle_m$ correspond to the eigenvalue difference

$$m_2 - m_1 = |\Delta m_{21}^2/2E|\,[\cos^2 2\theta(1 - N_e/N_{res})^2 + \sin^2 2\theta]^{1/2}. \tag{21.141}$$

We see that although the new term is certainly very small, being proportional to G_F, nevertheless since Δm^2 is very small also, a significant effect can occur. In particular, if it should happen that $N_e \approx N_{res}$ for some (θ, E), then θ_m will be 'maximal' ($\theta_m = \pi/4$), irrespective of the value of the original θ. This is called 'resonant mixing' (Mikheev and Smirnov 1985, 1986). It implies that the probability for a $\nu_e \to \nu_\mu$ flavour change could be greatly enhanced over the vacuum value, which is proportional to $\sin^2 2\theta_{e2}$. A point to note, also, is that the corresponding formulae for $\bar{\nu}_e$s are obtained by replacing N_e by $-N_e$; then, depending on the sign of $\Delta m^2 \cos 2\theta_{e2}$, resonant mixing can occur for one or the other of ν_e or $\bar{\nu}_e$ as they pass through matter, but not both. Similar considerations apply to the propagation of neutrinos through the earth, but we shall not pursue this here (see Nakamura and Petcov in Nakamura *et al.* 2010).

In the case of solar neutrinos, the effect of the above modifications is quite simple. For the highest energy neutrinos, $N_e \gg N_{res}$ at the centre of the sun, so that $\theta_m \sim \pi/2$ at production in the core, and the ν_e is in the heavier mass state $|2\rangle_m$. On the way to the surface of the Sun, N_e will decrease, and a point will be reached when $N_e = N_{res}$. Here the mass difference (21.141) reaches its minimum, and two limiting cases may be distinguished depending on the scale of the variation in the electron density, which has been assumed constant in (21.139)–(21.141). (i) If the density variation is slow enough that

at least one oscillation length fits into the resonant density region, then it can be shown that the state stays with state $|2\rangle_m$ ('adiabatic evolution') until it reaches the surface of the Sun, when $\theta_m \to \theta_{e2}$. The probability that the neutrino will survive to the earth is then (using (21.133)) $|\langle\nu_e|2\rangle_m|^2 = \sin^2\theta_{e2}$, which has a value of about 1/3. In the alternative limit, (ii), in which the oscillation length in matter is relatively large with respect to the scale of density variation, the state may 'jump' to the other mass state $|1\rangle_m$ ('extreme non-adiabatic evolution'), and then $|\langle\nu_e|1\rangle_m|^2 = \cos^2\theta_{e2}$. These are clearly extreme cases, and numerical work is required in the general case. However, the data from SNO and other water Cerenkov detectors are consistent with the first (adiabatic) alternative, and with the value $\sin^2\theta_{e2} \sim 1/3$. Note that the solar data imply that $(m_2^2 - m_1^2)\cos 2\theta_{e2} > 0$.

By contrast, for the lowest energy neutrinos we can take $\theta_m \approx \theta$, so that the neutrinos are produced in the state $\cos\theta_{e2}|1\rangle + \sin\theta_{e2}|2\rangle$, and propagate as in a vacuum, oscillating with maximum excursion $\sin^2 2\theta_{e2}$. The detectors average over many oscillations, giving a factor of 1/2, so that the survival probability for the low energy ν_es is $1 - \frac{1}{2}\sin^2 2\theta_{e2} \sim 5/9$. The Gallium experiments are sensitive to the lower energy neutrinos, and indeed record some 60–70% of the expected flux.

In summary, the solar neutrino data are consistent with the interpretation in terms of neutrino oscillations, as modified by the Wolfenstein-Mikheev-Smirnov (MSW) effect. A global solar + KamLAND analysis yields best fit values (Aharmim *et al.* 2010)

$$\theta_{e2} = 34.06^{+1.16}_{-0.84} \quad \Delta m^2_{21} = 7.59^{+0.20}_{-0.21} \times 10^{-5}\ \text{eV}^2. \tag{21.142}$$

21.4.5 Further developments

Despite the remarkable experimental progress in the studies of neutrino oscillations over the last decade, there still remain some basic gaps in our knowledge. Perhaps the most fundamental is the Dirac/Majorana nature of massive neutrinos. The most feasible (but very difficult) test is neutrinoless double β-decay ($0\nu\beta\beta$-decay), already touched on in section 20.3. As noted there, the amplitude is proportional to an average Majorana mass parameter $\langle m\rangle$. Experiments place a lower bound on the half-life for the decay, which translates into an upper bound on $\langle m\rangle$. The most stringent lower bounds on the half-lives have been obtained with decays of ^{76}Ge (Klapdor-Kleingrothaus *et al.* 2001), ^{130}Te (Andreotti *et al.* 2011) and ^{100}Mo (Arnold *et al.* 2006). Lower bounds on the half-lives range from 10^{24} to 10^{25} years, with corresponding upper bounds on $\langle m\rangle$ of the order of 0.5 eV. It should, however, be noted that some participants of the Heidelberg-Moscow experiment claimed the observation of $0\nu\beta\beta$ decay of ^{76}Ge with a half-life of $2.23^{+0.44}_{-0.31} \times 10^{25}$ years, from which they deduced $\langle m\rangle = 0.32 \pm 0.03$ eV (Klapdor-Kleingrothaus *et al.* 2006). The GERDA experiment (Ur *et al.* 2011) should be able to check this claim after one year of running. Other experiments currently running, or planned, will

push the bound on half-lives up to 10^{26}–10^{27} years, and the upper bound on $\langle m \rangle$ down to magnitudes of the order of a few times 10^{-2} eV.

A second crucial question concerns the magnitude of **CP**-violation effects in neutrino oscillations. We recall from (20.172) that this vanishes if $\sin\theta_{e3} = 0$. As we saw earlier, CHOOZ set a 90% CL limit $\sin^2 2\theta_{e3} < 0.17$. A non-zero value of $\sin^2 2\theta_{e3}$ has now been observed by two groups, both $\bar{\nu}_e$ disappearance experiments: the Daya Bay collaboration (An *et al.* 2012) and the RENO collaboration (Ahn *et al.* 2012). Their reported results were

$$\sin^2 2\theta_{e3} = 0.092 \pm 0.016 \pm 0.005 \quad \text{(Daya Bay)} \tag{21.143}$$

$$\sin^2 2\theta_{e3} = 0.113 \pm 0.013 \pm 0.019 \quad \text{(RENO)}, \tag{21.144}$$

in a 3-neutrino framework. For this value of $\sin\theta_{e3}$, it should be possible to detect a **CP**-violating difference in the probabilities for $\nu_\mu \to \nu_e$ and $\bar{\nu}_\mu \to \bar{\nu}_e$, and it may be enough to sustain leptogenesis models.

The value of $\sin\theta_{e3}$ is also relevant to the determination of the sign of Δm^2_{31}; we shall mention just one possibility. We have seen that the MSW effect for solar neutrinos implies that $m_2 > m_1$ (using the fact that $\cos\theta_{e2} > 0$), but the mass spectrum (for 3-neutrino mixing) could be ordered as $m_1 < m_2 < m_3$ ('normal spectrum') or as $m_3 < m_1 < m_2$ ('inverted spectrum'). We have ignored the terrestrial MSW effect, but it can be significant in long-baseline accelerator-based experiments, and could be exploited to determine the sign of $m_3 - m_1$. In the vacuum, the probability of the appearance of a ν_e in a ν_μ beam is given by (21.134) (in our customary effective 2-state mixing approximation). As in the solar case, these probabilities will be modified by the MSW effect, which will enhance (suppress) the appearance probability for neutrinos (antineutrinos) in the case of the normal spectrum, and vice versa for the inverted spectrum. Clearly if θ_{e3} were too small, the effect would be very hard to see, but the value in (21.143) and (21.144) makes this a realistic experiment; it formed part of the physics motivation for the NOνA experiment at Fermilab (Ayres *et al.* 2005). NOνA is a long-baseline neutrino oscillation experiment now under construction, which aims to detect the appearance of ν_e and $\bar{\nu}_e$ in the NuMI muon neutrino beam. The beam from Fermilab is directed 14 mrad off-axis to a detector 810 kn away; the neutrino energy is narrowly peaked around 2.2 GeV. NOνA will also have sensitivity to leptonic **CP**-violation.

Problems

21.1 Verify equation (21.34).

21.2 Verify equations (21.46) and (21.47).

20.3 Verify equations (21.48) and (21.49).

21.4 Verify equations (21.56).

21.5 Verify equations (21.89) and (21.90).

21.6 Verify equation (21.119).

21.7 Verify equation (21.126).

21.8 Verify equations (21.140) and (21.141).


Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

22

The Glashow–Salam–Weinberg Gauge Theory of Electroweak Interactions

22.1 Difficulties with the current–current and 'naive' IVB models

In chapter 20 we developed the 'V-A current–current' phenomenology of weak interactions. We saw that this gives a remarkably accurate account of a wide range of data – so much so, in fact, that one might well wonder why it should not be regarded as a fully-fledged theory. One good reason for wanting to do this would be in order to carry out calculations beyond the lowest order, which is essentially all we have used it for so far (with the significant exceptions of the GIM argument, and box diagrams in M-$\bar{\text{M}}$ mixing). Such higher-order calculations are indeed required by the precision attained in modern high energy experiments. But the electroweak theory of Glashow, Salam and Weinberg, now recognized as one of the pillars of the Standard Model, was formulated long before such precision measurements existed, under the impetus of quite compelling theoretical arguments. These had to do, mainly, with certain in-principle difficulties associated with the current–current model, if viewed as a 'theory'. Since we now believe that the GSW theory is the correct description of electroweak interactions up to currently tested energies, further discussions of these old issues concerning the current–current model might seem irrelevant. However, these difficulties do raise several important points of principle. An understanding of them provides valuable motivation for the GSW theory – and some idea of what is 'at stake' in regard to experiments relating to the Higgs sector, which has only recently begun to be explored (see section 22.8.3).

Before reviewing the difficulties, however, it is worth emphasizing once again a more positive motivation for a gauge theory of weak interactions (Glashow 1961). This is the remarkable 'universality' structure noted in chapter 20, not only as between different types of lepton, but also (within the context of CKM mixing) between the quarks and the leptons. This recalls very strongly the 'universality' property of QED, and the generalization of this property in the non-Abelian theories of chapter 13. A gauge theory would provide a natural framework for such universal couplings.

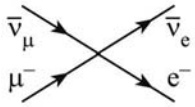


FIGURE 22.1
Current–current amplitude for $\bar{\nu}_\mu + \mu^- \to \bar{\nu}_e + e^-$.

22.1.1 Violations of unitarity

We have seen several examples, in chapter 20, in which cross sections were predicted to rise indefinitely as a function of the invariant variable s, which is the square of the total energy in the CM frame. We begin by showing why this is ultimately an unacceptable behaviour.

Consider the process (figure 22.1)

$$\bar{\nu}_\mu + \mu^- \to \bar{\nu}_e + e^- \tag{22.1}$$

in the current–current model, regarding it as fundamental interaction, treated to lowest order in perturbation theory. A similar process was discussed in chapter 20. Since the troubles we shall find occur at high energies, we can simplify the expressions by neglecting the lepton masses without altering the conclusions. In this limit the invariant amplitude is (problem 22.1), up to a numerical factor,

$$\mathcal{M} = G_F E^2 (1 + \cos\theta) \tag{22.2}$$

where E is the CM energy, and θ is the CM scattering angle of the e^- with respect to the direction of the incident μ^-. This leads to the following behaviour of the cross section (cf (20.83), remembering that $s = 4E^2$):

$$\sigma \sim G_F^2 E^2. \tag{22.3}$$

The dependence on E^2 is a consequence of the fact that G_F is not dimensionless, having the dimensions of $[M]^{-2}$. Its value is (Nakamura *et al.* 2010)

$$G_F = 1.16637(1) \times 10^{-5}\ \text{GeV}^{-2}. \tag{22.4}$$

The cross section has dimensions of $[L]^2 = [M]^{-2}$, but must involve G_F^2 which has dimension $[M]^{-4}$. It must also be relativistically invariant. At energies well above lepton masses, the only invariant quantity available to restore the correct dimensions to σ is s, the square of the CM energy E, so that $\sigma \sim G_F E^2$.

Consider now a partial wave analysis of this process. For spinless particles the total cross section may be written as a sum of partial wave cross sections

$$\sigma = \frac{4\pi}{k^2} \sum_J (2J+1)|f_J|^2 \tag{22.5}$$

where f_J is the partial wave amplitude for angular momentum J and k is the CM momentum. It is a consequence of *unitarity*, or flux conservation (see, for example, Merzbacher 1998, chapter 13), that the partial wave amplitude may be written in terms of a phase shift δ_J:

$$f_J = \mathrm{e}^{\mathrm{i}\delta_J} \sin \delta_J \tag{22.6}$$

so that

$$|f_J| \leq 1. \tag{22.7}$$

Thus the cross section in each partial wave is bounded by

$$\sigma_J \leq 4\pi(2J+1)/k^2 \tag{22.8}$$

which falls as the CM energy rises. By contrast, in (22.3) we have a cross section that rises with CM energy:

$$\sigma \sim E^2. \tag{22.9}$$

Moreover, since the amplitude (equation (22.2)) only involves $(\cos\theta)^0$ and $(\cos\theta)^1$ contributions, it is clear that this rise in σ is associated with only a few partial waves, and is not due to more and more partial waves contributing to the sum in σ. Therefore, at some energy E, the unitarity bound will be violated by this lowest-order (Born approximation) expression for σ.

This is the essence of the 'unitarity disease' of the current–current model. To fill in all the details, however, involves a careful treatment of the appropriate partial wave analysis for the case when all particles carry spin. We shall avoid those details. Instead we argue, again on dimensional grounds, that the dimensionless partial wave amplitude f_J (note the $1/k^2$ factor in (22.5)) must be proportional to $G_\mathrm{F}E^2$, which violates the bound (22.7) for CM energies

$$E \geq G_\mathrm{F}^{-1/2} \sim 300\mathrm{GeV}. \tag{22.10}$$

At this point the reader may recall a very similar-sounding argument made in section 11.8, which led to the same estimate of the 'dangerous' energy scale (22.10). In that case, the discussion referred to a hypothetical '4-fermi' interaction without the V–A structure, and it was concerned with renormalization rather than unitarity. The gamma-matrix structure is irrelevant to these issues, which ultimately have to do with the dimensionality of the coupling constant, in both cases. In fact, as we shall see, unitarity and renormalizability are actually rather closely related.

Faced with this unitarity difficulty, we appeal to the most successful theory we have, and ask: what happens in QED? We consider an apparently quite similar process, namely $\mathrm{e}^+\mathrm{e}^- \to \mu^+\mu^-$ in lowest order (figure 22.2). In chapter 8 the total cross section for this process, neglecting lepton masses, was found to be (see problem 8.18 and equation (9.87))

$$\sigma = 4\pi\alpha^2/3E^2 \tag{22.11}$$

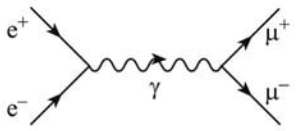


FIGURE 22.2
One-photon annihilation graph for $e^+e^- \to \mu^+\mu^-$.

which obediently falls with energy as required by unitarity. In this case the coupling constant α, analogous to G_F, is dimensionless, so that a factor E^2 is required in the denominator to give $\sigma \sim [L]^2$.

If we accept this clue from QED, we are led to search for a theory of weak interactions that involves a dimensionless coupling constant. Pressing the analogy with QED further will help us to see how one might arise. Fermi's current–current model was, as we said, motivated by the vector currents of QED. But in Fermi's case the currents interact directly with each other, whereas in QED they interact only indirectly via the mediation of the electromagnetic field. More formally, the Fermi current–current interaction has the 'four point' structure

$$`G_F(\bar{\hat{\psi}}\hat{\psi}) \cdot (\bar{\hat{\psi}}\hat{\psi})' \tag{22.12}$$

while QED has the 'three-point' (Yukawa) structure

$$`e\bar{\hat{\psi}}\hat{\psi}\hat{A}.' \tag{22.13}$$

Dimensional analysis easily shows, once again, that $[G_F] = M^{-2}$ while $[e] = M^0$. This strongly suggests that we should take Fermi's analogy further, and look for a weak interaction analogue of (22.13), having the form

$$`g\bar{\hat{\psi}}\hat{\psi}\hat{W}' \tag{22.14}$$

where $\hat{W}$ is a bosonic field. Dimensional analysis shows, of course, that $[g] = M^0$.

Since the weak currents are in fact vector-like, we must assume that the $\hat{W}$ fields are also vectors (spin-1) so as to make (22.14) Lorentz invariant. And because the weak interactions are plainly *not* long-range, like electromagnetic ones, the mass of the W quanta cannot be zero. So we are led to postulate the existence of a massive weak analogue of the photon, the 'intermediate vector boson' (IVB), and to suppose that weak interactions are mediated by the exchange of IVB's.

There is, of course, one further difference with electromagnetism, which is that the currents in β-decay, for example, carry charge (e.g. $\bar{\hat{\psi}}_e\gamma^\mu(1 - \gamma_5)\hat{\psi}_{\nu_e}$ creates negative charge or destroys positive charge). The 'companion'

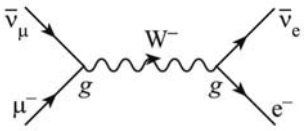


FIGURE 22.3
One-W$^-$ annihilation graph for $\bar{\nu}_\mu + \mu^- \to \bar{\nu}_e + e^-$.

hadronic current carries the opposite charge (e.g. $\bar{\hat{\psi}}_p \gamma_\mu (1 - r\gamma_5)\hat{\psi}_n$ destroys negative charge or creates positive charge), so as to make the total effective interaction charge-conserving, as required. It follows that the $\hat{W}$ fields must then be charged, so that expressions of the form (22.14) are neutral. Because both charge-raising and charge-lowering currents exist, we need both W$^+$ and W$^-$. The reaction (22.1), for example, is then conceived as proceeding via the Feynman diagram shown in figure 22.3, quite analogous to figure 22.2.

Because we also have weak neutral currents, we need a neutral vector boson as well, Z^0. In addition to all these, there is the familiar massless neutral vector boson, the photon. Despite the fact that they are *not* massless, the W$^\pm$ and Z^0 can be understood as gauge quanta, thanks to the symmetry-breaking mechanism explained in section 19.6. For the moment, however, we are going to follow a more scenic route, and accept (as Glashow did in 1961) that we are dealing with ordinary 'unsophisticated' massive vector particles, charged and uncharged.

We now investigate whether the IVB model can do any better with unitarity than the current–current model. The analysis will bear a close similarity to the discussion of the renormalizability of the model in section 19.1, and we shall take up that issue again in section 22.1.2.

The unitarity-violating processes turn out to be those involving *external* W particles. Consider, for example, the process

$$\nu_\mu + \bar{\nu}_\mu \to W^+ + W^- \tag{22.15}$$

proceeding via the graph shown in figure 22.4. The fact that this is experimentally a somewhat esoteric reaction is irrelevant for the subsequent argument: the proposed theory, represented by the IVB modification of the four-fermion model, will necessarily generate the amplitude shown in figure 22.4, and since this amplitude violates unitarity, the theory is unacceptable. The amplitude for this process is proportional to

$$\begin{aligned} \mathcal{M}_{\lambda_1\lambda_2} &= g^2 \epsilon_\mu^{-*}(k_2, \lambda_2)\epsilon_\nu^{+*}(k_1, \lambda_1)\bar{v}(p_2)\gamma^\mu(1-\gamma_5) \\ &\quad \times \frac{(\not{p}_1 - \not{k}_1 + m_\mu)}{(p_1 - k_1)^2 - m_\mu^2}\gamma^\nu(1-\gamma_5)u(p_1) \end{aligned} \tag{22.16}$$

where the $\epsilon^\pm$ are the polarization vectors of the W's: $\epsilon_\mu^{-*}(k_2, \lambda_2)$ is that

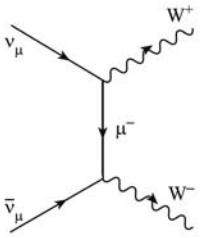


FIGURE 22.4
μ^--exchange graph for $\nu_\mu + \bar{\nu}_\mu \to \mathrm{W}^+ + \mathrm{W}^-$.

associated with the outgoing W^- with 4-momentum k_2 and polarization state λ_2, and similarly for ϵ_ν^{+*}.

To calculate the total cross section, we must form $|\mathcal{M}|^2$ and sum over the three states of polarization for each of the W's. To do this, we need the result

$$\sum_{\lambda=0,\pm1} \epsilon_\mu(k,\lambda)\epsilon_\nu^*(k,\lambda) = -g_{\mu\nu} + k_\mu k_\nu / M_{\mathrm{W}}^2 \tag{22.17}$$

already given in (19.19). Our interest will as usual be in the high-energy behaviour of the cross section, in which regime it is clear that the $k_\mu k_\nu / M_{\mathrm{W}}^2$ term in (22.17) will dominate the $g_{\mu\nu}$ term. It is therefore worth looking a little more closely at this term. From (19.17) and (19.18) we see that in a frame in which $k^\mu = (k^0, 0, 0, |\boldsymbol{k}|)$, the transverse polarization vectors $\epsilon^\mu(k, \lambda = \pm 1)$ involve no momentum dependence, which is in fact carried solely in the longitudinal polarization vector $\epsilon^\mu(k, \lambda = 0)$. We may write this as

$$\epsilon(k, \lambda = 0) = \frac{k^\mu}{M_{\mathrm{W}}} + \frac{M_{\mathrm{W}}}{(k^0 + |\boldsymbol{k}|)} \cdot (-1, \hat{\boldsymbol{k}}) \tag{22.18}$$

which at high energy tends to k^μ / M_{W}. Thus it is clear that it is the longitudinal polarization states which are responsible for the $k^\mu k^\nu$ parts of the polarization sum (12.21), and which will dominate real production of W's at high energy.

Concentrating therefore on the production of longitudinal W's, we are led to examine the quantity

$$\frac{g^4}{M_{\mathrm{W}}^4 (p_1 - k_1)^4} \mathrm{Tr}[\not{k}_2(1-\gamma_5)(\not{p}_1 - \not{k}_1)\ \not{k}_1\ \not{p}_1\ \not{k}_1(\not{p}_1 - \not{k}_1)\ \not{k}_2\ \not{p}_2] \tag{22.19}$$

where we have neglected m_μ, commuted the $(1-\gamma_5)$ factors through, and neglected neutrino masses, in forming $\sum_{\text{spins}} |\mathcal{M}_{00}|^2$. Retaining only the leading powers of energy, we find (see problem 22.2)

$$\sum_{\text{spins}} |\mathcal{M}_{00}|^2 \sim (g^4/M_{\mathrm{W}}^4)(p_1 \cdot k_2)(p_2 \cdot k_2) = (g^4/M_{\mathrm{W}}^4) E^4 (1 - \cos^2\theta) \tag{22.20}$$

where E is the CM energy and θ the CM scattering angle. We see that the (unsquared) amplitude must behave essentially as $g^2E^2/M_{\rm W}^2$, the quantity $g^2/M_{\rm W}^2$ effectively replacing $G_{\rm F}$ of the current–current model. The unitarity bound is violated for $E \geq M_{\rm W}/g \sim 300$ GeV, taking $g \sim e$.

Other unitarity-violating processes can easily be invented, and we have to conclude that the IVB model is, in this respect, no more fitted to be called a theory than was the four-fermion model. In the case of the latter, we argued that the root of the disease lay in the fact that $G_{\rm F}$ was not dimensionless, yet somehow this was not a good enough cure after all: perhaps (it is indeed so) 'dimensionlessness' is necessary but not sufficient (see the following section). Why is this? Returning to $\mathcal{M}_{\lambda_1,\lambda_2}$ for $\nu\bar{\nu} \to {\rm W}^+{\rm W}^-$ (equation (22.16)) and setting $\epsilon = k_\mu/M$ for the *longitudinal* polarization vectors, we see that we are involved with an effective amplitude

$$\frac{g^2}{M_{\rm W}^2}\bar{v}(p_2)\ \not{k}_2(1-\gamma_5)\frac{\not{p}_1 - \not{k}_1}{(p_1-k_1)^2}\ \not{k}_1(1-\gamma_5)u(p_1). \tag{22.21}$$

Using the Dirac equation $\not{p}_1 u(p_1) = 0$ and $p_1^2 = 0$, this can be reduced to

$$-\frac{g^2}{M_{\rm W}^2}\bar{v}(p_2)\ \not{k}_2(1-\gamma_5)u(p_1). \tag{22.22}$$

We see that the longitudinal ϵ's have brought in the factors $M_{\rm W}^{-2}$, which are 'compensated' by the factor $\not{k}_2$, and it is this latter factor which causes the rise with energy. The longitudinal polarization states have effectively reintroduced a dimensional coupling constant $g/M_{\rm W}$.

What happens in QED? We learnt in section 7.3 that, for real photons, the longitudinal state of polarization is absent altogether. We might well suspect, therefore, that since it was the longitudinal W's that caused the 'bad' high-energy behaviour of the IVB model, the 'good' high-energy behaviour of QED might have its origin in the absence of such states for photons. And this circumstance can, in its turn, be traced (cf section 7.3.1) to the *gauge invariance* property of QED.

Indeed, in section 8.6.3 we saw that in the analogue of (22.17) for photons (this time involving only the two transverse polarization states), the right-hand side could be taken to be just $-g_{\mu\nu}$, *provided that* the Ward identity (8.166) held, a condition directly following from gauge invariance.

We have arrived here at an important theoretical indication that what we really need is a *gauge theory* of the weak interactions, in which the W's are gauge quanta. It must, however, be a peculiar kind of gauge theory, since normally gauge invariance requires the gauge field quanta to be massless. However, we have already seen how this 'peculiarity' can indeed arise, if the local symmetry is spontaneously broken (chapter 19). But before proceeding to implement that idea, in the GSW theory, we discuss one further disease (related to the unitarity one) possessed by both current–current and IVB models – that of non-renormalizability.

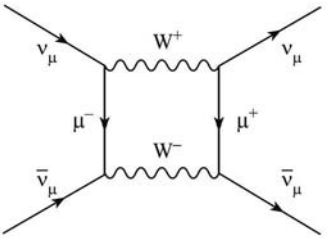


FIGURE 22.5
$O(g^4)$ contribution to $\nu_\mu\bar{\nu}_\mu \to \nu_\mu\bar{\nu}_\mu$.

22.1.2 The problem of non-renormalizability in weak interactions

The preceding line of argument about unitarity violations is open to the following objection. It is an argument conducted entirely within the framework of perturbation theory. What it shows, in fact, is simply that perturbation theory must fail, in theories of the type considered, at some sufficiently high energy. The essential reason is that the effective expansion parameter for perturbation theory is $EG_{\rm F}^{1/2}$. Since $EG_{\rm F}^{1/2}$ becomes large at high energy, arguments based on lowest-order perturbation theory are irrelevant. The objection is perfectly valid, and we shall take account of it by linking high-energy behaviour to the problem of renormalizability, rather than unitarity. We might, however, just note in passing that yet another way of stating the results of the previous two sections is to say that, for both the current–current and IVB theories, 'weak interactions become strong at energies of order 1 TeV'.

We gave an elementary introduction to renormalization in chapters 10 and 11 of volume 1. In particular, we discussed in some detail, in section 11.8, the difficulties that arise when one tries to do higher-order calculations in the case of a four-fermion interaction with the same form (apart from the V-A structure) as the current–current model. Its coupling constant, which we called $G_{\rm F}$, also had dimension $(\text{mass})^{-2}$. The 'non-renormalizable' problem was essentially that, as one approached the 'dangerous' energy scale (22.10), one needed to supply the values of an ever-increasing number of parameters from experiment, and the theory lost predictive power.

Does the IVB model fare any better? In this case, the coupling constant is dimensionless, just as in QED. 'Dimensionlessness' alone is not enough, it turns out: the IVB model is not renormalizable either. We gave an indication of why this is so in section 19.1, but we shall now be somewhat more specific, relating the discussion to the previous one about unitarity.

Consider, for example, the fourth-order processes shown in figure 22.5, for the IVB-mediated process $\nu_\mu\bar{\nu}_\mu \to \nu_\mu\bar{\nu}_\mu$. It seems plausible from the diagram that the amplitude must be formed by somehow 'sticking together' two copies

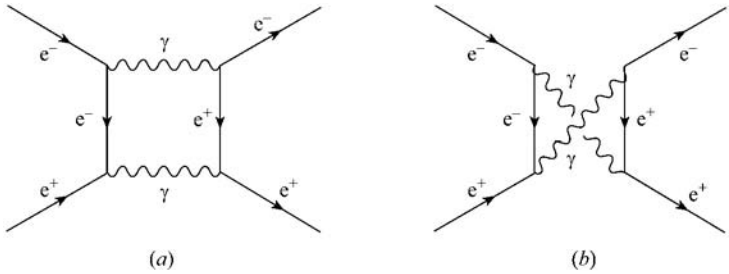


FIGURE 22.6
$O(e^4)$ contributions to $e^+e^- \to e^+e^-$.

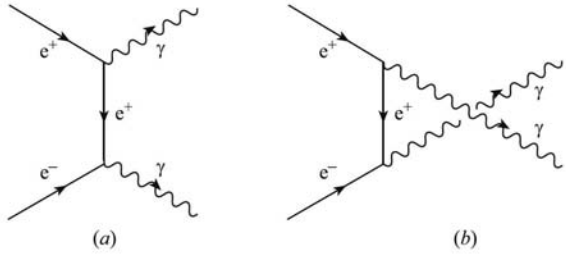


FIGURE 22.7
Lowest-order amplitudes for $e^+e^- \to \gamma\gamma$: (a) direct graph, (b) crossed graph.

of the tree graph shown in figure 22.4.[1] Now we saw that the high-energy behaviour of the amplitude $\nu\bar{\nu} \to W^+W^-$ (figure 22.4) grows as E^2, due to the k dependence of the longitudinal polarization vectors, and this turns out to produce, via figure 22.5, a non-renormalizable divergence, for the reason indicated in section 19.1 – namely, the 'bad' behaviour of the $k^\mu k^\nu / M_W^2$ factors in the W-propagators, at large k.

So it is plain that, once again, the blame lies with the longitudinal polarization states for the W's. Let us see how QED – a renormalizable theory – manages to avoid this problem. In this case, there are two box graphs, shown in figures 22.6. There are also two corresponding tree graphs, shown in figures 22.7(a) and (b). Consider, therefore mimicking for figures 22.7(a) and (b) the calculation we did for figure 22.4. We would obtain the leading high-energy behaviour by replacing the photon polarization vectors by the corresponding momenta, and it can be checked (problem 21.3) that when this replacement

[1]The reader may here usefully recall the discussion of unitarity for one-loop graphs in section 13.3.3.

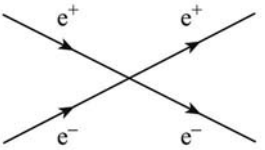


FIGURE 22.8
Four-point e^+e^- vertex.

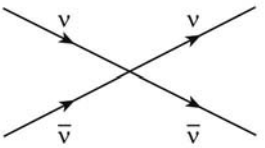


FIGURE 22.9
Four-point $\nu\bar{\nu}$ vertex.

is made for each photon the complete amplitude for the sum of figures 22.7(a) and (b) *vanishes*.

In physical terms, of course, this result was expected, since we knew in advance that it is always possible to choose polarization vectors for *real* photons such that they are purely transverse, so that no physical process can depend on a part of ϵ_μ proportional to k_μ. Nevertheless, the calculation is highly relevant to the question of renormalizing the graphs in figure 22.6. The photons in this process are not real external particles, but are instead virtual, internal ones. This has the consequence that we should in general include their longitudinal ($\epsilon_\mu \propto k_\mu$) states as well as the transverse ones (see section 13.3.3 for something similar in the case of unitarity for 1-loop diagrams). The calculation of problem 22.3 then suggests that these longitudinal states are harmless, provided that both contributions in figure 22.7 are included.

Indeed, the sum of these two box graphs for $e^+e^- \to e^+e^-$ is *not divergent*. If it were, an infinite counter term proportional to a four-point vertex $e^+e^- \to e^+e^-$ (figure 22.8) would have to be introduced, and the original QED theory, which of course lacks such a fundamental interaction, would not be renormalizable. This is exactly what *does* happen in the case of figure 22.5. The bad high-energy behaviour of $\nu\bar{\nu} \to W^+W^-$ translates into a divergence of figure 22.5 – and this time there is no 'crossed' amplitude to cancel it. This divergence entails the introduction of a new vertex, figure 22.9, not present in the original IVB theory. Thus the theory without this vertex is non-renormalizable – and if we include it, we are landed with a four-field pointlike vertex which is non-renormalizable, as in the Fermi (current–current) case.

Our presentation hitherto has emphasized the fact that, in QED, the bad high-energy behaviour is rendered harmless by a cancellation between contributions from figures 22.7(a) and (b) (or figures 22.6(a) and (b)). Thus one way to 'fix up' the IVB theory might be to hypothesize a new physical process, to be added to figure 22.4, in such a way that a cancellation occurred at high energies. The search for such high-energy cancellation mechanisms can indeed be pushed to a successful conclusion (Llewellyn Smith 1973), given sufficient ingenuity and, arguably, a little hindsight. However, we are in possession of a more powerful principle. In QED, we have already seen (section 8.6.2) that the vanishing of amplitudes when an ϵ_μ is replaced by the corresponding k_μ is due to *gauge invariance*: in other words, the potentially harmful longitudinal polarization states are in fact harmless in a gauge-invariant theory.

We have therefore arrived once more, after a somewhat more leisurely discussion than that of section 19.1, at the idea that we need a *gauge* theory of massive vector bosons, so that the offending $k^\mu k^\nu$ part of the propagator can be 'gauged away' as in the photon case. This is precisely what is provided by the 'spontaneously broken' gauge theory concept, as developed in chapter 19. There we saw that, taking the U(1) case for simplicity, the general expression for the gauge boson propagator in such a theory (in a 't Hooft gauge) is

$$\mathrm{i}\left[-g^{\mu\nu}+\frac{(1-\xi)k^\mu k^\nu}{k^2-\xi M_{\mathrm{W}}^2}\right]\Big/(k^2-M_{\mathrm{W}}^2+\mathrm{i}\epsilon) \tag{22.23}$$

where ξ is a gauge parameter. Our IVB propagator corresponds to the $\xi \to \infty$ limit, and with this choice of ξ all the troubles we have been discussing appear to be present. But for any finite ξ (for example $\xi = 1$) the high-energy behaviour of the propagator is actually $\sim 1/k^2$, the same as in the renormalizable QED case. This strongly suggests that such theories – in particular non-Abelian ones – are in fact renormalizable. 't Hooft's proof that they are ('t Hooft 1971b) triggered an explosion of theoretical work, as it became clear that, for the first time, it would be possible to make higher-order calculations for weak interaction processes using consistent renormalization procedures, of the kind that had worked so well for QED.

We now have all the pieces in place, and can proceed to introduce the GSW theory, based on the local gauge symmetry of SU(2) × U(1).

22.2 The SU(2) × U(1) electroweak gauge theory

22.2.1 Quantum number assignments; Higgs, W and Z masses

Given the preceding motivations for considering a gauge theory of weak interactions, the remaining question is this: what is the relevant symmetry

group of local phase transformations, i.e. the relevant *weak gauge group*? Several possibilities were suggested, but it is now very well established that the one originally proposed by Glashow (1961), subsequently treated as a spontaneously broken gauge symmetry by Weinberg (1967) and by Salam (1968), and later extended by other authors, produces a theory which is in remarkable agreement with currently known data. We shall not give a critical review of all the experimental evidence, but instead proceed directly to an outline of the GSW theory, introducing elements of the data at illustrative points.

An important clue to the symmetry group involved in the weak interactions is provided by considering the transitions induced by these interactions. This is somewhat analogous to discovering the multiplet structure of atomic levels and hence the representations of the rotation group, a prominent symmetry of the Schrödinger equation, by studying electromagnetic transitions. However, there is one very important difference between the 'weak multiplets' we shall be considering, and those associated with symmetries which are not spontaneously broken. We saw in chapter 12 how an unbroken non-Abelian symmetry leads to multiplets of states which are degenerate in mass, but in section 17.1 we learned that that result only holds provided the vacuum is left invariant under the symmetry transformation. When the symmetry is spontaneously broken, the vacuum is *not* invariant, and we must expect that the degenerate multiplet structure will then, in general, disappear completely. This is precisely the situation in the electroweak theory.

Nevertheless, as we shall see, essential consequences of the weak symmetry group – specifically, the relations it requires between otherwise unrelated masses and couplings – are accessible to experiment. Moreover, despite the fact that members of a multiplet of a global symmetry which is spontaneously broken will, in general, no longer have even approximately the same mass, the concept of a multiplet is still useful. This is because when the symmetry is made a *local* one, we shall find (in sections 22.2.2 and 22.2.3) that the associated gauge quanta still mediate interactions between members of a given symmetry multiplet, just as in the manifest local non-Abelian symmetry example of QCD. Now, the leptonic transitions associated with the weak charged currents are, as we saw in chapter 20, $\nu_{\rm e} \leftrightarrow {\rm e}, \nu_\mu \leftrightarrow \mu$ etc. This suggests that these pairs should be regarded as *doublets* under some group. Further we saw in section 20.7 how weak transitions involving charged quarks suggested a similar doublet structure for them also. The simplest possibility is therefore to suppose that, in both cases, a 'weak SU(2) group' is involved, called 'weak isospin'. We emphasize once more that this weak isospin is distinct from the hadronic isospin of chapter 12, which is part of $\mathrm{SU(3)_f}$. We use the symbols t, t_3 for the quantum numbers of weak isospin, and make the specific assignments for the leptonic fields

$$t = \frac{1}{2}, \qquad \begin{cases} t_3 = +1/2 \\ t_3 = -1/2 \end{cases} \quad \begin{pmatrix} \hat{\nu}'_{\rm e} \\ \hat{e}^- \end{pmatrix}_{\rm L}, \qquad \begin{pmatrix} \hat{\nu}'_\mu \\ \hat{\mu}^- \end{pmatrix}_{\rm L}, \qquad \begin{pmatrix} \hat{\nu}'_\tau \\ \hat{\tau}^- \end{pmatrix}_{\rm L} \tag{22.24}$$

where $\hat{e}_{\rm L} = \frac{1}{2}(1-\gamma_5)\hat{e}$ etc, and for the quark fields

$$t = \frac{1}{2}, \qquad \begin{cases} t_3 = +1/2 \\ t_3 = -1/2 \end{cases} \qquad \begin{pmatrix} \hat{u} \\ \hat{d}' \end{pmatrix}_{\rm L}, \qquad \begin{pmatrix} \hat{c} \\ \hat{s}' \end{pmatrix}_{\rm L}, \qquad \begin{pmatrix} \hat{t} \\ \hat{b}' \end{pmatrix}_{\rm L}. \quad (22.25)$$

As discussed in section 20.2.2, the subscript 'L' refers to the fact that only the left-handed chiral components of the fields enter, in consequence of the V–A structure. For this reason, the weak isospin group is referred to as $\mathrm{SU}(2)_{\rm L}$, to show that the weak isospin assignments and corresponding transformation properties apply only to these left-handed parts. Notice that, as anticipated for a spontaneously broken symmetry, these doublets all involve pairs of particles which are not mass degenerate. In (22.24) and (22.25), the primes indicate that these fields are related to the (unprimed) fields of definite mass by the unitary matrices **U** (for neutrinos) and **V** (for quarks), as discussed in sections 21.4.1 and 20.7.3 respectively.

Making this $\mathrm{SU}(2)_{\rm L}$ into a local phase invariance (following the logic of chapter 13) will entail the introduction of three gauge fields, transforming as a $t = 1$ multiplet (a triplet) under the group. Because (as with the ordinary $\mathrm{SU}(2)_{\rm f}$ of hadronic isospin) the members of a weak isodoublet differ by one unit of charge, the two gauge fields associated with transitions between doublet members will have charge ± 1. The quanta of these fields will, of course, be the now familiar $\mathrm{W}^{\pm}$ bosons mediating the charged current transitions, and associated with the weak isospin raising and lowering operators $t_{\pm}$. What about the third gauge boson of the triplet? This will be electrically neutral, and a very economical and appealing idea would be to associate this neutral vector particle with the photon, thereby *unifying* the weak and electromagnetic interactions. A model of this kind was originally suggested by Schwinger (1957). Of course, the W's must somehow acquire mass, while the photon remains massless. Schwinger arranged this by introducing appropriate couplings of the vector bosons to additional scalar and pseudoscalar fields. These couplings were arbitrary and no prediction of the W masses could be made. We now believe, following the arguments of the preceding section, that the W mass must arise via the spontaneous breakdown of a non-Abelian gauge symmetry, and as we saw in section 19.6, this *does* constrain the W mass.

Apart from the question of the W mass in Schwinger's model, we now know (see chapter 20) that there exist *neutral current* weak interactions, in addition to those of the charged currents. We must also include these in our emerging gauge theory, and an obvious suggestion is to have these currents mediated by the neutral member W^0 of the $\mathrm{SU}(2)_{\rm L}$ gauge field triplet. Such a scheme was indeed proposed by Bludman (1958), again pre-Higgs, so that W masses were put in 'by hand'. In this model, however, the neutral currents will have the same pure left-handed V–A structure as the charged currents: but, as we saw in chapter 20, the neutral currents are *not* pure V-A. Furthermore, the attractive feature of including the photon, and thus unifying weak and electromagnetic interactions, has been lost.

A key contribution was made by Glashow (1961); similar ideas were also advanced by Salam and Ward (1964). Glashow suggested enlarging the Schwinger–Bludman SU(2) schemes by inclusion of an additional U(1) gauge group, resulting in an '$\mathrm{SU(2)_L \times U(1)}$' group structure. The new Abelian U(1) group is associated with a weak analogue of hypercharge – 'weak hypercharge' – just as $\mathrm{SU(2)_L}$ was associated with 'weak isospin'. Indeed, Glashow proposed that the Gell-Mann–Nishijima relation for charges should also hold for these weak analogues, giving

$$eQ = e(t_3 + y/2) \tag{22.26}$$

for the electric charge Q (in units of e) of the t_3 member of a weak isomultiplet, assigned a weak hypercharge y. Clearly, therefore, the lepton doublets, $(\nu'_{\mathrm{e}}, \mathrm{e}^-)$, etc, then have $y = -1$, while the quark doublets (u, d'), etc, have $y = +\frac{1}{3}$. Now, when *this* group is gauged, everything falls marvellously into place: the charged vector bosons appear as before, but there are now *two* neutral vector bosons, which between them will be responsible for the weak neutral current processes, and for electromagnetism. This is exactly the piece of mathematics we went through in section 19.6, which we now appropriate as an important part of the Standard Model.

For convenience, we reproduce here the main results of section 19.6. The Higgs field $\hat{\phi}$ is an SU(2) doublet

$$\hat{\phi} = \begin{pmatrix} \hat{\phi}^+ \\ \hat{\phi}^0 \end{pmatrix} \tag{22.27}$$

with an assumed vacuum expectation value (in unitary gauge) given by

$$\langle 0|\hat{\phi}|0\rangle = \begin{pmatrix} 0 \\ v/\sqrt{2} \end{pmatrix}. \tag{22.28}$$

Fluctuations about this value are parametrized in this gauge by

$$\hat{\phi} = \begin{pmatrix} 0 \\ \frac{1}{\sqrt{2}}(v + \hat{H}) \end{pmatrix} \tag{22.29}$$

where $\hat{H}$ is the (physical) Higgs field. The Lagrangian for the sector consisting of the gauge fields and the Higgs fields is

$$\mathcal{L}_{\mathrm{G}\Phi} = (\hat{D}_\mu\hat{\phi})^\dagger(\hat{D}^\mu\hat{\phi}) + \mu^2\hat{\phi}^\dagger\hat{\phi} - \frac{\lambda}{4}(\hat{\phi}^\dagger\hat{\phi})^2 - \frac{1}{4}\hat{\boldsymbol{F}}_{\mu\nu}\cdot\hat{\boldsymbol{F}}^{\mu\nu} - \frac{1}{4}\hat{G}_{\mu\nu}\hat{G}^{\mu\nu}, \tag{22.30}$$

where $\hat{\boldsymbol{F}}_{\mu\nu}$ is the SU(2) field strength tensor (19.80) for the gauge fields $\hat{\boldsymbol{W}}^\mu$ and $\hat{G}_{\mu\nu}$ is the U(1) field strength tensor (19.81) for the gauge field B^μ, and $\hat{D}^\mu\hat{\phi}$ is given by (19.79). After symmetry breaking (i.e. the insertion of (22.29)

in (22.30)) the quadratic parts of (22.30) can be written in unitary gauge as (see problem 19.9)

$$\hat{\mathcal{L}}^{\text{free}}_{\text{G}\Phi} = \frac{1}{2}\partial_\mu \hat{H}\partial^\mu \hat{H} - \mu^2 \hat{H}^2 \tag{22.31}$$

$$- \frac{1}{4}(\partial_\mu \hat{W}_{1\nu} - \partial_\nu \hat{W}_{1\mu})(\partial^\mu \hat{W}_1^\nu - \partial^\nu \hat{W}_1^\mu) + \frac{1}{8}g^2 v^2 \hat{W}_{1\mu}\hat{W}_1^\mu \tag{22.32}$$

$$- \frac{1}{4}(\partial_\mu \hat{W}_{2\nu} - \partial_\nu \hat{W}_{2\mu})(\partial^\mu \hat{W}_2^\nu - \partial^\nu \hat{W}_2^\mu) + \frac{1}{8}g^2 v^2 \hat{W}_{2\mu}\hat{W}_2^\mu \tag{22.33}$$

$$- \frac{1}{4}(\partial_\mu \hat{Z}_\nu - \partial_\nu \hat{Z}_\mu)(\partial^\mu \hat{Z}^\nu - \partial^\nu \hat{Z}^\mu) + \frac{v^2}{8}(g^2 + g'^2)\hat{Z}_\mu \hat{Z}^\mu \tag{22.34}$$

$$- \frac{1}{4}\hat{F}_{\mu\nu}\hat{F}^{\mu\nu} \tag{22.35}$$

where

$$\hat{Z}^\mu = \cos\theta_{\text{W}} \hat{W}_3^\mu - \sin\theta_{\text{W}} \hat{B}^\mu, \tag{22.36}$$

$$\hat{A}^\mu = \sin\theta_{\text{W}} \hat{W}_3^\mu + \cos\theta_{\text{W}} \hat{B}^\mu, \tag{22.37}$$

and

$$\hat{F}^{\mu\nu} = \partial^\mu A^\nu - \partial^\nu A^\mu, \tag{22.38}$$

with

$$\cos\theta_{\text{W}} = g/(g^2 + g'^2)^{1/2}, \qquad \sin\theta_{\text{W}} = g'/(g^2 + g'^2)^{1/2}. \tag{22.39}$$

Feynman rules for the vector boson propagators (in unitary gauge) and couplings, and for the Higgs couplings, can be read off from (22.30), and are given in appendix Q.

Equations (22.31)–(22.35) give the tree-level masses of the Higgs boson and the gauge bosons: (22.31) tells us that the mass of the Higgs boson is

$$m_{\text{H}} = \sqrt{2}\mu = \sqrt{\lambda}\, v/\sqrt{2}, \tag{22.40}$$

where $v/\sqrt{2}$ is the (tree-level) Higgs vacuum value; (22.32) and (22.33) show that the charged W's have a mass

$$M_{\text{W}} = gv/2 \tag{22.41}$$

where g is the $\text{SU}(2)_{\text{L}}$ gauge coupling constant; (22.34) gives the mass of the Z^0 as

$$M_{\text{Z}} = M_{\text{W}}/\cos\theta_{\text{W}} \tag{22.42}$$

and (22.35) shows that the A^μ field describes a massless particle (to be identified with the photon).

Still unaccounted for are the *right-handed* chiral components of the fermion fields. There is at present no evidence for any weak interactions coupling to the right-handed field components, and it is therefore natural – and a basic assumption of the electroweak theory – that all 'R' components are singlets

TABLE 22.1
Weak isospin and hypercharge assignments.

	t	t_3	y	Q
$\nu'_{\mathrm{eL}},\ \nu'_{\mu\mathrm{L}},\ \nu'_{\tau\mathrm{L}}$	1/2	1/2	-1	0
$\nu'_{\mathrm{eR}},\ \nu'_{\mu\mathrm{R}},\ \nu'_{\tau\mathrm{R}}$	0	0	0	0
$\mathrm{e_L},\ \mu_{\mathrm{L}},\ \tau_{\mathrm{L}}$	1/2	-1/2	-1	-1
$\mathrm{e_R},\ \mu_{\mathrm{R}},\ \tau_{\mathrm{R}}$	0	0	-2	-1
$\mathrm{u_L},\ \mathrm{c_L},\ \mathrm{t_L}$	1/2	1/2	1/3	2/3
$\mathrm{u_R},\ \mathrm{c_R},\ \mathrm{t_R}$	0	0	4/3	2/3
$\mathrm{d'_L},\ \mathrm{s'_L},\ \mathrm{b'_L}$	1/2	-1/2	1/3	-1/3
$\mathrm{d'_R},\ \mathrm{s'_R},\ \mathrm{b'_R}$	0	0	-2/3	-1/3
ϕ^+	1/2	1/2	1	1
ϕ^0	1/2	-1/2	1	0

under the weak isospin group. Crucially, however, the 'R' components do interact via the U(1) field $\hat{B}^\mu$; it is this that allows electromagnetism to emerge free of parity-violating γ_5 terms, as we shall see. With the help of the weak charge formula (equation (22.26)), we arrive at the assignments shown in table 22.1.

We have included 'R' components for the neutrinos in the table. It is, however, fair to say that in the original Standard Model the neutrinos were taken to be massless, with no neutrino mixing. We have seen in chapter 20 that it is for many purposes an excellent approximation to treat the neutrinos as massless, except when discussing neutrino oscillations. We shall mention their masses again in section 22.5.2, but for the moment we proceed in the 'massless neutrinos' approximation. In this case, there are *no* 'R' components for neutrinos, and no neutrino mixing.

We can now proceed to write down the *currents* of the electroweak theory. We will show that these dynamical symmetry currents are precisely the same as the phenomenological currents of the current–current model developed in chapter 20. The new feature here is that – as in the electromagnetic case – the currents interact with each other by the exchange of a gauge boson, rather than directly.

22.2.2 The leptonic currents (massless neutrinos): relation to current–current model

We write the $\mathrm{SU(2)_L}\times\mathrm{U(1)}$ covariant derivative, in terms of the fields $\hat{\boldsymbol{W}}^\mu$ and $\hat{B}^\mu$ of section 19.6, as

$$\hat{D}^\mu = \partial^\mu + \mathrm{i}g\boldsymbol{\tau}\cdot\hat{\boldsymbol{W}}^\mu/2 + \mathrm{i}g'y\hat{B}^\mu/2 \qquad \text{on 'L' SU(2) doublets} \tag{22.43}$$

and as

$$\hat{D}^\mu = \partial^\mu + \mathrm{i}g'y\hat{B}^\mu/2 \qquad \text{on 'R' SU(2) singlets.} \tag{22.44}$$

The leptonic couplings to the gauge fields therefore arise from the 'gauge-covariantized' free leptonic Lagrangian:

$$\hat{\mathcal{L}}_{\text{lept}} = \sum_{f=\text{e},\mu,\tau} \bar{\hat{l}}_{f\text{L}} \text{i}\hat{\not{D}} \hat{l}_{f\text{L}} + \sum_{f=\text{e},\mu,\tau} \bar{\hat{l}}_{f\text{R}} \text{i}\hat{\not{D}} \hat{l}_{f\text{R}}, \tag{22.45}$$

where the $\hat{l}_{f\text{L}}$ are the left-handed doublets

$$\hat{l}_{f\text{L}} = \begin{pmatrix} \hat{\nu}_f \\ \hat{f}^- \end{pmatrix}_{\text{L}} \tag{22.46}$$

and $\hat{l}_{f\text{R}}$ are the singlets $\hat{l}_{\text{eR}} = \hat{e}_{\text{R}}$ etc.

Consider first the *charged leptonic currents*. The correct normalization for the charged fields is that $\hat{W}^\mu \equiv (\hat{W}_1^\mu - \text{i}\hat{W}_2^\mu)/\sqrt{2}$ destroys the W^+ or creates the W^- (cf (7.15)). The '$\boldsymbol{\tau} \cdot \hat{\boldsymbol{W}}/2$' terms can be written as

$$\boldsymbol{\tau} \cdot \hat{\boldsymbol{W}}^\mu/2 = \frac{1}{\sqrt{2}} \left\{ \tau_+ \frac{(\hat{W}_1^\mu - \text{i}\hat{W}_2^\mu)}{\sqrt{2}} + \tau_- \frac{(\hat{W}_1^\mu + \text{i}\hat{W}_2^\mu)}{\sqrt{2}} \right\} + \frac{\tau_3}{2} \hat{W}_3^\mu, \tag{22.47}$$

where $\tau_\pm = (\tau_1 \pm \text{i}\tau_2)/2$ are the usual raising and lowering operators for the doublets. Thus the 'f=e' contribution to the first term in (22.45) picks out the process $\text{e}^- \to \nu_\text{e} + \text{W}^-$ for example, with the result that the corresponding vertex is given by

$$-\frac{\text{i}g}{\sqrt{2}} \gamma^\mu \frac{(1-\gamma_5)}{2}. \tag{22.48}$$

The 'universality' of the single coupling constant 'g' ensures that (22.48) is also the amplitude for the $\mu - \nu_\mu - \text{W}$ and $\tau - \nu_\tau - \text{W}$ vertices. Thus the amplitude for the $\nu_\mu + \text{e}^- \to \mu^- + \nu_\text{e}$ process considered in section 20.8 is

$$\left\{ -\frac{\text{i}g}{\sqrt{2}} \bar{u}(\mu) \gamma_\mu \frac{(1-\gamma_5)}{2} u(\nu_\mu) \right\} \frac{\text{i}[-g^{\mu\nu} + k^\mu k^\nu / M_\text{W}^2]}{k^2 - M_\text{W}^2} \left\{ -\text{i}\frac{g}{\sqrt{2}} \bar{u}(\nu_\text{e}) \gamma_\nu \frac{(1-\gamma_5)}{2} u(\text{e}) \right\} \tag{22.49}$$

corresponding to the Feynman graph of figure 22.10.

For $k^2 \ll M_\text{W}^2$ we can replace the W-propagator by the constant value $g^{\mu\nu}/M_\text{W}^2$, leading to the amplitude

$$-\frac{\text{i}g^2}{8M_\text{W}^2} \bar{u}(\mu) \gamma_\mu (1-\gamma_5) u(\nu_\mu) \bar{u}(\nu_\text{e}) \gamma^\mu (1-\gamma_5) u(\text{e}), \tag{22.50}$$

which may be compared with the form we used in the current–current theory, equation (20.50). This comparison gives

$$\frac{G_\text{F}}{\sqrt{2}} = \frac{g^2}{8M_\text{W}^2}. \tag{22.51}$$

This is an important equation, giving the precise version, in the GSW theory,

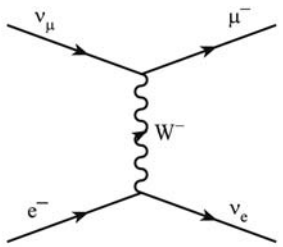


FIGURE 22.10
W-exchange process in $\nu_\mu + e^- \to \mu^- + \nu_e$.

of the qualitative relation $g^2/M_W^2 \sim G_F$ introduced following equation (22.20), and in volume 1, at equation (1.32).

Putting together (22.41) and (22.51) we can deduce

$$G_F/\sqrt{2} = 1/(2v^2) \tag{22.52}$$

so that from the known value (22.4) of G_F there follows the value of v:

$$v \simeq 246 \text{ GeV}. \tag{22.53}$$

Alternatively we may quote $v/\sqrt{2}$ (the vacuum value of the Higgs field):

$$v/\sqrt{2} \simeq 174 \text{ GeV}. \tag{22.54}$$

This parameter sets the scale of electroweak symmetry breaking, but as yet no theory is able to predict its value. It is related to the parameters λ, μ of (22.30) by $v/\sqrt{2} = \sqrt{2}\mu/\lambda^{1/2}$ (cf (17.98)).

In general, the charge-changing part of (22.45) can be written as

$$-\frac{g}{\sqrt{2}}\left\{\bar{\hat{\nu}}_e\gamma^\mu\frac{(1-\gamma_5)}{2}\hat{e} + \bar{\hat{\nu}}_\mu\gamma^\mu\frac{(1-\gamma_5)}{2}\hat{\mu} + \bar{\hat{\nu}}_\tau\gamma^\mu\frac{(1-\gamma_5)}{2}\hat{\tau}\right\}\hat{W}_\mu$$
$$+\text{hermitian conjugate}, \tag{22.55}$$

where $\hat{W}^\mu = (\hat{W}_1^\mu - i\hat{W}_2^\mu)/\sqrt{2}$. (22.55) has the form

$$-\hat{j}^\mu_{CC}(\text{leptons})\hat{W}_\mu - j^{\mu\dagger}_{CC}(\text{leptons})\hat{W}^\dagger_\mu \tag{22.56}$$

where the *leptonic weak charged current* $\hat{j}^\mu_{CC}(\text{leptons})$ is precisely that used in the current–current model (equation (20.38)), up to the usual factors of g's and $\sqrt{2}$'s. Thus the dynamical symmetry currents of the $SU(2)_L$ gauge theory are exactly the 'phenomenological' currents of the earlier current–current model. The Feynman rules for the lepton-W couplings (appendix Q) can be read off from (22.55).

Turning now to the *leptonic weak neutral current*, this will appear via the couplings to the Z^0, written as

$$-\hat{j}^{\mu}_{\rm NC}(\text{leptons})\hat{Z}_{\mu}. \tag{22.57}$$

Referring to (22.36) for the linear combination of $\hat{W}^{\mu}_3$ and $\hat{B}^{\mu}$ which represents $\hat{Z}^{\mu}$, we find (problem 22.4)

$$\hat{j}^{\mu}_{\rm NC}(\text{leptons}) = \frac{g}{\cos\theta_{\rm W}} \sum_l \bar{\hat{\psi}}_l \gamma^{\mu} \left[t^l_3 \left(\frac{1-\gamma_5}{2} \right) - \sin^2\theta_{\rm W} Q_l \right] \hat{\psi}_l, \tag{22.58}$$

where the sum is over the six lepton fields $\nu_{\rm e}, {\rm e}^-, \nu_{\mu}, \ldots \tau^-$. For the $Q = 0$ neutrinos with $t_3 = +\frac{1}{2}$,

$$\hat{j}^{\mu}_{\rm NC}(\text{neutrinos}) = \frac{g}{2\cos\theta_{\rm W}} \sum_l \bar{\hat{\nu}}_l \gamma^{\mu} \frac{(1-\gamma_5)}{2} \hat{\nu}_l, \tag{22.59}$$

where now $l = {\rm e}, \mu, \tau$. For the other (negatively charged) leptons, we shall have both L and R couplings from (22.58), and we can write

$$\hat{j}^{\mu}_{\rm NC}(\text{charged leptons}) = \frac{g}{\cos\theta_{\rm W}} \sum_{l={\rm e},\mu,\tau} \bar{\hat{l}} \gamma^{\mu} \left[c^l_{\rm L} \left(\frac{1-\gamma_5}{2} \right) + c^l_{\rm R} \left(\frac{1+\gamma_5}{2} \right) \right] \hat{l}, \tag{22.60}$$

where

$$c^l_{\rm L} = t^l_3 - \sin^2\theta_{\rm W} Q_l = -\frac{1}{2} + \sin^2\theta_{\rm W} \tag{22.61}$$

$$c^l_{\rm R} = -\sin^2\theta_{\rm W} Q_l = \sin^2\theta_{\rm W}. \tag{22.62}$$

As noted earlier, the Z^0 coupling is not pure 'V–A'. These relations (22.59)–(22.62) are exactly the ones given earlier, in (20.85)–(20.87); in particular, the couplings are independent of 'l' and hence exhibit lepton universality. The alternative notation

$$\hat{j}^{\mu}_{\rm NC}(\text{charged leptons}) = \frac{g}{2\cos\theta_{\rm W}} \sum_l \bar{\hat{l}} \gamma^{\mu} (g^l_{\rm V} - g^l_{\rm A}\gamma_5) \hat{l} \tag{22.63}$$

is often used, where

$$g^l_{\rm V} = -\frac{1}{2} + 2\sin^2\theta_{\rm W} \qquad g^l_{\rm A} = -\frac{1}{2}, \text{independent of } l. \tag{22.64}$$

Note that the $g^l_{\rm V}$ vanishes for $\sin^2\theta_{\rm W} = 0.25$. Again, the Feynman rules for lepton-Z couplings (appendix Q) are contained in (22.59) and (22.60).

As in the case of W-mediated charge-charging processes, Z^0-mediated processes reduce to the current–current form at low k^2. For example, the amplitude for ${\rm e}^-\mu^- \to {\rm e}^-\mu^-$ via Z^0 exchange (figure 22.11) reduces to

$$-\frac{{\rm i}g^2}{4\cos^2\theta_{\rm W} M^2_{\rm Z}} \quad \bar{u}({\rm e})\gamma_{\mu}[c^l_{\rm L}(1-\gamma_5) + c^l_{\rm R}(1+\gamma_5)]u({\rm e})\bar{u}(\mu)\gamma^{\mu} \times [c^l_{\rm L}(1-\gamma_5) + c^l_{\rm R}(1+\gamma_5)]u(\mu). \tag{22.65}$$

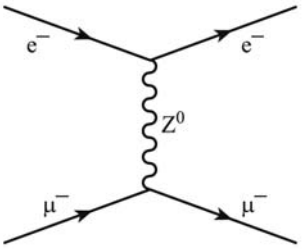


FIGURE 22.11
Z^0-exchange process in $e^-\mu^- \to e^-\mu^-$.

It is customary to define the parameter

$$\rho = M_W^2/(M_Z^2 \cos^2\theta_W), \tag{22.66}$$

which is unity at tree-level, in the absence of loop corrections. The ratio of factors in front of the $\bar{u}\ldots u$ expressions in (22.65) and (22.50) (i.e. 'neutral current process'/'charged current process') is then 2ρ.

We may also check the electromagnetic current in the theory, by looking for the piece that couples to $\hat{A}^\mu$. We find

$$\hat{j}^\mu_{\rm emag} = -g\sin\theta_W \sum_{l={\rm e},\mu,\tau} \bar{\hat{l}}\gamma^\mu \hat{l} \tag{22.67}$$

which allows us to identify the electromagnetic charge e as

$$e = g\sin\theta_W \tag{22.68}$$

as already suggested in (19.97) of chapter 19. Note that all the γ_5's cancel from (22.67), as is of course required.

22.2.3 The quark currents

The charge-changing quark currents, which are coupled to the $W^\pm$ fields, have a form very similar to that of the charged leptonic currents, except that the $t_3 = -\frac{1}{2}$ components of the L-doublets have to be understood as the flavour-mixed (weakly interacting) states

$$\begin{pmatrix} \hat{d}' \\ \hat{s}' \\ \hat{b}' \end{pmatrix}_{\rm L} = \begin{pmatrix} V_{\rm ud} & V_{\rm us} & V_{\rm ub} \\ V_{\rm cd} & V_{\rm cs} & V_{\rm cb} \\ V_{\rm td} & V_{\rm ts} & V_{\rm tb} \end{pmatrix} \begin{pmatrix} \hat{d} \\ \hat{s} \\ \hat{b} \end{pmatrix}_{\rm L}, \tag{22.69}$$

where $\hat{d}, \hat{s}$ and $\hat{b}$ are the strongly interacting fields with masses $m_{\rm d}, m_{\rm s}$ and $m_{\rm b}$, and the V-matrix is the CKM matrix used extensively in chapter 21. We

shall discuss this matrix further in section 22.5.2. Thus the *charge-changing weak quark current* is

$$\hat{j}^{\mu}_{\rm CC}(\text{quarks}) = \frac{g}{\sqrt{2}}\left\{\bar{\hat{u}}\gamma^{\mu}\frac{(1-\gamma_5)}{2}\hat{d}' + \bar{\hat{c}}\gamma^{\mu}\frac{(1-\gamma_5)}{2}\hat{s}' + \bar{\hat{t}}\gamma^{\mu}\frac{(1-\gamma_5)}{2}\hat{b}'\right\}, \tag{22.70}$$

which generalizes (20.90) to three generations and supplies the factor $g/\sqrt{2}$, as for the leptons.

The neutral currents are diagonal in flavour if the matrix V is unitary (see also section 22.5.2). Thus $\hat{j}^{\mu}_{\rm NC}(\text{quarks})$ will be given by the same expression as (20.103), except that now the sum will be over all six quark flavours. The *neutral weak quark current* is thus

$$\hat{j}^{\mu}_{\rm NC}(\text{quarks}) = \frac{g}{\cos\theta_{\rm W}}\sum_q \bar{\hat{q}}\gamma^{\mu}\left[c^q_{\rm L}\frac{(1-\gamma_5)}{2} + c^q_{\rm R}\frac{(1+\gamma_5)}{2}\right]\hat{q}, \tag{22.71}$$

where

$$c^q_{\rm L} = t^q_3 - \sin^2\theta_{\rm W}Q_q \tag{22.72}$$
$$c^q_{\rm R} = -\sin^2\theta_{\rm W}Q_q. \tag{22.73}$$

These expressions are exactly as given in (20.103)–(20.105). As for the charged leptons, we can alternatively write (22.71) as

$$\hat{j}^{\mu}_{\rm NC}(\text{quarks}) = \frac{g}{2\cos\theta_{\rm W}}\sum_q \bar{\hat{q}}\gamma^{\mu}(g^q_{\rm V} - g^q_{\rm A}\gamma_5)\hat{q}, \tag{22.74}$$

where

$$g^q_{\rm V} = t^q_3 - 2\sin^2\theta_{\rm W}Q_q \tag{22.75}$$
$$g^q_{\rm A} = t^q_3. \tag{22.76}$$

Before proceeding to discuss some simple phenomenological consequences, we remind the reader of one important feature of the Standard Model currents in general. Reading (22.24) and (22.25) together 'vertically', the leptons and quarks are grouped in three *generations*, each with two leptons and two quarks. The theoretical motivation for such family grouping is that *anomalies* are cancelled within each complete generation, as discussed in section 18.4.

22.3 Simple (tree-level) predictions

The theory as so far developed has just 4 parameters: the gauge couplings g and g', and the parameters λ and μ of the Higgs potential. The previous two

subsections show that all the couplings to fermions can be written in terms of the known quantities $G_{\rm F}$ and e (or α), and one free parameter which may be taken to be $\sin\theta_{\rm W}$. We noted in section 20.9 that, before the discovery of the W and Z particles, the then known neutrino data were consistent with a single value of $\theta_{\rm W}$ given by $\sin^2\theta_{\rm W} \simeq 0.23$. Using (22.51) and (22.68), it was then possible to predict the value of $M_{\rm W}$:

$$M_{\rm W} = \left(\frac{\pi\alpha}{\sqrt{2}G_{\rm F}}\right)^{1/2} \frac{1}{\sin\theta_{\rm W}} \simeq \frac{37.28}{\sin\theta_{\rm W}}\text{GeV} \simeq 77.73 \text{ GeV}. \tag{22.77}$$

Similarly, using (22.42) we predict

$$M_{\rm Z} = M_{\rm W}/\cos\theta_{\rm W} \simeq 88.58 \text{ GeV}. \tag{22.78}$$

These predictions of the theory (at lowest order) indicate the power of the underlying symmetry to tie together many apparently unrelated quantities, which are all determined in terms of only a few basic parameters. We now present a number of other simple tree-level predictions.

The width for $\mathrm{W}^- \to \mathrm{e}^- + \bar{\nu}_{\rm e}$ can be calculated using the vertex (22.48), with the result (problem 22.5)

$$\Gamma(\mathrm{W}^- \to \mathrm{e}^-\bar{\nu}_{\rm e}) = \frac{1}{12}\frac{g^2}{4\pi}M_{\rm W} = \frac{G_{\rm F}}{2^{1/2}}\frac{M_{\rm W}^3}{6\pi} \simeq 205 \text{ MeV}, \tag{22.79}$$

using (22.77). The widths to $\mu^-\bar{\nu}_\mu, \tau^-\bar{\nu}_\tau$ are the same. Neglecting CKM flavour mixing among the two energetically allowed quark channels $\bar{\rm u}\rm d$ and $\bar{\rm c}\rm s$, their widths would also be the same, apart from a factor of 3 for the different colour channels. The total W width for all these channels will therefore be about nine times the value in (22.79), i.e. 1.85 GeV, while the branching ratio for $\mathrm{W} \to \mathrm{e}\nu$ is

$$B(\mathrm{e}\nu) = \Gamma(\mathrm{W} \to \mathrm{e}\nu)/\Gamma(\text{total}) \simeq 11\%. \tag{22.80}$$

In making these estimates we have neglected all fermion masses.

The width for $\mathrm{Z}^0 \to \nu\bar{\nu}$ can be found from (22.79) by replacing $g/2^{1/2}$ by $g/2\cos\theta_{\rm W}$, and $M_{\rm W}$ by $M_{\rm Z}$, giving

$$\Gamma(\mathrm{Z}^0 \to \nu\bar{\nu}) = \frac{1}{24}\frac{g^2}{4\pi}\frac{M_{\rm Z}}{\cos^2\theta_{\rm W}} = \frac{G_{\rm F}}{2^{1/2}}\frac{M_{\rm Z}^3}{12\pi} \simeq 152 \text{ MeV}, \tag{22.81}$$

using (22.78). Charged lepton pairs couple with both $c_{\rm L}^l$ and $c_{\rm R}^l$ terms, leading (with neglect of lepton masses) to

$$\Gamma(\mathrm{Z}^0 \to l\bar{l}) = \left(\frac{|c_{\rm L}^l|^2 + |c_{\rm R}^l|^2}{6}\right)\frac{g^2}{4\pi}\frac{M_{\rm Z}}{\cos^2\theta_{\rm W}}. \tag{22.82}$$

The values $c_{\rm L}^\nu = \frac{1}{2}, c_{\rm R}^\nu = 0$ in (22.82) reproduce (22.81). With $\sin^2\theta_{\rm W} \simeq 0.23$, we find

$$\Gamma(\mathrm{Z}^0 \to l\bar{l}) \simeq 76.5 \text{ MeV}. \tag{22.83}$$

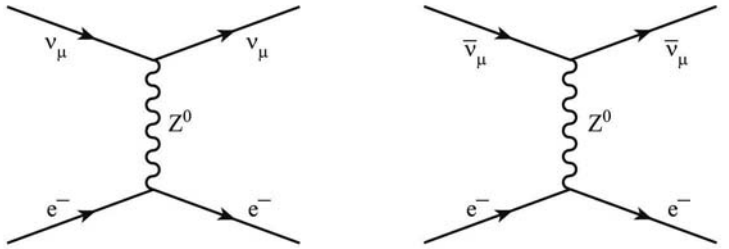


FIGURE 22.12
Neutrino-electron graphs involving Z^0 exchange.

Quark pairs couple as in (22.71), the GIM mechanism ensuring that all flavour-changing terms cancel. The total width to $u\bar{u}, d\bar{d}, c\bar{c}, s\bar{s}$ and $b\bar{b}$ channels (allowing 3 for colour and neglecting masses) is then 1538 MeV, producing an estimated total width of approximately 2.22 GeV. (QCD corrections will increase these estimates by a factor of order 1.1). The branching ratio to charged leptons is approximately 3.4%, to the three (invisible) neutrino channels 20.5%, and to hadrons (via hadronization of the $q\bar{q}$ channels) about 69.3%. In section 22.4.3 we shall see how a precise measurement of the total Z^0 width at LEP determined the number of light neutrinos to be 3.

Cross sections for neutrino–lepton scattering proceeding via Z^0 exchange can be calculated (for $k^2 \ll M_Z^2$) using the currents (22.59) and (22.60), and the method of section 20.5. Examples are

$$\nu_\mu e^- \to \nu_\mu e^- \tag{22.84}$$

and

$$\bar{\nu}_\mu e^- \to \bar{\nu}_\mu e^- \tag{22.85}$$

as shown in figure 22.12. Since the neutral current for the electron is not pure V–A, as was the charged current, we expect to see terms involving both $|c_L^l|^2$ and $|c_R^l|^2$, and possibly an interference term. The cross section for (22.84) is found to be ('t Hooft 1971c)

$$d\sigma/dy = (2G_F^2 E m_e/\pi)[|c_L^l|^2 + |c_R^l|^2(1-y)^2 - \frac{1}{2}(c_R^{l\,*} c_L^l + c_L^{l\,*} c_R^l) y m_e/E], \tag{22.86}$$

where E is the energy of the incident neutrino in the 'laboratory' system, and $y = (E - E')/E$ as before, where E' is the energy of the outgoing neutrino in the 'laboratory' system[2]. Equation (22.86) may be compared with the $\nu_\mu e^- \to \mu^- \nu_e$ (charged current) cross section of (20.84) by noting that $t = -2m_e E y$: the $|c_L^l|^2$ term agrees with the pure V–A result (20.84), while the $|c_R^l|^2$ term

[2]In the kinematics, lepton masses have been neglected wherever possible.

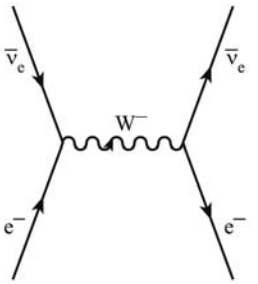


FIGURE 22.13
One-W annihilation graph in $\bar{\nu}_e e^- \to \bar{\nu}_e e^-$.

involves the same $(1-y)^2$ factor discussed for $\nu\bar{q}$ scattering in section 20.7.2. The interference term is negligible for $E \gg m_e$. The cross section for the antineutrino process (22.85) is found from (22.86) by interchanging c_L^l and c_R^l.

A third neutrino–lepton process is experimentally available,

$$\bar{\nu}_e e^- \to \bar{\nu}_e e^-, \tag{22.87}$$

the cross section for which was measured by Reines, Gurr and Sobel (1976), using electron antineutrinos from an 1800-MW fission reactor at Savannah River. In this case there is a single W intermediate state graph, shown in figure 22.13, to consider as well as the Z^0 one; the latter is similar to the right-hand graph in figure 22.12, but with $\bar{\nu}_\mu$ replaced by $\bar{\nu}_e$. The cross section for (22.87) turns out to be given by an expression of the form (22.86), but with the replacements

$$c_L^l \to \frac{1}{2} + \sin^2\theta_W, c_R^l \to \sin^2\theta_W. \tag{22.88}$$

Reines, Gurr and Sobel reported the result $\sin^2\theta_W = 0.29 \pm 0.05$.

We emphasize once more that all these cross sections are determined in terms of G_F, α and only one further parameter, $\sin^2\theta_W$. As mentioned in section 20.9, experimental fits to these predictions are reviewed by Commins and Bucksbaum (1983), Renton (1990) and Winter (2000).

Particularly precise determinations of the Standard Model parameters were made at the e^+e^- colliders, LEP and SLC. Consider the reaction $e^+e^- \to f\bar{f}$ where f is μ or τ, at energies where the lepton masses may be neglected in the final answers. In lowest order, the process is mediated by both γ-exchange and Z^0-exchange as shown in figure 22.14. Calculations of the cross section were made early on, by Budny (1973) for example. In modern notation, the differential cross section for the scattering of unpolarized e^- and e^+ is given by

$$\frac{d\sigma}{d\cos\theta} = \frac{\pi\alpha^2}{2s}[(1+\cos^2\theta)A + \cos\theta B] \tag{22.89}$$

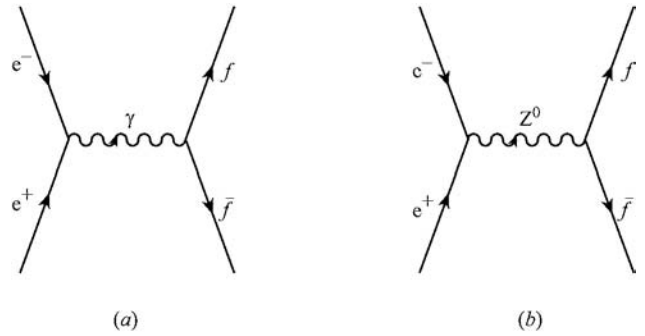


FIGURE 22.14
(a) One-γ and (b) one-W annihilation graphs in $e^+e^- \to f\bar{f}$.

where θ is the CM scattering angle of the final state lepton, $s = (p_{e^-} + p_{e^+})^2$, and

$$A = 1 + 2g_V^e g_V^f \mathrm{Re}\chi(s) + [(g_A^e)^2 + (g_V^e)^2][(g_A^f)^2 + (g_V^f)^2]|\chi(s)|^2 \quad (22.90)$$

$$B = 4g_A^e g_A^f \mathrm{Re}\chi(s) + 8g_A^e g_V^e g_A^f g_V^f |\chi(s)|^2 \quad (22.91)$$

$$\chi(s) = s/[4\sin^2\theta_W \cos^2\theta_W (s - M_Z^2 + i\Gamma_Z M_Z)]. \quad (22.92)$$

Notice that the term surviving when all the g's are set to zero, which is therefore the pure single photon contribution, is exactly as calculated in problem 8.18. The presence of the $\cos\theta$ term leads to the forward–backward asymmetry noted in that problem.

The forward–backward asymmetry A_{FB} may be defined as

$$A_{FB} \equiv (N_F - N_B)/(N_F + N_B), \quad (22.93)$$

where N_F is the number scattered into the forward hemisphere $0 \le \cos\theta \le 1$, and N_B that into the backward hemisphere $-1 \le \cos\theta \le 0$. Integrating (22.89) one easily finds

$$A_{FB} = 3B/8A. \quad (22.94)$$

For $\sin^2\theta_W = 0.25$ we noted after (22.64) that the g_V^l's vanish, so they are very small for $\sin^2\theta_W \simeq 0.23$. The effect is therefore controlled essentially by the first term in (22.91). At $\sqrt{s} = 29$ GeV, for example, the asymmetry is $A_{FB} \simeq -0.063$.

This asymmetry was observed in experiments with PETRA at DESY and with PEP at SLAC (see figure 8.20(b)). These measurements, made at energies well below the Z^0 peak, were the first indication of the presence of Z^0 exchange in e^+e^- collisions.

However, QED alone produces a small positive $A_{\rm FB}$, through interference between 1γ and 2γ annihilation processes (which have different charge conjugation parity), as well as between initial and final state bremsstrahlung corrections to figure 22.14(a). Indeed, *all* one-loop radiative effects must clearly be considered, in any comparison with modern high precision data.

At the CERN e^+e^- collider LEP, many such measurements were made 'on the Z peak', i.e. at $s = M_Z^2$ in the parametrization (22.92). In that case, $\mathrm{Re}\chi(s) = 0$, and (22.94) becomes (neglecting the photon contribution)

$$A_{\rm FB}(\mathrm{Z}^0 \text{ peak}) = \frac{3g_{\rm A}^{\rm e}g_{\rm V}^{\rm e}g_{\rm A}^{f}g_{\rm V}^{f}}{\{[(g_{\rm A}^{\rm e})^2 + g_{\rm V}^{e})^2][(g_{\rm A}^{f})^2 + (g_{\rm V}^{f})^2]\}}. \tag{22.95}$$

Another important asymmetry observable is that involving the difference of the cross sections for left- and right-handed incident electrons:

$$A_{\rm LR} \equiv (\sigma_{\rm L} - \sigma_{\rm R})/(\sigma_{\rm L} + \sigma_{\rm R}), \tag{22.96}$$

for which the tree-level prediction is

$$A_{\rm LR} = 2g_{\rm V}^{\rm e}g_{\rm A}^{\rm e}/[(g_{\rm V}^{\rm e})^2 + (g_{\rm A}^{\rm e})^2]. \tag{22.97}$$

A similar combination of the g's for the final state leptons can be measured by forming the 'L–R F–B' asymmetry

$$A_{\rm LR}^{\rm FB} = [(\sigma_{\rm LF} - \sigma_{\rm LB}) - (\sigma_{\rm RF} - \sigma_{\rm RB})]/(\sigma_{\rm R} + \sigma_{\rm L}) \tag{22.98}$$

for which the tree level prediction is

$$A_{\rm LR}^{\rm FB} = 2g_{\rm V}^{f}g_{\rm A}^{f}/[(g_{\rm V}^{f})^2 + (g_{\rm A}^{f})^2]. \tag{22.99}$$

The quantity on the right-hand side of (22.99) is usually denoted by A_f:

$$A_f = 2g_{\rm V}^{f}g_{\rm A}^{f}/[(g_{\rm V}^{f})^2 + (g_{\rm A}^{f})^2]. \tag{22.100}$$

The asymmetry $A_{\rm FB}$ is not, in fact, direct evidence for parity violation in $e^+e^- \to \mu^+\mu^-$, since we see from (22.90) and (22.91) that it is even under $g_{\rm A}^{l} \to -g_{\rm A}^{l}$, whereas a true parity-violating effect would involve terms odd (linear) in $g_{\rm A}^{l}$. However, electroweak-induced parity violation effects in an apparently electromagnetic process were observed in a remarkable experiment by Prescott *et al.* (1978). Longitudinally polarized electrons were inelastically scattered from deuterium, and the flux of scattered electrons was measured for incident electrons of definite helicity. An asymmetry between the results, depending on the helicities, was observed – a clear signal for parity violation. This was the first demonstration of parity-violating effects in an 'electromagnetic' process; the corresponding value of $\sin^2\theta_{\rm W}$ was in agreement with that determined from ν data.

We now turn to some of the main experimental evidence, beginning with the discoveries of the $\mathrm{W}^{\pm}$ and Z^0 1983.

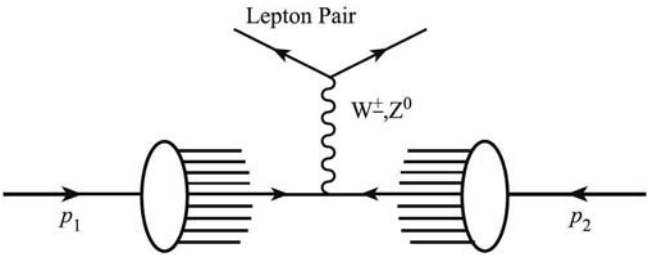


FIGURE 22.15
Parton model amplitude for $W^{\pm}$ or Z^0 production in $\bar{p}p$ collisions.

22.4 The discovery of the $W^{\pm}$ and Z^0 at the CERN $p\bar{p}$ collider

22.4.1 Production cross sections for W and Z in $p\bar{p}$ colliders

The possibility of producing the predicted $W^{\pm}$ and Z^0 particles was the principal motivation for transforming the CERN SPS into a $p\bar{p}$ collider using the stochastic cooling technique (Rubbia *et al.* 1977, Staff of the CERN $\bar{p}p$ project 1981). Estimates of W and Z^0 production in $\bar{p}p$ collisions may be obtained (see, for example, Quigg 1977) from the parton model, in a way analogous to that used for the Drell–Yan process in section 9.4 with γ replaced by W or Z^0, as shown in figure 22.15 (cf figure 9.11), and for two-jet cross sections in section 14.3.2. As in (14.51), we denote by $\hat{s}$ the subprocess invariant

$$\hat{s} = (x_1 p_1 + x_2 p_2)^2 = x_1 x_2 s \tag{22.101}$$

for massless partons. With $\hat{s}^{1/2} = M_{\mathrm{W}} \sim 80$ GeV, and $s^{1/2}$ =630 GeV for the $\bar{p}p$ collider energy, we see that the x's are typically ∼0.13, so that the valence q's in the proton and $\bar{q}$'s in the antiproton will dominate (at $\sqrt{s} = 1.8$ TeV, appropriate to the Fermilab Tevatron, $x \simeq 0.04$ and the sea quarks contribute). The parton model cross section $p\bar{p} \to W^{\pm}+$ anything is then (setting $V_{\mathrm{ud}} = 1$ and all other $V_{ij} = 0$)

$$\sigma(p\bar{p} \to W^{\pm} + X) = \frac{1}{3}\int_0^1 dx_1 \int_0^1 dx_2 \hat{\sigma}(x_1, x_2) \left\{ \begin{array}{l} u(x_1)\bar{d}(x_2) + \bar{d}(x_1)u(x_2) \\ \bar{u}(x_1)d(x_2) + d(x_1)\bar{u}(x_2) \end{array} \right\} \tag{22.102}$$

where the $\frac{1}{3}$ is the same colour factor as in the Drell–Yan process, and the subprocess cross section $\hat{\sigma}$ for $q\bar{q} \to W^{\pm} + X$ is (neglecting the $W^{\pm}$ width)

$$\hat{\sigma} = 4\pi^2 \alpha (1/4 \sin^2 \theta_{\mathrm{W}}) \delta(\hat{s} - M_{\mathrm{W}}^2) \tag{22.103}$$

$$= \pi 2^{1/2} G_{\mathrm{F}} M_{\mathrm{W}}^2 \delta(x_1 x_2 s - M_{\mathrm{W}}^2). \tag{22.104}$$

QCD corrections to (22.102) must as usual be included. Leading logarithms will make the distributions Q^2-dependent, and they should be evaluated at $Q^2 = M_{\rm W}^2$. There will be further ($O(\alpha_{\rm s}^2)$) corrections, which are often accounted for by a multiplicative factor 'K', which is of order 1.5–2 at these energies. $O(\alpha_{\rm s}^2)$ calculations are presented in Hamberg *et al.* (1991) and by van der Neerven and Zijlstra (1992); see also Ellis *et al.* (1996) section 9.4. The total cross section for production of $\rm W^+$ and $\rm W^-$ at $\sqrt{s}$ =630 GeV is then of order 6.5 nb, while a similar calculation for the $\rm Z^0$ gives about 2 nb. Multiplying these by the branching ratios gives

$$\sigma(\mathrm{p\bar{p}} \to \mathrm{W} + \mathrm{X} \to \mathrm{e}\nu\mathrm{X}) \simeq 0.7 \text{ nb} \tag{22.105}$$

$$\sigma(\mathrm{p\bar{p}} \to \mathrm{Z}^0 + \mathrm{X} \to \mathrm{e}^+\mathrm{e}^-\mathrm{X}) \simeq 0.07 \text{ nb} \tag{22.106}$$

at $\sqrt{s}$ =630 GeV.

The total cross section for $\rm p\bar{p}$ is about 70 mb at these energies: hence (22.105) represents $\sim 10^{-8}$ of the total cross section, and (22.106) is 10 times smaller. The rates could, of course, be increased by using the $\rm q\bar{q}$ modes of W and $\rm Z^0$, which have bigger branching ratios. But the detection of these is very difficult, being very hard to distinguish from conventional two-jet events produced via the mechanism discussed in section 14.3.2, which has a cross section some 10^3 higher than (22.105). W and $\rm Z^0$ would appear as slight shoulders on the edge of a very steeply falling invariant mass distribution, similar to that shown in figure 9.12, and the calorimetric jet energy resolution capable of resolving such an effect is hard to achieve. Thus despite the unfavourable branching ratios, the leptonic modes provide the better signatures, as discussed further in section 22.4.3.

22.4.2 Charge asymmetry in $\rm W^{\pm}$ decay

At energies such that the simple valence quark picture of (22.102) is valid, the $\rm W^+$ is created in the annihilation of a left-handed u quark from the proton and a right-handed $\rm \bar{d}$ quark from the $\rm \bar{p}$ (neglecting fermion masses). In the $\rm W^+ \to e^+\nu_e$ decay, a right-handed $\rm e^+$ and left-handed $\nu_{\rm e}$ are emitted. Referring to figure 22.16, we see that angular momentum conservation allows $\rm e^+$ production parallel to the direction of the antiproton, but forbids it parallel to the direction of the proton. Similarly, in $\rm W^- \to e^-\bar{\nu}_e$, the $\rm e^-$ is emitted preferentially parallel to the proton (these considerations are exactly similar to those mentioned in section 20.7.2 with reference to νq and $\bar{\nu}$q scattering). The actual distribution has the form $\sim (1 + \cos\theta_{\rm e}^*)^2$, where $\theta_{\rm e}^*$ is the angle, in the rest frame of the W, between the $\rm e^-$ and the p (for $\rm W^- \to e^-\bar{\nu}_e$) or the $\rm e^+$ and the $\rm \bar{p}$ (for $\rm W^+ \to e^+\nu_e$).

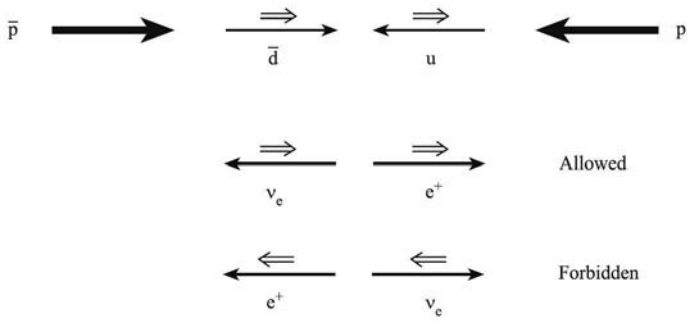


FIGURE 22.16
Preferred direction of leptons in W^+ decay.

22.4.3 Discovery of the $W^{\pm}$ and Z^0 at the $p\bar{p}$ collider, and their properties

As already indicated in section 22.4.1, the best signatures for W and Z production in $p\bar{p}$ collisions are provided by the leptonic modes

$$p\bar{p} \to W^{\pm}X \to e^{\pm}\nu X \tag{22.107}$$

$$p\bar{p} \to Z^0X \to e^+e^-X. \tag{22.108}$$

Reaction (22.107) has the larger cross section, by a factor of 10 (cf (22.105) and (22.106)), and was observed first (UA1, Arnison *et al.* 1983a; UA2, Banner *et al.* 1983). However, the kinematics of (22.108) is simpler and so the Z^0 discovery (UA1, Arnison *et al.* (1983b); UA2, Bagnaia *et al.* 1983) will be discussed first.

The signature for (22.108) is an isolated, and approximately back-to-back, e^+e^- pair with invariant mass peaked around 90 GeV (cf (22.78)). Very clean events can be isolated by imposing a modest transverse energy cut – the e^+e^- pairs required are coming from the decay of a massive relatively slowly moving Z^0. Figure 22.17 shows the transverse energy distribution of a candidate Z^0 event from the first UA2 sample. Figure 22.18 shows (Geer 1986) the invariant mass distribution for a later sample of 14 UA1 events in which both electrons have well measured energies, together with the Breit–Wigner resonance curve appropriate to $M_Z = 93$ GeV/c^2, with experimental mass resolution folded in. The UA1 result for the Z^0 mass was

$$M_Z = 93.0 \pm 1.4(\text{stat}) \pm 3.2(\text{syst.}) \ \text{GeV}. \tag{22.109}$$

The corresponding UA2 result (DiLella 1986), based on 13 well measured pairs, was

$$M_Z = 92.5 \pm 1.3(\text{stat.}) \pm 1.5(\text{syst.}) \ \text{GeV}. \tag{22.110}$$

In both cases the systematic error reflects the uncertainty in the absolute calibration of the calorimeter energy scale. Clearly the agreement with (22.78)

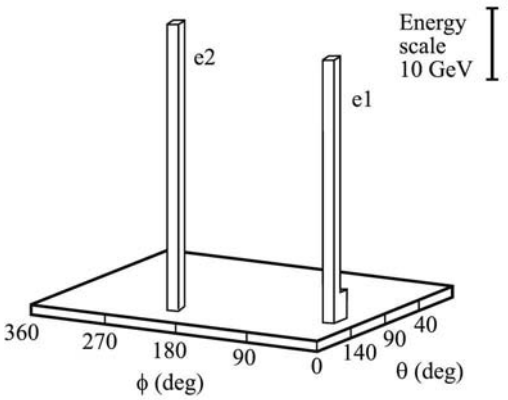


FIGURE 22.17
The cell transverse energy distribution for a $Z^0 \to e^+e^-$ event (UA2, Bagnaia *et al.* 1983) in the θ and ϕ plane, where θ and ϕ are the polar and azimuth angles relative to the beam axis.

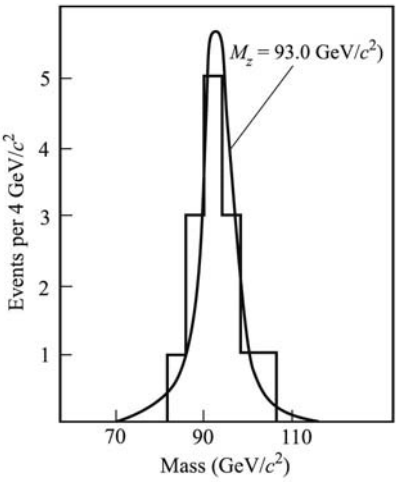


FIGURE 22.18
Invariant mass distribution for 14 well measured $Z^0 \to e^+e^-$ decays (UA1). Figure reprinted with permission from S Geer in *High Energy Physics 1985, Proc. Yale Theoretical Advanced Study Institute*, eds M J Bowick and F Gursey; copyright 1986 World Scientific Publishing Company.

is good, but there is a suggestion that the tree-level prediction is on the low side. Indeed, loop corrections adjust (22.78) to a value $M_{\mathrm{Z}}^{\mathrm{th}} \simeq 91.19$ GeV, in excellent agreement with the current experimental value (Nakamura *et al.* 2010).

The total Z^0 width Γ_{Z} is an interesting quantity. If we assume that, for any fermion family additional to the three known ones, only the neutrinos are significantly less massive than $M_{\mathrm{Z}}/2$, we have

$$\Gamma_{\mathrm{Z}} \simeq (2.5 + 0.16\Delta N_\nu)\ \mathrm{GeV} \tag{22.111}$$

from section 22.3, where ΔN_ν is the number of additional light neutrinos (i.e. beyond $\nu_{\mathrm{e}}, \nu_\mu$ and ν_τ) which contribute to the width through the process $\mathrm{Z}^0 \to \nu\bar{\nu}$. Thus (22.111) can be used as an important measure of the number of such neutrinos (i.e. generations) if Γ_{Z} can be determined accurately enough. The mass resolution of the $\mathrm{\bar{p}p}$ experiments was of the same order as the total expected Z^0 width, so that (22.111) could not be used directly. The advent of LEP provided precision checks on (22.111); at the cost of departing from the historical development, we show data from DELPHI (Abreu *et al.* 1990, Abe 1991) in figure 22.19, which established $N_\nu = 3$.

We turn now to the $\mathrm{W}^{\pm}$. In this case an invariant mass plot is impossible, since we are looking for the $\mathrm{e}\nu$ ($\mu\nu$) mode, and cannot measure the ν's. However, it is clear that – as in the case of $\mathrm{Z}^0 \to \mathrm{e}^+\mathrm{e}^-$ decay – slow moving massive W's will emit isolated electrons with high transverse energy. Further, such electrons should be produced in association with large *missing* transverse energy (corresponding to the ν's), which can be measured by calorimetry, and which should balance the transverse energy of the electrons. Thus electrons of high E_{T} accompanied by balancing high missing E_{T} (i.e. similar in magnitude to that of the e^- but opposite in azimuth) were the signatures used for the early event samples (UA1, Arnison *et al.* 1983a; UA2, Banner *et al.* 1983).

The determination of the mass of the W is not quite so straightforward as that of the Z, since we cannot construct directly an invariant mass plot for the $\mathrm{e}\nu$ pair: only the missing transverse momentum (or energy) can be attributed to the ν, since some unidentified longitudinal momentum will always be lost down the beam pipe. In fact, the distribution of events in p_{eT}, the magnitude of the transverse momentum of the e^-, should show a pronounced peaking towards the maximum kinematically allowed value, which is $p_{\mathrm{eT}} \approx \frac{1}{2}M_{\mathrm{W}}$, as may be seen from the following argument. Consider the decay of a W at rest (figure 22.20). We have $|\boldsymbol{p}_{\mathrm{e}}| = \frac{1}{2}M_{\mathrm{W}}$ and $|\boldsymbol{p}_{\mathrm{eT}}| = \frac{1}{2}M_{\mathrm{W}}\sin\theta \equiv p_{\mathrm{eT}}$. Thus the transverse momentum distribution is given by

$$\frac{\mathrm{d}\sigma}{\mathrm{d}p_{\mathrm{eT}}} = \frac{\mathrm{d}\sigma}{\mathrm{d}\cos\theta}\left|\frac{\mathrm{d}\cos\theta}{\mathrm{d}p_{\mathrm{eT}}}\right| = \frac{\mathrm{d}\sigma}{\mathrm{d}\cos\theta}\left(\frac{2p_{\mathrm{eT}}}{M_{\mathrm{W}}}\right)\left(\frac{1}{4}M_{\mathrm{W}}^2 - p_{\mathrm{eT}}^2\right)^{-1/2}, \tag{22.112}$$

and the last (Jacobian) factor in (22.112) produces a strong peaking towards $p_{\mathrm{eT}} = \frac{1}{2}M_{\mathrm{W}}$. This peaking will be smeared by the width, and transverse motion, of the W. Early determinations of M_{W} used (22.112), but sensitivity

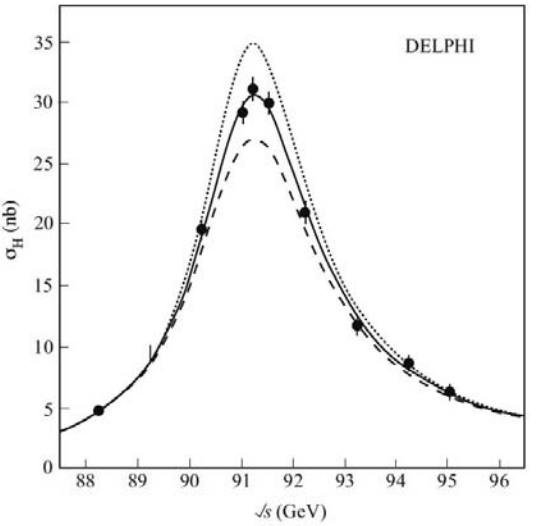


FIGURE 22.19
The cross-section for $e^+e^- \to$ hadrons around the Z^0 mass (DELPHI, 1990). The dotted, continuous and dashed lines are the predictions of the Standard Model assuming two, three and four massless neutrino species respectively. Figure reprinted with permission from K Abe in *Proc. 25th Int. Conf. on High Energy Physics* eds K K Phua and Y Yamaguchi; copyright 1991 World Scientific Publishing Company.

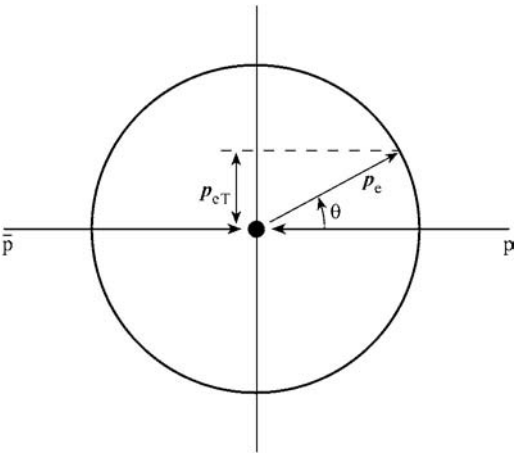


FIGURE 22.20
Kinematics of $W \to e\nu$ decay.

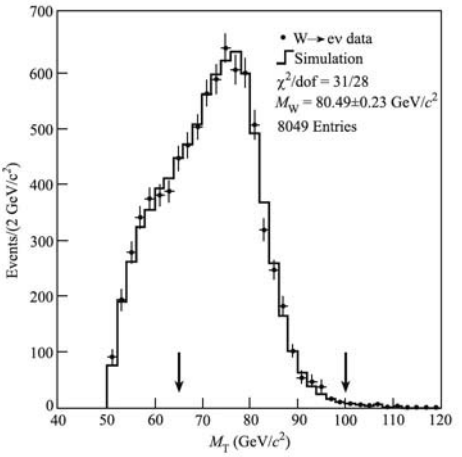


FIGURE 22.21
$W \to e\nu$ transverse mass distribution measured by the CDF collaboration. Figure reprinted with permission from F Abe *et al.* (CDF Collaboration) *Phys. Rev.* D **52** 4784 (1995). Copyright 1995 by the American Physical Society.

to the transverse momentum of the W can be much reduced (Barger *et al.* 1983) by considering instead the distribution in 'transverse mass', defined by

$$M_{\mathrm{T}}^2 = (E_{\mathrm{eT}} + E_{\nu\mathrm{T}})^2 - (\boldsymbol{p}_{\mathrm{eT}} + \boldsymbol{p}_{\nu\mathrm{T}})^2 \simeq 2p_{\mathrm{eT}}p_{\nu\mathrm{T}}(1 - \cos\phi), \qquad (22.113)$$

where ϕ is the azimuthal separation between p_{eT} and $p_{\nu\mathrm{T}}$. Here $E_{\nu\mathrm{T}}$ and $\boldsymbol{p}_{\mathrm{T}}$ are the neutrino transverse energy and momentum, measured from the missing transverse energy and momentum obtained from the global event reconstruction. This inclusion of additional measured quantities improves the precision as compared with the Jacobian peak method, using (22.112). A Monte Carlo simulation was used to generate M_{T} distributions for different values of M_{W}, and the most probable value was found by a maximum likelihood fit. The quoted results were

$$\text{UA1 (Geer 1986):}\quad M_{\mathrm{W}} = 83.5 \pm^{1.1}_{1.0}\,(\text{stat.}) \pm 2.8(\text{syst.})\ \text{GeV} \qquad (22.114)$$

$$\text{UA2 (DiLella 1986):}\quad M_{\mathrm{W}} = 81.2 \pm 1.1(\text{stat.}) \pm 1.3(\text{syst.})\ \text{GeV} \qquad (22.115)$$

the systematic errors again reflecting uncertainty in the absolute energy scale of the calorimeters. The two experiments also quoted (Geer 1986, DiLella 1986)

$$\left.\begin{array}{ll} \text{UA1} & \Gamma_{\mathrm{W}} < 6.5\ \text{GeV} \\ \text{UA2} & \Gamma_{\mathrm{W}} < 7.0\ \text{GeV} \end{array}\right\} 90\%\ \text{c.l.} \qquad (22.116)$$

Once again, the agreement between the experiments, and of both with (22.77), is good, the predictions again being on the low side. Loop corrections adjust

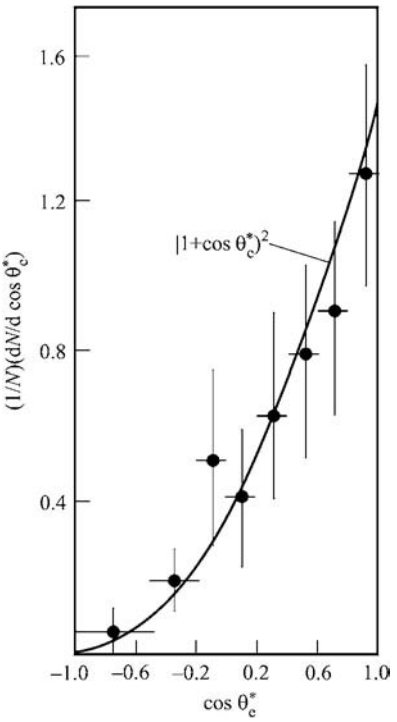


FIGURE 22.22
The W decay angular distribution of the emission angle θ_e^* of the positron (electron) with respect to the antiproton (proton) beam direction, in the rest frame of the W, for a total of 75 events; background subtracted and acceptance corrected (Arnison *et al.* 1986).

(22.77) to $M_W \simeq 80.38$ GeV (Nakamura *et al.* 2010). We show in figure 22.21 a later determination of M_W by the CDF collaboration (Abe *et al.* 1995a).

The W and Z mass values may be used together with (22.42) to obtain $\sin^2\theta_W$ via

$$\sin^2\theta_W = 1 - M_W^2/M_Z^2. \tag{22.117}$$

The weighted average of UA(1) and UA(2) yielded

$$\sin^2\theta_W = 0.212 \pm 0.022 \text{ (stat.)}. \tag{22.118}$$

Radiative corrections have in general to be applied, but one renormalization scheme (see section 22.6) promotes (22.117) to a definition of the renormalized $\sin^2\theta_W$ to all orders in perturbation theory. Using this scheme and quoted values of M_W, M_Z (Nakamura *et al.* 2010) one finds $\sin^2\theta_W \simeq 0.223$.

Finally, figure 22.22 shows (Arnison *et al.* 1986) the angular distribution of the charged lepton in $W \to e\nu$ decay (see section 22.4.2); θ_e^* is the $e^+(e^-)$ angle

in the W rest frame, measured with respect to a direction parallel (antiparallel) to the $\bar{\mathrm{p}}$(p) beam. The expected form $(1+\cos\theta_{\mathrm{e}}^{*})^2$ is followed very closely.

In summary, we may say that the early discovery experiments provided remarkably convincing confirmation of the principal expectations of the GSW theory, as outlined in section 22.3.

We now consider some further aspects of the theory.

22.5 Fermion masses

22.5.1 One generation

The fact that the $\mathrm{SU}(2)_{\mathrm{L}}$ gauge group acts only on the L components of the fermion fields immediately appears to create a fundamental problem as far as the *masses* of these particles are concerned; we mentioned this briefly at the end of section 19.6. Let us recall first that the standard way to introduce the interactions of gauge fields with matter fields (e.g. fermions) is via the covariant derivative replacement

$$\partial^{\mu} \to D^{\mu} \equiv \partial^{\mu} + \mathrm{i}g\boldsymbol{\tau}\cdot\boldsymbol{W}^{\mu}/2 \tag{22.119}$$

for SU(2) fields $\boldsymbol{W}^{\mu}$ acting on $t = 1/2$ doublets. Now it is a simple exercise (compare problem 18.3) to check that the ordinary 'kinetic' part of a free Dirac fermion does not mix the L and R components of the field:

$$\bar{\hat{\psi}}\ \not{\partial}\hat{\psi} = \bar{\hat{\psi}}_{\mathrm{R}}\ \not{\partial}\psi_{\mathrm{R}} + \bar{\hat{\psi}}_{\mathrm{L}}\ \not{\partial}\psi_{\mathrm{L}}. \tag{22.120}$$

Thus we can in principle contemplate 'gauging' the L and the R components differently. Of course, in the case of QCD (cf (18.39)) the replacement $\not{\partial} \to \not{D}$ was made equally in each term on the right-hand side of (22.120). But this was because QCD conserves parity, and must therefore treat L and R components the same. Weak interactions are parity violating, and the $\mathrm{SU}(2)_{\mathrm{L}}$ covariant derivative acts only in the *second* term of (22.120). On the other hand, a Dirac mass term has the form

$$-m(\bar{\hat{\psi}}_{\mathrm{L}}\hat{\psi}_{\mathrm{R}} + \bar{\hat{\psi}}_{\mathrm{R}}\hat{\psi}_{\mathrm{L}}) \tag{22.121}$$

(see equation (18.41) for example), and it precisely *couples* the L and R components. It is easy to see that if only $\hat{\psi}_{\mathrm{L}}$ is subject to a transformation, then (22.121) is not invariant. Thus mass terms for Dirac fermions will *explicitly* break $\mathrm{SU}(2)_{\mathrm{L}}$. The same is also true for Majorana fermions which might describe the neutrinos.

This kind of explicit breaking of the gauge symmetry cannot be tolerated, in the sense that it will lead, once again, to violations of unitarity, and then of

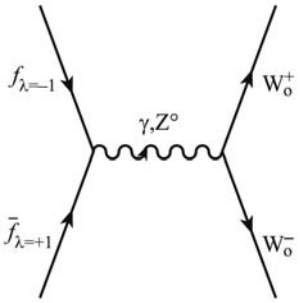


FIGURE 22.23
One-Z^0 and one-γ annihilation contribution to $f_{\lambda=-1}\bar{f}_{\lambda=1} \to W_0^+ W_0^-$.

renormalizability. Consider, for example, a fermion–antifermion annihilation process of the form

$$f\bar{f} \to W_0^+ W_0^-, \tag{22.122}$$

where the subscript indicates the $\lambda = 0$ (longitudinal) polarization state of the $W^{\pm}$. We studied such a reaction in section 22.1.1 in the context of unitarity violations (in lowest-order perturbation theory) for the IVB model. Appelquist and Chanowitz (1987) considered first the case in which 'f' is a lepton with $t = \frac{1}{2}, t_3 = -\frac{1}{2}$ coupling to W's, Z^0 and γ with the usual $SU(2)_L \times U(1)$ couplings, but having an explicit (Dirac) mass m_f. They found that in the 'right' helicity channels for the leptons ($\lambda = +1$ for $\bar{f}, \lambda = -1$ for f) the bad high energy behaviour associated with a fermion-exchange diagram of the form of figure 22.4 was *cancelled* by that of the diagrams shown in figure 22.23. The sum of the amplitudes tends to a constant as s (or E^2) $\to \infty$. Such cancellations are a feature of gauge theories, as we indicated at the end of section 22.1.2, and represent one aspect of the renormalizability of the theory. But suppose, following Appelquist and Chanowitz (1987), we examine channels involving the 'wrong' helicity component, for example $\lambda = +1$ for the fermion f. Then it is found that the cancellation no longer occurs, and we shall ultimately have a 'non-renormalizable' problem on our hands, all over again.

An estimate of the energy at which this will happen can be made by recalling that the 'wrong' helicity state participates only by virtue of a factor (m_f/energy) (recall section 20.2.2), which here we can take to be $m_f/\sqrt{s}$. The typical bad high energy behaviour for an amplitude $\mathcal{M}$ was $\mathcal{M} \sim G_F s$, which we expect to be modified here to

$$\mathcal{M} \sim G_F s m_f/\sqrt{s} \sim G_F m_f \sqrt{s}. \tag{22.123}$$

The estimate obtained by Appelquist and Chanowitz differs only by a factor of $\sqrt{2}$. Attending to all the factors in the partial wave expansion gives the result

that the unitarity bound will be saturated at $E = E_f$ (TeV) $\sim \pi/m_f$ (TeV). Thus for $m_t \sim 175$ GeV, $E_t \sim 18$ TeV. This would constitute a serious flaw in the theory, even though the breakdown occurs at energies beyond those currently reachable.

However, in a theory with spontaneous symmetry breaking, there is a way of giving fermion masses without introducing an explicit mass term in the Lagrangian. Consider the electron, for example, and let us hypothesize a 'Yukawa'–type coupling between the electron-type SU(2) doublet

$$\hat{l}_{\rm eL} = \begin{pmatrix} \hat{\nu}_{\rm e} \\ \hat{\rm e}^- \end{pmatrix}_{\rm L}, \tag{22.124}$$

the Higgs doublet $\hat{\phi}$, and the R-component of the electron field:

$$\hat{\mathcal{L}}^{\rm e}_{\rm Yuk} = -g_{\rm e}(\bar{\hat{l}}_{\rm eL}\hat{\phi}\hat{e}_{\rm R} + \bar{\hat{e}}_{\rm R}\hat{\phi}^\dagger \hat{l}_{\rm eL}). \tag{22.125}$$

In each term of (22.125), the two $\mathrm{SU(2)_L}$ doublets are 'dotted together' so as to form an $\mathrm{SU(2)_L}$ scalar, which multiplies the $\mathrm{SU(2)_L}$ scalar R-component. Thus (22.125) is $\mathrm{SU(2)_L}$-invariant, and the symmetry is preserved, at the Lagrangian level, by such a term. But now insert just the vacuum value (22.28) of $\hat{\phi}$ into (22.125): we find the result

$$\hat{\mathcal{L}}^{\rm e}_{\rm Yuk}({\rm vac}) = -g_{\rm e}\frac{v}{\sqrt{2}}(\bar{\hat{e}}_{\rm L}\hat{e}_{\rm R} + \bar{\hat{e}}_{\rm R}\hat{e}_{\rm L}) \tag{22.126}$$

which is exactly a (Dirac) mass of the form (22.121), allowing us to make the identification

$$m_{\rm e} = g_{\rm e}v/\sqrt{2}. \tag{22.127}$$

When oscillations about the vacuum value are considered via the replacement (22.29), the term (22.125) will generate a coupling between the electron and the Higgs fields of the form

$$-g_{\rm e}\bar{\hat{e}}\hat{e}\hat{H}/\sqrt{2} = -(m_{\rm e}/v)\bar{\hat{e}}\hat{e}\hat{H} \tag{22.128}$$

$$= -(gm_{\rm e}/2M_{\rm W})\bar{\hat{e}}\hat{e}\hat{H}. \tag{22.129}$$

The presence of such a coupling, if present for the process $f\bar{f} \to \mathrm{W}_0^+\mathrm{W}_0^-$ considered earlier, will mean that, in addition to the f-exchange graph analogous to figure 22.4 and the annihilation graphs of figure 22.23, a further graph shown in figure 22.24, must be included. The presence of the fermion mass in the coupling to H suggests that this graph might be just what is required to cancel the 'bad' high energy behaviour found in (22.123) – and by this time the reader will not be surprised to be told that this is indeed the case.

At first sight it might seem that this stratagem will only work for the $t_3 = -\frac{1}{2}$ components of doublets, because of the form of $\langle 0|\hat{\phi}|0\rangle$. But we learned in section 12.1.3 that if a pair of states $\begin{pmatrix} u \\ d \end{pmatrix}$ forming an SU(2)

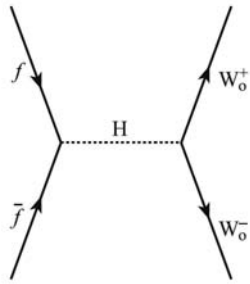


FIGURE 22.24
One-H annihilation graph.

doublet transform by

$$\begin{pmatrix} u \\ d \end{pmatrix}' = \mathrm{e}^{-\mathrm{i}\boldsymbol{\alpha}\cdot\boldsymbol{\tau}/2} \begin{pmatrix} u \\ d \end{pmatrix}, \tag{22.130}$$

then the charge conjugate states $\mathrm{i}\tau_2 \begin{pmatrix} u^* \\ d^* \end{pmatrix}$ transform in exactly the same way. Thus if, in our case, $\hat{\phi}$ is the SU(2) doublet

$$\hat{\phi} = \begin{pmatrix} \frac{1}{\sqrt{2}}(\hat{\phi}_1 - \mathrm{i}\hat{\phi}_2) \equiv \hat{\phi}^+ \\ \frac{1}{\sqrt{2}}(\hat{\phi}_3 - \mathrm{i}\hat{\phi}_4) \equiv \hat{\phi}^0 \end{pmatrix}, \tag{22.131}$$

then the charge conjugate field

$$\hat{\phi}_{\mathbf{C}} \equiv \mathrm{i}\tau_2\hat{\phi}^* = \begin{pmatrix} \frac{1}{\sqrt{2}}(\hat{\phi}_3 + \mathrm{i}\hat{\phi}_4) \\ -\frac{1}{\sqrt{2}}(\hat{\phi}_1 + \mathrm{i}\hat{\phi}_2) \end{pmatrix} \equiv \begin{pmatrix} \bar{\hat{\phi}}^0 \\ -\hat{\phi}^- \end{pmatrix} \tag{22.132}$$

is also an SU(2) doublet, transforming in just the same way as $\hat{\phi}$. ((22.131) and (22.132) may be thought of as analogous to the $(\mathrm{K}^+, \mathrm{K}^0)$ and $(\bar{\mathrm{K}}^0, \mathrm{K}^-)$ isospin doublets in $\mathrm{SU(3)_f}$). Note that the vacuum value (22.28) will now appear in the upper component of (22.132). With the help of $\hat{\phi}_{\mathbf{C}}$ we can write down another SU(2)-invariant coupling in the $\nu_\mathrm{e} - \mathrm{e}$ sector, namely

$$-g_{\nu_\mathrm{e}}(\bar{\hat{l}}_{\mathrm{eL}}\hat{\phi}_{\mathbf{C}}\hat{\nu}_{\mathrm{eR}} + \bar{\hat{\nu}}_{\mathrm{eR}}\hat{\phi}^\dagger_{\mathbf{C}}\hat{l}_{\mathrm{eL}}), \tag{22.133}$$

assuming now the existence of the field $\hat{\nu}_{\mathrm{eR}}$. In the Higgs vacuum (22.28), (22.133) then yields

$$-(g_{\nu_\mathrm{e}}v/\sqrt{2})(\bar{\hat{\nu}}_{\mathrm{eL}}\hat{\nu}_{\mathrm{eR}} + \bar{\hat{\nu}}_{\mathrm{eR}}\hat{\nu}_{\mathrm{eL}}) \tag{22.134}$$

which is precisely a (Dirac) mass for the neutrino, if we set $g_{\nu_\mathrm{e}}v/\sqrt{2} = m_{\nu_\mathrm{e}}$.

It is clearly possible to go on like this, and arrange for all the fermions, quarks as well as leptons, to acquire a mass by the same 'mechanism'. We will look more closely at the quarks in the next section. But one must admit to a certain uneasiness concerning the enormous difference in magnitudes represented by the couplings $g_{\nu_e}, \ldots g_e, \ldots g_t$. If $m_{\nu_e} < 1$ eV then $g_{\nu_e} < 10^{-11}$, while $g_t \sim 1$! Besides, whereas the use of the Higgs field 'mechanism' in the W–Z sector is quite economical, in the present case it seems rather unsatisfactory simply to postulate a different 'g' for each fermion–Higgs interaction. This does appear to indicate that we are dealing here with a 'phenomenological model', once more, rather than a 'theory'.

As far as the neutrinos are concerned, however, there is another possibility, already discussed in sections 7.5.2, 20.3 and 21.4.1, which is that they could be Majorana (not Dirac) fermions. In this case, rather than the four degrees of freedom (ν_{eL}, ν_{eR}, and their antiparticles) which exist for massive Dirac particles, only two possibilities exist for neutrinos, which we may take to be ν_{eL} and ν_{eR}. With these, it is certainly possible to construct a Dirac-type mass term of the form (22.134). But since, after all, the ν_{eR} component has been assigned zero quantum members both for $SU(2)_L$ W-interactions and for U(1) B-interactions (see table 22.1), we could consider economically dropping it altogether, making do with just the ν_{eL} component.

Suppose, then, that we keep only the field $\hat{\nu}_{eL}$. We need to form a mass term for it. The charge-conjugate field is defined by (see (7.151))

$$(\hat{\nu}_{eL})_{\mathbf{C}} = i\gamma^2\gamma_0\bar{\hat{\nu}}_{eL}^{T} = i\gamma^2\hat{\nu}_L^{\dagger T}, \tag{22.135}$$

and we know that the charge-conjugate field transforms under Lorentz transformations in the same way as the original field. So we can use $(\hat{\nu}_{eL})_{\mathbf{C}}$ to form a Lorentz invariant

$$\overline{(\hat{\nu}_{eL})_{\mathbf{C}}}\ \nu_{eL} \tag{22.136}$$

which has mass dimension M^3. Hence we may write a mass term for $\hat{\nu}_{eL}$ in the form

$$-\frac{1}{2}m_M[\overline{(\hat{\nu}_{eL})_{\mathbf{C}}}\,\hat{\nu}_{eL} + \bar{\hat{\nu}}_{eL}(\hat{\nu}_{eL})_{\mathbf{C}}] \tag{22.137}$$

where the $\frac{1}{2}$ is conventional. Written out in more detail, we have

$$\overline{(\hat{\nu}_{eL})_{\mathbf{C}}}\,\hat{\nu}_{eL} = \hat{\nu}_{eL}^{T}(-i\gamma^{2\dagger}\gamma_0)\hat{\nu}_{eL} = \hat{\nu}_{eL}^{T}i\gamma^2\gamma_0\hat{\nu}_{eL}, \tag{22.138}$$

in the representation (20.14). Now

$$i\gamma^2\gamma_0 = \begin{pmatrix} -i\sigma_2 & 0 \\ 0 & i\sigma_2 \end{pmatrix}. \tag{22.139}$$

But since $\hat{\nu}_{eL}$ is an L-chiral field, only its 2 lower components are present (cf (20.26)) and (22.138) is effectively

$$\overline{(\hat{\nu}_{eL})_{\mathbf{C}}}\,\hat{\nu}_{eL} = \hat{\nu}_{eL}^{T}(i\sigma_2)\hat{\nu}_{eL}. \tag{22.140}$$

This is just the form of the mass term for a Majorana field, as we saw in equation (7.159). The two formalisms are equivalent.

As noted in section 21.4.1, the mass term (22.137) is not invariant under a global U(1) phase transformation

$$\hat{\nu}_{\mathrm{eL}} \to \mathrm{e}^{-\mathrm{i}\alpha}\hat{\nu}_{\mathrm{eL}} \tag{22.141}$$

which would correspond to lepton number (if accompanied by a similar transformation for the electron fields): the Majorana mass term violates lepton number conservation.

There is a further interesting aspect to (22.140) which is that, since two $\hat{\nu}_{\mathrm{eL}}$ operators appear rather than a $\hat{\nu}_{\mathrm{e}}$ and a $\hat{\nu}_{\mathrm{e}}^{\dagger}$ (which would lead to L_{e} conservation), the (t, t_3) quantum numbers of the term are (1,1). This means that we cannot form an $\mathrm{SU}(2)_{\mathrm{L}}$ invariant with it, using only the Standard Model Higgs $\hat{\phi}$, since the latter has $t = \frac{1}{2}$ and cannot combine with the (1,1) operator to form a singlet. Thus we cannot make a 'tree-level' Majorana mass by the mechanism of Yukawa coupling to the Higgs field, followed by symmetry breaking.

However, we could generate suitable 'effective' operators via loop corrections, perhaps, much as we generated an effective operator representing an anomalous magnetic moment interaction in QED (cf section 11.7). But whatever it is, the operator would have to violate lepton number conservation, which is actually conserved by all the Standard Model interactions. Thus such an effective operator could not be generated in perturbation theory. It could arise, however, as a low energy limit of a theory defined at a higher mass scale, as the current–current model is the low energy limit of the GSW one. The typical form of such operator we need, in order to generate a term $\hat{\nu}_{\mathrm{eL}}^{\mathrm{T}}\mathrm{i}\sigma_2\hat{\nu}_{\mathrm{eL}}$, is

$$-\frac{g_{\mathrm{eM}}}{M}(\bar{\hat{l}}_{\mathrm{eL}}\hat{\phi}_{\mathbf{C}})^{\mathrm{T}}\mathrm{i}\sigma_2(\hat{\phi}_{\mathbf{C}}^{\dagger}\hat{l}_{\mathrm{eL}}). \tag{22.142}$$

Note, most importantly, that the operator '$(l\phi)(\phi l)$' in (22.142) has mass dimension *five*, which is why we introduced the factor M^{-1} in the coupling; it is indeed a non-renormalizable effective interaction, just like the current–current one. We may interpret M as the mass scale at which 'new physics' enters, in the spirit of the discussion in section 11.7. Suppose, for the sake of argument, this was $M \sim 10^{16}$ GeV (a scale typical of Grand Unified Theories). After symmetry breaking, then, (22.142) will generate the required Majorana mass term, with

$$m_{\mathrm{M}} \sim g_{\mathrm{eM}}\frac{v^2}{M} \sim g_{\mathrm{eM}}10^{-2}\ \mathrm{eV}. \tag{22.143}$$

Thus an effective coupling of 'natural' size $g_{\mathrm{eM}} \sim 0.1$ emerges from this argument, if indeed the mass of the ν_{e} is of order 10^{-3}eV.

A more specific model can be constructed in which a relation of the form (22.143) can arise naturally. Suppose $\hat{\nu}_{\mathrm{R}}$ is an R-type neutrino field which is an $\mathrm{SU}(2) \times \mathrm{U}(1)$ singlet, and which has a gauge-invariant Yukawa coupling

to the Higgs field, of the form (22.133). Then the Yukawa and the mass terms $\hat{\nu}_{\rm R}$ are

$$\mathcal{L}_{Y,R} = -g_{\rm R}(\bar{\hat{\ell}}_{\rm eL}\hat{\phi}_{\mathbf{C}}\hat{\nu}_{\rm R} + \bar{\hat{\nu}}_{\rm R}\hat{\phi}^{\dagger}_{\mathbf{C}}\hat{\ell}_{\rm eL}) - \frac{1}{2}m_{\rm R}[\overline{(\hat{\nu}_{\rm R})_{\mathbf{C}}}\,\hat{\nu}_{\rm R} + \text{h.c.}]. \qquad (22.144)$$

Then, in the Higgs vacuum the first term in (22.144) becomes

$$-m_{\rm D}(\bar{\hat{\nu}}_{\rm eL}\hat{\nu}_{\rm R} + \bar{\hat{\nu}}_{\rm R}\hat{\nu}_{\rm eL}) \qquad (22.145)$$

where $m_{\rm D} = g_{\rm R}v/\sqrt{2}$. The term (22.145) couples the fields $\hat{\nu}_{\rm R}$ and $\hat{\nu}_{\rm eL}$, so that we need to do a diagonalization to find the true mass eigenvalues and eigenstates. The combined mass terms from (22.144) and (22.145) can be written as

$$-\frac{1}{2}\,\overline{(\hat{N}_{\rm L})_{\mathbf{C}}}\,\mathbf{M}\,\hat{N}_{\rm L} + \text{h.c.} \qquad (22.146)$$

where

$$\hat{N}_{\rm L} \equiv \begin{pmatrix} \hat{\nu}_{\rm eL} \\ (\hat{\nu}_{\rm R})_{\mathbf{C}} \end{pmatrix}, \quad \mathbf{M} = \begin{pmatrix} 0 & m_{\rm D} \\ m_{\rm D} & m_{\rm R} \end{pmatrix}. \qquad (22.147)$$

CP invariance would imply that the parameters $m_{\rm R}$ and $m_{\rm D}$ are real, as we will assume, for simplicity.

Suppose now that $m_{\rm D} \ll m_{\rm R}$. Then the eigenvalues of $\mathbf{M}$ are approximately

$$m_1 \approx m_{\rm R}, \quad m_2 \approx -m_{\rm D}^2/m_{\rm R}. \qquad (22.148)$$

The apparently troubling minus sign can be absorbed into the mixing parameters. Thus one eigenvalue is (by assumption) very large compared to $m_{\rm D}$, and one is very much smaller. The vanishing of the first element in $\mathbf{M}$ ensures that the lepton number violating term (22.137) is characterized by a large mass scale $m_{\rm R}$. It may be natural to assume that $m_{\rm D}$ is a 'typical' quark or lepton mass term, which would then imply that m_2 of (22.148) is very much lighter than that – as appears to be true for the neutrinos. This is the famous 'see-saw' mechanism of Minkowski (1977), Gell-Mann *et al.* (1979), Yanagida (1979) and Mohapatra and Senjanovic (1980, 1981). If in fact $m_{\rm R} \sim 10^{16}$Gev, we recover an estimate for m_2 which is similar to that in (22.143). It is worth emphasizing that the Majorana nature of the massive neutrinos is an essential part of the see-saw mechanism.

These considerations are tending to take us 'beyond the Standard Model', so we shall not pursue them at any greater length. Instead, we must now generalize the discussion of fermion masses to the three-generation case.

22.5.2 Three-generation mixing

We introduce three doublets of left-handed quark fields

$$\hat{q}_{\rm L1} = \begin{pmatrix} \hat{u}_{\rm L1} \\ \hat{d}_{\rm L1} \end{pmatrix}, \qquad \hat{q}_{\rm L2} = \begin{pmatrix} \hat{u}_{\rm L2} \\ \hat{d}_{\rm L2} \end{pmatrix}, \qquad \hat{q}_{\rm L3} = \begin{pmatrix} \hat{u}_{\rm L3} \\ \hat{d}_{\rm L3} \end{pmatrix} \qquad (22.149)$$

and the corresponding six singlets

$$\hat{u}_{\mathrm{R}1}, \hat{d}_{\mathrm{R}1}, \hat{u}_{\mathrm{R}2}, \hat{d}_{\mathrm{R}2}, \hat{u}_{\mathrm{R}3}, \hat{d}_{\mathrm{R}3}, \tag{22.150}$$

which transform in the now familiar way under $\mathrm{SU}(2)_L \times \mathrm{U}(1)$. The $\hat{u}$-fields correspond to the $t_3 = +\frac{1}{2}$ components of $\mathrm{SU}(2)_{\mathrm{L}}$, the $\hat{d}$ ones to the $t_3 = -\frac{1}{2}$ components, and to their 'R' partners. The labels 1, 2 and 3 refer to the family number; for example, with no mixing at all, $\hat{u}_{\mathrm{L}1} = \hat{u}_{\mathrm{L}}, \hat{d}_{\mathrm{L}1} = \hat{d}_{\mathrm{L}}$, etc. We have to consider what is the most general $\mathrm{SU}(2)_{\mathrm{L}} \times \mathrm{U}(1)$–invariant interaction between the Higgs field (assuming we can still get by with only one) and these various fields. Apart from the symmetry, the only other theoretical requirement is renormalizability – for, after all, if we drop this we might as well abandon the whole motivation for the 'gauge' concept. This implies (as in the discussion of the Higgs potential $\hat{V}$) that we cannot have terms like $(\bar{\hat{\psi}}\hat{\psi}\hat{\phi})^2$ appearing – which would have a coupling with dimensions $(\mathrm{mass})^{-4}$ and would be non-renormalizable. In fact the only renormalizable Yukawa coupling is of the form '$\bar{\hat{\psi}}\hat{\psi}\hat{\phi}$', which has a dimensionless coupling (as in the g_{e} and $g_{\nu_{\mathrm{e}}}$ of (22.125) and (22.133)). However, there is no *a priori* requirement for it to be 'diagonal' in the weak interaction family index i. The allowed generalization of (22.125) and (22.133) is therefore an interaction of the form (summing on repeated indices)

$$\hat{\mathcal{L}}_{\psi\phi} = a_{ij}\bar{\hat{q}}_{\mathrm{L}i}\hat{\phi}_{\mathbf{C}}\hat{u}_{\mathrm{R}j} + b_{ij}\bar{\hat{q}}_{\mathrm{L}i}\hat{\phi}\hat{d}_{\mathrm{R}j} + \mathrm{h.c.} \tag{22.151}$$

where

$$\hat{q}_{\mathrm{L}i} = \begin{pmatrix} \hat{u}_{\mathrm{L}i} \\ \hat{d}_{\mathrm{L}i} \end{pmatrix}, \tag{22.152}$$

and a sum on the family indices i and j (from 1 to 3) in (22.151) is assumed. After symmetry breaking, using the gauge (22.29), we find (problem 22.6)

$$\hat{\mathcal{L}}_{f\phi} = -\left(1 + \frac{\hat{H}}{v}\right)[\bar{\hat{u}}_{\mathrm{L}i}m^{\mathrm{u}}_{ij}\hat{u}_{\mathrm{R}j} + \bar{\hat{d}}_{\mathrm{L}i}m^{\mathrm{d}}_{ij}\hat{d}_{\mathrm{R}j} + \mathrm{h.c.}], \tag{22.153}$$

where the 'mass matrices' are

$$m^{\mathrm{u}}_{ij} = -\frac{v}{\sqrt{2}}a_{ij}, \qquad m^{\mathrm{d}}_{ij} = -\frac{v}{\sqrt{2}}b_{ij}. \tag{22.154}$$

Although we have not indicated it, the m^{u} and m^{d} matrices could involve a 'γ_5' part as well as a '1' part in Dirac space. It can be shown (Weinberg 1973, Feinberg *et al.* 1959) that m^{u} and m^{d} can both be made Hermitean, γ_5-free, and diagonal by making four separate unitary transformations on the 'generation triplets'

$$\hat{u}_{\mathrm{L}} = \begin{pmatrix} \hat{u}_{\mathrm{L}1} \\ \hat{u}_{\mathrm{L}2} \\ \hat{u}_{\mathrm{L}3} \end{pmatrix}, \qquad \hat{d}_{\mathrm{L}} = \begin{pmatrix} \hat{d}_{\mathrm{L}1} \\ \hat{d}_{\mathrm{L}2} \\ \hat{d}_{\mathrm{L}3} \end{pmatrix}, \mathrm{etc} \tag{22.155}$$

via

$$\hat{u}_{\mathrm{L}\alpha} = (U_{\mathrm{L}}^{(\mathrm{u})})_{\alpha i}\hat{u}_{\mathrm{L}i}, \qquad \hat{u}_{\mathrm{R}\alpha} = (U_{\mathrm{R}}^{(\mathrm{u})})_{\alpha i}\hat{u}_{\mathrm{R}i} \tag{22.156}$$

$$\hat{d}_{\mathrm{L}\alpha} = (U_{\mathrm{L}}^{(\mathrm{d})})_{\alpha i}\hat{d}_{\mathrm{L}i}, \qquad \hat{d}_{\mathrm{R}\alpha} = (U_{\mathrm{R}}^{(\mathrm{d})})_{\alpha i}\hat{d}_{\mathrm{R}i}. \tag{22.157}$$

In this notation, 'α' is the index of the 'mass diagonal' basis, and 'i' is that of the 'weak interaction' basis.[3] Then (22.153) becomes

$$\hat{\mathcal{L}}_{qH} = -\left(1+\frac{\hat{H}}{v}\right)[m_{\mathrm{u}}\bar{\hat{u}}\hat{u} + \ldots + m_{\mathrm{b}}\bar{\hat{b}}\hat{b}]. \tag{22.158}$$

Rather remarkably, we can still manage with only the one Higgs field. It couples to each fermion with a strength proportional to the mass of that fermion, divided by M_{W}.

Now consider the $\mathrm{SU}(2)_{\mathrm{L}}\times\mathrm{U}(1)$ gauge invariant interaction part of the Lagrangian. Written out in terms of the 'weak interaction' fields $\hat{u}_{\mathrm{L,R}i}$ and $\hat{d}_{\mathrm{L,R}i}$ (cf (22.43) and (22.44)), it is

$$\begin{aligned}\hat{\mathcal{L}}_{f,W,B} &= \mathrm{i}(\bar{\hat{u}}_{\mathrm{L}j}, \bar{\hat{d}}_{\mathrm{L}j})\gamma^{\mu}(\partial_{\mu} + \mathrm{i}g\boldsymbol{\tau}\cdot\hat{\boldsymbol{W}}_{\mu}/2 + \mathrm{i}g'y\hat{B}_{\mu}/2)\begin{pmatrix}\hat{u}_{\mathrm{L}j}\\ \hat{d}_{\mathrm{L}j}\end{pmatrix} \\ &\quad + \mathrm{i}\bar{\hat{u}}_{\mathrm{R}j}\gamma^{\mu}(\partial_{\mu} + \mathrm{i}g'y\hat{B}_{\mu}/2)\hat{u}_{\mathrm{R}j} + \mathrm{i}\bar{\hat{d}}_{\mathrm{R}j}\gamma^{\mu}(\partial_{\mu} + \mathrm{i}g'y\hat{B}_{\mu}/2)\hat{d}_{\mathrm{R}j}\end{aligned} \tag{22.159}$$

where a sum on j is understood. This now has to be rewritten in terms of the mass-eigenstate fields $\hat{u}_{\mathrm{L,R}\alpha}$ and $\hat{d}_{\mathrm{L,R}\alpha}$.

Problem 22.7 shows that the neutral current part of (22.159) is diagonal in the mass basis, provided the U matrices of (22.156) and (22.157) are unitary; that is, the neutral current interactions do not change the flavour of the physical (mass eigenstate) quarks. The charged current processes, however, involve the *non*–diagonal matrices τ_1 and τ_2 in (22.159), and this spoils the argument used in problem 22.7. Indeed, using (22.47) we find that the charged current piece is

$$\begin{aligned}\hat{\mathcal{L}}_{\mathrm{CC}} &= -\frac{g}{\sqrt{2}}(\bar{\hat{u}}_{\mathrm{L}j}, \bar{\hat{d}}_{\mathrm{L}j})\gamma_{\mu}\tau_{+}\hat{W}_{\mu}\begin{pmatrix}\hat{u}_{\mathrm{L}j}\\ \hat{d}_{\mathrm{L}j}\end{pmatrix} + \mathrm{h.c.} \\ &= -\frac{g}{\sqrt{2}}\bar{\hat{u}}_{\mathrm{L}j}\gamma^{\mu}\hat{d}_{\mathrm{L}j}\hat{W}_{\mu} + \mathrm{h.c.} \\ &= -\frac{g}{\sqrt{2}}\bar{\hat{u}}_{\mathrm{L}\alpha}[(U_{\mathrm{L}}^{(\mathrm{u})})_{\alpha j}(U_{\mathrm{L}}^{(\mathrm{d})\dagger})_{j\beta}]\gamma^{\mu}\hat{d}_{\mathrm{L}\beta}\hat{W}_{\mu} + \mathrm{h.c.},\end{aligned} \tag{22.160}$$

where the matrix

$$V_{\alpha\beta} \equiv [U_{\mathrm{L}}^{(\mathrm{u})}U_{\mathrm{L}}^{(\mathrm{d})\dagger}]_{\alpha\beta} \tag{22.161}$$

is not diagonal, though it is unitary. This is the CKM matrix (Cabibbo

[3] So, for example, $\hat{u}_{\mathrm{L}\alpha=\mathrm{t}} \equiv \hat{t}_{\mathrm{L}}, \hat{d}_{\mathrm{L}\alpha=\mathrm{s}} \equiv \hat{s}_{\mathrm{L}}$, etc.

1963, Kobayashi and Maskawa 1973), originally introduced by Kobayashi and Maskawa in the context of their three-generation extension of the then-developing Standard Model, in order to provide room for **CP** violation within the SU(2) × U(1) gauge theory framework. The interaction (22.160) then has the form

$$-\frac{g}{\sqrt{2}}\hat{W}_\mu[\bar{\hat{u}}_{\rm L}\gamma^\mu\hat{d}'_{\rm L}+\bar{\hat{c}}_{\rm L}\gamma^\mu\hat{s}'_{\rm L}+\bar{\hat{t}}_{\rm L}\gamma^\mu\hat{b}'_{\rm L}]+{\rm h.c.}, \tag{22.162}$$

where

$$\begin{pmatrix}\hat{d}'_{\rm L}\\ \hat{s}'_{\rm L}\\ \hat{b}'_{\rm L}\end{pmatrix}=\begin{pmatrix}V_{\rm ud}&V_{\rm us}&V_{\rm ub}\\ V_{\rm cd}&V_{\rm cs}&V_{\rm cb}\\ V_{\rm td}&V_{\rm ts}&V_{\rm tb}\end{pmatrix}\begin{pmatrix}\hat{d}_{\rm L}\\ \hat{s}_{\rm L}\\ \hat{b}_{\rm L}\end{pmatrix}, \tag{22.163}$$

with the phenomenology described in the previous chapter.

An analysis similar to the above can be carried out in the leptonic sector. We would then have leptonic flavour mixing in charged current processes, via the leptonic analogue of the CKM matrix, namely the PMNS matrix (Pontecorvo 1957, 1958, 1967; Maki, Nakagawa and Sakata 1962); this is the matrix whose elements are probed in neutrino oscillations, as we saw in chapter 21.

22.6 Higher-order corrections

The Z^0 mass

$$M_{\rm Z}=91.1876\pm0.0021\ {\rm GeV} \tag{22.164}$$

has been determined from the Z-lineshape scan at LEP1 (Schael *et al.* 2006). The W mass is (Nakamura *et al.* 2010)

$$M_{\rm W}=80.399\pm0.023\ {\rm GeV}. \tag{22.165}$$

The asymmetry parameter $A_{\rm e}$ (see (22.100)) is (Abe *et al.* (2000))

$$A_{\rm e}=0.15138\pm0.00216 \tag{22.166}$$

from measurements at SLD. These are just three examples from the table of 36 observables listed in the review of the electroweak model by Erler and Langacker in Nakamura *et al.* (2010). Such remarkable precision is a triumph of machine design and experimental art – and it is the reason why we need a renormalizable electroweak theory. The overall fit to the data, including higher-order corrections, is generally very good, as quoted by Erler and Langacker with $\chi^2/{\rm d.o.f}=43.0/44$. One of the few discrepancies is a 2.7σ deviation in the Z-pole forward-backward asymmetry $A_{\rm FB}^{(0,{\rm b})}$ from LEP1; another is a 2.5σ deviation in the muon anomalous magnetic moment, $g_\mu-2$. This strong numerical consistency lends impressive support to the belief that we

are indeed dealing with a *renormalizable spontaneously broken gauge theory*: renormalizable, because no extra parameters, not in the original Lagrangian, have had to be introduced; a gauge theory, because the fermion and gauge boson couplings obey the relations imposed by the local SU(2) × U(1) symmetry; and spontaneously broken because the same symmetry is not seen in the particle spectrum (consider the mass separation in the t-b doublet, for instance).

In fact, one can turn this around, in more than one way. First, one crucially important element in the theory – the Higgs boson – has a mass m_{H} which is largely unconstrained by theory (see section 22.8.2), and it is therefore a parameter in the fits. Some information about m_{H} can therefore be gained by seeing how the fits vary with m_{H}. Actually, we shall see in equation (22.181) that the dependence on m_{H} is only logarithmic – it acts rather like a cut-off, so the fits are not very sensitive to m_{H}. The 90 % central confidence range from all precision data is given by Erler and Langacker as 55 GeV$< m_{\mathrm{H}} <$ 135 GeV. By contrast, some loop corrections are proportional to the square of the top mass (see (22.180)) and consequently very tight bounds could be placed on m_{t} via its *virtual* presence (i.e. in loops, for example as shown in figure 22.25) before its *real* presence was confirmed, as we shall discuss shortly and in section 22.7. Secondly, it is still entirely possible that very careful analysis of small discrepancies between precision data and electroweak predictions may indicate the presence of 'new physics'.

After all this (and earlier) emphasis on the renormalizability of the electroweak theory, and the introduction to one-loop calculations in QED at the end of volume 1, the reader perhaps has a right to expect, now, an exposition of loop corrections in the electroweak theory. But the fact is that this is a very complicated and technical story, requiring quite a bit more formal machinery, which would be outside the intended scope of this book (suitable references include Altarelli *et al.* 1989, especially the pedagogical account by Consoli *et al.* 1989; and the equally approachable lectures by Hollik 1991). Instead, we want to touch on just a few of the simpler and more important aspects of one-loop corrections, especially insofar as they have phenomenological implications.

As we have seen, we obtain cut-off independent results from loop corrections in a renormalizable theory by taking the values of certain parameters – those appearing in the original Lagrangian – from experiment, according to a well-defined procedure ('renormalization scheme'). In the electroweak case, the parameters in the Lagrangian are

$$\text{gauge couplings } g, g' \tag{22.167}$$

$$\text{Higgs potential parameters } \lambda, \mu^2 \tag{22.168}$$

$$\text{Higgs–fermion Yukawa couplings } g_f \tag{22.169}$$

$$\text{CKM angles } \theta_{12}, \theta_{13}, \theta_{23};\ \text{phase } \delta \tag{22.170}$$

$$\text{PMNS angles } \theta_{\mathrm{e}2}, \theta_{\mathrm{e}3}, \theta_{\mu 3},\ \text{phase } \delta^{\nu}\ (+\ \alpha_{21}, \alpha_{31}?). \tag{22.171}$$

The fermion masses and mixings, and the Higgs mass, can be separated off, leaving g, g' and one combination of λ and μ^2 (for instance, the tree-level vacuum value v). These three parameters are usually replaced by the equivalent and more convenient set

$$\alpha \quad \text{(Bouchiendra } et\ al.\ 2011); \tag{22.172}$$

$$G_{\rm F} \quad \text{(Marciano and Sirlin 1988, van Ritbergen and Stuart 1999),} \tag{22.173}$$

(see also Nir 1989, Pak and Czarnecki 2008, Chitwood *et al.* 2007, and Barezyk *et al.* 2008); and

$$M_{\rm Z} \quad \text{(Schael } et\ al.\ 2006). \tag{22.174}$$

These are, of course, related to g, g' and v; for example, at tree-level

$$\alpha = g^2 g'^2/(g^2+g'^2)4\pi, \quad M_{\rm Z} = \frac{1}{2} v\sqrt{g^2+g'^2}, \quad G_{\rm F} = \frac{1}{\sqrt{2}v^2}, \tag{22.175}$$

but these relations become modified in higher order. The renormalized parameters will 'run' in the way described in chapters 15 and 16; the running of α, for example, has been observed directly, as noted in section 11.5.

After renormalization one can derive radiatively corrected values for physical quantities in terms of the set (22.172)–(22.174) (together with $m_{\rm H}$ and the fermion masses and mixings). But a renormalization scheme has to be specified, at any finite order (though in practice the differences are very small). One conceptually simple scheme is the 'on-shell' one (Sirlin 1980, 1984; Kennedy *et al.* 1989; Kennedy and Lynn 1989; Bardin *et al.* 1989; Hollik 1990; for reviews see Langacker 1995). In this scheme, the tree-level formula

$$\sin^2\theta_{\rm W} = 1 - M_{\rm W}^2/M_{\rm Z}^2 \tag{22.176}$$

is promoted into a *definition* of the renormalized $\sin^2\theta_{\rm W}$ to all orders in perturbation theory, it being then denoted by $s_{\rm W}^2$:

$$s_{\rm W}^2 = 1 - M_{\rm W}^2/M_{\rm Z}^2 \approx 0.223. \tag{22.177}$$

The radiatively corrected value for $M_{\rm W}$ is then

$$M_{\rm W}^2 = \frac{(\pi\alpha/\sqrt{2}G_{\rm F})}{s_{\rm W}^2(1-\Delta r)} \tag{22.178}$$

where Δr includes the radiative corrections relating $\alpha, \alpha(M_{\rm Z}), G_{\rm F}, M_{\rm W}$ and $M_{\rm Z}$. Another scheme is the modified minimal subtraction ($\overline{\rm MS}$) scheme (appendix O) which introduces the quantity $\sin^2\hat{\theta}_{\rm W}(\mu) \equiv \hat{g}'^2(\mu)/[\hat{g}'^2(\mu)+\hat{g}^2(\mu)]$ where the couplings $\hat{g}$ and $\hat{g}'$ are defined in the $\overline{\rm MS}$ scheme and μ is chosen to be $M_{\rm Z}$ for most electroweak processes. Attention is then focused on $\hat{s}_{\rm Z}^2 \equiv \sin^2\hat{\theta}_{\rm W}(M_{\rm Z})$. This is the scheme used by Erler and Langacker in Nakamura *et al* (2010).

We shall continue here with the scheme defined by (22.177). We cannot go into detail about all the contributions to Δr, but we do want to highlight two features of the result – which are surprising, important phenomenologically, and related to an interesting symmetry. It turns out (Consoli *et al.* 1989, Hollik 1991) that the leading terms in Δr have the form

$$\Delta r = \Delta r_0 - \frac{(1 - s^2_{\rm W})}{s^2_{\rm W}} \Delta\rho + (\Delta r)_{\rm rem}. \tag{22.179}$$

In (22.179), $\Delta r_0 = 1 - \alpha/\alpha(M_{\rm Z})$ is due to the running of α, and has the value $\Delta r_0 = 0.0664(2)$ (see section 11.5.3). $\Delta\rho$ is given by (Veltman 1977)

$$\Delta\rho = \frac{3G_{\rm F}(m^2_{\rm t} - m^2_{\rm b})}{8\pi^2\sqrt{2}}, \tag{22.180}$$

while the 'remainder' $(\Delta r)_{\rm rem}$ contains a significant term proportional to $\ln(m_{\rm t}/m_{\rm Z})$, and a contribution from the Higgs boson which is (for $m_{\rm H} \gg M_{\rm W}$)

$$(\Delta r)_{\rm rem,H} \approx \frac{\sqrt{2}G_{\rm F}M^2_{\rm W}}{16\pi^2}\frac{11}{3}\left[\ln\left(\frac{m^2_{\rm H}}{M^2_{\rm W}}\right) - \frac{5}{6}\right]. \tag{22.181}$$

As the notation suggests, $\Delta\rho$ is a leading contribution to the parameter ρ introduced in (22.66). As explained there, it measures the strength of neutral current processes relative to charged current ones. $\Delta\rho$ is then a radiative correction to ρ. It turns out that, to good approximation, electroweak radiative corrections in $\mathrm{e^+e^-} \to \mathrm{Z^0} \to f\bar{f}$ can be included by replacing the fermionic couplings $g^f_{\rm V}$ and $g^f_{\rm A}$(see (22.64), (22.75) and (22.76)) by

$$\bar{g}^f_{\rm V} = \sqrt{\rho_f}(t^{(f)}_3 - 2Q_f\kappa_f s^2_{\rm W}) \tag{22.182}$$

and

$$\bar{g}^f_{\rm A} = \sqrt{\rho_f}t^{(f)}_3, \tag{22.183}$$

together with corrections to the $\mathrm{Z^0}$-propagator. The corrections have the form (in the on-shell scheme) $\rho_f \approx 1 + \Delta\rho$ (of equation (22.180)) and $\kappa_f \approx 1 + \frac{s^2_{\rm W}}{(1-s^2_{\rm W})}\Delta\rho$, for $f \neq \mathrm{b, t}$. For the b-quark there is an additional contribution coming from the presence of the virtual top quark in vertex corrections to $\mathrm{Z} \to \mathrm{b\bar{b}}$ (Akhundov *et al.* 1986, Beenakker and Hollik 1988).

The running of α in Δr_0 is expected, but (22.180) and (22.181) contain surprising features. As regards (22.180), it is associated with top-bottom quark loops in vacuum polarization amplitudes, of the kind discussed for $\bar{\Pi}^{[2]}_\gamma$ in section 11.5, but this time in weak boson propagators. In the QED case, referring to equation (11.39) for example, we saw that the contribution of heavy fermions '$(|q^2| \ll m^2_f)$' was suppressed, appearing as $O(|q^2|/m^2_f)$. In such a situation (which is the usual one) the heavy particles are said to 'decouple'. But the correction (22.180) is quite different, the fermion masses being in the

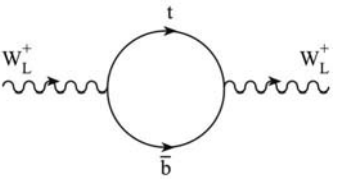


FIGURE 22.25
t - $\bar{\text{b}}$ vacuum polarization contribution.

numerator. Clearly, with a large value m_{t}, this can make a relatively big difference. This is why some precision measurements are surprisingly sensitive to the value of m_{t}, in the range near (as we now know) the physical value. Secondly, as regards the dependence on m_{H}, we might well have expected it to involve m_{H}^2 in the numerator if we considered the typical divergence of a scalar particle in a loop (we shall return to this after discussing (22.180)). Δr would then have been very sensitive to m_{H}, but in fact the sensitivity is only logarithmic.

We can understand the appearance of the fermion masses (squared) in the numerator of (22.180) as follows. The shift $\Delta\rho$ is associated with vector boson vacuum polarization contributions, for example the one shown in figure 22.25. Consider in particular the contribution from the longitudinal polarization components of the W's. As we have seen, these components are nothing but three of the four Higgs components which the $\text{W}^{\pm}$ and Z^0 'swallowed' to become massive. But the couplings of these 'swallowed' Higgs fields to fermions are determined by just the same Higgs-fermion Yukawa couplings as we introduced to generate the fermion masses via spontaneous symmetry breaking. Hence we expect the fermion loops to contribute (to these longitudinal W states) something of order $g_f^2/4\pi$ where g_f is the Yukawa coupling. Since $g_f \sim m_f/v$ (see (22.127)) we arrive at an estimate $\sim m_f^2/4\pi v^2 \sim G_F m_f^2/4\pi$ as in (22.180). An important message is that *particles which acquire their mass spontaneously do not 'decouple'.*

But we now have to explain why $\Delta\rho$ in (22.180) would vanish if $m_{\text{t}}^2 = m_{\text{b}}^2$, and why only $\ln m_{\text{H}}^2$ appears in (22.181). Both these facts are related to a symmetry of the assumed minimal Higgs sector which we have not yet discussed. Let us first consider the situation at tree level, where $\rho = 1$. It may be shown (Ross and Veltman 1975) that $\rho = 1$ is a natural consequence of having the symmetry broken by an SU(2)_{L} doublet Higgs field (rather than a triplet, say) – or indeed by any number of doublets. The nearness of the measured ρ parameter to 1 is, in fact, good support for the hypothesis that there are only doublet Higgs fields. Problem 22.8 explores a simple model with a Higgs field in the triplet representation.

At tree level, it is simplest to think of ρ in connection with the mass ratio (22.66). To see the significance of this, let us go back to the Higgs-gauge field

Lagrangian $\hat{\mathcal{L}}_{G\Phi}$ of (22.30) which produced the gauge boson masses. With the doublet Higgs of the form (22.131), it is a striking fact that the Higgs potential only involves the highly symmetrical combination of fields

$$\hat{\phi}_1^2 + \hat{\phi}_2^2 + \hat{\phi}_3^2 + \hat{\phi}_4^2, \tag{22.184}$$

as does the vacuum condition (17.102). This suggests that there may be some extra symmetry in (22.30) which is special to the doublet structure. But of course, to be of any interest, this symmetry has to be present in the $(\hat{D}_\mu\hat{\phi})^\dagger(\hat{D}^\mu\hat{\phi})$ term as well.

The nature of this symmetry is best brought out by introducing a change of notation for Higgs doublet $\hat{\phi}^+$ and $\hat{\phi}^0$: instead of (22.131), we now write (cf (18.70))

$$\hat{\phi} = \begin{pmatrix} (\hat{\pi}_2 + \mathrm{i}\hat{\pi}_1)/\sqrt{2} \\ (\hat{\sigma} - \mathrm{i}\hat{\pi}_3)/\sqrt{2} \end{pmatrix} \tag{22.185}$$

while the $\hat{\phi}_{\mathbf{C}}$ field of (22.132) becomes

$$\hat{\phi}_{\mathbf{C}} = \begin{pmatrix} (\hat{\sigma} + \mathrm{i}\hat{\pi}_3)/\sqrt{2} \\ -(\hat{\pi}_2 - \mathrm{i}\hat{\pi}_1)/\sqrt{2} \end{pmatrix}. \tag{22.186}$$

We then find that these can be written as

$$\hat{\phi} = \frac{1}{\sqrt{2}}(\hat{\sigma} + \mathrm{i}\boldsymbol{\tau}\cdot\hat{\boldsymbol{\pi}})\begin{pmatrix} 0 \\ 1 \end{pmatrix}, \qquad \hat{\phi}_{\mathbf{C}} = \frac{1}{\sqrt{2}}(\hat{\sigma} + \mathrm{i}\boldsymbol{\tau}\cdot\hat{\boldsymbol{\pi}})\begin{pmatrix} 1 \\ 0 \end{pmatrix}. \tag{22.187}$$

Consider now the covariant $\mathrm{SU}(2)_\mathrm{L}\times\mathrm{U}(1)$ derivative acting on $\hat{\phi}$, as in (22.30), and suppose to begin with that $g' = 0$. Then

$$\begin{aligned} \hat{D}_\mu\hat{\phi} &= \frac{1}{\sqrt{2}}(\partial_\mu + \mathrm{i}g\boldsymbol{\tau}\cdot\hat{\boldsymbol{W}}_\mu/2)(\hat{\sigma} + \mathrm{i}\boldsymbol{\tau}\cdot\hat{\boldsymbol{\pi}})\begin{pmatrix} 0 \\ 1 \end{pmatrix} \\ &= \frac{1}{\sqrt{2}}\Big\{\partial_\mu\hat{\sigma} + \mathrm{i}\boldsymbol{\tau}\cdot\partial_\mu\hat{\boldsymbol{\pi}} + \mathrm{i}\frac{g}{2}\hat{\sigma}\boldsymbol{\tau}\cdot\hat{\boldsymbol{W}}_\mu \\ &\qquad - \frac{g}{2}[\hat{\boldsymbol{\pi}}\cdot\hat{\boldsymbol{W}}_\mu + \mathrm{i}\boldsymbol{\tau}\cdot\hat{\boldsymbol{W}}_\mu\times\hat{\boldsymbol{\pi}}]\Big\}\begin{pmatrix} 0 \\ 1 \end{pmatrix} \end{aligned} \tag{22.188}$$

using $\tau_i\tau_j = \delta_{ij} + \mathrm{i}\epsilon_{ijk}\tau_k$. Now the vacuum choice (22.28) corresponds to $\hat{\sigma} = v, \hat{\boldsymbol{\pi}} = 0$, so that when we form $(\hat{D}_\mu\hat{\phi})^\dagger(\hat{D}^\mu\hat{\phi})$ from (22.188) we will get just

$$\frac{1}{2}(0,1)\left\{\frac{g^2}{4}v^2(\boldsymbol{\tau}\cdot\hat{\boldsymbol{W}}_\mu)(\boldsymbol{\tau}\cdot\hat{\boldsymbol{W}}^\mu)\right\}\begin{pmatrix} 0 \\ 1 \end{pmatrix} = \frac{1}{2}M_\mathrm{W}^2\hat{\boldsymbol{W}}_\mu\cdot\hat{\boldsymbol{W}}^\mu \tag{22.189}$$

with $M_\mathrm{W} = gv/2$ as usual. The condition $g' = 0$ corresponds (cf (22.39)) to $\theta_\mathrm{W} = 0$, and thus to $\hat{W}_{3\mu} = \hat{Z}_\mu$, and so (22.189) says that in the limit of $g' \to 0$, $M_\mathrm{W} = M_\mathrm{Z}$, as expected if $\cos\theta_\mathrm{W} = 1$. It is clear from (22.188) that the three components $\hat{\boldsymbol{W}}_\mu$ are treated on a precisely equal footing by the

Higgs field (22.185), and indeed the notation suggests that $\hat{\boldsymbol{W}}_\mu$ and $\hat{\boldsymbol{\pi}}$ should perhaps be regarded as some kind of *new* triplets.

It is straightforward to calculate $(\hat{D}_\mu\hat{\phi})^\dagger(\hat{D}^\mu\hat{\phi})$ from (22.188); one finds (problem 22.9)

$$\begin{aligned}(D_\mu\hat{\phi})^\dagger D^\mu\hat{\phi} &= \frac{1}{2}(\partial_\mu\hat{\sigma})^2+\frac{1}{2}(\partial_\mu\hat{\boldsymbol{\pi}})^2-\frac{g}{2}\partial_\mu\hat{\sigma}\hat{\boldsymbol{\pi}}\cdot\hat{\boldsymbol{W}}^\mu \\ &\quad+\frac{g}{2}\hat{\sigma}\partial_\mu\hat{\boldsymbol{\pi}}\cdot\hat{\boldsymbol{W}}^\mu+\frac{g}{2}\partial_\mu\hat{\boldsymbol{\pi}}\cdot(\hat{\boldsymbol{\pi}}\times\hat{\boldsymbol{W}}^\mu) \\ &\quad+\frac{g^2}{8}\hat{\boldsymbol{W}}_\mu^{\,2}(\hat{\sigma}^2+\hat{\boldsymbol{\pi}}^2). \end{aligned} \tag{22.190}$$

This expression now reveals what the symmetry is: (22.190) is invariant under global SU(2) transformations under which $\hat{\boldsymbol{W}}_\mu$ and $\hat{\boldsymbol{\pi}}$ are vectors – that is

$$\left.\begin{array}{c}\hat{\boldsymbol{W}}_\mu\to\hat{\boldsymbol{W}}_\mu+\boldsymbol{\epsilon}\times\hat{\boldsymbol{W}}_\mu \\ \hat{\boldsymbol{\pi}}\to\hat{\boldsymbol{\pi}}+\boldsymbol{\epsilon}\times\hat{\boldsymbol{\pi}} \\ \hat{\sigma}\to\hat{\sigma}\end{array}\right\}. \tag{22.191}$$

This is why, from the term $\hat{\boldsymbol{W}}_\mu^{\,2}\hat{\sigma}^2$, all three W fields have the same mass in this $g'\to 0$ limit.

If we now reinstate g', and use (22.36) and (22.37) to write $\hat{W}_{3\mu}$ and $\hat{B}_\mu$ in terms of the physical fields $\hat{Z}_\mu$ and $\hat{A}_\mu$ as in (19.96), (22.188) becomes

$$\begin{aligned}\frac{1}{\sqrt{2}}\Bigg\{&\partial_\mu+\mathrm{i}g\frac{\tau_1}{2}\hat{W}_{1\mu}+\mathrm{i}g\frac{\tau_2}{2}\hat{W}_{2\mu}+\mathrm{i}g\frac{\tau_3}{2}\frac{\hat{Z}_\mu}{\cos\theta_{\mathrm{W}}}+\mathrm{i}g\sin\theta_{\mathrm{W}}\left(\frac{1+\tau_3}{2}\right)\hat{A}_\mu \\ &-\frac{\mathrm{i}g}{\cos\theta_{\mathrm{W}}}\sin^2\theta_{\mathrm{W}}\left(\frac{1+\tau_3}{2}\right)\hat{Z}_\mu\Bigg\}(\hat{\sigma}+\mathrm{i}\boldsymbol{\tau}\cdot\hat{\boldsymbol{\pi}})\begin{pmatrix}0\\1\end{pmatrix}. \end{aligned} \tag{22.192}$$

We see from (22.192) that $g'\neq 0$ has two effects. First, there is a '$\boldsymbol{\tau}\cdot\hat{\boldsymbol{W}}$'–like term, as in (22.188), except that the '$\hat{W}_3$' part of it is now $\hat{Z}/\cos\theta_{\mathrm{W}}$. In the vacuum $\hat{\sigma}=v,\hat{\boldsymbol{\pi}}=0$ which simply means that the mass of the Z is $M_{\mathrm{Z}}=M_{\mathrm{W}}/\cos\theta_{\mathrm{W}}$ i.e. $\rho=1$; and this relation is preserved under 'rotations' of the form (22.191), since they do not mix $\hat{\boldsymbol{\pi}}$ and $\hat{\sigma}$. Hence this mass relation (and $\rho=1$) is a consequence of the global SU(2) symmetry of the interactions and the vacuum under (22.191), and of the relations (22.36) and (22.37) which embody the requirement of a massless photon.

On the other hand, there are additional terms in (22.192) which single out the 'τ_3' component, and therefore break this global SU(2). These terms vanish as $g'\to 0$, and do not contribute at tree level, but we expect that they will cause $O(g'^2)$ corrections to $\rho=1$ at the one loop level.

None of the above, however, yet involves the quark masses, and the question of why $m_{\mathrm{t}}^2-b_{\mathrm{b}}^2$ appears in the numerator in (22.180). We can now answer this question. Consider a typical mass term, of the form discussed in section 22.5.2, for a quark doublet of the i^{th} family

$$\hat{\mathcal{L}}_m=-g_+(\bar{\hat{u}}_{\mathrm{L}i}\bar{\hat{d}}_{\mathrm{L}i})\hat{\phi}_{\mathbf{C}}\hat{u}_{\mathrm{R}i}-g_-(\bar{\hat{u}}_{\mathrm{L}i}\bar{\hat{d}}_{\mathrm{L}i})\hat{\phi}\hat{d}_{\mathrm{R}i}. \tag{22.193}$$

Using (22.185) and (22.186), this can be written as

$$\begin{aligned}\hat{\mathcal{L}}_m &= \frac{-g_+}{\sqrt{2}}(\bar{\hat{u}}_{\mathrm{L}i}\bar{\hat{d}}_{\mathrm{L}i})(\hat{\sigma}+\mathrm{i}\boldsymbol{\tau}\cdot\hat{\boldsymbol{\pi}})\begin{pmatrix}\hat{u}_{\mathrm{R}i}\\0\end{pmatrix}-\frac{g_-}{\sqrt{2}}(\bar{\hat{u}}_{\mathrm{L}i}\bar{\hat{d}}_{\mathrm{L}i})(\hat{\sigma}+\mathrm{i}\boldsymbol{\tau}\cdot\hat{\boldsymbol{\pi}})\begin{pmatrix}0\\\hat{d}_{\mathrm{R}i}\end{pmatrix}\\ &= -\frac{(g_++g_-)}{2\sqrt{2}}(\bar{\hat{u}}_{\mathrm{L}i}\bar{\hat{d}}_{\mathrm{L}i})(\hat{\sigma}+\mathrm{i}\boldsymbol{\tau}\cdot\hat{\boldsymbol{\pi}})\begin{pmatrix}\hat{u}_{\mathrm{R}i}\\\hat{d}_{\mathrm{R}i}\end{pmatrix}\\ &\quad -\frac{(g_+-g_-)}{2\sqrt{2}}(\bar{\hat{u}}_{\mathrm{L}i}\bar{\hat{d}}_{\mathrm{L}i})(\hat{\sigma}+\mathrm{i}\boldsymbol{\tau}\cdot\hat{\boldsymbol{\pi}})\tau_3\begin{pmatrix}\hat{u}_{\mathrm{R}i}\\\hat{d}_{\mathrm{R}i}\end{pmatrix}. \qquad (22.194)\end{aligned}$$

Consider now a simultaneous (infinitesimal) global SU(2) transformation on the two doublets $(\hat{u}_{\mathrm{L}i},\hat{d}_{\mathrm{L}i})^{\mathrm{T}}$ and $(\hat{u}_{\mathrm{R}i},\hat{d}_{\mathrm{R}i})^{\mathrm{T}}$:

$$\begin{pmatrix}\hat{u}_{\mathrm{L}i}\\\hat{d}_{\mathrm{L}i}\end{pmatrix}\to(1-\mathrm{i}\boldsymbol{\epsilon}\cdot\boldsymbol{\tau}/2)\begin{pmatrix}\hat{u}_{\mathrm{L}i}\\\hat{d}_{\mathrm{L}i}\end{pmatrix},\qquad\begin{pmatrix}\hat{u}_{\mathrm{R}i}\\\hat{d}_{\mathrm{R}i}\end{pmatrix}\to(1-\mathrm{i}\boldsymbol{\epsilon}\cdot\boldsymbol{\tau}/2)\begin{pmatrix}\hat{u}_{\mathrm{R}i}\\\hat{d}_{\mathrm{R}i}\end{pmatrix}. \qquad (22.195)$$

Under (22.195), the first term of (22.194) becomes (to first order in $\boldsymbol{\epsilon}$)

$$-\frac{(g_++g_-)}{2\sqrt{2}}(\bar{\hat{u}}_{\mathrm{L}i}\bar{\hat{d}}_{\mathrm{L}i})[\hat{\sigma}+\mathrm{i}\boldsymbol{\tau}\cdot(\hat{\boldsymbol{\pi}}+\hat{\boldsymbol{\pi}}\times\boldsymbol{\epsilon})]\begin{pmatrix}\hat{u}_{\mathrm{R}i}\\\hat{d}_{\mathrm{R}i}\end{pmatrix}. \qquad (22.196)$$

From (22.196) we see that if, at the same time as (22.195), we *also* make the transformation of $\boldsymbol{\pi}$ given in (22.191), then this first term in $\hat{\mathcal{L}}_m$ will be invariant under these combined transformations. The second term in (22.194), however, will not be invariant under (22.195), but only under transformations with $\epsilon_1=\epsilon_2=0,\epsilon_3\neq 0$. We conclude that the global SU(2) symmetry of (22.191), which was responsible for $\rho=1$ at the tree level, can be extended also to the quark sector; but – because the $g_\pm$ in (22.193) are proportional to the masses of the quark doublet – this symmetry is explicitly broken by the quark mass difference. This is why a t–$\bar{\mathrm{b}}$ loop in a W vacuum polarization correction can produce the 'non-decoupled' contribution (22.180) to ρ, which grows as $m_{\mathrm{t}}^2-m_{\mathrm{b}}^2$ and produces quite detectable shifts from the tree-level predictions, given the accuracy of the data.

Returning to (22.195), the transformation on the L-components is just the same as a standard $\mathrm{SU}(2)_{\mathrm{L}}$ transformation, except that it is global; so the gauge interactions of the quarks obey this symmetry also. As far as the R-components are concerned, they are totally decoupled in the gauge dynamics, and we are free to make the transformation (22.195) if we wish. The resulting complete transformation, which does the same to both the L and R components, is a non-chiral one – in fact it is precisely an ordinary 'isospin' transformation of the type

$$\begin{pmatrix}\hat{u}_i\\\hat{d}_i\end{pmatrix}\to(1-\mathrm{i}\boldsymbol{\epsilon}\cdot\boldsymbol{\tau}/2)\begin{pmatrix}\hat{u}_i\\\hat{d}_i\end{pmatrix}. \qquad (22.197)$$

The reader will recognize that the mathematics here is exactly the same as that in section 18.3, involving the SU(2) of isospin in the σ-model. This

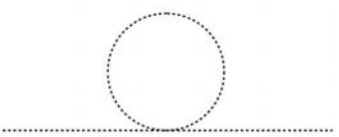


FIGURE 22.26
One-boson self-energy graph in $(\hat{\phi}^\dagger\hat{\phi})^2$.

analysis of the symmetry of the Higgs (or a more general symmetry breaking sector) was first given by Sikivie *et al.* (1980). The isospin-SU(2) is frequently called 'custodial SU(2)' since it 'protects' $\rho = 1$.

What about the *absence* of $m_{\rm H}^2$ corrections? Here the position is rather more subtle. Without the Higgs particle H the theory is non-renormalizable, and hence one might expect to see some radiative correction becoming very large ($O(m_{\rm H}^2)$) as one tried to 'banish' H from theory by sending $m_{\rm H} \to \infty$ ($m_{\rm H}$ would be acting like a cut-off). The reason is that in such a $(\hat{\phi}^\dagger\hat{\phi})^2$ theory, the simplest loop we meet is that shown in figure 22.18, and it is easy to see by counting powers, as usual, that it diverges as the square of the cut-off. This loop contributes to the Higgs self-energy, and will be renormalized by taking the value of the coefficient of $\hat{\phi}^\dagger\hat{\phi}$ in (22.30) from experiment. We will return to this particular detail in section 22.8.1.

Even without a Higgs contribution however, it turns out that the electroweak theory is renormalizable at the one-loop level if the fermion masses are zero (Veltman 1968,1970). Thus one suspects that the large $m_{\rm H}^2$ effects will not be so dramatic after all. In fact, calculation shows (Veltman 1977; Chanowitz *et al.* 1978, 1979) that one-loop radiative corrections to electroweak observables grow at most like $\ln m_{\rm H}^2$ for large $m_{\rm H}$. While there are finite corrections which are approximately $O(m_{\rm H}^2)$ for $m_{\rm H}^2 \ll M_{\rm W,Z}^2$, for $m_{\rm H}^2 \gg M_{\rm W,Z}^2$ the $O(m_{\rm H}^2)$ pieces cancel out from all observable quantities[4], leaving only $\ln m_{\rm H}^2$ terms. This is just what we have in (22.181), and it means, unfortunately, that the sensitivity of the data to this important parameter of the Standard Model is only logarithmic. Fits to data typically give $m_{\rm H}$ in the region of 90 GeV at the minimum of the χ^2 curve, but the error (which is not simple to interpret) is of the order of 25 GeV.

At the two-loop level, the expected $O(m_{\rm H}^4)$ behaviour becomes $O(m_{\rm H}^2)$ instead (van der Bij and Veltman 1984, van der Bij 1984) – and of course appears (relative to the one-loop contributions) with an additional factor of $O(\alpha)$. This relative insensitivity of the radiative corrections to $m_{\rm H}$, in the limit of large $m_{\rm H}$, was discovered by Veltman (1977) and called a 'screening' phenomenon by him: for large $m_{\rm H}$ (which also means, as we have seen, large λ) we have an effectively strongly interacting theory whose principal effects are

[4]Apart from the $\hat{\phi}^\dagger\hat{\phi}$ coefficient! See section 22.8.1.

screened off from observables at lower energy. It was shown by Einhorn and Wudka (1989) that this screening is also a consequence of the (approximate) isospin-SU(2) symmetry we have just discussed in connection with (22.180). Phenomenologically, the upshot is that it was unfortunately very difficult to get an accurate handle on the value of $m_{\rm H}$ from fits to the precision data. With the top quark, the situation was very different.

22.7 The top quark

Having drawn attention to the relative sensitivity of radiative connections to loops containing virtual top quarks, it is worth devoting a little space to a 'backward glance' at the year immediately prior to the discovery of the t-quark (Abe *et al.* 1994a, b, 1995b, Abachi *et al.* 1995b) at the CDF and D0 detectors at FNAL's Tevatron, in $\mathrm{p}-\bar{\mathrm{p}}$ collisions at $E_{\rm cm} = 1.8$ TeV.

The W and Z particles were, as we have seen, discovered in 1983 and at that time, and for some years subsequently, the data were not precise enough to be sensitive to virtual t-effects. In the late 1980's and early 1990's, LEP at CERN and SLC at Stanford began to produce new and highly accurate data which did allow increasingly precise predictions to be made for the top quark mass, $m_{\rm t}$. Thus a kind of race began, between experimentalists searching for the real top, and theorists fitting ever more precise data to get tighter and tighter limits on $m_{\rm t}$, from its virtual effects.

In fact, by the time of the actual experimental discovery of the top quark, the experimental error in $m_{\rm t}$ was just about the same as the theoretical one (and – of course – the central values were consistent). Thus, in their May 1994 review of the electroweak theory (contained in Montanet *et al.* 1994, p 1304ff) Erler and Langacker gave the result of a fit to all electroweak data as

$$m_{\rm t} = 169 \pm^{16}_{18} \pm^{17}_{20} \text{ GeV}, \tag{22.198}$$

the central figure and first error being based on $m_{\rm H} = 300$ GeV, the second (+) error assuming $m_{\rm H} = 1000$ GeV and the second (−) error assuming $m_{\rm H} = 60$ GeV.[5] At about the same time, Ellis *et al.* (1994) gave the extraordinarily precise value

$$m_{\rm t} = 162 \pm 9 \text{ GeV} \tag{22.199}$$

without any assumption for $m_{\rm H}$.

A month or so earlier, the CDF collaboration (Abe *et al.* 1994a,b) announced 12 events consistent with the hypothesis of production of a $\mathrm{t\bar{t}}$ pair, and on this hypothesis the mass was found to be

$$m_{\rm t} = 174 \pm 10 \pm^{13}_{12} \text{ GeV}, \tag{22.200}$$

[5] The relatively small effect of large variations in $m_{\rm H}$ illustrates the lack of sensitivity to virtual Higgs effects, noted in the preceding section.

and this was followed by nine similar events from D0 (Abachi *et al.* 1995a). By February 1995 both groups had amassed more data and the discovery was announced (Abe *et al.* 1995b, Abachi *et al.* 1995b). The 2010 experimental value for m_t is 173.1±1.3 GeV (Nakamura *et al.* 2010) as compared to the value predicted by fits to the electroweak data of 173.2±1.3 GeV. This represents an extraordinary triumph for both theory and experiment. It is surely remarkable how the quantum fluctuations of a yet-to-be-detected new particle could pin down its mass so precisely. It seems hard to deny that Nature has indeed made use of the subtle intricacies of a spontaneously broken non-Abelian gauge theory.

One feature of the 'real' top events in particularly noteworthy. Unlike the mass of the other quarks, m_t is greater than M_W, and this means that it can decay to b + W via *real* W emission:

$$t \to W^+ + b. \tag{22.201}$$

In contrast, the b quark itself decays by the usual *virtual* W processes. Now we have seen that the virtual process is supressed by $\sim 1/M_W^2$ if the energy release (as in the case of b-decay) is well below M_W. But the real process (22.201) suffers no such suppression and proceeds very much faster. In fact (problem 22.10) the top quark lifetime from (22.201) is estimated to be $\sim 4 \times 10^{-25}$ s! This is quite similar to the lifetime of the W^+ itself, via $W^+ \to e^+\nu_e$ for example. Consider now the production of a $t\bar{t}$ pair in the collision between two partons. As the t and $\bar{t}$ separate, the strong interactions which should eventually 'hadronize' them will not play a role until they are ~ 1 fm apart. But if they are travelling close to the speed of light, they can only travel some 10^{-16} m before decaying. Thus t's tend to decay before they experience the confining QCD interactions, a point we also made in section 1.2. Instead, the hadronization is associated with the b quark, which has a more typical weak lifetime ($\sim 1.5 \times 10^{-12}$ s). By the same token, this fast decay of the t quark means that there will be no detectable $t\bar{t}$ 'toponium', bound by QCD.

With the t quark safely real, the Higgs boson was the one remaining missing particle in the Standard Model complement, and its discovery was of the utmost importance. We end this book with a brief review of Higgs physics and the experiments leading to the probable discovery of this long-awaited particle in 2012.

22.8 The Higgs sector

> It is worth noting that an essential feature of the type of theory which has been described in this note is the prediction of incomplete multiplets of scalar and vector bosons.
>
> —P W Higgs (1964)

22.8.1 Introduction

The Lagrangian for an *unbroken* $\mathrm{SU(2)_L \times U(1)}$ gauge theory of vector bosons and fermions is rather simple and elegant, all the interactions being determined by just two Lagrangian parameters g and g' in a 'universal' way. All the particles in this hypothetical world are, however, massless. In the real world, while the electroweak interactions are undoubtedly well described by the $\mathrm{SU(2)_L \times U(1)}$ theory, neither the mediating gauge quanta (apart from the photon) nor the fermions are massless. They must acquire mass in some way that does not break the gauge symmetry of the Lagrangian, or else the renormalizability of the theory is destroyed, and its remarkable empirical success (at a level which includes loop corrections) would be some kind of freak accident. In chapter 19 we discussed how such a breaking of a gauge symmetry does happen, dynamically, in a superconductor. In that case 'electron pairing' was a crucial ingredient. In particle physics, while a lot of effort has gone into examining various analogous 'dynamical symmetry breaking' theories, none has yet emerged as both theoretically compelling and phenomenologically viable. However, a simple count of the number of degrees of freedom in a massive vector field, as opposed to a massless one, indicates that *some* additional fields must be present in order to give mass to the originally massless gauge bosons. And so, in the Standard Model, it is simply *assumed*, following the original ideas of Higgs and others (Higgs 1984, Englert and Brout 1964, Guralnik *et al.* 1964; Higgs 1966) that a suitable scalar ('Higgs') field exists, with a potential which causes the ground state (the vacuum) to break the symmetry spontaneously. Furthermore, rather than (as in BCS theory) obtaining the fermion mass gaps dynamically, they too are put in 'by hand' via Yukawa-like couplings to the Higgs field.

It has to be admitted that this part of the Standard Model appears to be the least satisfactory. Consider the Higgs couplings, which are listed in appendix Q, section Q.2.3. While the couplings of the Higgs field to the gauge fields are all determined by the gauge symmetry, the Higgs self-couplings (trilinear and quadrilinear) are not gauge interactions and are unrelated to anything else in the theory. Likewise, the Yukawa-like fermion couplings are not gauge interactions either, and they are both unconstrained and uncomfortably different in orders of magnitude. True, all these are renormalizable couplings – but this basically means that their values are not calculable and have all to be taken from experiment.

Such considerations may indicate that the 'Higgs Sector' of the Standard Model is on a somewhat different footing from the rest of it – a commonly held view, indeed. Perhaps it should be regarded as more a 'phenomenology' than a 'theory', much as the current–current model was. In this connection, we may mention a point which has long worried many theorists. In section 22.6 we noted that figure 22.26 gives a quadratically divergent ($O(\Lambda^2)$) and positive contribution to the $\hat{\phi}^\dagger\hat{\phi}$ term in the Lagrangian, at one loop order. This term would ordinarily, of course, be just the mass term of the scalar field. But in

the Higgs case, the matter is much more delicate. The whole phenomenology depends on the renormalized coefficient having a *negative* value, triggering the spontaneous breaking of the symmetry. This means that the $O(\Lambda^2)$ one-loop correction must be cancelled by the 'bare' mass term $\frac{1}{2}m_{\mathrm{H},0}^2\hat{\phi}^\dagger\hat{\phi}$ so as to achieve a negative coefficient of order $-v^2$. This cancellation between $m_{\mathrm{H},0}^2$ and Λ^2 will have to be very precise indeed if Λ – the scale of 'new physics' – is very high, as is commonly assumed (say 10^{16} GeV).

The reader may wonder why attention should *now* be drawn to this particular piece of renormalization: aren't all divergences handled this way? In a sense they are, but the fact is that this is the first case we have had in which we have to cancel a *quadratic* divergence. The other mass-corrections have all been logarithmic, for which there is nothing like such a dramatic 'fine-tuning' problem. There is a good reason for this in the case of the electron mass, which we remarked on in section 11.2. Chiral symmetry forces self-energy corrections for fermions to be proportional to their mass, and hence to contain only logarithms of the cut-off. Similarly, gauge invariance for the vector bosons prohibits any $O(\Lambda^2)$ connections in perturbation theory. But there is no symmetry, within the Standard Model, which 'protects' the coefficient of $\hat{\phi}^\dagger\hat{\phi}$ in this way. It is hard to understand what can be stopping it from being of order Λ^2, if we take the apparently reasonable point of view that the Standard Model will ultimately fail at some scale Λ where new physics enters. Thus the difficulty is: why is the empirical parameter v 'shielded' from the presumed high scale of new physics? This 'problem' is often referred to as the 'hierarchy problem', or the 'fine-tuning problem'. We stress again that we are dealing here with an absolutely crucial symmetry-breaking term, which one would really like to understand far better.

Of course, the problem would go away if the scale Λ were as low as, say a few TeV. As we shall see in the next section this happens to be, not accidentally, the same scale at which the Standard Model ceases to be a perturbatively calculable theory. Various possibilities have been suggested for the kind of physics that might enter at energies of a few TeV. For example, 'technicolour' models (Peskin 1997) regard the Higgs field as a composite of some new heavy fermions, rather like the BCS-pairing idea referred to earlier. A second possibility is supersymmetry (Aitchison 2007), in which there is a 'protective' symmetry operating, since scalar fields can be put alongside fermions in supermultiplets, and benefit from the protection enjoyed by the fermions. A third possibility is that of 'large' extra dimensions (Antoniadis 2002).

These undoubtedly fascinating ideas obviously take us well beyond our proper subject to which we must now return. Whatever may lie 'beyond' it, the Lagrangian of the Higgs sector of the Standard Model leads to many perfectly definite predictions which may be confronted with experiment, as we shall briefly discuss in section 22.8.3 (for a full account see Dawson *et al.* 1990, and for more compact ones see Ellis, Stirling and Webber 1996, chapter 11, and the review by Bernardi, Carena and Junk in Beringer *et al.* 2012). The

elucidation of the mechanism of gauge symmetry breaking is undoubtedly of the greatest importance to particle physics: quite apart from the $\mathrm{SU}(2)_\mathrm{L}\times\mathrm{U}(1)$ theory, very many of the proposed theories which go 'beyond the Standard Model' face a similar 'mass problem', and generally appeal to some variant of the 'Higgs mechanism' to deal with it.

As Higgs noted in the final paragraph of his 2-page Letter (Higgs 1964), an essential feature of the spontaneous symmetry breaking mechanism, in a gauge theory, is the appearance of incomplete multiplets of both scalar and vector bosons. Let us just rehearse this once more, in the $\mathrm{SU}(2)\times\mathrm{U}(1)$ case. We started with 4 massless gauge fields, belonging to an SU(2) triplet and a U(1) singlet; and, in addition, 4 scalar fields of equal mass, in an SU(2) doublet. After symmetry breaking, three massive vector bosons emerged, leaving the photon massless. In the scalar sector, three of the scalars became the longitudinal components of the three massive vector bosons, and one lone massive scalar field survived, all that remained of the original scalar doublet. Its mass is a free parameter of the theory, being given by $m_\mathrm{H}=\sqrt{2}\mu=\sqrt{\lambda}v/\sqrt{2}$. The discovery – or otherwise – of this *Higgs boson* has therefore been a vital goal in particle physics for over forty years. Before turning to experiment, however, we want to mention some theoretical considerations concerning m_H by way of orientation.

22.8.2 Theoretical considerations concerning m_H

The coupling constant λ, which determines m_H given the known value of v, is unfortunately undetermined in the Standard Model. However, some quite strong theoretical arguments suggest that m_H cannot be arbitrarily large.

Like all coupling constants in a renormalizable theory, λ must 'run'. For the $(\hat{\phi}^\dagger\hat{\phi})^2$ interaction of (22.30), a one-loop calculation of the β-function leads to

$$\lambda(E)=\frac{\lambda(v)}{1-\frac{3\lambda(v)}{8\pi^2}\ln(E/v)}. \tag{22.202}$$

Like QED, this theory is not asymptotically free: the coupling increases with the scale E. In fact, the theory becomes non-perturbative at the scale E^* such that

$$E^*\sim v\exp\left(\frac{8\pi^2}{3\lambda(v)}\right). \tag{22.203}$$

Note that this is exponentially sensitive to the 'low-energy' coupling constant $\lambda(v)$ – and that E^* decreases rapidly as $\lambda(v)$ increases. But (see (22.40)) m_H is essentially proportional to $\lambda^{1/2}(v)$. Hence as m_H increases, non-perturbative behaviour sets in increasingly early. Suppose we say that we should like perturbative behaviour to be maintained up to an energy scale Λ. Then we require

$$m_H<v\left[\frac{4\pi^2}{3\ln(\Lambda/v)}\right]^{1/2}. \tag{22.204}$$

For $\Lambda \sim 10^{16}$ GeV, this gives $m_{\rm H} < 160$ GeV. On the other hand, if the non-perturbative regime sets in at 1TeV, then the bound on $m_{\rm H}$ is weaker, $m_H < 750$ GeV.

This is an oversimplified argument for various reasons, though the essential point is correct. An important omission is the contribution of the top quark to the running of $\lambda(E)$. A more refined version (Hambye and Riesselmann 1997) concludes that for $m_{\rm H} < 180$ GeV the perturbative regime could extend all the way to the Planck mass, $\sim 10^{19}$ GeV.

There is another, independent, argument which suggests that $m_{\rm H}$ cannot be too large. We have previously considered violations of unitarity by the lowest-order diagrams for certain processes (see chapter 21 and section 22.6). As we saw, in a non-gauge theory with massive vector bosons, such violations are associated with the longitudinal polarization states of the bosons, which carry factors proportional to the 4-momentum k^μ (see (22.18)). In a gauge theory, strong cancellations in the high energy behaviour occur between different lowest-order diagrams. This behaviour is characteristic of gauge theories (Llewellyn Smith 1973, Cornwall *et al.* 1974), and is related to their renormalizability. One process of this sort which we did not yet consider, however, is that in which two longitudinally polarized W's scatter from each other. A considerable number of diagrams (7 in all) contribute to this process, in leading order : exchange of γ, Z and Higgs particles, together with the W–W self interaction. When all these are added up the high-energy behaviour of the total amplitude turns out to be proportional to λ, the Higgs coupling constant (see for example Ellis, Stirling and Webber 1996, chapter 8). This at first sight unexpected result can be understood as follows. The longitudinal components of the W's arise from the '$\partial^\mu\hat{\phi}$' parts in (22.30) (compare equation (19.48) in the U(1) case), which produce k^μ factors. Thus the scattering of longitudinal W's is effectively the scattering of the 3 Goldstone bosons in the complex Higgs doublet. These bosons have self interactions arising from the $\lambda(\hat{\phi}^\dagger\hat{\phi})^2$ Higgs potential, for which the Feynman amplitude is just proportional to λ. Now, although such a constant term obviously cannot violate unitarity as the energy increases (as happened in the other cases), it can do so if λ itself is too big – and since $\lambda \propto m_{\rm H}^2$, this puts a bound on $m_{\rm H}$. A constant amplitude is pure $J = 0$ and so, in order of magnitude, we expect unitarity to imply $\lambda < 1$. In terms of standard quantities,

$$\lambda = m_{\rm H}^2 G_{\rm F}/\sqrt{2}, \tag{22.205}$$

and so we expect

$$m_{\rm H} < G_{\rm F}^{-1/2} \sim 300 \text{ GeV}, \tag{22.206}$$

an energy scale we have seen several times before. A more refined analysis (Lee *et al.* 1977a,b) gives

$$m_{\rm H} < \left(\frac{8\sqrt{2}\pi}{3G_{\rm F}}\right)^{1/2} \approx 1\text{TeV}. \tag{22.207}$$

Like the preceding argument, this one does not say that $m_{\rm H}$ *must* be less than some fixed number. Rather, it says that if $m_{\rm H}$ gets bigger than a certain value, perturbation theory will fail, or 'new physics' will enter. It is, in fact, curiously reminiscent of the original situation with the four-fermion current–current interaction itself (compare (22.10) with (22.206)). Perhaps this is a clue that we may eventually need to replace the Higgs phenomenology. At all events, this line of reasoning seems to imply that the Higgs boson will either be found at a mass well below 1 TeV, or else some electroweak interactions will become effectively strong with new physical consequences. This 'no lose' situation provided powerful motivation for the construction of the LHC.

There is also an interesting *lower* bound on the Higgs mass, which is derived from the requirement of vacuum stability. If the Higgs mass is sufficiently lighter than the top quark mass, the top quark loop contribution to the running of the quartic coupling $\lambda(E)$ can cause the coupling to go negative at large energy scales (Cabibbo *et al.* 1979). This would imply that, at such scales, the effective scalar potential of the Standard Model would be unbounded below at large absolute values of the field, and there would no longer be a stable ground state (vacuum). This can be tolerated if the lifetime of the metastable vacuum is less that the age of the Universe (see Isidori *et al.* 2001, and references cited therein). A re-examination of the issue by Elias-Miro *et al.* (2012) showed that the Standard Model vacuum would become unstable at scales around the Planck mass, for $m_{\rm H} < 130$ GeV. For $m_{\rm H} \sim 125$ GeV, instability occurs at scales of order 10^{10} GeV, but the lifetime is greater than the age of the Universe. Of course, new physics may enter well before such a scale. It is nevertheless intriguing that a Higgs mass in this region may have implications for the physics of the early Universe.

We now consider some simple aspects of Higgs boson production and decay processes at collider energies, as predicted by the Standard Model, and conclude with the experiments leading to the probable Higgs boson discovery in 2012.

22.8.3 Higgs boson searches and the 2012 discovery

We begin by considering the main production and decay modes. The existing lower bound on $m_{\rm H}$ established at LEP (LEP 2003)

$$m_{\rm H} \geq 114.4 \text{ GeV (95\% Confidence Level)} \tag{22.208}$$

already excluded many possibilities in both production and decay. Subsequent searches were carried out at the hadron colliders. At both the Tevatron and the LHC, the dominant parton-level production mechanism is 'gluon fusion' via an intermediate top quark loop as shown in figure 22.27 (Georgi *et al.* 1978, Glashow *et al.* 1978, Stange *et al.* 1994a,b). The intermediate t quark dominates, because the Higgs couplings to fermions are proportional to the fermion mass. Since the gluon probability distribution rises rapidly at small x values, which are probed at larger collider energy $\sqrt{s}$, the cross section for

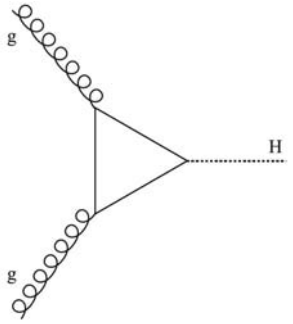


FIGURE 22.27
Higgs boson production process by 'gluon fusion'.

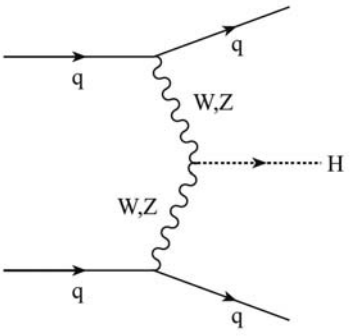


FIGURE 22.28
Higgs boson production process by 'vector boson fusion'.

this process (which is the same for pp and $p\bar{p}$ colliders) will rise with energy. At the Tevatron with $\sqrt{s} = 1.96$ TeV, the cross section ranges from about 1 pb for $m_H \simeq 100$ GeV to 0.2 pb for $m_H \simeq 200$ GeV. At an LHC energy of $\sqrt{s} = 7$ TeV, the cross section is about 25 pb for $m_H \simeq 100$ GeV and 0.1 pb for $m_H \simeq 700$ GeV, rising to about 70 pb and 1 pb respectively at $\sqrt{s} = 14$ TeV (Dittmaier *et al.* 2011). These numbers include QCD corrections, which increase the parton-level cross sections by a factor of about 2.

The next largest cross sections, some ten times smaller than the gluon fusion process, are for Higgs production via 'vector boson fusion' ($qq' \to qq'H$, see figure 22.28) and for associated production of a Higgs boson with a vector boson ($q\bar{q} \to WH, ZH$, see figure 22.29).

These processes involve the trilinear Higgs couplings to the vector bosons, which are proportional to their masses (see appendix Q). At the LHC, the first of these cross sections is somewhat larger than the second for $m_H < 130$ GeV,

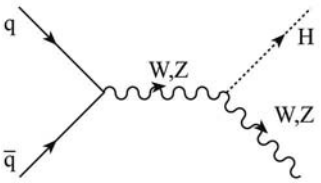


FIGURE 22.29
Higgs boson production in association with W or Z.

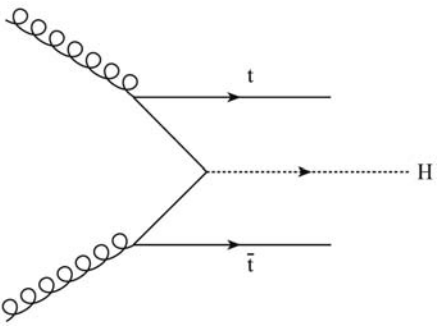


FIGURE 22.30
Higgs boson production in association with a t $\bar{\text{t}}$ pair.

while the order is reversed at the Tevatron because the initial state is p$\bar{\text{p}}$. A fourth production possibility, at a significantly smaller rate, is 'associated production with top quarks' as shown in figure 22.30, for example. Figure 22.31 (taken from Ellis, Stirling and Webber 1996) shows the cross sections for the various production processes as a function of m_{H}, for pp collisions at $\sqrt{s} = 14$ TeV. Updated calculations (including QCD and electroweak corrections) are described in reports by Dittmaier *et al.* (2011, 2012), which present the results of a very large-scale theoretical effort.

The Higgs boson must be detected via its decays. For $m_{\text{H}} < 135$ GeV, decays to fermion–antifermion pairs dominate, of which b$\bar{\text{b}}$ has the largest branching ratio because of the larger value of m_{b}; the decay to $\tau^+\tau^-$ is roughly an order of magnitude smaller. The width of $\text{H} \to f\bar{f}$ is easily calculated to lowest order and is (problem 22.11)

$$\Gamma(\text{H} \to f\bar{f}) = \frac{CG_{\text{F}} m_f^2 m_{\text{H}}}{4\pi\sqrt{2}} \left(1 - \frac{4m_f^2}{m_{\text{H}}^2}\right)^{3/2}, \tag{22.209}$$

where the colour factor C is 3 for quarks and 1 for leptons. For such m_{H} values, $\Gamma(\text{H} \to f\bar{f})$ is less than 5 MeV, and the total decay width is less than

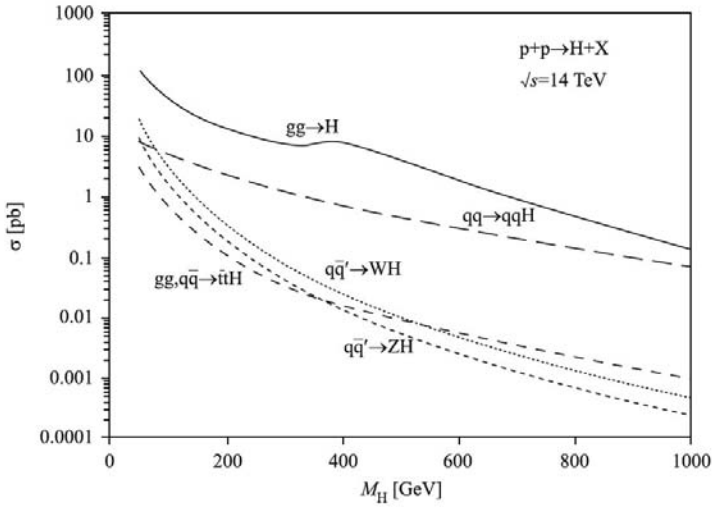


FIGURE 22.31
Higgs boson production cross sections in pp collisions at the LHC (figure from R K Ellis, W J Stirling and B R Webber *QCD and Collider Physics* 1996, courtesy Cambridge University Press).

10 MeV. QCD corrections are largely accounted for by replacing m_f^2 in the first factor on the right-hand side of (22.209), which arises from the Higgs-fermion Yukawa coupling, by the $\overline{\rm MS}$ running mass value $\overline{m}_f{}^2(m_{\rm H})$.

However, the large rate for the process $gg \to H \to b\bar{b}$ has to compete against a very large background from the inclusive production of pp (or $p\bar{p}$) $\to b\bar{b}+X$ via the strong interaction. The Higgs signal can be separated from such a background by using subleading decay modes such as $H \to \gamma\gamma$. The Higgs boson's coupling to photons is induced by quark triangle loops (figure 22.32) or a W loop. In a similar way, the associated production modes $W^{\pm}H$, ZH, allow use of the leptonic W and Z decays to reject QCD backgrounds.

Decays to a pair of vector bosons are also important. The tree-level width for $H \to W^+W^-$ is (problem 22.11)

$$\Gamma(H \to W^+W^-) = \frac{G_F m_H^3}{8\pi\sqrt{2}} \left(1 - \frac{4M_W^2}{m_H^2}\right)^{1/2} \left(1 - \frac{4M_W^2}{m_H^2} + 12\frac{M_W^4}{m_H^4}\right), \tag{22.210}$$

and the width for $H \to ZZ$ is the same with $M_W \to M_Z$ and a factor of $\frac{1}{2}$ to allow for the two identical bosons in the final state. These widths rise rapidly with m_H, reaching $\Gamma \sim 1$ GeV when $m_H \sim 200$ GeV. Even for values of m_H below the physical W^+W^- and ZZ thresholds, H can still decay through modes mediated by virtual bosons, via the off-shell decays $H \to WW^*$ and $H \to ZZ^*$.

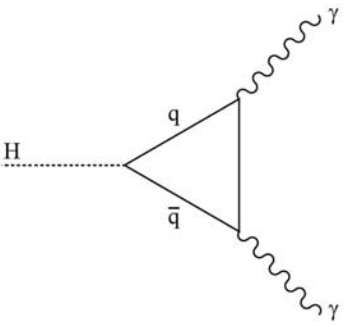


FIGURE 22.32
Higgs boson decay via quark triangle.

Figure 22.33, taken from Ellis, Stirling and Webber (1996), shows the complete set of phenomenologically relevant Higgs branching ratios for a Higgs boson with $m_{\mathrm{H}} < 200$ GeV. Updated results for SM Higgs branching ratios are reported in Dittmaier *et al.* (2012).

We turn now to the experiments. The Tevatron $\mathrm{p\bar{p}}$ collider at Fermilab operated at $\sqrt{s} = 1.96$ TeV until its shutdown in 2011. Higgs searches were conducted by two experiments, CDF and D0, which each collected approximately 10 fb^{-1} of data with the capability of seeing a Higgs signal in the mass range 90–185 GeV. The analyses searched for a Higgs boson produced through gluon fusion, in association with a vector boson, and through vector boson fusion. The decays $\mathrm{H} \to \mathrm{b\bar{b}}$, $\mathrm{H} \to \mathrm{W^+W^-}$, $\mathrm{H} \to \mathrm{ZZ}$, $\mathrm{H} \to \tau^+\tau^-$ and $\mathrm{H} \to \gamma\gamma$ were all studied.

The LHC is a pp collider at CERN which started running in 2010. The two general purpose detectors ATLAS ('A Toroidal LHC ApparatuS') and CMS ('Compact Muon Solenoid') were designed to study physics at the TeV scale, and in particular to search for the Higgs boson. In 2011, the LHC delivered to ATLAS and CMS up to 5.1 fb^{-1} of integrated luminosity of pp collisions at $\sqrt{s} = 7$ TeV. In 2012 the CMS energy was increased to 8 TeV, and by July 2012 up to 5.9 fb^{-1} of further data was delivered. At the LHC, the main Higgs boson production processes are the same as at the Tevatron, but as mentioned above vector boson fusion is more important than production in association with W or Z, or with $\mathrm{t\bar{t}}$. The LHC experiments are sensitive to Higgs bosons of much higher mass than the Tevatron experiments, ranging from the LEP bound (22.209) up to about 600 GeV. The same decay channels were studied as at the Tevatron.

By early 2012, the ATLAS and CMS experiments had excluded an m_{H} value in the interval 129 GeV to 539 GeV at the 95% CL, and the mass region 120–130 GeV was under intensive study, excesses of events having been

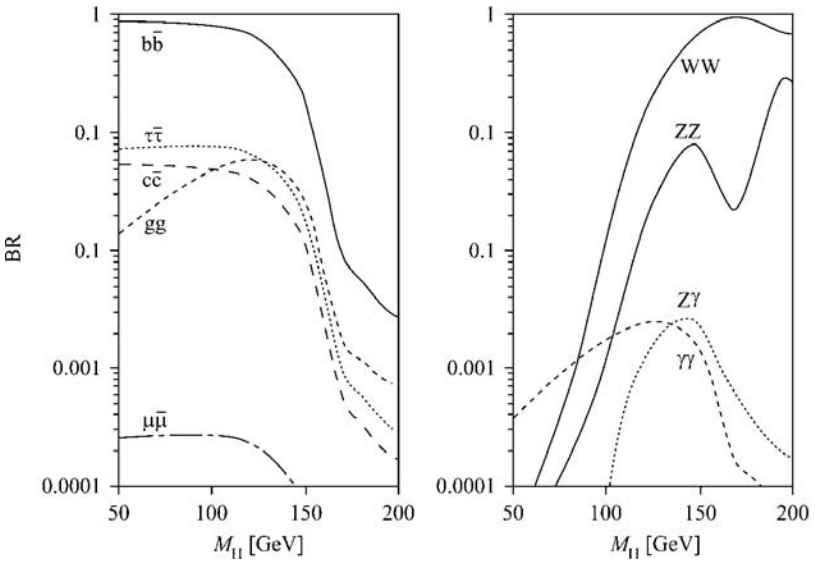


FIGURE 22.33
Branching ratios of the Higgs boson (figure from R K Ellis, W J Stirling and B R Webber *QCD and Collider Physics* 1996, courtesy Cambridge University Press).

reported by both experiments in the region 124–126 GeV (Aad *et al.* 2012a, Chatrchyan *et al.* 2012a). Then, on July 4, 2012, the ATLAS and CMS collaborations simultaneously announced the observation (at a significance greater than 5σ) of a new boson with a mass in the range 125–126 GeV and with properties compatible with those of a SM Higgs boson. These results (updated) are reported in Aad *et al.* (2012b) and Chatrchyan *et al.* (2012b). The crucial channels in the discovery were the decay modes $H \to \gamma\gamma$ and $H \to ZZ^* \to 4$ leptons, both of which provide a high-resolution invariant mass for fully reconstructed candidates. The cover illustration for Volume 1 of this book (copyright CERN) shows a candidate $\gamma\gamma$ event recorded by CMS, and that for volume 2 (copyright CERN) shows a candidate four muon event recorded by ATLAS. The channel $H \to WW^* \to \ell\nu\ell\nu$ is equally sensitive but has low resolution. The ATLAS result for the mass of the boson was (Aad *et al.* 2012b)

$$126.0 \pm 0.4(\text{stat}) \pm 0.4(\text{syst.}) \text{ GeV} \tag{22.211}$$

and the CMS result was (Chatrchyan *et al.* 2012b)

$$125.3 \pm 0.4(\text{stat}) \pm 0.5(\text{syst.}) \text{ GeV}. \tag{22.212}$$

At about the same time, the CDF and D0 collaborations at the Tevatron reported the combined results of their searches for a SM Higgs boson produced in association with a W or a Z boson, and subsequently decaying to a $b\bar{b}$ pair. The data corresponded to an integrated luminosity of 9.7 fb^{-1}. An excess of events was observed in the mass range 120–135 GeV, at a significance of 3σ, which was interpreted as evidence for a new particle, consistent with the SM Higgs boson (Aaltonen *et al.* 2012). This provided the first strong indication for the decay of the new particle to a fermion–antifermion pair at a rate consistent with the SM prediction.

Is the particle discovered by the ATLAS and CMS collaborations the Higgs boson of the Standard Model? The decay to two photons implies that its spin cannot be unity (Landau 1948, Yang 1950), but spin-0 has not yet been established. Even so, this already implies that the particle is different from all the other SM particles. The decay modes $\gamma\gamma$, ZZ^*, WW^* have been observed by ATLAS and CMS, and $b\bar{b}$ by CDF/D0. The $\tau^+\tau^-$ mode has not yet been seen. A measure of the compatibility of the observed boson with the SM Higgs boson is provided by the best-fit value of the common signal strength parameter μ defined by

$$\mu = \sigma \cdot \mathrm{BR}/(\sigma \cdot \mathrm{BR})_{\mathrm{SM}} \tag{22.213}$$

where σ is the boson production cross section and BR is the branching ratio of the boson to the observed final state. 'SM' denotes the SM prediction, so that the value $\mu = 1$ is the SM hypothesis. ATLAS reported a best-fit μ-value of $\mu = 1.4 \pm 0.3$ for $m_{\mathrm{H}} = 126$ GeV; the μ-values for the individual channels were all within one standard deviation (s.d.) of unity. CMS reported a best-fit value of $\mu = 0.87 \pm 0.23$ at $m_{\mathrm{H}} = 125.5$ GeV, and again the individual values in the observed channels were within 1 s.d. of unity. The conclusion is that these results are consistent, within uncertainties, with the predictions for the SM Higgs boson.

We end this book with a discovery which opens a new era in particle physics, in which the electroweak symmetry-breaking (Higgs) sector will be rigorously tested. The aim will be to measure the couplings of the new boson to the other SM particles with increasing accuracy, so as to reveal possible deviations from the SM values. The level of precision required to provide clear pointers to physics beyond the SM may be very high (see for example Gupta *et al.* 2012). The LHC will continue running until early 2013, when it will be shut down for machine improvements needed to allow operation at $\sqrt{s} = 14$ TeV and higher luminosity; beyond that, the High Luminosity LHC is planned to begin data-taking in 2022. However, just as the discovery of the W and Z bosons at the CERN $p\bar{p}$ collider was followed by precision studies at the e^+e^- colliders LEP and SLC, a lepton collider is likely to be needed on the next stage of this fundamental exploration.

Problems

22.1

(a) Using the representation for $\boldsymbol{\alpha}, \beta$ and γ_5 introduced in section 20.2.2 (equation (20.14)), massless particles are described by spinors of the form

$$u = E^{1/2} \begin{pmatrix} \phi_+ \\ \phi_- \end{pmatrix} \qquad \text{(normalized to } u^\dagger u = 2E\text{)}$$

where $\boldsymbol{\sigma} \cdot \hat{\boldsymbol{p}} \phi_\pm = \pm \phi_\pm, \hat{\boldsymbol{p}} = \boldsymbol{p}/|\boldsymbol{p}|$. Find the explicit form of u for the case $\hat{\boldsymbol{p}} = (\sin\theta, 0, \cos\theta)$.

(b) Consider the process $\bar{\nu}_\mu + \mu^- \to \bar{\nu}_\mathrm{e} + \mathrm{e}^-$, discussed in section 22.1, in the limit in which all masses are neglected. The amplitude is proportional to

$$G_\mathrm{F} \bar{v}(\bar{\nu}_\mu, \mathrm{R}) \gamma_\mu (1 - \gamma_5) u(\mu^-, \mathrm{L}) \bar{u}(\mathrm{e}^-, \mathrm{L}) \gamma^\mu (1 - \gamma_5) v(\bar{\nu}_\mathrm{e}, \mathrm{R})$$

where we have explicitly indicated the appropriate helicities R or L (note that, as explained in section 20.2.2, $(1 - \gamma_5)/2$ is the projection operator for a right-handed antineutrino). In the CM frame, let the initial μ^- momentum be $(0, 0, E)$ and the final e^- momentum be $E(\sin\theta, 0, \cos\theta)$. Verify that the amplitude is proportional to $G_\mathrm{F} E^2 (1 + \cos\theta)$. (*Hint:* evaluate the 'easy' part $\bar{v}(\bar{\nu}_\mu)\gamma_\mu(1-\gamma_5)u(\mu^-)$ first; this will show that the components $\mu = 0, z$ vanish, so that only the $\mu = x, y$ components of the dot product need to be calculated.)

22.2 Verify equation (22.20).

22.3 Check that when the polarization vector of each photon in figures 22.7(a) and (b) is replaced by the corresponding photon momentum, the sum of these two amplitudes vanishes.

22.4 By identifying the part of (22.45) which has the form (22.57), derive (22.58).

22.5 Using the vertex (22.48), verify (22.79).

22.6 Insert (22.29) into (22.151) to derive (22.153).

22.7 Verify that the neutral current part of (22.159) is diagonal in the 'mass' basis.

22.8 Suppose that the Higgs field is a triplet of $\mathrm{SU}(2)_\mathrm{L}$ rather than a doublet; and suppose that its vacuum value is

$$\langle 0 | \hat{\phi} | 0 \rangle = \begin{pmatrix} 0 \\ 0 \\ f \end{pmatrix}$$

in the gauge in which it is real. The non-vanishing component has $t_3 = -1$, using

$$t_3 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & -1 \end{pmatrix}$$

in the 'angular-momentum-like' basis. Since we want the charge of the vacuum to be zero, and we have $Q = t_3 + y/2$, we must assign $y(\hat{\phi}) = 2$. So the covariant derivative on $\hat{\phi}$ is

$$(\partial_\mu + \mathrm{i}g\boldsymbol{t} \cdot \hat{\boldsymbol{W}}_\mu + \mathrm{i}g'\hat{B}_\mu)\hat{\phi},$$

where

$$t_1 = \begin{pmatrix} 0 & \frac{1}{\sqrt{2}} & 0 \\ \frac{1}{\sqrt{2}} & 0 & \frac{1}{\sqrt{2}} \\ 0 & \frac{1}{\sqrt{2}} & 0 \end{pmatrix}, \quad t_2 = \begin{pmatrix} 0 & \frac{-\mathrm{i}}{\sqrt{2}} & 0 \\ \frac{\mathrm{i}}{\sqrt{2}} & 0 & \frac{-\mathrm{i}}{\sqrt{2}} \\ 0 & \frac{\mathrm{i}}{\sqrt{2}} & 0 \end{pmatrix}$$

and t_3 is as above (it is easy to check that these three matrices do satisfy the required SU(2) commutation relations $[t_1, t_2] = \mathrm{i}t_3$). Show that the photon and Z fields are still given by (22.36) and (22.37), with the same $\sin\theta_\mathrm{W}$ as in (22.39), but that now

$$M_\mathrm{Z} = \sqrt{2}M_\mathrm{W}/\cos\theta_\mathrm{W}.$$

What is the value of the parameter ρ in this model?

22.9 Use (22.188) to verify (22.190).

22.10 Calculate the lifetime of the top quark to decay via $\mathrm{t} \to \mathrm{W}^+ + \mathrm{b}$.

22.12 Using the Higgs couplings given in appendix Q, verify (22.209) and (22.210).


Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

M

Group Theory

M.1 Definition and simple examples

A group $\mathcal{G}$ is a set of elements $(a, b, c, \ldots)$ with a law for combining any two elements a, b so as to form their ordered 'product' ab, such that the following four conditions hold:

(i) For every $a, b \in \mathcal{G}$, the product $ab \in \mathcal{G}$ (the symbol '$\in$' means 'belongs to', or 'is a member of').

(ii) The law of combination is associative, i.e.

$$(ab)c = a(bc). \tag{M.1}$$

(iii) $\mathcal{G}$ contains a unique identity element, e, such that for all $a \in \mathcal{G}$,

$$ae = ea = a. \tag{M.2}$$

(iv) For all $a \in \mathcal{G}$, there is a unique inverse element, a^{-1}, such that

$$aa^{-1} = a^{-1}a = e. \tag{M.3}$$

Note that in general the law of combination is not commutative – i.e. $ab \neq ba$; if it is commutative, the group is *Abelian*; if not, it is *non-Abelian*. Any finite set of elements satisfying the conditions (i)–(iv) forms a finite group, the *order* of the group being equal to the number of elements in the set. If the set does not have a finite number of elements it is an infinite group.

As a simple example, the set of four numbers (1, i, –1, –i) form a finite Abelian group of order 4, with the law of combination being ordinary multiplication. The reader may check that each of (i)–(iv) is satisfied, with e taken to be the number 1, and the inverse being the algebraic reciprocal. A second group of order 4 is provided by the matrices

$$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}, \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix}, \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}, \tag{M.4}$$

with the combination law being matrix multiplication, 'e' being the first (unit) matrix, and the inverse being the usual matrix inverse. Although matrix

multiplication is not commutative in general, it happens to be so for these particular matrices. In fact, the way these four matrices multiply together is (as the reader can verify) exactly the same as the way the four numbers (1, i, −1, −i) (in that order) do. Further, the correspondence between the elements of the two groups is 'one to one': that is, if we label the two sets of group elements by (e, a, b, c) and (e', a', b', c'), we have the correspondences $e \leftrightarrow e'$, $a \leftrightarrow a'$, $b \leftrightarrow b'$, $c \leftrightarrow c'$. Two groups with the same multiplication structure, and with a one-to-one correspondence between their elements, are said to be *isomorphic*. If they have the same multiplication structure but the correspondence is not one-to-one, they are *homomorphic*.

M.2 Lie groups

We are interested in *continuous groups* – that is, groups whose elements are labelled by a number of continuously variable real parameters $\alpha_1, \alpha_2, \ldots, \alpha_r$: $g(\alpha_1, \alpha_2, \ldots, \alpha_r) \equiv g(\boldsymbol{\alpha})$. In particular, we are concerned with various kinds of 'coordinate transformations' (not necessarily space-time ones, but including also 'internal' transformations such as those of SU(3)). For example, rotations in three dimensions form a group, whose elements are specified by three real parameters (e.g. two for defining the axis of the rotation, and one for the angle of rotation about that axis). Lorentz transformations also form a group, this time with six real parameters (three for 3-D rotations, three for pure velocity transformations). The matrices of SU(3) are specified by the values of eight real parameters. By convention, parametrizations are arranged in such a way that $g(\mathbf{0})$ is the identity element of the group. For a continuous group, condition (i) takes the form

$$g(\boldsymbol{\alpha})g(\boldsymbol{\beta}) = g(\boldsymbol{\gamma}(\boldsymbol{\alpha}, \boldsymbol{\beta})), \tag{M.5}$$

where the parameters $\boldsymbol{\gamma}$ are continuous functions of the parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. A more restrictive condition is that $\boldsymbol{\gamma}$ should be an *analytic* function of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$; if this is the case, the group is a *Lie group*.

The analyticity condition implies that if we are given the form of the group elements in the neighbourhood of any one element, we can 'move out' from that neighbourhood to other nearby elements, using the mathematical procedure known as 'analytic continuation' (essentially, using a power series expansion); by repeating the process, we should be able to reach all group elements which are 'continuously connected' to the original element. The simplest group element to consider is the identity, which we shall now denote by I. Lie proved that the properties of the elements of a Lie group which can be reached continuously from the identity I are determined from elements lying in the neighbourhood of I.

M.3 Generators of Lie groups

Consider (following Lichtenberg 1970, chapter 5) a group of transformations defined by

$$x_i' = f_i(x_1, x_2, \ldots, x_N; \alpha_1, \alpha_2, \ldots, \alpha_r), \tag{M.6}$$

where the x_i's ($i = 1, 2, \ldots, N$) are the 'coordinates' on which the transformations act, and the α's are the (real) parameters of the transformations. By convention, $\boldsymbol{\alpha} = \mathbf{0}$ is the identity transformation, so

$$x_i = f_i(\boldsymbol{x}, \mathbf{0}). \tag{M.7}$$

A transformation in the neighbourhood of the identity is then given by

$$\mathrm{d}x_i = \sum_{\nu=1}^{r} \frac{\partial f_i}{\partial \alpha_\nu} \mathrm{d}\alpha_\nu, \tag{M.8}$$

where the $\{\mathrm{d}\alpha_\nu\}$ are infinitesimal parameters, and the partial derivative is understood to be evaluated at the point $(\boldsymbol{x}, \mathbf{0})$.

Consider now the change in a function $F(\boldsymbol{x})$ under the infinitesimal transformation (M.8). We have

$$\begin{aligned} F \to F + \mathrm{d}F &= F + \sum_{i=1}^{N} \frac{\partial F}{\partial x_i} \mathrm{d}x_i \\ &= F + \sum_{i=1}^{N} \left[\sum_{\nu=1}^{r} \frac{\partial f_i}{\partial \alpha_\nu} \mathrm{d}\alpha_\nu \right] \frac{\partial F}{\partial x_i} \\ &\equiv \{1 - \sum_{\nu=1}^{r} \mathrm{d}\alpha_\nu \mathrm{i}\hat{X}_\nu\} F, \end{aligned} \tag{M.9}$$

where

$$\hat{X}_\nu \equiv \mathrm{i} \sum_{i=1}^{N} \frac{\partial f_i}{\partial \alpha_\nu} \frac{\partial}{\partial x_i} \tag{M.10}$$

is a *generator of infinitesimal transformations*[1]. Note that in (M.10) ν runs from 1 to r, so there are as many generators as there are parameters labelling the group elements. Finite transformations are obtained by 'exponentiating' the quantity in braces in (M.9) (compare (12.30)):

$$\hat{U}(\boldsymbol{\alpha}) = \exp\{-\mathrm{i}\boldsymbol{\alpha} \cdot \hat{\boldsymbol{X}}\}, \tag{M.11}$$

[1] Clearly there is lot of 'convention' (the sign, the i) in the definition of $\hat{X}_\nu$. It is chosen for convenient consistency with familiar generators, for example those of SO(3) (see section M.4.1).

where we have written $\sum_{\nu=1}^{r} \alpha_\nu \hat{X}_\nu = \boldsymbol{\alpha} \cdot \hat{\boldsymbol{X}}$.

An important theorem states that the commutator of any two generators of a Lie group is a linear combination of the generators:

$$[\hat{X}_\lambda, \hat{X}_\mu] = c^\nu_{\lambda\mu} \hat{X}_\nu, \tag{M.12}$$

where the constants $c^\nu_{\lambda\mu}$ are complex numbers called the *structure constants* of the group; a sum over ν from 1 to r is understood on the right-hand side. The commutation relations (M.12) are called the *algebra* of the group.

M.4 Examples

M.4.1 SO(3) and three-dimensional rotations

Rotations in three dimensions are defined by

$$\boldsymbol{x}' = R\,\boldsymbol{x}, \tag{M.13}$$

where R is a real 3×3 matrix such that the length of $\boldsymbol{x}$ is preserved, i.e. $\boldsymbol{x}'^{\mathrm{T}}\boldsymbol{x}' = \boldsymbol{x}^{\mathrm{T}}\boldsymbol{x}$. This implies that $R^{\mathrm{T}}R = I$, so that R is an orthogonal matrix. It follows that

$$1 = \det(R^{\mathrm{T}}R) = \det R^{\mathrm{T}} \det R = (\det R)^2, \tag{M.14}$$

and so $\det R = \pm 1$. Those R's with $\det R = -1$ include a parity transformation ($\boldsymbol{x}' = -\boldsymbol{x}$), which is not continuously connected to the identity. Those with $\det R = 1$ are 'proper rotations', and they form the elements of the group SO(3): the *S*pecial *O*rthogonal group in *3* dimensions.

An R close to the identity matrix I can be written as $R = I + \delta R$ where

$$(I + \delta R)^{\mathrm{T}}(I + \delta R) = I. \tag{M.15}$$

Expanding this out to first order in δR gives

$$\delta R^{\mathrm{T}} = -\delta R, \tag{M.16}$$

so that δR is an antisymmetric 3×3 matrix (compare (12.19)). We may parametrize δR as

$$\delta R = \begin{pmatrix} 0 & \epsilon_3 & -\epsilon_2 \\ -\epsilon_3 & 0 & \epsilon_1 \\ \epsilon_2 & -\epsilon_1 & 0 \end{pmatrix}, \tag{M.17}$$

and an infinitesimal rotation is then given by

$$\boldsymbol{x}' = \boldsymbol{x} - \boldsymbol{\epsilon} \times \boldsymbol{x}, \tag{M.18}$$

(compare (12.64)), or

$$\mathrm{d}x_1 = -\epsilon_2 x_3 + \epsilon_3 x_2, \quad \mathrm{d}x_2 = -\epsilon_3 x_1 + \epsilon_1 x_3, \quad \mathrm{d}x_3 = -\epsilon_1 x_2 + \epsilon_2 x_1. \tag{M.19}$$

Thus in (M.8), identifying $\mathrm{d}\alpha_1 \equiv \epsilon_1$, $\mathrm{d}\alpha_2 \equiv \epsilon_2$, $\mathrm{d}\alpha_3 \equiv \epsilon_3$, we have

$$\frac{\partial f_1}{\partial \alpha_1} = 0, \quad \frac{\partial f_1}{\partial \alpha_2} = -x_3, \quad \frac{\partial f_1}{\partial \alpha_3} = x_2, \quad \text{etc.} \tag{M.20}$$

The generators (M.10) are then

$$\left.\begin{array}{rcll} \hat{X}_1 & = & \mathrm{i}x_3\frac{\partial}{\partial x_2} & - \mathrm{i}x_2\frac{\partial}{\partial x_3} \\ \hat{X}_2 & = & \mathrm{i}x_1\frac{\partial}{\partial x_3} & - \mathrm{i}x_3\frac{\partial}{\partial x_1} \\ \hat{X}_3 & = & \mathrm{i}x_2\frac{\partial}{\partial x_1} & - \mathrm{i}x_1\frac{\partial}{\partial x_2} \end{array}\right\} \tag{M.21}$$

which are easily recognized as the quantum-mechanical angular momentum operators

$$\hat{\boldsymbol{X}} = \boldsymbol{x} \times -\mathrm{i}\boldsymbol{\nabla}, \tag{M.22}$$

which satisfy the *SO(3) algebra*

$$[\hat{X}_i, \hat{X}_j] = \mathrm{i}\epsilon_{ijk}\hat{X}_k. \tag{M.23}$$

The action of finite rotations, parametrized by $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \alpha_3)$, on functions F is given by

$$\hat{U}(\boldsymbol{\alpha}) = \exp\{-\mathrm{i}\boldsymbol{\alpha} \cdot \hat{\boldsymbol{X}}\}. \tag{M.24}$$

The operators $\hat{U}(\boldsymbol{\alpha})$ form a group which is isomorphic to SO(3). The structure constants of SO(3) are $\mathrm{i}\epsilon_{ijk}$, from (M.23).

M.4.2 SU(2)

We write the infinitesimal SU(2) transformation (acting on a general complex two-component column vector) as (cf (12.27))

$$\begin{pmatrix} q_1' \\ q_2' \end{pmatrix} = (1 + \mathrm{i}\boldsymbol{\epsilon} \cdot \boldsymbol{\tau}/2) \begin{pmatrix} q_1 \\ q_2 \end{pmatrix}, \tag{M.25}$$

so that

$$\begin{aligned} \mathrm{d}q_1 &= \frac{\mathrm{i}\epsilon_3}{2} q_1 + \left(\frac{\mathrm{i}\epsilon_1}{2} + \frac{\epsilon_2}{2}\right) q_2 \\ \mathrm{d}q_2 &= \frac{-\mathrm{i}\epsilon_3}{2} q_2 + \left(\frac{\mathrm{i}\epsilon_1}{2} - \frac{\epsilon_2}{2}\right) q_1. \end{aligned} \tag{M.26}$$

Then (with $\mathrm{d}\alpha_1 \equiv \epsilon_1$ etc.)

$$\frac{\partial f_1}{\partial \alpha_1} = \frac{\mathrm{i}q_2}{2}, \quad \frac{\partial f_1}{\partial \alpha_2} = \frac{q_2}{2}, \quad \frac{\partial f_1}{\partial \alpha_3} = \frac{\mathrm{i}q_1}{2}, \tag{M.27}$$

$$\frac{\partial f_2}{\partial \alpha_1} = \frac{\mathrm{i}q_1}{2},\ \frac{\partial f_2}{\partial \alpha_2} = -\frac{q_2}{2},\ \frac{\partial f_2}{\partial \alpha_3} = -\frac{\mathrm{i}q_2}{2}, \tag{M.28}$$

and (from (M.10))

$$\hat{X}_1' = -\frac{1}{2}\left\{q_2\frac{\partial}{\partial q_1} + q_1\frac{\partial}{\partial q_2}\right\} \tag{M.29}$$

$$\hat{X}_2' = \frac{\mathrm{i}}{2}\left\{q_2\frac{\partial}{\partial q_1} - q_1\frac{\partial}{\partial q_2}\right\} \tag{M.30}$$

$$\hat{X}_3' = \frac{1}{2}\left\{-q_1\frac{\partial}{\partial q_1} + q_2\frac{\partial}{\partial q_2}\right\}. \tag{M.31}$$

It is an interesting exercise to check that the commutation relations of the $\hat{X}_i'$'s are exactly the same as those of the $\hat{X}_i$'s in (M.23). The two groups are therefore said to have the same algebra, with the same structure constants, and they are in fact isomorphic in the vicinity of their respective identity elements. They are not the same for 'large' transformations, however, as we discuss in section M.7.

M.4.3 SO(4): The special orthogonal group in four dimensions

This is the group whose elements are 4×4 matrices S such that $S^{\mathrm{T}}S = I$, where I is the 4×4 unit matrix, with the condition $\det S = +1$. The Euclidean (length)^2 $x_1^2 + x_2^2 + x_3^2 + x_4^2$ is left invariant under SO(4) transformations. Infinitesimal SO(4) transformations are characterized by the 4-D analogue of those for SO(3), namely by 4×4 real antisymmetric matrices δS, which have 6 real parameters. We choose to parametrize δS in such a way that the Euclidean 4-vector $(\boldsymbol{x}, x_4)$ is transformed to (cf (18.76) and (18.77))

$$\begin{aligned} \boldsymbol{x}' &= \boldsymbol{x} - \boldsymbol{\epsilon} \times \boldsymbol{x} - \boldsymbol{\eta} x_4, \\ x_4' &= x_4 + \boldsymbol{\eta} \cdot \boldsymbol{x}, \end{aligned} \tag{M.32}$$

where $\boldsymbol{x} = (x_1, x_2, x_3)$ and $\boldsymbol{\eta} = (\eta_1, \eta_2, \eta_3)$. Note that the first three components transform by (M.18) when $\boldsymbol{\eta} = 0$, so that SO(3) is a *subgroup* of SO(4). The six generators are (with $\mathrm{d}\alpha_1 \equiv \epsilon_1$ etc.)

$$\hat{X}_1 = \mathrm{i}x_3\frac{\partial}{\partial x_2} - \mathrm{i}x_2\frac{\partial}{\partial, x_3}, \tag{M.33}$$

and similarly for $\hat{X}_2$ and $\hat{X}_3$ as in (M.21), together with (defining $\mathrm{d}\alpha_4 = \eta_1$ etc.)

$$\hat{X}_4 = \mathrm{i}\left(-x_4\frac{\partial}{\partial x_1} + x_1\frac{\partial}{\partial x_4}\right) \tag{M.34}$$

$$\hat{X}_5 = \mathrm{i}\left(-x_4\frac{\partial}{\partial x_2} + x_2\frac{\partial}{\partial x_4}\right) \tag{M.35}$$

$$\hat{X}_6 = \mathrm{i}\left(-x_4\frac{\partial}{\partial x_3} + x_3\frac{\partial}{\partial x_4}\right). \tag{M.36}$$

Relabelling these last three generators as $\hat{Y}_1 \equiv \hat{X}_4,\ \hat{Y}_2 \equiv \hat{X}_5,\ \hat{Y}_3 \equiv \hat{X}_6$, we find the following algebra:

$$[\hat{X}_i, \hat{X}_j] = \mathrm{i}\epsilon_{ijk}\hat{X}_k \tag{M.37}$$

$$[\hat{X}_i, \hat{Y}_j] = \mathrm{i}\epsilon_{ijk}\hat{Y}_k \tag{M.38}$$

$$[\hat{Y}_i, \hat{Y}_j] = \mathrm{i}\epsilon_{ijk}\hat{X}_k, \tag{M.39}$$

together with

$$[\hat{X}_1, \hat{Y}_1] = [\hat{X}_2, \hat{Y}_2] = [\hat{X}_3, \hat{Y}_3] = 0. \tag{M.40}$$

(M.37) confirms that the three generators controlling infinitesimal transformations among the first three components $\boldsymbol{x}$ obey the angular momentum commutation relations. (M.37)–(M.40) constitute the algebra of SO(4).

This algebra may be simplified by introducing the linear combinations

$$\hat{M}_i = \frac{1}{2}(\hat{X}_i + \hat{Y}_i) \tag{M.41}$$

$$\hat{N}_i = \frac{1}{2}(\hat{X}_i - \hat{Y}_i), \tag{M.42}$$

which satisfy

$$[\hat{M}_i, \hat{M}_j] = \mathrm{i}\epsilon_{ijk}\hat{M}_k \tag{M.43}$$

$$[\hat{N}_i, \hat{N}_j] = \mathrm{i}\epsilon_{ijk}\hat{N}_k \tag{M.44}$$

$$[\hat{M}_i, \hat{N}_j] = 0. \tag{M.45}$$

From (M.43)–(M.45) we see that, in this form, the six generators have separated into two sets of three, each set obeying the algebra of SO(3) (or of SU(2)), and commuting with the other set. They therefore behave like two *independent* angular momentum operators. The algebra (M.43)–(M.45) is referred to as SU(2) × SU(2).

M.4.4 The Lorentz group

In this case the quadratic form left invariant by the transformation is the Minkowskian one $(x^0)^2 - \boldsymbol{x}^2$ (see appendix D of volume 1). We may think of infinitesimal Lorentz transformations as corresponding physically to ordinary infinitesimal 3-D rotations, together with infinitesimal pure velocity transformations ('boosts'). The basic 4-vector then transforms by

$$\left.\begin{aligned} x^{0\prime} &= x^0 - \boldsymbol{\eta}\cdot\boldsymbol{x} \\ \boldsymbol{x}' &= \boldsymbol{x} - \boldsymbol{\epsilon}\times\boldsymbol{x} - \boldsymbol{\eta}x^0 \end{aligned}\right\} \tag{M.46}$$

where $\boldsymbol{\eta}$ is now the infinitesimal velocity parameter (the reader may check that $(x^0)^2 - \boldsymbol{x}^2$ is indeed left invariant by (M.46), to first order in $\boldsymbol{\epsilon}$ and $\boldsymbol{\eta}$). The six generators are then $\hat{X}_1, \hat{X}_2, \hat{X}_3$ as in (M.21), together with

$$\hat{K}_1 = -\mathrm{i}\left(x^1\frac{\partial}{\partial x^0} + x^0\frac{\partial}{\partial x^1}\right) \tag{M.47}$$

$$\hat{K}_2 = -\mathrm{i}\left(x^2\frac{\partial}{\partial x^0} + x^0\frac{\partial}{\partial x^2}\right) \tag{M.48}$$

$$\hat{K}_3 = -\mathrm{i}\left(x^3\frac{\partial}{\partial x^0} + x^0\frac{\partial}{\partial x^3}\right). \tag{M.49}$$

The corresponding algebra is

$$[\hat{X}_i, \hat{X}_j] = \mathrm{i}\epsilon_{ijk}\hat{X}_k \tag{M.50}$$

$$[\hat{X}_i, \hat{K}_j] = \mathrm{i}\epsilon_{ijk}\hat{K}_k \tag{M.51}$$

$$[\hat{K}_i, \hat{K}_j] = -\mathrm{i}\epsilon_{ijk}\hat{X}_k. \tag{M.52}$$

Note the minus sign on the right-hand side of (M.52) as compared with (M.39).

M.4.5 SU(3)

A general infinitesimal SU(3) transformation may be written as (cf (12.71) and (12.72))

$$\begin{pmatrix} q_1 \\ q_2 \\ q_3 \end{pmatrix}' = \left(1 + \mathrm{i}\frac{1}{2}\boldsymbol{\eta}\cdot\boldsymbol{\lambda}\right)\begin{pmatrix} q_1 \\ q_2 \\ q_3 \end{pmatrix}, \tag{M.53}$$

where there are now 8 of these $\boldsymbol{\eta}$'s, $\boldsymbol{\eta} = (\eta_1, \eta_2, \ldots, \eta_8)$, and the λ-matrices are the Gell-Mann matrices

$$\lambda_1 = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix},\ \lambda_2 = \begin{pmatrix} 0 & -\mathrm{i} & 0 \\ \mathrm{i} & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix},\ \lambda_3 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 0 \end{pmatrix} \tag{M.54}$$

$$\lambda_4 = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix},\ \lambda_5 = \begin{pmatrix} 0 & 0 & -\mathrm{i} \\ 0 & 0 & 0 \\ \mathrm{i} & 0 & 0 \end{pmatrix},\ \lambda_6 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix} \tag{M.55}$$

$$\lambda_7 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & -\mathrm{i} \\ 0 & \mathrm{i} & 0 \end{pmatrix},\ \lambda_8 = \begin{pmatrix} \frac{1}{\sqrt{3}} & 0 & 0 \\ 0 & \frac{1}{\sqrt{3}} & 0 \\ 0 & 0 & -\frac{2}{\sqrt{3}} \end{pmatrix}. \tag{M.56}$$

In this parametrization the first three of the eight generators $\hat{G}_r$ $(r = 1, 2, \ldots, 8)$ are the same as $\hat{X}_1', \hat{X}_2', \hat{X}_3'$ of (M.29)–(M.30). The others may be constructed as usual from (M.10); for example,

$$\hat{G}_5 = \frac{\mathrm{i}}{2}\left(q_3\frac{\partial}{\partial q_1} - q_1\frac{\partial}{\partial q_3}\right),\ \hat{G}_7 = \frac{\mathrm{i}}{2}\left(q_3\frac{\partial}{\partial q_2} - q_2\frac{\partial}{\partial q_3}\right). \tag{M.57}$$

The SU(3) algebra is found to be

$$[\hat{G}_a, \hat{G}_b] = \mathrm{i}f_{abc}\hat{G}_c, \tag{M.58}$$

where a, b and c each run from 1 to 8. The structure constants are $\mathrm{i}f_{abc}$, and the non-vanishing $f's$ are as follows:

$$f_{123} = 1,\ f_{147} = 1/2,\ f_{156} = -1/2,\ f_{246} = 1/2,\ f_{257} = 1/2 \tag{M.59}$$

$$f_{345} = 1/2,\ f_{367} = -1/2,\ f_{458} = \sqrt{3}/2,\ f_{678} = \sqrt{3}/2. \tag{M.60}$$

Note that the f's are antisymmetric in all pairs of indices (Carruthers (1966) chapter 2).

M.5 Matrix representations of generators, and of Lie groups

We have shown how the generators $\hat{X}_1, \hat{X}_2, \ldots, \hat{X}_r$ of a Lie group can be constructed as differential operators, understood to be acting on functions of the 'coordinates' to which the transformations of the group refer. These generators satisfy certain commutation relations, the Lie algebra of the group. For any given Lie algebra, it is also possible to find sets of *matrices* $X_1, X_2, \ldots, X_r$ (without hats) which satisfy the same commutation relations as the $\hat{X}_\nu$'s – that is, they have the same algebra. Such matrices are said to form a (matrix) representation of the Lie algebra, or equivalently of the generators. The idea is familiar from the study of angular momentum in quantum mechanics (Schiff 1968, section 27), where the entire theory may be developed from the commutation relations (with $\hbar = 1$)

$$[\hat{J}_i, \hat{J}_j] = \mathrm{i}\epsilon_{ijk}\hat{J}_k \tag{M.61}$$

for the angular momentum operators $\hat{J}_i$, together with the physical requirement that the $\hat{J}_i$'s (and the matrices representing them) must be Hermitian. In this case the matrices are of the form (in quantum-mechanical notation)

$$\left(J_i^{(J)}\right)_{M'_J M_J} \equiv \langle JM'_J|\hat{J}_i|JM_J\rangle, \tag{M.62}$$

where $|JM_J\rangle$ is an eigenstate of $\hat{\boldsymbol{J}}^2$ and of $\hat{J}_3$ with eigenvalues $J(J+1)$ and M_J respectively. Since M_J and M'_J each run over the $2J+1$ values defined by $-J \le M_J,\ M'_J \le J$, the matrices $J_i^{(J)}$ are of dimension $(2J+1) \times (2J+1)$. Clearly, since the generators of SU(2) have the same algebra as (M.61), an identical matrix representation may be obtained for them; these matrices were denoted by $T_i^{(T)}$ in section 12.1.2. It is important to note that J (or T) can take an infinite sequence of values $J = 0, 1/2, 1, 3/2, \ldots$, corresponding physically to various 'spin' magnitudes. Thus there are infinitely many sets of three matrices $(J_1^{(J)}, J_2^{(J)}, J_3^{(J)})$ all with the same commutation relations as (M.61).

A similar method for obtaining matrix representations of Lie algebras may be followed in other cases. In physical terms, the problem amounts to finding a correct labelling of the base states, analogous to $|JM\rangle$. In the latter case, the quantum number J specifies each different representation. The reason it does so is because (as should be familiar) the corresponding operator $\hat{\boldsymbol{J}}^2$ commutes with every generator:

$$[\hat{\boldsymbol{J}}^2, \hat{J}_i] = 0. \tag{M.63}$$

Such an operator is called a *Casimir operator*, and by a lemma due to Schur (Hammermesh 1962, pages 100–101) it must be a multiple of the unit operator. The numerical value it has is different for each different representation, and may therefore be used to characterize a representation (namely as '$J = 0$', '$J = 1/2$', etc.).

In general, more than one such operator is needed to characterize a representation completely. For example, in SO(4), the two operators $\hat{\boldsymbol{M}}^2$ and $\hat{\boldsymbol{N}}^2$ commute with all the generators, and take values $M(M+1)$ and $N(N+1)$ respectively, where $M, N = 0, 1/2, 1, \ldots$. Thus the labelling of the matrix elements of the generators is the same as it would be for two independent particles, one of spin M and the other of spin N. For given M, N the matrices are of dimension $[(2M+1)+(2N+1)] \times [(2M+1)+(2N+1)]$. The number of Casimir operators required to characterize a representation is called the *rank* of the group (or the algebra). This is also equal to the number of independent mutually commuting generators (though this is by no means obvious). Thus SO(4) is a rank two group, with two commuting generators $\hat{M}_3$ and $\hat{N}_3$; so is SU(3), since $\hat{G}_3$ and $\hat{G}_8$ commute. Two Casimir operators are therefore required to characterize the representations of SU(3), which may be taken to be the 'quadratic' one

$$\hat{C}_2 \equiv \hat{G}_1^2 + \hat{G}_2^2 + \ldots + \hat{G}_8^2, \tag{M.64}$$

together with a 'cubic' one

$$\hat{C}_3 \equiv d_{abc} \hat{G}_a \hat{G}_b \hat{G}_c, \tag{M.65}$$

where the coefficients d_{abc} are defined by the relation

$$\{\lambda_a, \lambda_b\} = \frac{4}{3}\delta_{ab} I + 2 d_{abc} \lambda_c, \tag{M.66}$$

and are symmetric in all pairs of indices (they are tabulated in Carruthers 1966, table 2.1). In practice, for the few SU(3) representations that are actually required, it is more common to denote them (as we have in the text) by their dimensionality, which for the cases **1** (singlet), **3** (triplet), $\mathbf{3}^*$ (antitriplet), **8** (octet) and **10** (decuplet) is in fact a unique labelling. The values of $\hat{C}_2$ in these representations are

$$\hat{C}_2(\mathbf{1}) = 0,\ \hat{C}_2(\mathbf{3},\ \mathbf{3}^*) = 4/3,\ \hat{C}_2(\mathbf{8}) = 3,\ \hat{C}_2(\mathbf{10}) = 6. \tag{M.67}$$

Having characterized a given representation by the eigenvalues of the Casimir operator(s), a further labelling is then required to characterize the states within a given representation (the analogue of the eigenvalue of $\hat{J}_3$ for angular momentum). For SO(4) these further labels may be taken to be the eigenvalues of $\hat{M}_3$ and $\hat{N}_3$; for SU(3) they are the eigenvalues of $\hat{G}_3$ and $\hat{G}_8$ – i.e. those corresponding to the third component of isospin and hypercharge, in the flavour case (see figures 12.3 and 12.4).

In the case of groups whose elements are themselves matrices, such as SO(3), SO(4), SU(2), SU(3), and the Lorentz group, one particular representation of the generators may always be obtained by considering the general form of a matrix in the group which is infinitesimally close to the unit element. In a suitable parametrization, we may write such a matrix as

$$1 + \mathrm{i} \sum_{\nu=1}^{r} \epsilon_\nu X_\nu^{(\mathcal{G})}, \tag{M.68}$$

where $(\epsilon_1, \epsilon_2, \ldots, \epsilon_r)$ are infinitesimal parameters, and $(X_1^{(\mathcal{G})}, X_2^{(\mathcal{G})}, \ldots, X_r^{(\mathcal{G})})$ are matrices representing the generators of the (matrix) group $\mathcal{G}$. This is exactly the same procedure we followed for SU(2) in section 12.1.1, where we found from (12.26) that the three $X_\nu^{(\mathrm{SU}(2))}$'s were just $\boldsymbol{\tau}/2$, satisfying the SU(2) algebra. Similarly, in section 12.2 we saw that the eight SU(3) $X_\nu^{(\mathrm{SU}(3))}$'s were just $\boldsymbol{\lambda}/2$, satisfying the SU(3) algebra. These particular two representations are called the *fundamental* representations of the SU(2) and SU(3) algebras, respectively; they are the representations of lowest dimensionality. For SO(3), the three $X_\nu^{(\mathrm{SO}(3))}$'s are (from (M.17))

$$\begin{aligned} X_1^{(\mathrm{SO}(3))} &= \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & -\mathrm{i} \\ 0 & \mathrm{i} & 0 \end{pmatrix} \\ X_2^{(\mathrm{SO}(3))} &= \begin{pmatrix} 0 & 0 & \mathrm{i} \\ 0 & 0 & 0 \\ -\mathrm{i} & 0 & 0 \end{pmatrix} \\ X_3^{(\mathrm{SO}(3))} &= \begin{pmatrix} 0 & -\mathrm{i} & 0 \\ \mathrm{i} & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \end{aligned} \tag{M.69}$$

which are the same as the 3×3 matrices $T_i^{(1)}$ of (12.48):

$$\left(T_i^{(1)}\right)_{jk} = -\mathrm{i}\epsilon_{ijk}. \tag{M.70}$$

The matrices $\tau_i/2$ and $T_i^{(1)}$ correspond to the values $J = 1/2,\ J = 1$, respectively, in angular momentum terms.

It is not a coincidence that the coefficients on the right-hand side of (M.70) are (minus) the SO(3) structure constants. One representation of a Lie algebra

is always provided by a set of matrices $\{X_\nu^{(\mathrm{R})}\}$ whose elements are defined by

$$\left(X_\lambda^{(\mathrm{R})}\right)_{\mu\nu} = -c_{\lambda\mu}^{\nu}, \tag{M.71}$$

where the c's are the structure constants of (M.12), and each of μ, ν, λ runs from 1 to r. Thus these matrices are of dimensionality $r \times r$, where r is the number of generators. That this prescription works is due to the fact that the generators satisfy the *Jacobi identity*

$$[\hat{X}_\lambda, [\hat{X}_\mu, \hat{X}_\nu]] + [\hat{X}_\mu, [\hat{X}_\nu, \hat{X}_\lambda]] + [\hat{X}_\nu, [\hat{X}_\lambda, \hat{X}_\mu]] = 0. \tag{M.72}$$

Using (M.12) to evaluate the commutators, and the fact that the generators are independent, we obtain

$$c_{\mu\nu}^{\alpha} c_{\lambda\alpha}^{\beta} + c_{\nu\lambda}^{\alpha} c_{\mu\alpha}^{\beta} + c_{\lambda\mu}^{\alpha} c_{\nu\alpha}^{\beta} = 0. \tag{M.73}$$

The reader may fill in the steps leading from here to the desired result:

$$\left(X_\lambda^{(\mathrm{R})}\right)_{\nu\alpha} \left(X_\mu^{(\mathrm{R})}\right)_{\alpha\beta} - \left(X_\mu^{(\mathrm{R})}\right)_{\nu\alpha} \left(X_\lambda^{(\mathrm{R})}\right)_{\alpha\beta} = c_{\lambda\mu}^{\alpha} \left(X_\alpha^{(\mathrm{R})}\right)_{\nu\beta}. \tag{M.74}$$

(M.74) is precisely the $(\nu\beta)$ matrix element of

$$[X_\lambda^{(\mathrm{R})}, X_\mu^{(\mathrm{R})}] = c_{\lambda\mu}^{\alpha} X_\alpha^{(\mathrm{R})}, \tag{M.75}$$

showing that the $X_\mu^{(\mathrm{R})}$'s satisfy the group algebra (M.12), as required. The representation in which the generators are represented by (minus) the structure constants, in the sense of (M.71), is called the *regular* or *adjoint* representation.

Having obtained any particular matrix representation $\boldsymbol{X}^{(\mathrm{P})}$ of the generators of a group $\mathcal{G}$, a corresponding *matrix representation of the group elements* can be obtained by exponentiation, via

$$D^{(\mathrm{P})}(\boldsymbol{\alpha}) = \exp\{\mathrm{i}\boldsymbol{\alpha} \cdot \boldsymbol{X}^{(\mathrm{P})}\}, \tag{M.76}$$

where $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_r)$ (see (12.31) and (12.49) for SU(2), and (12.74) and (12.81) for SU(3)). In the case of the groups whose elements are matrices, exponentiating the generators $\boldsymbol{X}^{(\mathcal{G})}$ just recreates the general matrices of the group, so we may call this the 'self-representation': the one in which the group elements are represented by themselves. In the more general case (M.76), the crucial property of the matrices $D^{(\mathrm{P})}(\boldsymbol{\alpha})$ is that they obey the same group combination law as the elements of the group $\mathcal{G}$ they are representing: that is, if the group elements obey

$$g(\boldsymbol{\alpha})g(\boldsymbol{\beta}) = g(\boldsymbol{\gamma}(\boldsymbol{\alpha}, \boldsymbol{\beta})), \tag{M.77}$$

then

$$D^{(\mathrm{P})}(\boldsymbol{\alpha})D^{(\mathrm{P})}(\boldsymbol{\beta}) = D^{(\mathrm{P})}(\boldsymbol{\gamma}(\boldsymbol{\alpha}, \boldsymbol{\beta})). \tag{M.78}$$

It is a rather remarkable fact that there are certain, say, 10×10 matrices which multiply together in exactly the same way as the rotation matrices of SO(3).

M.6 The Lorentz group

Consideration of matrix representations of the Lorentz group provides insight into the equations of relativistic quantum mechanics, for example the Dirac equation. Consider the infinitesimal Lorentz transformation (M.46). The 4×4 matrix corresponding to this may be written in the form

$$1 + \mathrm{i}\boldsymbol{\epsilon}\cdot\boldsymbol{X}^{(\mathrm{LG})} - \mathrm{i}\boldsymbol{\eta}\cdot\boldsymbol{K}^{(\mathrm{LG})}, \tag{M.79}$$

where

$$X_1^{(\mathrm{LG})} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -\mathrm{i} \\ 0 & 0 & \mathrm{i} & 0 \end{pmatrix} \quad \text{etc}, \tag{M.80}$$

(as in (M.69) but with an extra border of 0's), and

$$\begin{aligned} K_1^{(\mathrm{LG})} &= \begin{pmatrix} 0 & -\mathrm{i} & 0 & 0 \\ -\mathrm{i} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \\ K_2^{(\mathrm{LG})} &= \begin{pmatrix} 0 & 0 & -\mathrm{i} & 0 \\ 0 & 0 & 0 & 0 \\ -\mathrm{i} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \\ K_3^{(\mathrm{LG})} &= \begin{pmatrix} 0 & 0 & 0 & -\mathrm{i} \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ -\mathrm{i} & 0 & 0 & 0 \end{pmatrix}. \end{aligned} \tag{M.81}$$

In (M.80) and (M.81) the matrices are understood to be acting on the four-component vector

$$\begin{pmatrix} x^0 \\ x^1 \\ x^2 \\ x^3 \end{pmatrix}. \tag{M.82}$$

It is straightforward to check that the matrices $X_i^{(\mathrm{LG})}$ and $K_i^{(\mathrm{LG})}$ satisfy the algebra (M.50)–(M.52) as expected.

An important point to note is that the matrices $K_i^{(\mathrm{LG})}$, in contrast to $X_i^{(\mathrm{LG})}$ or $X_i^{(\mathrm{SO}(3))}$, and to the corresponding matrices of SU(2) and SU(3), are *not* Hermitian. A theorem states that only the generators of *compact* Lie groups can be represented by finite-dimensional Hermitian matrices. Here 'compact' means that the domain of variation of all the parameters is bounded (none exceeds a given positive number p in absolute magnitude) and closed

(the limit of every convergent sequence of points in the set also lies in the set). For the Lorentz group, the limiting velocity c is not included (the γ-factor goes to infinity), and so the group is *non-compact*.

In a general representation of the Lorentz group, the generators X_i, K_i will obey the algebra (M.50)–(M.52). Let us introduce the combinations

$$\boldsymbol{P} \equiv \frac{1}{2}(\boldsymbol{X} + \mathrm{i}\boldsymbol{K}) \tag{M.83}$$

$$\boldsymbol{Q} \equiv \frac{1}{2}(\boldsymbol{X} - \mathrm{i}\boldsymbol{K}). \tag{M.84}$$

Then the algebra becomes

$$[P_i, P_j] = \mathrm{i}\epsilon_{ijk}P_k \tag{M.85}$$

$$[Q_i, Q_j] = \mathrm{i}\epsilon_{ijk}Q_k \tag{M.86}$$

$$[P_i, Q_j] = 0, \tag{M.87}$$

which are apparently the same as (M.43)–(M.45). We can see from (M.81) that the matrices $\mathrm{i}\boldsymbol{K}^{(\mathrm{LG})}$ *are* Hermitian, and the same is in fact true in a general finite-dimensional representation. So we can appropriate standard angular momentum theory to set up the representations of the algebra of the $\boldsymbol{P}$'s and $\boldsymbol{Q}$'s – namely, they behave just like two independent (mutually commuting) angular momenta. The eigenvalues of $\boldsymbol{P}^2$ are of the form $P(P+1)$, for $P = 0, 1/2, \ldots$, and similarly for $\boldsymbol{Q}^2$; the eigenvalues of P_3 are M_P where $-P \leq M_P \leq P$, and similarly for Q_3.

Consider the particular case where the eigenvalue of $\boldsymbol{Q}^2$ is zero ($Q = 0$), and the value of P is $1/2$. The first condition implies that the $\boldsymbol{Q}$'s are identically zero, so that

$$\boldsymbol{X} = \mathrm{i}\boldsymbol{K} \tag{M.88}$$

in this representation, while the second condition tells us that

$$\boldsymbol{P} = \frac{1}{2}(\boldsymbol{X} + \mathrm{i}\boldsymbol{K}) = \frac{1}{2}\boldsymbol{\sigma}, \tag{M.89}$$

the familiar matrices for spin-1/2. We label this representation by the values of P (1/2) and Q (0) (these are the eigenvalues of the two Casimir operators). Then using (M.88) and (M.89) we find

$$\boldsymbol{X}^{(\frac{1}{2},0)} = \frac{1}{2}\boldsymbol{\sigma} \tag{M.90}$$

and

$$\boldsymbol{K}^{(\frac{1}{2},0)} = -\frac{\mathrm{i}}{2}\boldsymbol{\sigma}. \tag{M.91}$$

Now recall that the general infinitesimal Lorentz transformation has the form

$$1 + \mathrm{i}\boldsymbol{\epsilon} \cdot \boldsymbol{X} - \mathrm{i}\boldsymbol{\eta} \cdot \boldsymbol{K}. \tag{M.92}$$

In the present case this becomes

$$1 + \mathrm{i}\boldsymbol{\epsilon} \cdot \boldsymbol{\sigma}/2 - \boldsymbol{\eta} \cdot \boldsymbol{\sigma}/2. \tag{M.93}$$

These matrices are of dimension 2×2, and act on two-component spinors, which therefore transform under an infinitesimal Lorentz transformation by (cf (4.19) and (4.42))

$$\phi' = (1 + \mathrm{i}\boldsymbol{\epsilon} \cdot \boldsymbol{\sigma}/2 - \boldsymbol{\eta} \cdot \boldsymbol{\sigma}/2)\phi. \tag{M.94}$$

We say that ϕ 'transforms as the $(1/2, 0)$ representation of the Lorentz group'. The '$1+\mathrm{i}\boldsymbol{\epsilon}\cdot\boldsymbol{\sigma}/2$' part is the familiar (infinitesimal) rotation matrix for spinors, first met in section 4.4; it exponentiates to give $\exp(\mathrm{i}\boldsymbol{\alpha} \cdot \boldsymbol{\sigma}/2)$ for finite rotations. The '$-\boldsymbol{\eta} \cdot \boldsymbol{\sigma}/2$' part shows how such a spinor transforms under a pure (infinitesimal) velocity transformation. The finite transformation law is

$$\phi' = \exp(-\boldsymbol{\vartheta} \cdot \boldsymbol{\sigma}/2)\phi \tag{M.95}$$

where the three real parameters $\boldsymbol{\vartheta} = (\vartheta_1, \vartheta_2, \vartheta_3)$ specify the direction and magnitude of the boost.

There is, however, a second two-dimensional representation, which is characterized by the labelling $P = 0, Q = 1/2$, which we denote by (0, 1/2). In this case, the previous steps yield

$$\boldsymbol{X}^{(\frac{1}{2},0)} = \frac{1}{2}\boldsymbol{\sigma} \tag{M.96}$$

as before, but

$$\boldsymbol{K}^{(0,\frac{1}{2})} = \frac{\mathrm{i}}{2}\boldsymbol{\sigma}. \tag{M.97}$$

So the corresponding two-component spinor χ transforms by (cf (4.19) and (4.42))

$$\chi' = (1 + \mathrm{i}\boldsymbol{\epsilon} \cdot \boldsymbol{\sigma}/2 + \boldsymbol{\eta} \cdot \boldsymbol{\sigma}/2)\chi. \tag{M.98}$$

We see that ϕ and χ behave the same under rotations, but 'oppositely' under boosts.

These transformation laws are exactly what we used in section 4.1.2 when discussing the behaviour of the Dirac wavefunction ψ under Lorentz transformations, where ψ is put together from one ϕ and one χ via

$$\psi = \begin{pmatrix} \phi \\ \chi \end{pmatrix}, \tag{M.99}$$

and describes a *massive* spin-1/2 particle according to the equations

$$\begin{aligned} E\phi &= \boldsymbol{\sigma} \cdot \boldsymbol{p}\phi + m\chi \\ E\chi &= -\boldsymbol{\sigma} \cdot \boldsymbol{p}\chi + m\phi, \end{aligned} \tag{M.100}$$

consistent with the representation (3.40) of the Dirac matrices.

M.7 The relation between SU(2) and SO(3)

We have seen (sections M.4.1 and M.4.2) that the algebras of these two groups are identical. So the groups are isomorphic in the vicinity of their respective identity elements. Furthermore, matrix representations of one algebra automatically provide representations of the other. Since exponentiating these infinitesimal matrix transformations produces matrices representing group elements corresponding to finite transformations in both cases, it might appear that the groups are fully isomorphic. But actually they are not, as we shall now discuss.

We begin by re-considering the parameters used to characterize elements of SO(3) and SU(2). A general 3-D rotation is described by the SO(3) matrix $R(\hat{\boldsymbol{n}}, \theta)$, where $\hat{\boldsymbol{n}}$ is the axis of the rotation and θ is the angle of rotation. For example,

$$R(\hat{z}, \theta) = \begin{pmatrix} \cos\theta & \sin\theta & 0 \\ -\sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{pmatrix}. \tag{M.101}$$

On the other hand, we can write the general SU(2) matrix $\boldsymbol{V}$ in the form

$$\boldsymbol{V} = \begin{pmatrix} a & b \\ -b^* & a^* \end{pmatrix}, \tag{M.102}$$

where $|a|^2+|b|^2 = 1$ from the unit determinant condition. It therefore depends on three real parameters, the choice of which we are now going to examine in more detail than previously. In (12.32) we wrote $\boldsymbol{V}$ as $\exp(\mathrm{i}\boldsymbol{\alpha}\cdot\boldsymbol{\tau}/2)$, which certainly involves three real parameters $\alpha_1, \alpha_2, \alpha_3$; and below (12.35) we proposed, further, to write $\boldsymbol{\alpha} = \hat{\boldsymbol{n}}\theta$, where θ is an angle and $\hat{\boldsymbol{n}}$ is a unit vector. Then, since (as the reader may verify)

$$\exp(\mathrm{i}\theta\boldsymbol{\tau}\cdot\hat{\boldsymbol{n}}/2) = \cos\theta/2 + \mathrm{i}\boldsymbol{\tau}\cdot\hat{\boldsymbol{n}}\sin\theta/2, \tag{M.103}$$

it follows that this latter parametrization corresponds to writing, in (M.102),

$$a = \cos\theta/2 + \mathrm{i}n_z\sin\theta/2, \quad b = (n_y + \mathrm{i}n_x)\sin\theta/2, \tag{M.104}$$

with $n_x^2 + n_y^2 + n_z^2 = 1$. Clearly the condition $|a|^2 + |b|^2 = 1$ is satisfied, and one can convince oneself that the full range of a and b is covered if $\theta/2$ lies between 0 and π (in particular, it is not necessary to extend the range of $\theta/2$ so as to include the interval π to 2π, since the corresponding region of a, b can be covered by changing the orientation of $\hat{\boldsymbol{n}}$, which has not been constrained in any way). It follows that the parameters $\boldsymbol{\alpha}$ satisfy $\boldsymbol{\alpha}^2 \leq 4\pi^2$; that is, the space of the α's is the interior, and surface, of a sphere of radius 2π, as shown in figure M.1.

What about the parameter space of SO(3)? In this case, the same parameters $\hat{\boldsymbol{n}}$ and θ specify a rotation, but now θ (rather than $\theta/2$) runs from 0 to π.

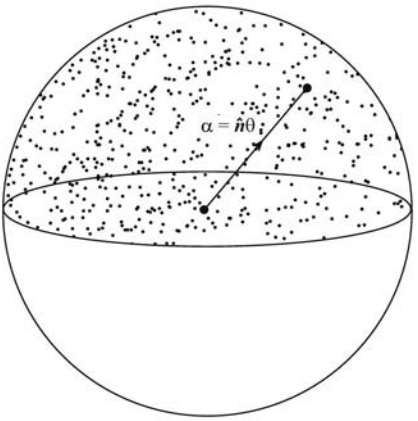


FIGURE M.1
The parameter spaces of SO(3) and SU(2): the whole sphere is the parameter space of SU(2), the upper (stippled) hemisphere that of SO(3).

However, we may allow the range of θ to extend to 2π, by taking advantage of the fact that

$$R(\hat{\boldsymbol{n}}, \pi + \theta) = R(-\hat{\boldsymbol{n}}, \theta). \tag{M.105}$$

Thus if we agree to limit $\hat{\boldsymbol{n}}$ to directions in the upper hemisphere of figure M.1, for 3-D rotations, we can say that the whole sphere represents the parameter space of SU(2), but that of SO(3) is provided by the upper half only.

Now let us consider the correspondence – or *mapping* – between the matrices of SO(3) and SU(2): we want to see if it is one-to-one. The notation strongly suggests that the matrix $\boldsymbol{V}(\hat{\boldsymbol{n}}, \theta) \equiv \exp(\mathrm{i}\theta\hat{\boldsymbol{n}} \cdot \tau/2)$ of SU(2) corresponds to the matrix $R(\hat{\boldsymbol{n}}, \theta)$ of SO(3), but the way it actually works has a subtlety.

We form the quantity $\boldsymbol{x} \cdot \boldsymbol{\tau}$, and assert that

$$\boldsymbol{x}' \cdot \boldsymbol{\tau} = \boldsymbol{V}(\hat{\boldsymbol{n}}, \theta)\, \boldsymbol{x} \cdot \boldsymbol{\tau}\, \boldsymbol{V}^{\dagger}(\hat{\boldsymbol{n}}, \theta), \tag{M.106}$$

where $\boldsymbol{x}' = R(\hat{\boldsymbol{n}}, \theta)\boldsymbol{x}$. We can easily verify (M.106) for the special case $R(\hat{z}, \theta)$, using (M.101); the general case follows with more labour (but the general infinitesimal case should by now be a familiar manipulation). (M.106) establishes a precise mapping between the elements of SU(2) and those of SO(3), but it is not one-to-one (i.e. not an isomorphism), since plainly $\boldsymbol{V}$ can always be replaced by $-\boldsymbol{V}$ and $\boldsymbol{x}'$ will be unchanged, and hence so will the associated SO(3) matrix $R(\hat{\boldsymbol{n}}, \theta)$. It is therefore a homomorphism.

Next, we prove a little theorem to the effect that the identity element e of a group $\mathcal{G}$ must be represented by the unit matrix of the representation: $D(e) = I$. For, let $D(a), D(e)$ represent the elements a, e of $\mathcal{G}$. Then $D(ae) =$

$D(a)D(e)$ by the fundamental property (M.78) of representation matrices. On the other hand, $ae = a$ by the property of e. So we have $D(a) = D(a)D(e)$, and hence $D(e) = I$.

Now let us return to the correspondence between SU(2) and SO(3). $\boldsymbol{V}(\hat{\boldsymbol{n}}, \theta)$ corresponds to $R(\hat{\boldsymbol{n}}, \theta)$, but can an SU(2) matrix be said to provide a valid representation of SO(3)? Consider the case $\boldsymbol{V}(\hat{n} = \hat{z}, \theta = 2\pi)$. From (M.103) this is equal to

$$\begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix}, \tag{M.107}$$

but the corresponding rotation matrix, from (M.101), is the identity matrix. Hence our theorem is violated, since (M.107) is plainly not the identity matrix of SU(2). Thus the SU(2) matrices can not be said to represent rotations, in the strict sense. Nevertheless, spin-1/2 particles certainly do exist, so Nature appears to make use of these 'not quite' representations! The SU(2) identity element is $\boldsymbol{V}(\hat{n} = \hat{z}, \theta = 4\pi)$, confirming that the rotational properties of a spinor are quite other than those of a classical object.

In fact, two and only two distinct elements of SU(2), namely

$$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix}, \tag{M.108}$$

correspond to the identity element of SO(3) in the correspondence (M.106) – just as, in general, $\boldsymbol{V}$ and $-\boldsymbol{V}$ correspond to the same SO(3) element $R(\hat{\boldsymbol{n}}, \theta)$, as we saw. The failure to be a true representation is localized simply to a sign: we may indeed say that, up to a sign, SU(2) matrices provide a representation of SO(3). If we 'factor out' this sign, the groups are isomorphic. A more mathematically precise way of saying this is given in Jones (1990, chapter 8).

N

Geometrical Aspects of Gauge Fields

N.1 Covariant derivatives and coordinate transformations

Let us go back to the U(1) case, equations (13.4)–(13.7). There, the introduction of the (gauge) covariant derivative D^μ produced an object, $D^\mu\psi(x)$, which transformed like $\psi(x)$ under local U(1) phase transformations, unlike the ordinary derivative $\partial^\mu\psi(x)$ which acquired an 'extra' piece when transformed. This followed from simple calculus, of course – but there is a slightly different way of thinking about it. The derivative involves not only $\psi(x)$ at the point x, but also ψ at the infinitesimally close, but different, point $x+\mathrm{d}x$; and the transformation law of $\psi(x)$ involves $\alpha(x)$, while that of $\psi(x+\mathrm{d}x)$ would involve the different function $\alpha(x+\mathrm{d}x)$. Thus we may perhaps expect something to 'go wrong' with the transformation law for the gradient.

To bring out the geometrical analogy we are seeking, let us write $\psi = \psi_\mathrm{R} + \mathrm{i}\psi_\mathrm{I}$ and $\alpha(x) = q\chi(x)$ so that (13.3) becomes (cf (2.64))

$$\begin{aligned} \psi'_\mathrm{R}(x) &= \cos\alpha(x)\psi_\mathrm{R}(x) - \sin\alpha(x)\psi_\mathrm{I}(x) \\ \psi'_\mathrm{I}(x) &= \sin\alpha(x)\psi_\mathrm{R}(x) + \cos\alpha(x)\psi_\mathrm{I}(x). \end{aligned} \tag{N.1}$$

If we think of $\psi_\mathrm{R}(x)$ and $\psi_\mathrm{I}(x)$ as being the components of a 'vector' $\vec{\psi}(x)$ along the $\vec{e}_\mathrm{R}$ and $\vec{e}_\mathrm{I}$ axes, respectively, then (N.1) would represent the components of $\vec{\psi}(x)$ as referred to new axes $\vec{e}_\mathrm{R}{}'$ and $\vec{e}_\mathrm{I}{}'$, which have been rotated by $-\alpha(x)$ about an axis in the direction $\vec{e}_\mathrm{R}\times\vec{e}_\mathrm{I}$ (i.e. normal to the $\vec{e}_\mathrm{R}-\vec{e}_\mathrm{I}$ plane), as shown in figure N.1. Other such 'vectors' $\vec{\phi}_1(x), \vec{\phi}_2(x), \ldots$ (i.e. other wavefunctions for particles of the same charge q) *when evaluated at the same point* x will have 'components' transforming the same as (N.1) under the axis rotation $\vec{e}_\mathrm{R}, \vec{e}_\mathrm{I} \to \vec{e}_\mathrm{R}{}', \vec{e}_\mathrm{I}{}'$. But the components of the vector $\vec{\psi}(x+\mathrm{d}x)$ will behave differently. The transformation law (N.1) when written at $x+\mathrm{d}x$ will involve $\alpha(x+\mathrm{d}x)$, which (to first order in $\mathrm{d}x$) is $\alpha(x)+\partial_\mu\alpha(x)\mathrm{d}x^\mu$. Thus for $\psi'_\mathrm{R}(x+\mathrm{d}x)$ and $\psi'_\mathrm{I}(x+\mathrm{d}x)$ the rotation angle is $\alpha(x) + \partial_\mu\alpha(x)dx^\mu$ rather than $\alpha(x)$. Now comes the key step in the analogy: we may think of the additional angle $\partial_\mu\alpha(x)\mathrm{d}x^\mu$ as coming about because, in going from x to $x+\mathrm{d}x$, the coordinate basis vectors $\vec{e}_\mathrm{R}$ and $\vec{e}_\mathrm{I}$ have been rotated through $+\partial_\mu\alpha(x)\mathrm{d}x^\mu$

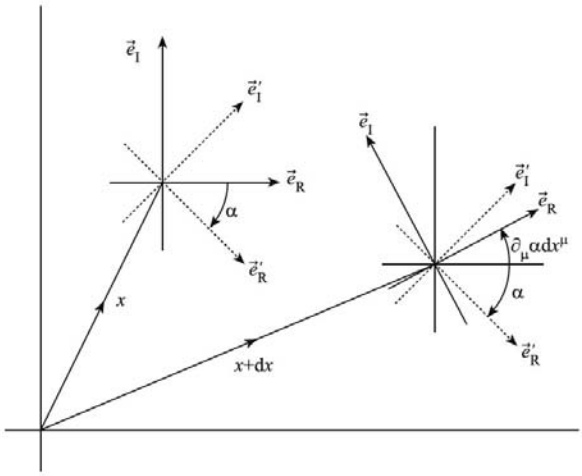


FIGURE N.1
Geometrical analogy for a U(1) gauge transformation.

(see figure N.2)! But that would mean that our 'naive' approach to rotations of the derivative of $\vec{\psi}(x)$ amounts to using one set of axes at x, and another at $x + \mathrm{d}x$, which is likely to lead to 'trouble'. Consider now an elementary example (from Schutz 1988, chapter 5) where just this kind of problem arises, namely the use of polar coordinate basis vectors $\vec{e}_r$ and $\vec{e}_\theta$, which point in the r and θ directions respectively. We have, as usual,

$$x = r\cos\theta, \quad y = r\sin\theta \tag{N.2}$$

and in a (real!) Cartesian basis $\mathrm{d}\vec{r}$ is given by

$$\mathrm{d}\vec{r} = \mathrm{d}x\ \vec{i} + \mathrm{d}y\ \vec{j}. \tag{N.3}$$

Using (N.2) in (N.3) we find

$$\begin{aligned} \mathrm{d}\vec{r} &= (\mathrm{d}r\cos\theta - r\sin\theta\ \mathrm{d}\theta)\vec{i} + (\mathrm{d}r\sin\theta + r\cos\theta\ \mathrm{d}\theta)\vec{j} \\ &= \mathrm{d}r\ \vec{e}_r + \mathrm{d}\theta\ \vec{e}_\theta \end{aligned} \tag{N.4}$$

where

$$\vec{e}_r = \cos\theta\,\vec{i} + \sin\theta\,\vec{j}, \vec{e}_\theta = -r\sin\theta\,\vec{i} + r\cos\theta\,\vec{j}. \tag{N.5}$$

Plainly, $\vec{e}_r$ and $\vec{e}_\theta$ change direction (and even magnitude, for $\vec{e}_\theta$) as we move about in the $x - y$ plane, as shown in figure N.2. So at each point (r, θ) we have *different* axes $\vec{e}_r, \vec{e}_\theta$.

Now suppose that we wish to describe a vector field $\vec{V}$ in terms of $\vec{e}_r$ and $\vec{e}_\theta$ via

$$\vec{V} = V^r\vec{e}_r + V^\theta\vec{e}_\theta \equiv V^\alpha\vec{e}_\alpha \text{ (sum on } \alpha = r, \theta), \tag{N.6}$$

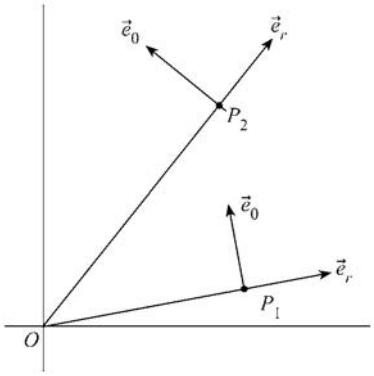


FIGURE N.2
Changes in the basis vectors $\vec{e}_r$ and $\vec{e}_\theta$ of polar coordinates.

and that we are also interested in the derivatives of $\vec{V}$, in this basis. Let us calculate $\frac{\partial \vec{V}}{\partial r}$, for example, by brute force:

$$\frac{\partial \vec{V}}{\partial r} = \frac{\partial V^r}{\partial r}\vec{e}_r + \frac{\partial V^\theta}{\partial r}\vec{e}_\theta + V^r\frac{\partial \vec{e}_r}{\partial r} + V^\theta\frac{\partial \vec{e}_\theta}{\partial r} \tag{N.7}$$

where we have included the derivatives of $\vec{e}_r$ and $\vec{e}_\theta$ to allow for the fact that *these vectors are not constant*. From (N.5) we easily find

$$\frac{\partial \vec{e}_r}{\partial r} = 0, \quad \frac{\partial \vec{e}_\theta}{\partial r} = -\sin\theta\,\vec{i} + \cos\theta\,\vec{j} = \frac{1}{r}\vec{e}_\theta, \tag{N.8}$$

which allows the last two terms in (N.7) to be evaluted. Similarly, we can calculate $\frac{\partial \vec{V}}{\partial \theta}$. In general, we may write these results as

$$\frac{\partial \vec{V}}{\partial q^\beta} = \frac{\partial V^\alpha}{\partial q^\beta}\vec{e}_\alpha + V^\alpha\frac{\partial \vec{e}_\alpha}{\partial q^\beta} \tag{N.9}$$

where $\beta = 1, 2$ with $q^1 = r, q^2 = \theta$, and $\alpha = r, \theta$.

In the present case, we were able to calculate $\partial\vec{e}_\alpha/\partial q^\beta$ explicitly from (N.5), as in (N.8). But whatever the nature of the coordinate system, $\partial\vec{e}_\alpha/\partial q^\beta$ is some vector and must be expressible as a linear combination of the basis vectors via an expression of the form

$$\frac{\partial \vec{e}_\alpha}{\partial q^\beta} = \Gamma^\gamma{}_{\alpha\beta}\vec{e}_\gamma \tag{N.10}$$

where the repeated index γ is summed over as usual ($\gamma = r, \theta$). Inserting (N.10) into (N.9) and interchanging the 'dummy' (i.e. summed over) indices

α and γ gives finally

$$\frac{\partial \vec{V}}{\partial q^\beta} = \left(\frac{\partial V^\alpha}{\partial q^\beta} + \Gamma^\alpha_{\ \gamma\beta} V^\gamma \right) \vec{e}_\alpha. \tag{N.11}$$

This is a very important result: it shows that, whereas the components of $\vec{V}$ in the basis $\vec{e}_\alpha$ are just V^α, the components of the derivative of $\vec{V}$ are not simply $\partial V^\alpha / \partial q^\beta$, but *contain an additional term*: the 'components of the derivative of a vector' are not just the 'derivatives of the components of the vector'.

Let us abbreviate $\partial / \partial q^\beta$ to ∂_β; then (N.11) tells us that in the $\vec{e}_\alpha$ basis, as used in (N.11), the components of the ∂_β derivative of $\vec{V}$ are

$$\partial_\beta V^\alpha + \Gamma^\alpha_{\ \gamma\beta} V^\gamma \equiv D_\beta V^\alpha. \tag{N.12}$$

The expression (N.12) is called the 'covariant derivative' of V^α within the context of the mathematics of general coordinate systems: it is denoted (as in (N.12)) by $D_\beta V^\alpha$ or, often, by $V^\alpha_{\ ;\beta}$ (in the latter notation, $\partial_\beta V^\alpha$ is $V^\alpha_{\ ,\beta}$). The most important property of $D_\beta V^\alpha$ is its transformation character under general coordinate transformations. Crucially, it transforms as a *tensor* T^α_β (see appendix D of volume 1) with the indicated 'one up, one down' indices; we shall not prove this here, referring instead to Schutz (1988), for example. This property is the reason for the name 'covariant derivative', meaning in this case essentially that it transforms the way its indices would have you believe it should. By contrast, and despite appearances, $\partial_\beta V^\alpha$ by itself does *not* transform as a 'T^α_β' tensor, and in a similar way $\Gamma^\alpha_{\ \gamma\beta}$ is *not* a '$T^\alpha_{\gamma\beta}$'-type tensor; only the combined object $D_\beta V^\alpha$ is a 'T^α_β'.

This circumstance is highly reminiscent of the situation we found in the case of gauge transformations. Consider the simplest case, that of U(1), for which $D_\mu \psi = \partial_\mu \psi + \mathrm{i} q A_\mu \psi$. The quantity $D_\mu \psi$ transforms under a gauge transformation in the same way as ψ itself, but $\partial_\mu \psi$ does not. There is thus a close analogy between the 'good' transformation properties of $D_\beta V^\alpha$ and of $D_\mu \psi$. Further, the structure of $D_\mu \psi$ is very similar to that of $D_\beta V^\alpha$. There are two pieces, the first of which is the straightforward derivative, while the second involves a new field (Γ or A) and is also proportional to the original field. The 'i' of course is a big difference, showing that in the gauge symmetry case the transformations mix the real and imaginary parts of the wavefunction, rather than actual spatial components of a vector.

Indeed, the analogy is even closer in the non-Abelian – e.g. local SU(2) – case. As we have seen, $\partial^\mu \psi^{(\frac{1}{2})}$ does not transform as an SU(2) isospinor because of the extra piece involving $\partial^\mu \boldsymbol{\epsilon}$; nor do the gauge fields $\boldsymbol{W}^\mu$ transform as pure $T = 1$ states, also because of a $\partial^\mu \boldsymbol{\epsilon}$ term. But the gauge covariant combination $(\partial^\mu + \mathrm{i} g \boldsymbol{\tau} \cdot \boldsymbol{W}^\mu / 2) \psi^{(\frac{1}{2})}$ does transform as an isospinor under local SU(2) transformations, the two 'extra' $\partial^\mu \boldsymbol{\epsilon}$ pieces cancelling each other out.

There is a useful way of thinking about the two contributions to $D_\beta V^\alpha$ (or $D_\mu \psi$). Let us multiply (N.12) by $\mathrm{d}q^\beta$ and sum over β so as to obtain

$$DV^\alpha \equiv \partial_\beta V^\alpha \mathrm{d}q^\beta + \Gamma^\alpha_{\ \gamma\beta} V^\gamma \mathrm{d}q^\beta. \tag{N.13}$$

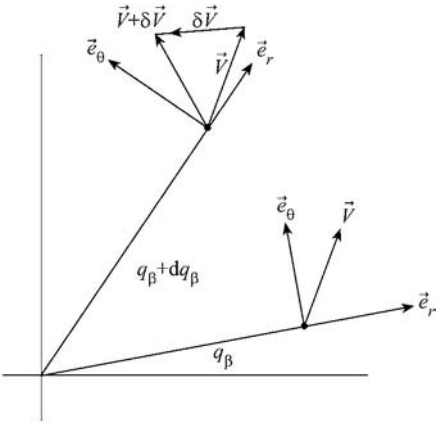


FIGURE N.3
Parallel transport of a vector $\vec{V}$ in a polar coordinate basis.

The first term on the right-hand side of (N.13) is $\frac{\partial V^\alpha}{\partial q^\beta}\mathrm{d}q^\beta$ which is just the conventional differential $\mathrm{d}V^\alpha$, representing the change in V^α in moving from q^β to $q^\beta + \mathrm{d}q^\beta$: $\mathrm{d}V^\alpha = [V^\alpha(q^\beta + \mathrm{d}q^\beta) - V^\alpha(q^\beta)]$. Again, despite appearances, the quantities $\mathrm{d}V^\alpha$ do not form the components of a vector, and the reason is that $V^\alpha(q^\beta + \mathrm{d}q^\beta)$ are components with respect to axes at $q^\beta + \mathrm{d}q^\beta$, while $V^\alpha(q^\beta)$ are components with respect to *different* axes at q^β. To form a 'good' differential DV^α, transforming as a vector, we must subtract quantities defined in the *same* coordinate system. This means that we need some way of 'carrying' $V^\alpha(q^\beta)$ to $q^\beta + \mathrm{d}q^\beta$, while keeping it somehow 'the same' as it was at q^β.

A reasonable definition of such a 'preserved' vector field is one that is unchanged in length, and has the same orientation relative to the axes at $q^\beta + \mathrm{d}q^\beta$ as it had relative to the axes at q^β (see figure N.3). In other words, $\vec{V}$ is 'dragged around' with the changing coordinate frame, a process called *parallel transport*. Such a definition of 'no change' of course implies that change *has* occurred, in general, with respect to the *original* axes at q^β. Let us denote by δV^α the difference between the components of $\vec{V}$ after parallel transport to $q^\beta + \mathrm{d}q^\beta$, and the components of $\vec{V}$ at q^β (see figure N.3). Then a reasonable definition of the 'good' differential of V^α would be $V^\alpha(q^\beta + \mathrm{d}q^\beta) - (V^\alpha(q^\beta) + \delta V^\alpha) = \mathrm{d}V^\alpha - \delta V^\alpha$. We interpret this as the covariant differential DV^α of (N.13), and accordingly, make the identification

$$\delta V^\alpha = -\Gamma^\alpha{}_{\gamma\beta} V^\gamma \mathrm{d}q^\beta. \tag{N.14}$$

On this interpretation, then, the coefficients $\Gamma^\alpha{}_{\gamma\beta}$ connect the components of a vector at one point with its components at a nearby point, after the vector

has been carried by 'parallel transport' from one point to the other; they are often called 'connection coefficients', or just 'the connection'.

In an analogous way we can write, in the U(1) gauge case,

$$\begin{aligned} D\psi \equiv D^\mu \psi \mathrm{d}x_\mu &= \partial^\mu \psi \mathrm{d}x_\mu + \mathrm{i}eA^\mu \psi \mathrm{d}x_\mu \\ &\equiv \mathrm{d}\psi - \delta\psi \end{aligned} \tag{N.15}$$

with

$$\delta\psi = -\mathrm{i}eA^\mu \psi \mathrm{d}x_\mu. \tag{N.16}$$

Equation (N.16) has a very similar structure to (N.14), suggesting that the electromagnetic potential A^μ might well be referred to as a 'gauge connection', as indeed it is in some quarters. Equations (N.15) and (N.16) generalize straightforwardly for $D\psi^{(\frac{1}{2})}$ and $\delta\psi^{(\frac{1}{2})}$.

We can relate (N.16) in a very satisfactory way to our original discussion of electromagnetism as a gauge theory in chapter 2, and in particular to (2.83). For transport restricted to the three spatial directions, (N.16) reduces to

$$\delta\psi(\boldsymbol{x}) = \mathrm{i}e\boldsymbol{A}\cdot\mathrm{d}\boldsymbol{x}\psi(x). \tag{N.17}$$

However, the solution (2.83) gives

$$\psi(\boldsymbol{x}) = \exp\left(\mathrm{i}e\int_{-\infty}^{\boldsymbol{x}} \boldsymbol{A}\cdot\mathrm{d}\boldsymbol{\ell}\right)\psi(\boldsymbol{A}=\boldsymbol{0},\boldsymbol{x}), \tag{N.18}$$

replacing q by e. So

$$\begin{aligned} &\psi(\boldsymbol{x}+\mathrm{d}\boldsymbol{x}) \\ &= \exp\left(\mathrm{i}e\int_{-\infty}^{\boldsymbol{x}+\mathrm{d}\boldsymbol{x}} \boldsymbol{A}\cdot\mathrm{d}\boldsymbol{\ell}\right)\psi(\boldsymbol{A}=\boldsymbol{0},\boldsymbol{x}+\mathrm{d}\boldsymbol{x}) \\ &= \exp\left(\mathrm{i}e\int_{\boldsymbol{x}}^{\boldsymbol{x}+\mathrm{d}\boldsymbol{x}} \boldsymbol{A}\cdot\mathrm{d}\boldsymbol{\ell}\right)\exp\left(\mathrm{i}e\int_{-\infty}^{\boldsymbol{x}} \boldsymbol{A}\cdot\mathrm{d}\boldsymbol{\ell}\right)\psi(\boldsymbol{A}=\boldsymbol{0},\boldsymbol{x}+\mathrm{d}\boldsymbol{x}) \\ &\approx (1+\mathrm{i}e\boldsymbol{A}\cdot\mathrm{d}\boldsymbol{x})\exp\left(\mathrm{i}e\int_{-\infty}^{\boldsymbol{x}} \boldsymbol{A}\cdot\mathrm{d}\boldsymbol{\ell}\right)[\psi(\boldsymbol{A}=\boldsymbol{0},\boldsymbol{x})+\boldsymbol{\nabla}\psi(\boldsymbol{A}=\boldsymbol{0},\boldsymbol{x})\cdot\mathrm{d}\boldsymbol{x}] \\ &\approx \psi(\boldsymbol{x})+\mathrm{i}e\boldsymbol{A}\cdot\mathrm{d}\boldsymbol{x}\psi(\boldsymbol{x})+\exp\left(\mathrm{i}e\int_{-\infty}^{\boldsymbol{x}} \boldsymbol{A}\cdot\mathrm{d}\boldsymbol{\ell}\right)\boldsymbol{\nabla}\psi(\boldsymbol{A}=\boldsymbol{0},\boldsymbol{x})\cdot\mathrm{d}\boldsymbol{x}, \end{aligned} \tag{N.19}$$

to first order in $\mathrm{d}\boldsymbol{x}$. On the right-hand side of (N.19) we see (i) the change $\delta\psi$ of (N.17), due to 'parallel transport' as prescribed by the gauge connection $\boldsymbol{A}$, and (ii) the change in ψ viewed as a function of $\boldsymbol{x}$, in the absence of $\boldsymbol{A}$. The solution (N.18) gives, in fact, the 'integrated' form of the small displacement law (N.19).

At this point the reader might object, going back to the $\vec{e}_r, \vec{e}_\theta$ example, that we had made a lot of fuss about nothing: after all, no one forced us

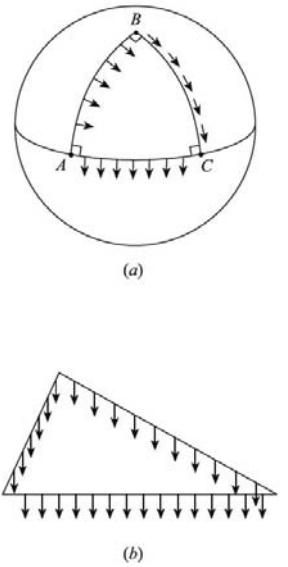


FIGURE N.4
Parallel transport (a) round a curved triangle on the surface of a sphere (b) round a triangle in a flat plane.

to use the $\vec{e}_r, \vec{e}_\theta$ basis, and if we had simply used the $\vec{i}, \vec{j}$ basis (which is constant throughout the plane) we would have had no such 'trouble'. This is a fair point, provided that we somehow knew that we are really doing physics in a 'flat' space, such as the Euclidean plane. But suppose instead that our two-dimensional space was the surface of a sphere. Then, an intuitively plausible definition of parallel transport is shown in figure N.4(a), in which transport is carried out around a closed path consisting of three great circle arcs $A \to B, B \to C, C \to A$, with the rule that at each stage the vector is drawn 'as parallel as possible' to the previous one. It is clear from the figure that the vector we end up with at A, after this circuit, is no longer parallel to the vector we started with; in fact, it has rotated by $\pi/2$ in this example, in which $\frac{1}{8}$th of the surface area of the unit sphere is enclosed by the triangle ABC. By contrast, the parallel transport of a vector round a flat triangle in the Euclidean plane leads to no such net change in the vector (figure N.4(b)).

It seems reasonable to suppose that the information about whether the space we are dealing with is 'flat' or 'curved' is contained in the connection $\Gamma^\alpha{}_{\gamma\beta}$. In a similar way, in the gauge case the analogy we have built up so far would lead us to expect that there are potentials A^μ which are somehow 'flat' ($\boldsymbol{E} = \boldsymbol{B} = \boldsymbol{0}$) and others which represent 'curvature' (non-zero $\boldsymbol{E}, \boldsymbol{B}$). This is what we discuss next.

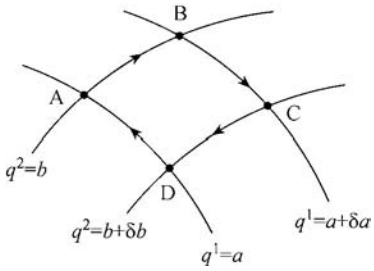


FIGURE N.5
Closed loop $ABCD$ in $q^1 - q^2$ space.

N.2 Geometrical curvature and the gauge field strength tensor

Consider a small closed loop in our (possibly curved) two-dimensional space – see figure N.5 – whose four sides are the coordinate lines $q^1 = a, q^1 = a + \delta a, q^2 = b, q^2 = b + \delta b$. We want to calculate the net change (if any) in δV^α as we parallel transport $\vec{V}$ around the loop. The change along $A \to B$ is

$$\begin{aligned}(\delta V^\alpha)_{AB} &= -\int_{q^2=b,q^1=a}^{q^2=b,q^1=a+\delta a} \Gamma^\alpha{}_{\gamma 1} V^\gamma dq^1 \\ &\approx -\delta a \Gamma^\alpha{}_{\gamma 1}(a,b) V^\gamma(a,b) \end{aligned} \tag{N.20}$$

to first order in δa, while that along $C \to D$ is

$$\begin{aligned}(\delta V^\alpha)_{CD} &= -\int_{q^2=b+\delta b,q^1=a+\delta a}^{q^2=b+\delta b,q^1=a} \Gamma^\alpha{}_{\gamma 1} V^\gamma dq^1 \\ &= +\int_{q^2=b+\delta b,q^1=a}^{q^2=b+\delta b,q^1=a+\delta a} \Gamma^\alpha{}_{\gamma 1} V^\gamma dq^1. \\ &\approx \delta a \Gamma^\alpha_{\gamma 1}(a,b+\delta b) V^\gamma(a,b+\delta b). \end{aligned} \tag{N.21}$$

Now

$$\Gamma^\alpha{}_{\gamma 1}(a,b+\delta b) \approx \Gamma^\alpha{}_{\gamma 1}(a,b) + \delta b \frac{\partial \Gamma^\alpha{}_{\gamma 1}}{\partial q^2} \tag{N.22}$$

and, remembering that we are parallel-transporting $\vec{V}$,

$$V^\gamma(a,b+\delta b) \approx V^\gamma(a,b) - \Gamma^\gamma{}_{\delta 2} V^\delta \delta b. \tag{N.23}$$

Combining (N.20) and (N.21) to lowest order, we find

$$(\delta V^\alpha)_{AB} + (\delta V^\alpha)_{CD} \approx \delta a \delta b \left[\frac{\partial \Gamma^\alpha{}_{\gamma 1}}{\partial q^2} V^\gamma - \Gamma^\alpha{}_{\gamma 1} \Gamma^\gamma{}_{\delta 2} V^\delta \right] \tag{N.24}$$

or, interchanging dummy indices γ and δ in the last term,

$$(\delta V^\alpha)_{AB} + (\delta V^\alpha)_{CD} \approx \delta a \delta b \left[\frac{\partial \Gamma^\alpha{}_{\gamma 1}}{\partial q^2} - \Gamma^\alpha{}_{\delta 1} \Gamma^\delta{}_{\gamma 2} \right] V^\gamma. \tag{N.25}$$

Similarly,

$$(\delta V^\alpha)_{BC} + (\delta V^\alpha)_{DA} \approx \delta a \delta b \left[-\frac{\partial \Gamma^\alpha{}_{\gamma 2}}{\partial q^1} + \Gamma^\alpha{}_{\delta 2} \Gamma^\delta{}_{\gamma 1} \right] V^\gamma, \tag{N.26}$$

and so the net change around the whole small loop is

$$(\delta V^\alpha)_{ABCD} \approx \delta a \delta b \left[\frac{\partial \Gamma^\alpha{}_{\gamma 1}}{\partial q^2} - \frac{\partial \Gamma^\alpha{}_{\gamma 2}}{\partial q^1} + \Gamma^\alpha{}_{\delta 2} \Gamma^\delta{}_{\gamma 1} - \Gamma^\alpha{}_{\delta 1} \Gamma^\delta{}_{\gamma 2} \right] V^\gamma. \tag{N.27}$$

The indices '1' and '2' appear explicitly because the loop was chosen to go along these directions. In general, (N.27) would take the form

$$(\delta V^\alpha)_{loop} \approx \left[\frac{\partial \Gamma^\alpha{}_{\gamma\beta}}{\partial q^\sigma} - \frac{\partial \Gamma^\alpha{}_{\gamma\sigma}}{\partial q^\beta} + \Gamma^\alpha{}_{\delta\sigma} \Gamma^\delta{}_{\gamma\beta} - \Gamma^\alpha{}_{\delta\beta} \Gamma^\delta{}_{\gamma\sigma} \right] V^\gamma \mathrm{d}A^{\beta\sigma} \tag{N.28}$$

where $\mathrm{d}A^{\beta\sigma}$ is the area element. The quantity in brackets in (N.28) is the *Reimann curvature tensor* $R^\alpha{}_{\gamma\beta\sigma}$ (up to a sign, depending on conventions), which can clearly be calculated once the connection coefficients are known. A flat space is one for which all components $R^\alpha{}_{\gamma\beta\sigma} = 0$; the reader may verify that this is the case for our polar basis $\vec{e}_r, \vec{e}_\theta$ in the Euclidean plane. A non-zero value for any component of $R^\alpha{}_{\gamma\beta\sigma}$ means the space is curved.

We now follow exactly similar steps to calculate the net change in $\delta\psi$ as given by (N.16), around the small two-dimensional rectangle defined by the coordinate lines $x_1 = a, x_1 = a + \delta a, x_2 = b, x_2 = b + \delta b$, labelled as in figure N.5 but with q^1 replaced by x_1 and q^2 by x_2. Then

$$(\delta\psi)_{AB} = -\mathrm{i}eA^1(a,b)\psi(a,b)\delta a \tag{N.29}$$

and

$$\begin{aligned}
(\delta\psi)_{CD} &= +\mathrm{i}eA^1(a, b+\delta b)\psi(a, b+\delta b)\delta a \\
&\approx \mathrm{i}e \left(A^1(a,b) + \frac{\partial A^1}{\partial x_2}\delta b \right) [\psi(a,b) - \mathrm{i}eA^2(a,b)\psi(a,b)\delta b]\delta a \\
&\approx \mathrm{i}eA^1(a,b)\psi(a,b)\delta a \\
&\quad + \mathrm{i}e \left[\frac{\partial A^1}{\partial x_2}\psi(a,b) - \mathrm{i}eA^1(a,b)A^2(a,b)\psi(a,b) \right] \delta a \delta b.
\end{aligned} \tag{N.30}$$

Combining (N.29) and (N.30) we find

$$(\delta\psi)_{AB} + (\delta\psi)_{CD} \approx \left[\mathrm{i}e\frac{\partial A^1}{\partial x_2}\psi + e^2 A^1 A^2 \psi \right] \delta a \delta b. \tag{N.31}$$

Similarly,

$$(\delta\psi)_{BC} + (\delta\psi)_{DA} \approx \left[-\mathrm{i}e\frac{\partial A^2}{\partial x_1}\psi - e^2 A^1 A^2 \psi\right] \delta a \delta b, \tag{N.32}$$

with the result that the net change around the loop is

$$(\delta\psi)_{ABCD} \approx \mathrm{i}e\left(\frac{\partial A^1}{\partial x_2} - \frac{\partial A^2}{\partial x_1}\right)\psi \delta a \delta b. \tag{N.33}$$

For a general loop, (N.33) is replaced by

$$\begin{aligned}(\delta\psi)_{loop} &= \mathrm{i}e\left(\frac{\partial A^\mu}{\partial x_\nu} - \frac{\partial A^\nu}{\partial x_\mu}\right)\psi \mathrm{d}x_\mu \mathrm{d}x_\nu \\ &= -\mathrm{i}eF^{\mu\nu}\psi \mathrm{d}x_\mu \mathrm{d}x_\nu \end{aligned} \tag{N.34}$$

where $F^{\mu\nu} = \partial^\mu A^\nu - \partial^\nu A^\mu$ is the familiar field strength tensor of QED.

The analogy we have been pursuing would therefore suggest that $F^{\mu\nu} = 0$ indicates 'no physical effect', while $F^{\mu\nu} \neq 0$ implies the presence of a physical effect. Indeed, when A^μ has the 'pure gauge' form $A^\mu = \partial^\mu \chi$ the associated $F^{\mu\nu}$ is zero; this is because such an A^μ can clearly be reduced to zero by a gauge transformation (and also, consistently, because $(\partial^\mu\partial^\nu - \partial^\nu\partial^\mu)\chi = 0$). If A^μ is not expressible as the 4-gradient of a scalar, then $F^{\mu\nu} \neq 0$ and an electromagnetic field is present, analogous to the spatial curvature revealed by $R^\alpha{}_{\gamma\beta\sigma} \neq 0$. Once again, there is a satisfying consistency between this 'geometrical' viewpoint and the discussion of the Aharonov-Bohm effect in Section 2.6. As in our remarks at the end of the previous section, and equations (N.17)–(N.19), equation (2.83) can be regarded as the integrated form of (N.34), for spatial loops. Transport round such a loop results in a non-trivial net phase change if non-zero $\boldsymbol{B}$ flux is enclosed, and this can be observed.

From this point of view there is undoubtedly a strong conceptual link between Einstein's theory of gravity and quantum gauge theories. In the former, matter (or energy) is regarded as the source of curvature of space-time, causing the space-time axes themselves to vary from point to point, and determining the trajectories of massive particles; in the latter, charge is the source of curvature in an 'internal' space (the complex ψ-plane, in the U(1) case), a curvature which we call an electromagnetic field, and which has observable physical effects.

The reader may consider repeating, for the local SU(2) case, the closed-loop transport calculation of (N.29)–(N.33). For this calculation, the place of the Abelian vector potential is taken by the matrix-valued non-Abelian potential $A^\mu = \boldsymbol{\tau}/2 \cdot \boldsymbol{A}^\mu$. It will lead to the expression for the non-Abelian field strength tensor as calculated in section 13.1.2.

O

Dimensional Regularization

After combining propagator denominators of the form $(p^2 - m^2 + \mathrm{i}\epsilon)^{-1}$ by Feynman parameters (cf (10.40)), and shifting the origin of the loop momentum to complete the square (cf (10.42) and (11.16)), all one-loop Feynman integrals may be reduced to evaluating an integral of the form

$$I_d(\Delta, n) \equiv \int \frac{\mathrm{d}^d k}{(2\pi)^d} \frac{1}{[k^2 - \Delta + \mathrm{i}\epsilon]^n}, \tag{O.1}$$

or to a similar integral with factors of k (such as $k_\mu k_\nu$) in the numerator. We consider (O.1) first.

For our purposes, the case of physical interest is $d = 4$, and n is commonly 2 (e.g. in one-loop self-energies). Power-counting shows that (O.1) diverges as $k \to \infty$ for $d \geq 2n$. The idea behind *dimensional regularization* ('t Hooft and Veltman 1972) is to treat d as a variable parameter, taking values smaller than $2n$, so that (O.1) converges and can be evaluated explicitly as a function of d (and of course the other variables, including n).[1] Then the nature of the divergence as $d \to 4$ can be exposed (much as we did with the cut-off procedure in section 10.3), and dealt with by a suitable renormalization scheme. The crucial advantage of dimensional regularization is that it preserves gauge invariance, unlike the simple cut-off regularization we used in chapters 10 and 11.

We write

$$I_d = \frac{1}{(n-1)!} \left(\frac{\partial}{\partial \Delta}\right)^{n-1} \int \frac{\mathrm{d}^d k}{(2\pi)^d} \frac{1}{[k^2 - \Delta + \mathrm{i}\epsilon]}. \tag{O.2}$$

The d dimensions are understood as one time-like dimension k^0, and $d-1$ spacelike dimensions. We begin by 'Euclideanizing' the integral, by setting $k^0 = \mathrm{i}k^e$ with k^e real. Then the Minkowskian square k^2 becomes $-(k^e)^2 - \boldsymbol{k}^2 \equiv -k_{\mathrm{E}}^2$, and $\mathrm{d}^d k$ becomes $\mathrm{i}\mathrm{d}^d k_{\mathrm{E}}$, so that now

$$I_d = \frac{-\mathrm{i}}{(n-1)!} \left(\frac{\partial}{\partial \Delta}\right)^{n-1} \int \frac{\mathrm{d}^d k_{\mathrm{E}}}{(2\pi)^d} \frac{1}{(k_{\mathrm{E}}^2 + \Delta)}; \tag{O.3}$$

the '$\mathrm{i}\epsilon$' may be understood as included in Δ. The integral is evaluated by

[1] We concentrate here on ultraviolet divergences, but infrared ones (such as those met in section 14.4.2) can be dealt with too, by choosing d larger than $2n$.

introducing the following way of writing $(k_{\rm E}^2+\Delta)^{-1}$:

$$(k_{\rm E}^2+\Delta)^{-1}=\int_0^\infty {\rm d}\beta {\rm e}^{-\beta(k_{\rm E}^2+\Delta)}, \tag{O.4}$$

which leads to

$$I_d=\frac{1}{(n-1)!}\left(\frac{\partial}{\partial\Delta}\right)^{n-1}\int_0^\infty {\rm d}\beta\int\frac{{\rm d}^dk_{\rm E}}{(2\pi)^d}{\rm e}^{-\beta(k_{\rm E}^2+\Delta)}. \tag{O.5}$$

The interchange of the orders of the β and $k_{\rm E}$ integrations is permissible since I_d is convergent. The $k_{\rm E}$ integrals are, in fact, a series of Gaussians:

$$\begin{aligned}\int\frac{{\rm d}^dk_{\rm E}}{(2\pi)^d}{\rm e}^{-\beta(k_{\rm E}^2+\Delta)} &= {\rm e}^{-\beta\Delta}\left\{\prod_{j=1}^d\int\frac{{\rm d}k_j}{(2\pi)}{\rm e}^{-\beta k_j^2}\right\}\\ &= \frac{{\rm e}^{-\beta\Delta}}{(2\pi)^d}\left(\frac{\pi}{\beta}\right)^{d/2}.\end{aligned} \tag{O.6}$$

Hence

$$\begin{aligned}I_d &= \frac{-{\rm i}}{(n-1)!}\frac{1}{(4\pi)^{d/2}}\left(\frac{\partial}{\partial\Delta}\right)^{n-1}\int{\rm d}\beta{\rm e}^{-\beta\Delta}\beta^{-d/2}\\ &= \frac{-{\rm i}}{(n-1)!}\frac{(-1)^{n-1}}{(4\pi)^{d/2}}\int{\rm d}\beta{\rm e}^{-\beta\Delta}\beta^{n-(d/2)-1}.\end{aligned} \tag{O.7}$$

The last integral can be written in terms of Euler's integral for the *gamma function* $\Gamma(z)$ defined by (see, for example, Boas 1983, chapter 11)

$$\Gamma(z)=\int_0^\infty x^{z-1}{\rm e}^{-x}{\rm d}x. \tag{O.8}$$

Since $\Gamma(n)=(n-1)!$, it is convenient to write (O.8) entirely in terms of Γ functions as

$$I_d={\rm i}\frac{(-1)^n}{(4\pi)^{d/2}}\frac{\Gamma(n-d/2)}{\Gamma(n)}\Delta^{(d/2)-n}. \tag{O.9}$$

Equation (O.9) gives an explicit definition of I_d which can be used for any value of d, not necessarily an integer. As a function of z, $\Gamma(z)$ has isolated poles (see appendix F of volume 1) at $z=0,-1,-2,\ldots$. The behaviour near $z=0$ is given by

$$\Gamma(z)=\frac{1}{z}-\gamma+O(z), \tag{O.10}$$

where γ is the Euler–Mascheroni constant having the value $\gamma\approx 0.5772$. Using

$$z\Gamma(z)=\Gamma(z+1), \tag{O.11}$$

we find the behaviour near $z = -1$:

$$\begin{aligned}\Gamma(-1+z) &= \frac{-1}{1-z}\Gamma(z) \\ &= -[\frac{1}{z} + 1 - \gamma + O(z)]; \end{aligned} \tag{O.12}$$

similarly

$$\Gamma(-2+z) = \frac{1}{2}[\frac{1}{z} + \frac{3}{2} - \gamma + O(z)]. \tag{O.13}$$

Consider now the case $n = 2$, for which $\Gamma(n - d/2)$ in (O.9) will have a pole at $d = 4$. Setting $d = 4 - \epsilon$, the divergent behaviour is given by

$$\Gamma(2 - d/2) = \frac{2}{\epsilon} - \gamma + O(\epsilon) \tag{O.14}$$

from (O.10). $I_d(\Delta, 2)$ is then given by

$$I_d(\Delta, 2) = \frac{\mathrm{i}}{(4\pi)^{2-\epsilon/2}} \Delta^{-\epsilon/2} \left[\frac{2}{\epsilon} - \gamma + O(\epsilon)\right]. \tag{O.15}$$

When $\Delta^{-\epsilon/2}$ and $(4\pi)^{-2+\epsilon/2}$ are expanded in powers of ϵ, for small ϵ, the terms linear in ϵ will produce terms independent of ϵ when multiplied by the ϵ^{-1} in the bracket of (O.15). Using $x^\epsilon \approx 1 + \epsilon \ln x + O(\epsilon^2)$ we find

$$I_d(\Delta, 2) = \frac{\mathrm{i}}{(4\pi)^2} \left[\frac{2}{\epsilon} - \gamma + \ln 4\pi - \ln \Delta + O(\epsilon)\right]. \tag{O.16}$$

Another source of ϵ-dependence arises from the fact (see problem 15.7) that a gauge coupling which is dimensionless in $d = 4$ dimensions will acquire mass dimension $\mu^{\epsilon/2}$ in $d = 4 - \epsilon$ dimensions (check this!). A vacuum polarization loop with two powers of the coupling will then contain a factor μ^ϵ. When expanded in powers of ϵ, this will convert the $\ln \Delta$ in (O.16) to $\ln(\Delta/\mu^2)$.

Renormalization schemes will subtract the explicit pole pieces (which diverge as $\epsilon \to 0$), but may also include in the subtraction certain finite terms as well. For example, in the 'minimal subtraction' (MS) scheme, one subtracts just the pole pieces; in the 'modified minimal subtraction' or $\overline{\mathrm{MS}}$ ('emm-ess-bar') scheme (Bardeen *et al.* 1978) one subtracts the pole and the '$-\gamma + \ln 4\pi$' piece.

The change from one scheme 'A' to another 'B' must involve a finite renormalization of the form (Ellis *et al.* 1966, section 2.5)

$$\alpha_{\mathrm{s}}^{\mathrm{B}} = \alpha_{\mathrm{s}}^{\mathrm{A}}(1 + A_1 \alpha_{\mathrm{s}}^{\mathrm{A}} + \ldots). \tag{O.17}$$

Note that this implies that the first two coefficients of the β function are unchanged under this transformation, so that they are scheme-independent. Subsequent coefficients are scheme-dependent, as is the QCD parameter Λ

introduced in section 15.3. From (15.54) the two corresponding values of Λ are related by

$$\ln\left(\frac{\Lambda_{\rm B}}{\Lambda_{\rm A}}\right) = \frac{1}{2}\int_{\alpha_{\rm s}^{\rm A}(|q^2|)}^{\alpha_{\rm s}^{\rm B}(|q^2|)} \frac{{\rm d}x}{\beta_0 x^2(1+\ldots)} \tag{O.18}$$

$$= \frac{A_1}{2\beta_0} \tag{O.19}$$

where we have taken $|q^2| \to \infty$ in (O.18) since the left-hand side is independent of $|q^2|$. Hence the relationship between the Λ's in different schemes is determined by the one-loop calculation which gives A_1 in (O.19). For example, changing from MS to $\overline{\rm MS}$ gives (problem 15.8)

$$\Lambda^2_{\overline{\rm MS}} = \Lambda^2_{\rm MS}\exp(\ln 4\pi - \gamma), \tag{O.20}$$

as the reader may check.

Finally, consider the integral

$$I_d^{\mu\nu}(\Delta, n) \equiv \int \frac{{\rm d}^d k}{(2\pi)^d}\frac{k^\mu k^\nu}{[k^2 - \Delta + {\rm i}\epsilon]^n}. \tag{O.21}$$

From Lorentz covariance this must be proportional to the only second-rank tensor available, namely $g^{\mu\nu}$:

$$I_d^{\mu\nu} = A g^{\mu\nu}. \tag{O.22}$$

The constant 'A' can be determined by contracting both sides of (O.21) with $g_{\mu\nu}$, using $g^{\mu\nu}g_{\mu\nu} = d$ in d-dimensions. So

$$\begin{aligned}
A &= \frac{1}{d}\int \frac{{\rm d}^d k}{(2\pi)^d}\frac{k^2}{(k^2 - \Delta + {\rm i}\epsilon)^n} \\
&= \frac{1}{d}\left\{\int \frac{{\rm d}^d k}{(2\pi)^d}\frac{1}{(k^2 - \Delta + {\rm i}\epsilon)^{n-1}} + \Delta\int \frac{{\rm d}^d k}{(2\pi)^d}\frac{1}{(k^2 - \Delta + {\rm i}\epsilon)^n}\right\} \\
&= \frac{{\rm i}(-1)^n}{(4\pi)^{d/2}}\frac{\Delta^{(d/2)-n+1}}{d}\left\{\frac{-\Gamma(n-1-d/2)}{\Gamma(n-1)} + \frac{\Gamma(n-d/2)}{\Gamma(n)}\right\} \\
&= \frac{{\rm i}(-1)^n}{(4\pi)^{d/2}}\frac{\Delta^{(d/2)-n+1}}{d}\frac{\Gamma(n-1-d/2)}{\Gamma(n)}\{-n + (n-d/2)\} \\
&= \frac{{\rm i}(-1)^{n-1}\Delta^{(d/2)-n+1}}{(4\pi)^{d/2}}\frac{1}{2}\frac{\Gamma(n-1-d/2)}{\Gamma(n)}.
\end{aligned} \tag{O.23}$$

Using these results, one can show straightforwardly that the gauge-non-invariant part of (11.18) – i.e. the piece in braces – vanishes. With the technique of dimensional regularization, starting from a gauge-invariant formulation of the theory the renormalization programme can be carried out while retaining manifest gauge invariance.

P

Grassmann Variables

In the path integral representation of quantum amplitudes (chapter 16) the fields are regarded as classical functions. Matrix elements of time-ordered products of bosonic operators could be satisfactorily represented (see the discussion following (16.79)). But something new is needed to represent, for example, the time-ordered product of two fermionic operators: there must be a sign difference between the two orderings, since the fermionic operators *anticommute*. Thus it seems that to represent amplitudes involving fermionic operators by path integrals we must think in terms of 'classical' anticommuting variables.

Fortunately, the necessary mathematics was developed by Grassmann in 1855, and applied to quantum amplitudes by Berezin (1966). Any two *Grassmann numbers* θ_1, θ_2 satisfy the fundamental relation

$$\theta_1\theta_2 + \theta_2\theta_1 = 0, \tag{P.1}$$

and of course

$$\theta_1^2 = \theta_2^2 = 0. \tag{P.2}$$

Grassmann numbers can be added and subtracted in the ordinary way, and muliplied by ordinary numbers. For our application, the essential thing we need to be able to do with Grassmann numbers is to integrate over them. It is natural to think that, as with ordinary numbers and functions, integration would be some kind of inverse of differentiation. So let us begin with differentiation.

We define

$$\frac{\partial(a\theta)}{\partial\theta} = a, \tag{P.3}$$

where a is any ordinary number, and

$$\frac{\partial}{\partial\theta_1}(\theta_1\theta_2) = \theta_2\,; \tag{P.4}$$

then necessarily

$$\frac{\partial}{\partial\theta_2}(\theta_1\theta_2) = -\theta_1. \tag{P.5}$$

Consider now a function of one such variable, $f(\theta)$. An expansion of f in powers of θ terminates after only two terms because of the property (P.2):

$$f(\theta) = a + b\theta. \tag{P.6}$$

So

$$\frac{\partial f(\theta)}{\partial \theta} = b, \tag{P.7}$$

but also

$$\frac{\partial^2 f}{\partial \theta^2} = 0 \tag{P.8}$$

for any such f. Hence the operator $\partial/\partial\theta$ has no inverse (think of the matrix analogue $A^2 = 0$: if A^{-1} existed, we could deduce $0 = A^{-1}(A^2) = (A^{-1}A)A = A$ for all A). Thus we must approach Grassmann integration other than via an inverse of differentiation.

We only need to consider integrals over the complete range of θ, of the form

$$\int \mathrm{d}\theta f(\theta) = \int \mathrm{d}\theta (a + b\theta). \tag{P.9}$$

Such an integral should be linear in f; thus it must be a linear function of a and b. One further property fixes its value: we require the result to be *invariant under translations of* θ *by* $\theta \to \theta + \eta$, where η is a Grassmann number. This property is crucial to manipulations made in the path integral formalism, for instance in 'completing the square' manipulations similar to those in section 16.3, but with Grassmann numbers. So we require

$$\int \mathrm{d}\theta (a + b\theta) = \int \mathrm{d}\theta ([a + b\eta] + b\theta). \tag{P.10}$$

This has changed the constant (independent of θ) term, but left the linear term unchanged. The only linear function of a and b which behaves like this is a multiple of b, which is conventionally taken to be simply b. Thus we define

$$\int \mathrm{d}\theta (a + b\theta) = b, \tag{P.11}$$

which means that integration is in some sense the same as differentiation!

When we integrate over products of different θ's, we need to specify a convention about the order in which the integrals are to be performed. We adopt the convention

$$\int \mathrm{d}\theta_1 \int \mathrm{d}\theta_2 \, \theta_2 \theta_1 = 1\,; \tag{P.12}$$

that is, the innermost integral is done first, then the next, and so on.

Since our application will be to Dirac fields, which are complex-valued, we need to introduce complex Grassmann numbers, which are built out of real and imaginary parts in the usual way (this would not be necessary for Majorana fermions). Thus we may define

$$\psi = \frac{1}{\sqrt{2}}(\theta_1 + \mathrm{i}\theta_2), \quad \psi^* = \frac{1}{\sqrt{2}}(\theta_1 - \mathrm{i}\theta_2), \tag{P.13}$$

and then

$$-\mathrm{i}\mathrm{d}\psi \mathrm{d}\psi^* = \mathrm{d}\theta_1 \mathrm{d}\theta_2. \tag{P.14}$$

It is convenient to define complex conjugation to include reversing the order of quantities:

$$(\psi\chi)^* = \chi^*\psi^*. \tag{P.15}$$

Then (P.14) is consistent under complex conjugation.

We are now ready to evaluate some Gaussian integrals over Grassmann variables, which is essentially all we need in the path integral formalism. We begin with

$$\begin{aligned}\int\int \mathrm{d}\psi^*\mathrm{d}\psi\, \mathrm{e}^{-b\psi^*\psi} &= \int\int \mathrm{d}\psi^*\mathrm{d}\psi(1-b\psi^*\psi) \\ &= \int\int \mathrm{d}\psi^*\mathrm{d}\psi(1+b\psi\psi^*) = b. \end{aligned}\tag{P.16}$$

Note that the analogous integral with ordinary variables is

$$\int\int \mathrm{d}x\mathrm{d}y\, \mathrm{e}^{-b(x^2+y^2)/2} = 2\pi/b. \tag{P.17}$$

The important point here is that, in the Grassman case, b appears with a positive, rather than a negative, power. On the other hand, if we insert a factor $\psi\psi^*$ into the integrand in (P.16), we find that it becomes

$$\int\int \mathrm{d}\psi^*\mathrm{d}\psi\, \psi\psi^*(1+b\psi\psi^*) = \int\int \mathrm{d}\psi^*\mathrm{d}\psi\, \psi\psi^* = 1, \tag{P.18}$$

and the insertion has effectively produced a factor b^{-1}. This effect of an insertion is the same in the 'ordinary variables' case:

$$\int\int \mathrm{d}x\mathrm{d}y(x^2+y^2)/2\, \mathrm{e}^{-b(x^2+y^2)/2} = 2\pi/b^2. \tag{P.19}$$

Now consider a Gaussian integral involving two different Grassmann variables:

$$\int \mathrm{d}\psi_1^*\mathrm{d}\psi_1\mathrm{d}\psi_2^*\mathrm{d}\psi_2\, \mathrm{e}^{-\psi^{*\mathrm{T}}M\psi}, \tag{P.20}$$

where

$$\psi = \begin{pmatrix} \psi_1 \\ \psi_2 \end{pmatrix}, \tag{P.21}$$

and M is a 2×2 matrix, whose entries are ordinary numbers. The only terms which survive the integration are those which, in the expansion of the exponential, contain each of $\psi_1^*,\ \psi_1,\ \psi_2^*$ and ψ_2 exactly once. These are the terms

$$\frac{1}{2}\left[M_{11}M_{22}(\psi_1^*\psi_1\psi_2^*\psi_2 + \psi_2^*\psi_2\psi_1^*\psi_1) + M_{12}M_{21}(\psi_1^*\psi_2\psi_2^*\psi_1 + \psi_2^*\psi_1\psi_1^*\psi_2)\right]. \tag{P.22}$$

To integrate (P.22) conveniently, according to the convention (P.12), we need to re-order the terms into the form $\psi_2\psi_2^*\psi_1\psi_1^*$; this produces

$$(M_{11}M_{22} - M_{12}M_{21})(\psi_2\psi_2^*\psi_1\psi_1^*), \tag{P.23}$$

and the integral (P.20) is therefore just

$$\int\int \mathrm{d}\psi_1^*\mathrm{d}\psi\mathrm{d}\psi_2^*\mathrm{d}\psi_2\, \mathrm{e}^{-\psi^{*\mathrm{T}}M\psi} = \det M. \tag{P.24}$$

The reader may show, or take on trust, the obvious generalization to N independent complex Grassmann variables $\psi_1,\ \psi_2,\ \psi_3,\ \ldots,\ \psi_N$. This result is sufficient to establish the assertion made in section 16.4 concerning the integral (16.90), when written in 'discretized' form.

We may contrast (P.24) with an analogous result for two ordinary complex numbers z_1, z_2. In this case we consider the integral

$$\int\int \mathrm{d}z_1^*\mathrm{d}z_1\mathrm{d}z_2^*\mathrm{d}z_2\, \mathrm{e}^{-z^*Hz}, \tag{P.25}$$

where z is a two-component column matrix with elements z_1 and z_2. We take the matrix H to be Hermitian, with positive eigenvalues b_1 and b_2. Let H be diagonalized by the unitary transformation

$$\begin{pmatrix} z_1' \\ z_2' \end{pmatrix} = U \begin{pmatrix} z_1 \\ z_2 \end{pmatrix}, \tag{P.26}$$

with $UU^\dagger = I$. Then

$$\mathrm{d}z_1'\mathrm{d}z_2' = \det U\, \mathrm{d}z_1\mathrm{d}z_2, \tag{P.27}$$

and so

$$\mathrm{d}z_1'\mathrm{d}z_1'^*\mathrm{d}z_2'\mathrm{d}z_2'^* = \mathrm{d}z_1\mathrm{d}z_1^*\mathrm{d}z_2\mathrm{d}z_2^*, \tag{P.28}$$

since $|\det U|^2 = 1$. The integral (P.25) then becomes

$$\int \mathrm{d}z_1'\mathrm{d}z_1'^*\mathrm{e}^{-b_1 z_1'^* z_1'} \int \mathrm{d}z_2'\mathrm{d}z_2'^*\mathrm{e}^{-b_2 z_2'^* z_2'}, \tag{P.29}$$

the integrals converging provided $b_1, b_2 > 0$. Next, setting $z_1 = (x_1 + \mathrm{i}y_1)/\sqrt{2}$, $z_2 = (x_2 + \mathrm{i}y_2)/\sqrt{2}$, (P.29) can be evaulated using (P.17), and the result is proportional to $(b_1 b_2)^{-1}$, which is the *inverse* of the determinant of the matrix H, when diagonalized. Thus – compare (P.16) and (P.17) – Gaussian integrals over complex Grassmann variables are proportional to the determinant of the matrix in the exponent, while those over ordinary complex variables are proportional to the inverse of the determinant.

Returning to integrals of the form (P.20), consider now a two-variable (both complex) analogue of (P.18):

$$\int \mathrm{d}\psi_1^*\mathrm{d}\psi_1\mathrm{d}\psi_2^*\mathrm{d}\psi_2\, \psi_1\psi_2^*\, \mathrm{e}^{-\psi^{*\mathrm{T}}M\psi}. \tag{P.30}$$

This time, only the term $\psi_1^*\psi_2$ in the expansion of the exponential will survive the integration, and the result is just $-M_{12}$. By exploring a similar integral (still with the term $\psi_1\psi_2^*$) in the case of three complex Grassmann variables, the reader should be convinced that the general result is

$$\prod_i \int\!\!\!\!\!\!\sum \mathrm{d}\psi_i^* \mathrm{d}\psi_i \, \psi_k \psi_l^* \, \mathrm{e}^{-\psi^{*\mathrm{T}} M \psi} = (M^{-1})_{kl} \, \mathrm{det} M. \tag{P.31}$$

With this result we can make plausible the fermionic analogue of (16.87), namely

$$\langle \Omega | T \left\{ \psi(x_1) \bar{\psi}(x_2) \right\} | \Omega \rangle = \frac{\int \mathcal{D}\bar{\psi}\mathcal{D}\psi \, \psi(x_1)\bar{\psi}(x_2) \mathrm{exp}[-\int \mathrm{d}^4 x_\mathrm{E} \bar{\psi}(\mathrm{i} \not{\partial} - m)\psi]}{\int \mathcal{D}\bar{\psi}\mathcal{D}\psi \mathrm{exp}[-\int\!\!\!\!\!\!\sum^4 x_\mathrm{E} \bar{\psi}(\mathrm{i} \not{\partial} - m)\psi]}; \tag{P.32}$$

note that $\bar{\psi}$ and ψ^* are unitarily equivalent. The denominator of this expression is[1] $\mathrm{det}(\mathrm{i} \not{\partial} - m)$, while the numerator is this same determinant multiplied by the inverse of the operator $(\mathrm{i} \not{\partial} - m)$; but this is just $(\not{p} - m)^{-1}$ in momentum space, the familiar Dirac propagator.

[1] The reader may interpret this as a finite-dimensional determinant, after discretization.


Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

Q

Feynman Rules for Tree Graphs in QCD and the Electroweak Theory

Q.1 QCD

Q.1.1 External particles

Quarks

The SU(3) colour degree of freedom is not written explicitly; the spinors have 3 (colour) × 4 (Dirac) components. For each fermion or antifermion line entering the graph include the spinor

$$u(p,s) \qquad \text{or} \qquad v(p,s) \tag{Q.1}$$

and for spin-$\frac{1}{2}$ particles leaving the graph the spinor

$$\bar{u}(p',s') \qquad \text{or} \qquad \bar{v}(p',s'), \tag{Q.2}$$

as for QED.

Gluons

Besides the spin-1 polarization vector, external gluons also have a 'colour polarization' vector $a^c (c = 1, 2, \ldots, 8)$ specifying the particular colour state involved. For each gluon line entering the graph include the factor

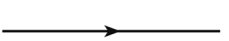
$$\epsilon_\mu(k,\lambda)\, a^c \tag{Q.3}$$

and for gluons leaving the graph the factor

$$\epsilon^*_\mu(k',\lambda')\, a^{c*}. \tag{Q.4}$$

Q.1.2 Propagators

Quark



$$= \frac{\mathrm{i}}{\not{p} - m} = \mathrm{i}\frac{\not{p} + m}{p^2 - m^2}. \tag{Q.5}$$

Gluon

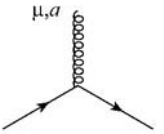
$$= \frac{\mathrm{i}}{k^2}\left(-g^{\mu\nu} + (1-\xi)\frac{k^\mu k^\nu}{k^2}\right)\delta^{ab} \qquad \text{(Q.6)}$$

for a general ξ gauge. Calculations are usually performed in Lorentz or Feynman gauge with $\xi = 1$ and gluon propagator equal to

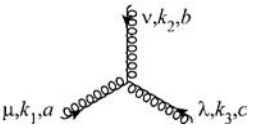
$$= \mathrm{i}\frac{(-g^{\mu\nu})\delta^{ab}}{k^2}. \qquad \text{(Q.7)}$$

Here a and b run over the 8 colour indices $1, 2, \ldots, 8$.

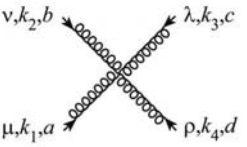
Q.1.3 Vertices



$$-\mathrm{i}g_\mathrm{s}\frac{\lambda^a}{2}\gamma_\mu$$



$$-g_\mathrm{s} f_{abc}[g_{\mu\nu}(k_1-k_2)_\lambda + g_{\nu\lambda}(k_2-k_3)_\mu + g_{\lambda\mu}(k_3-k_1)_\nu]$$



$$-\mathrm{i}g_\mathrm{s}^2[f_{abe}f_{cde}(g_{\mu\lambda}g_{\nu\rho} - g_{\mu\rho}g_{\nu\lambda}) + f_{ade}f_{bce}(g_{\mu\nu}g_{\lambda\rho} - g_{\mu\lambda}g_{\nu\rho}) + f_{ace}f_{dbe}(g_{\mu\rho}g_{\nu\lambda} - g_{\mu\nu}g_{\lambda\rho})]$$

It is important to remember that the rules given above are only adequate for tree diagram calculations in QCD (see section 13.3.3).

Q.2 The electroweak theory

For tree graph calculations, it is convenient to use the U gauge Feynman rules (sections 19.5 and 19.6) in which no unphysical particles appear. These U gauge rules are given below for the leptons $l = (\mathrm{e}, \mu, \tau)$, $\nu_l = (\nu_\mathrm{e}, \nu_\mu, \nu_\tau)$; for

the $t_3 = +1/2$ quarks denoted by f, where $\mathrm{f} = \mathrm{u}, \mathrm{c}, \mathrm{t}$; and for the $t_3 = -1/2$ CKM-mixed quarks denoted by f′ where $\mathrm{f}' = \mathrm{d}', \mathrm{s}', \mathrm{b}'$.

Note that for simplicity we do not include neutrino flavour mixing.

Q.2.1 External particles

Leptons and quarks

For each fermion or antifermion line entering the graph include the spinor

$$u(p, s) \qquad \text{or} \qquad v(p, s) \tag{Q.8}$$

and for spin-$\frac{1}{2}$ particles leaving the graph the spinor

$$\bar{u}(p', s') \qquad \text{or} \qquad \bar{v}(p', s'). \tag{Q.9}$$

Vector bosons

For each vector boson line entering the graph include the factor

$$\epsilon_\mu(k, \lambda) \tag{Q.10}$$

and for vector bosons leaving the graph the factor

$$\epsilon^*_\mu(k', \lambda'). \tag{Q.11}$$

Q.2.2 Propagators

Leptons and quarks

$$= \frac{\mathrm{i}}{\not{p} - m} = \mathrm{i}\frac{\not{p} + m}{p^2 - m^2}. \tag{Q.12}$$

Vector bosons (U gauge)

$$= \frac{\mathrm{i}}{k^2 - M_{\mathrm{V}}^2}(-g_{\mu\nu} + k_\mu k_\nu/m_{\mathrm{V}}^2) \tag{Q.13}$$

where ‘V’ stands for either ‘W’ (the W-boson) or ‘Z’ (the Z^0).

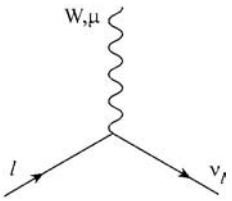
Higgs particle

$$= \frac{\mathrm{i}}{p^2 - m_{\mathrm{H}}^2} \tag{Q.14}$$

Q.2.3 Vertices

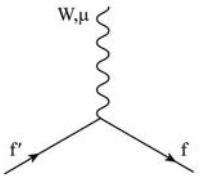
Charged current weak interactions

Leptons



$$-\mathrm{i}\frac{g}{\sqrt{2}}\gamma_\mu\frac{1-\gamma_5}{2}$$

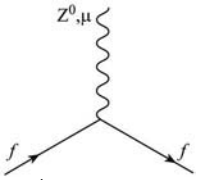
Quarks



$$-\mathrm{i}\frac{g}{\sqrt{2}}\gamma_\mu\frac{1-\gamma_5}{2}V_{\mathrm{f\,f'}}$$

Neutral current weak interactions (no neutrino mixing)

Fermions



$$\frac{-\mathrm{i}g}{\cos\theta_\mathrm{W}}\gamma_\mu\left(c_\mathrm{L}^f\frac{1-\gamma_5}{2}+c_\mathrm{R}^f\frac{1+\gamma_5}{2}\right),$$

where

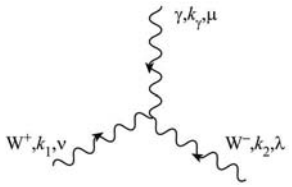
$$c_\mathrm{L}^f = t_3^f - \sin^2\theta_\mathrm{W}Q_f \tag{Q.15}$$

$$c_\mathrm{R}^f = -\sin^2\theta_\mathrm{W}Q_f, \tag{Q.16}$$

and f stands for any fermion.

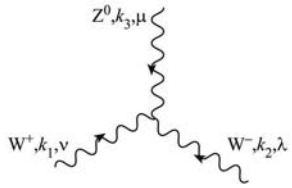
Vector boson couplings

(i) Trilinear couplings:
$\gamma\mathrm{W}^+\mathrm{W}^-$ vertex



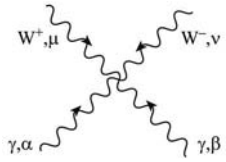
$$\mathrm{i}e[g_{\nu\lambda}(k_1 - k_2)_\mu + g_{\lambda\mu}(k_2 - k_\gamma)_\nu + g_{\mu\nu}(k_\gamma - k_1)_\lambda]$$

$\mathrm{Z}^0\mathrm{W}^+\mathrm{W}^-$ vertex

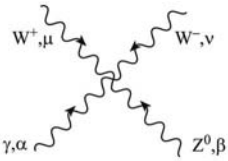


$$\mathrm{i}g\cos\theta_\mathrm{W}[g_{\nu\lambda}(k_1 - k_2)_\mu + g_{\lambda\mu}(k_2 - k_3)_\nu + g_{\mu\nu}(k_3 - k_1)_\lambda]$$

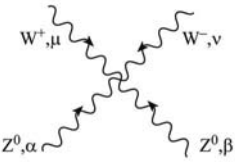
(ii) Quadrilinear couplings:



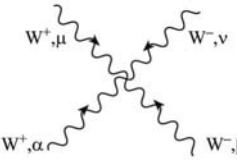
$$-\mathrm{i}e^2(2g_{\alpha\beta}g_{\mu\nu} - g_{\alpha\mu}g_{\beta\nu} - g_{\alpha\nu}g_{\beta\mu})$$



$$-\mathrm{i}eg\cos\theta_\mathrm{W}(2g_{\alpha\beta}g_{\mu\nu} - g_{\alpha\mu}g_{\beta\nu} - g_{\alpha\nu}g_{\beta\mu})$$



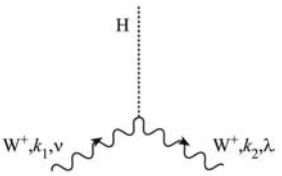
$$-\mathrm{i}g^2\cos^2\theta_\mathrm{W}(2g_{\alpha\beta}g_{\mu\nu} - g_{\alpha\mu}g_{\beta\nu} - g_{\alpha\nu}g_{\beta\mu})$$



$$\mathrm{i}g^2(2g_{\mu\alpha}g_{\nu\beta} - g_{\mu\beta}g_{\alpha\nu} - g_{\mu\nu}g_{\alpha\beta})$$

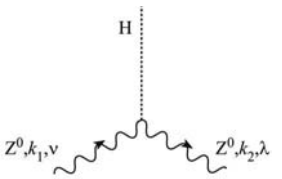
Higgs couplings

(i) Trilinear couplings
HW^+W^- vertex



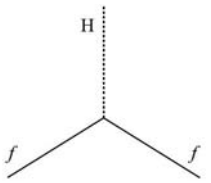
$$igM_{\mathrm{W}}g_{\nu\lambda}$$

HZ^0Z^0 vertex



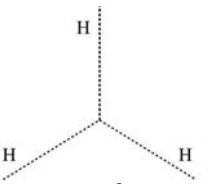
$$\frac{ig}{\cos\theta_{\mathrm{W}}}M_{\mathrm{Z}}g_{\nu\lambda}$$

Fermion Yukawa couplings (fermion mass m_f)



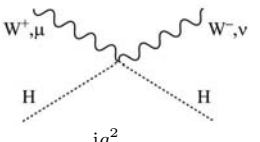
$$-i\frac{g}{2}\frac{m_f}{M_{\mathrm{W}}}$$

Trilinear self-coupling



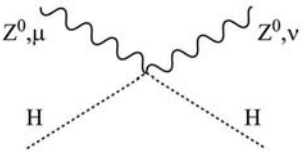
$$-i\frac{3m_{\mathrm{H}}^2 g}{2M_{\mathrm{W}}}$$

(ii) Quadrilinear couplings:
HHW^+W^- vertex



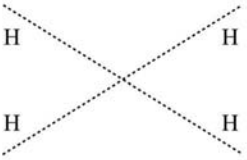
$$\frac{ig^2}{2}g_{\mu\nu}$$

HHZZ vertex



$$\frac{\mathrm{i}g^2}{2\cos^2\theta_{\mathrm{W}}}g_{\mu\nu}$$

Quadrilinear self-coupling



$$-\frac{\mathrm{i}3m_{\mathrm{H}}^2g^2}{4M_{\mathrm{W}}^2}$$


Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

References

Aad G *et al.* 2012a (ATLAS Collaboration) *Phys. Lett.* B **710** 49
——2012b *Phys. Lett.* B **716** 1
Aaij R *et al.* 2012 (LHCb Collaboration) *Phys. Rev. Lett.* **108** 111602
Aaltonen T *et al.* 2008 (CDF Collaboration) *Phys. Rev. Lett.* **100** 121802
——2012 (CDF and D0 Collaborations) *Phys. Rev. Lett.* **109** 071804
Abachi S *et al.* 1995a (D0 Collaboration) *Phys. Rev. Lett.* **74** 2422
——1995b *Phys. Rev. Lett.* **74** 2632
Abashian A *et al.* 2001 (Belle Collaboration) *Phys. Rev. Lett.* **86** 2509
Abbiendi G *et al.* 2001 *Eur. Phys. J.* C **20** 601
Abdurashitov J N *et al.* 2009 *Phys. Rev.* C **80** 015807
Abe F *et al.* 1994a (CDF Collaboration) *Phys. Rev.* D **50** 2966
——1994b *Phys. Rev. Lett.* **73** 225
——1995a *Phys. Rev.* D **52** 4784
——1995b *Phys. Rev. Lett.* **74** 2626
Abe K 1991 *Proc. 25th Int. Conf. on High Energy Physics* eds K K Phua and Y Yamaguchi (Singapore: World Scientific) p 33
Abe K *et al.* 2000 *Phys. Rev. Lett.* **84** 5945
——2011 (T2K Collaboration) *Phys. Rev. Lett.* **107** 041801
Abe S *et al.* 2008 (KamLAND Collaboration) *Phys. Rev. Lett.* **100** 221803
Abramovicz H *et al.* 1982a *Z. Phys.* C **12** 289
——1982b *Z. Phys.* C **13** 199
——1983 *Z. Phys.* C **17** 283
Abreu P *et al.* 1990 *Phys. Lett.* B **242** 536
Adamson P *et al* 2011 (MINOS Collaboration) *Phys. Rev. Lett.* **106** 181801
Adler S L 1963 *Phys. Rev.* **143** 1144
——1965 *Phys. Rev. Lett.* **14** 1051
——1969 *Phys. Rev.* **177** 2426
——1970 *Lectures on Elementary Particles and Quantum Field Theory (Proceedings of the Brandeis Summer Institute)* vol 1, ed S Deser *et al.* (Boston, MA: MIT)
Adler S L and Bardeen W A 1969 *Phys. Rev.* **182** 1517
Aharmim B *et al.* 2010 (SNO Collaboration) *Phys. Rev.* C **81** 055504
Ahmad Q R *et al.* 2001 (SNO Collaboration) *Phys. Rev. Lett.* **87** 071301
——2002 *Phys. Rev. Lett.* **89** 011301
Ahn J K *et al.* 2012 (RENO Collaboration) *Phys. Rev. Lett.* **108** 191802
Ahn M H *et al.* 2006 (K2K Collaboration) *Phys. Rev.* D **74** 072003
Aitchison I J R 2007 *Supersymmetry in Particle Physics An Elementary Introduction* (Cambridge: Cambridge University Press)
Aitchison I J R *et al.* 1995 *Phys. Rev.* B **51** 6531
Akhundov A A *et al.* 1986 *Nucl. Phys.* B **276** 1
Akrawy M Z *et al.* 1990 (OPAL Collaboration) *Phys. Lett.* B **235** 389

Alavi-Harati A *et al.* 2003 (KTeV Collaboration) *Phys. Rev.* D **67** 012005; *ibid.* D **70** 079904 (erratum)
Ali A and Kramer G 2011 *Eur. Phys. J.* H **36** 245
Allaby J *et al.* 1988 *J. Phys. C: Solid State Phys.* **38** 403
Allasia D *et al.* 1984 *Phys. Lett.* B **135** 231
——1985 *Z. Phys.* C **28** 321
Allton C R *et al.* 1995 *Nucl. Phys.* B **437** 641
——2002 (UKQCD Collaboration) *Phys. Rev.* D **65** 054502
Alper B *et al.* 1973 *Phys. Lett.* B **44** 521
Altarelli G 1982 *Phys. Rep.* **81** 1
Altarelli G and Parisi G 1977 *Nucl. Phys.* B **126** 298
Altarelli G *et al.* 1978a *Nucl. Phys.* B **143** 521
——1978b *Nucl. Phys.* B **146** 544(E)
——1979 *Nucl. Phys.* B **157** 461
——1989 Z. Phys. *at LEP-1* CERN 89-08 (Geneva)
Altman M *et al.* 2005 *Phys. Lett.* B **616** 174
Amaudruz P *et al.* 1992 (NMC Collaboration) *Phys. Lett.* B **295** 159
An F P *et al.* 2012 (Daya Bay Collaboration) *Phys. Rev. Lett.* **108** 171803
Anderson P W 1963 *Phys. Rev.* **130** 439
Andreotti E *et al.* (CUORCINO Collaboration) 2011 *Astropart. Phys.* **34** 822
Antoniadis I 2002 *2001 European School of High Energy Physics* ed N Ellis and J March-Russell CERN 2002-002 (Geneva) pp 301ff
Apollonio M *et al.* 2003 (CHOOZ Collaboration) *Eur. Phys. J.* C **27** 331
Appel J A *et al.* 1986 *Z. Phys.* C **30** 341
Appelquist T and Chanowitz M S 1987 *Phys. Rev. Lett.* **59** 2405
Arnison G *et al.* 1983a *Phys. Lett.* B **122** 103
——1983b *Phys. Lett.* B **126** 398
——1984 *Phys. Lett.* B **136** 294
——1985 *Phys. Lett.* B **158** 494
——1986 *Phys. Lett.* B **166** 484
Arnold R *et al.* (NEMO Collaboration) 2006 *Nucl. Phys.* A **765** 483
Attwood D *et al.* 2001 *Phys. Rev.* D **63** 036005
Aubert B *et al.* 2001 (BaBar Collaboration) *Phys. Rev. Lett.* **86** 2515
——2004 *Phys. Rev. Lett.* **93** 131801
——2007a *Phys. Rev. Lett.* **99** 171803
——2007b *Phys. Rev.* D **76** 102004
——2007c *Phys. Rev. Lett.* **98** 211802
——2008 *Phys. Rev.* D **78** 034023
——2010 *Phys. Rev. Lett.* **105** 121801
Aubin C *et al.* 2004 *Phys. Rev.* D **70** 09505
Ayres D S (NOνA Collaboration) 1995 arXiv:hep-ex/0503053
Bagnaia P *et al.* 1983 *Phys. Lett.* B **129** 130
——1984 *Phys. Lett.* B **144** 283
Bahcall J N *et al.* 1968 *Phys. Rev. Lett.* **20** 1209
——2001 *Astrophys. J.* **555** 990
——2005 *ibid.* **621** L85
Baikov P A *et al.* 2008 *Phys. Rev. Lett.* **101** 012002
——2009 *Nucl. Phys. Proc. Suppl.* **189** 49
Bailin D 1982 *Weak Interactions* (Bristol: Adam Hilger)

Banks T *et al.* 1976 *Phys. Rev.* D **13** 1043
Banner M *et al.* 1982 *Phys. Lett.* B **118** 203
——1983 *Phys. Lett.* B **122** 476
Barber D P *et al.* 1979 *Phys. Rev. Lett.* **43** 830
Bardeen J, Cooper L N and Schrieffer J R 1957 *Phys. Rev.* **108** 1175
Bardeen W A, Fritzsch H and Gell-Mann M 1973 in *Scale and Conformal Symmetry in Hadron Physics* ed R Gatto (New York: Wiley) pp 139-151
Bardeen W A *et al.* 1978 *Phys. Rev.* D **18** 3998
Bardin D Yu *et al.* 1989 *Z. Phys.* C **44** 493
Barezyk A *et al.* 2008 (FAST Collaboration) *Phys. Lett.* B **663** 172
Barger V *et al.* 1983 *Z. Phys.* C **21** 99
Barr G D *et al.* 1993 (NA31 Collaboration) *Phys. Lett.* B **317** 233
Bartel *et al.* 1986 (JADE Collaboration) *Z. Phys.* C **33** 23
Batley J R *et al.* 2002 (NA48 Collaboration) *Phys. Lett.* B **544** 97
Beenakker W and Hollik W 1988 *Z. Phys.* C **40** 569
Belavin A A *et al.* 1975 *Phys. Lett.* B **59** 85
Bell J S and Jackiw R 1969 *Nuovo Cimento* A **60** 47
Benvenuti A C *et al.* 1989 (BCDMS Collaboration) *Phys. Lett.* B **223** 485
Berends F A *et al.* 1981 *Phys. Lett.* B **103** 124
Berezin F A 1966 *The Method of Second Quantisation* (New York: Academic)
Bergsma F *et al.* 1983 *Phys. Lett.* B **122** 465
Beringer J *et al.* 2012 (Partial Data Group) *Phys. Rev.* D **86** 010001
Berman S M, Bjorken J D and Kogut J B 1971 *Phys. Rev.* D **4** 3388
Bernard C W, Golterman M F and Shamir Y 2006 *Phys. Rev.* D **73** 114511
Bernreuther W and Wetzel W 1982 *Nucl. Phys.* B **197** 128
Bernstein J 1974 *Rev. Mod. Phys.* **46** 7
Bethke S 2009 *Eur. Phys. J.* C **64** 689
Bethke S *et al.* 1988 (JADE Collaboration) *Phys. Lett.* B **213** 235
Bettini A 2008 *Introduction to Elementary Particle Physics* (Cambridge: Cambridge University Press)
Bigi I I and Sanda A I 2000 *CP Violation* (Cambridge: Cambridge University Press)
Bijnens J *et al.* 1996 *Phys. Lett.* B **374** 2010
Binney J J *et al.* 1992 *The Modern Theory of Critical Phenomena* (Oxford: Clarendon)
Bjorken J D 1973 *Phys. Rev.* D**8** 4098
Bjorken J D and Glashow S L 1964 *Phys. Lett.* **11** 255
Blatt J M 1964 *Theory of Superconductivity* (New York: Academic)
Bloch F and Nordsieck A 1937 *Phys. Rev.* **52** 54
Bludman S A 1958 *Nuovo Cimento* **9** 443
Boas M L 1983 *Mathematical Methods in the Physical Sciences* (New York: Wiley)
Bogoliubov N N 1947 *J. Phys. USSR* **11** 23
——1958 *Nuovo Cimento* **7** 794
Bogoliubov N N *et al.* 1959 *A New Method in the Theory of Superconductivity* (New York: Consultants Bureau, Inc.)
Bouchiat C C *et al.* 1972 *Phys. Lett.* B **38** 519
Bouchiendra R *et al.* 2011 *Phys. Rev. Lett.* **106** 080801
Branco G C *et al.* 1999 *CP Violation* (Oxford: Oxford University Press)
Brandelik R *et al.* 1979 *Phys. Lett.* B **86** 243
Brodsky S J, Lepage G P and Mackenzie P B 1983 *Phys. Rev.* D **28** 228

Budny R 1975 *Phys. Lett.* B **55** 227
Buras A J *et al.* 1994 *Phys. Rev.* D **50** 3433
Büsser F W *et al.* 1972 *Proc. XVI Int. Conf. on High Energy Physics (Chicago, IL)* vol 3 (Batavia: FNAL)
——1973 *Phys. Lett.* B **46** 471
Cabibbo N 1963 *Phys. Rev. Lett.* **10** 531
Cabibbo N *et al.* 1979 *Nucl. Phys.* B **158** 295
Cacciari M *et al.* 2008 *JHEP* **0804** 063
Callan C G 1970 *Phys. Rev.* D **2** 1541
Callan C, Coleman S, Wess J, and Zumino B 1969 *Phys. Rev.* **177** 2247
Campbell J M *et al.* 2007 *Rept. on Prog. in Phys.* **70** 89
Carruthers P A 1966 *Introduction to Unitary Symmetry* (New York: Wiley)
Carter A B and Sanda A I 1980 *Phys. Rev. Lett.* **45** 952
——1981 *Phys. Rev.* D **23** 1567
Caswell W E 1974 *Phys. Rev. Lett.* **33** 244
Catani S *et al.* 1991 *Phys. Lett.* B **269** 432
——1993 *Nucl. Phys.* B **406** 187
Ceccucci A *et al.* 2010 in Nakamura K *et al.* 2010
Celmaster W and Gonsalves R J 1980 *Phys. Rev. Lett.* **44** 560
Chadwick J 1932 *Proc. R. Soc.* A **136** 692
Chanowitz M *et al.* 1978 *Phys. Lett.* B **78** 285
——1979 *Nucl. Phys.* B **153** 402
Chao Y *et al.* 2005 (Belle Collaboration) *Phys. Rev.* D **71** 031502
Charles J *et al.* 2005 (CKMfitter Group) *Eur. Phys. J.* C **41** 1
Chatrchyan S *et al.* 2012a (CMS Collaboration) *Phys. Lett.* B **710** 26
——2012b *Phys. Lett.* B **716** 30
Chau L L and Keung W Y 1984 *Phys. Rev. Lett.* **53** 1802
Chen M-S and Zerwas P 1975 *Phys. Rev.* D **12** 187
Cheng T-P and Li L-F 1984 *Gauge Theory of Elementary Particle Physics* (Oxford: Clarendon)
Chetyrkin K G *et al.* 1979 *Phys. Lett.* B **85** 277
——1997 *Phys. Rev. Lett.* **79** 2184
Chitwood D B *et al.* 2007 (MuLan Collaboration) *Phys. Rev. Lett.* **99** 032001
Christenson J H *et al* 1964 *Phys. Rev. Lett.* **13** 138
Cleveland B T *et al.* 1998 *Astrophys. J.* **496** 505
Cohen A G *et al.* 2009 *Phys. Lett.* B **678** 191
Colangelo G *et al.* 2001 *Nucl. Phys.* B **603** 125
Coleman S 1985 *Aspects of Symmetry* (Cambridge: Cambridge University Press)
——1966 *J. Math. Phys.* **7** 787
Coleman S and Gross D J 1973 *Phys. Rev. Lett.* 31 851
Coleman S, Wess J and Zumino B 1969 *Phys. Rev.* **177** 2239
Collins J C and Soper D E 1987 *Annu. Rev. Nucl. Part. Sci.* **37** 383
——1998 *Phys. Lett.* B **438** 184
Collins P D B and Martin A D 1984 *Hadron Interactions* (Bristol: Adam Hilger)
Combridge B L *et al.* 1977 *Phys. Lett.* B **70** 234
Combridge B L and Maxwell C J 1984 *Nucl. Phys.* B **239** 429

Commins E D and Bucksbaum P H 1983 *Weak Interactions of Quarks and Leptons* (Cambridge: Cambridge University Press)
Consoli M *et al.* 1989 Z. Phys. *at LEP-I* ed G Altarelli *et al.* CERN 89-08 (Geneva)
Cooper L N 1956 *Phys. Rev.* **104** 1189
Cornwall J M *et al* 1974 *Phys. Rev.* D 10 1145
Cowan C L *et al.* 1956 *Science* **124** 103
Czakon M 2005 *Nucl. Phys.* B **710** 485
Dalitz R H 1953 *Phil. Mag.* **44** 1068
——1965 *High Energy Physics* ed C de Witt and M Jacob (New York: Gordon and Breach)
Danby G *et al.* 1962 *Phys. Rev. Lett.* **9** 36
Davies C T H *et al.* 2008 (HPQCD Collaboration) *Phys. Rev.* D **78** 114507
Davies C T H *et al.* 2004 (HPQCD, UKQCD, MILC and Fermilab Collaborations) *Phys. Rev. Lett.* **92** 022001
Davis R 1955 *Phys. Rev.* **97** 766
——1964 *Phys. Rev. Lett.* **12** 303
Davis R *et al.* 1968 *Phys. Rev. Lett.* **20** 1205
Dawson S *et al.* 1990 *The Higgs Hunters Guide* (Reading, MA: Addison-Wesley)
de Groot J G H *et al.* 1979 *Z. Phys.* C **1** 143
DiLella L 1985 *Annu. Rev. Nucl. Part. Sci.* **35** 107
——1986 *Proc. Int. Europhysics Conf. on High Energy Physics, Bari, Italy, July 1985* eds L Nitti and G Preparata (Bari: Laterza) pp 761ff
Dine M and Sapirstein J 1979 *Phys. Rev. Lett.* **43** 668
Dirac P A M 1931 *Proc. R. Soc.* A **133** 60
Dittmaier S *et al.* 2011 (LHC Higgs Cross section Working Group Collaboration) *Handbook of LHC Higgs Cross sections: 1. Inclusive Observables* CERN-2011-002 (arXiv:1101.0593 [hep-ph])
——2012 *Handbook of LHC Higgs Cross Sections: 2. Differential Distributions* CERN-2012-002 (arXiv:1201.3084 [hep-ph])
Dokshitzer Yu L 1977 *Sov. Phys.JETP* **46** 641
Donoghue J F, Golowich E and Holstein B R 1992 *Dynamics of the Standard Model* (Cambridge: Cambridge University Press)
Duke D W and Owens J F 1984 *Phys. Rev.* D **30** 49
Dürr S *et al.* (Budapest-Marseille-Wuppertal Collaboration) *Science* **322** 1224
Eden R J, Landshoff P V, Olive D I and Polkinghorne J C 1966 *The Analytic S-Matrix* (Cambridge: Cambridge University Press)
Eichten E *et al.* 1980 *Phys. Rev.* D **21** 203
Einhorn M B and Wudka J 1989 *Phys. Rev.* D **39** 2758
Eitel K *et al.* 2005 *Nucl. Phys. (Proc. Suppl.)* B **143** 197
Elias-Miro J *et al.* 2012 *Phys. Lett.* B **709** 222
Ellis J *et al.* 1976 *Nucl. Phys.* B **111** 253
——1977 Erratum *ibid.* B **130** 516
——1994 *Phys. Lett.* B **333** 118
Ellis R K, Stirling W J and Webber B R 1996 *QCD and Collider Physics* (Cambridge: Cambridge University Press)
Ellis S D and Soper D E 1993 *Phys. Rev.* D **48** 3160
Ellis S D *et al.* 2008 *Prog. Part. Nucl. Phys.* **60** 484
Englert F and Brout R 1964 *Phys. Rev. Lett.* **13** 321

Enz C P 1992 *A Course on Many-Body Theory Applied to Solid-State Physics (World Scientific Lecture Notes in Physics 11)* (Singapore: World Scientific)
——2002 *No Time to be Brief* (Oxford: Oxford University Press)
Fabri E and Picasso L E 1966 *Phys. Rev. Lett.* **16** 408
Faddeev L D and Popov V N 1967 *Phys. Lett.* B **25** 29
Feinberg G *et al* 1959 *Phys. Rev. Lett.* **3** 527, especially footnote 9
Fermi E 1934a *Nuovo Cimento* **11** 1
——1934b *Z. Phys.* **88** 161
Feynman R P 1963 *Acta Phys. Polon.* **26** 697
——1977 in *Weak and Electromagnetic Interactions at High Energies* ed R Balian and C H Llewellyn Smith (Amsterdam: North-Holland) p 121
Feynman R P and Gell-Mann M 1958 *Phys. Rev.* **109** 193
Feynman R P and Hibbs A R 1965 *Quantum Mechanics and Path Intergrals* (New York: McGraw-Hill)
Fritzsch H and Gell-Mann M 1972 *Proc. XVI Int. Conf. on High Energy Physics, Batavia IL* eds J D Jackson and R G Roberts, pp 135-165
Fritzsch H, Gell-Mann M and Leutwyler H 1973 *Phys. Lett.* B **47** 365
Fukuda Y *et al.* 1998 *Phys. Rev. Lett.* **81** 1562
Fukugita M and Yanagida T 1986 *Phys. Lett.* B **174** 45
Gaillard M K and Lee B W 1974 *Phys. Rev.* D **10** 897
Gamow G and Teller E 1936 *Phys. Rev.* **49** 895
Gasser J and Leutwyler H 1982 *Phys. Rep.* **87** 77
——1984 *Ann. Phys.* **158** 142
——1985 *Nucl. Phys.* B **250** 465
Geer S 1986 *High Energy Physics 1985, Proc. Yale Theoretical Advanced Study Institute* eds Bowick M J and Gursey F (Singapore: World Scientific)
Gell-Mann M 1961 *California Institute of Technology Report* CTSL-20 (reprinted in Gell-Mann and Ne'eman 1964)
Gell-Mann M *et al.* 1979 *Supergravity* ed D Freedman and P van Nieuwenhuizen (Amsterdam: North-Holland) p 315
Gell-Mann M and Levy M 1960 *Nuovo Cimento* **16** 705
Gell-Mann M and Low F E 1954 *Phys. Rev.* **95** 1300
Gell-Mann M and Ne'eman 1964 *The Eightfold Way* (New York: Benjamin)
Gell-Mann M and Pais A 1955 *Phys. Rev.* **97** 1387
Georgi H *et al.* 1978 *Phys. Rev. Lett.* **40** 692
Georgi H and Politzer H D 1974 *Phys. Rev.* D **9** 416
Gibbons L K 1993 (E731 Collaboration) *Phys. Rev. Lett.* **70** 1203
Ginsparg P and Wilson K G 1982 *Phys. Rev.* D **25** 25
Ginzburg V I and Landau L D 1950 *Zh. Eksp. Teor. Fiz.* **20** 1064
Giri A *et al.* 2003 *Phys. Rev.* D **68** 054018
Glashow S L 1961 *Nucl. Phys.* **22** 579
Glashow S L *et al.* 1978 *Phys. Rev.* D **18** 1724
Glashow S L, Iliopoulos J and Maiani L 1970 *Phys. Rev.* D **2** 1285
Goldberger M L and Treiman S B 1958 *Phys. Rev.* **95** 1300
Goldhaber M *et al* 1958 *Phys. Rev.* **109** 1015
Goldstone J 1961 *Nuovo Cimento* **19** 154
Goldstone J, Salam A and Weinberg S 1962 *Phys. Rev.* **127** 965
Gorishnii S G and Larin S A 1986 *Phys. Lett.* **172** 109
Gorishnii S G *et al.* 1991 *Phys. Lett.* B **259** 144

Gorkov L P 1959 *Zh. Eksp. Teor. Fiz.* **36** 1918
Gottschalk T and Sivers D 1980 *Phys. Rev.* D **21** 102
Gray A *et al* 2005 (HPQCD and UKQCD Collaborations) *Phys. Rev.* D **72** 094507
Greenberg O W 1964 *Phys. Rev. Lett.* **13** 598
Gribov V N and Lipatov L N 1972 *Sov. J. Nucl. Phys.* **15** 438
Gronau M 1991 *Phys. Lett.* B **265** 389
Gronau M and London D 1990 *Phys. Rev. Lett.* **65** 3381
Gross D J and Llewellyn Smith C H 1969 *Nucl. Phys.* B **14** 337
Gross D J and Wilczek F 1973 *Phys. Rev. Lett.* **30** 1343
——1974 *Phys. Rev.* D **9** 980
Grossman Y *et al.* 2005 *Phys. Rev.* D **72** 031501
Gupta R S *et al.* 2012 *How well do we need to measure the Higgs boson couplings?* arXiv:1206:3560 [hep-ph]
Guralnik G S *et al.* 1964 *Phys. Rev. Lett.* **13** 585
——1968 *Advances in Particle Physics* vol 2, ed R Cool and R E Marshak (New York: Interscience) pp 567ff
Haag R 1958 *Phys. Rev.* **112** 669
Hagiwara K *et al.* 2002 *Phys. Rev.* D **66** 010001
Halzen F and Martin A D 1984 *Quarks and Leptons* (New York: Wiley)
Hamberg R *et al* 1991 *Nucl. Phys.* B **359** 343
Hambye T and Reisselmann K 1997 *Phys. Rev.* D **55** 7255
Hammermesh M 1962 *Group Theory and its Applications to Physical Problems* (Reading, MA: Addison-Wesley)
Han M Y and Nambu Y 1965 *Phys. Rev.* B **139** 1066
Harrison P F and Quinn H R 1998 *The BaBar physics book: Physics at an asymmetric B factory* SLAC-R-0504
Hasenfratz P *et al.* 1998 *Phys. Lett.* B **427** 125
Hasenfratz P and Niedermayer F 1994 *Nucl. Phys.* B **414** 785
Hasert F J *et al.* 1973 *Phys. Lett.* B **46** 138
Heisenberg W 1932 *Z. Phys.* **77** 1
Higgs P W 1964 *Phys. Rev. Lett.* **13** 508
——1966 *Phys. Rev.* **145** 1156
Hirata K S *et al.* 1989 *Phys. Rev. Lett.* **63** 16
Höcker A *et al.* 2001 *Eur. Phys. J.* C **21** 225
Hollik W 1990 *Fortsch. Phys.* **38** 165
——1991 *1989 CERN-JINR School of Physics* CERN 91-07 (Geneva) p 50ff
Hornbostel K, Lepage G P and Morningstar C 2003 *Phys. Rev.* D **67** 034023
Hosaka J *et al.* 2006 *Phys. Rev.* D **73** 112001
Hughes R J 1980 *Phys. Lett.* B **97** 246
——1981 *Nucl. Phys.* B **186** 376
Isgur N and Wise M B 1989 *Phys. Lett.* B **232** 113
——1990 *Phys. Lett.* B **237** 527
Isidori G *et al.* 2001 *Nucl. Phys.* B **609** 387
Jackiw R 1972 *Lectures in Current Algebra and its Applications* ed S B Treiman, R Jackiw and D J Gross (Princeton, NJ: Princeton University Press) pp 97–254
Jacob M and Landshoff P V 1978 *Phys. Rep.* C **48** 285
Jarlskog C 1985 *Phys. Rev. Lett.* **55** 1039
Jones D R T 1974 *Nucl. Phys.* B **75** 531
Jones H F 1990 *Groups, Representations and Physics* (Bristol: IOP Publishing)

Kabir P K 1968 *The CP Puzzle: Strange Decays of the Neutral Kaon* (London and New York: Academic Press)
Kadanoff L P 1977 *Rev. Mod. Phys.* **49** 267
Kaplan D B 1992 *Phys. Lett.* B **288** 342
Kayser B 1981 *Phys. Rev.* D **24** 110
Kennedy D C *et al.* 1989 *Nucl. Phys.* B **321** 83
Kennedy D C and Lynn B W 1989 *Nucl. Phys.* B **322** 1
Kibble T W B 1967 *Phys. Rev.* **155** 1554
Kim K J and Schilcher K 1978 *Phys. Rev.* D **17** 2800
Kinoshita T 1962 *J. Math. Phys.* **3** 650
Kittel C 1987 *Quantum Theory of Solids* second revised printing (New York: Wiley)
Klapdor-Kleingrothaus H V *et al.* (Heidelberg-Moscow Collaboration) 2001 *Eur. Phys. J.* A **12** 147
——2006 *Mod. Phys. Lett.* A **21** 1547
Klein O 1948 *Nature* **161** 897
Kluth S 2006 *Rept. on Prog. in Phys.* **69** 1771
Kobayashi M 2009 *Rev. Mod. Phys.* **81** 1019
Kobayashi M and Maskawa K 1973 *Prog. Theor. Phys.* **49** 652
Kogut J B and Susskind L 1975 *Phys. Rev.* D **11** 395
Kramer G and Lampe B 1987 *Z. Phys.* C **34** 497
Krastev P I and Petcov S T 1988 *Phys. Lett.* B **205** 84
Kugo T and Ojima I 1979 *Prog. Theor. Phys. Suppl.* **66** 1
Kunszt Z and Piétarinen E 1980 *Nucl. Phys.* B **164** 45
Kusaka A *et al.* 2007 (Belle Collaboration) *Phys. Rev. Lett.* **98** 221602
Kuzmin V A, Rubakov V A and Shaposhnikov M E 1985 *Phys. Lett.* B **155** 36
Landau L D 1948 *Dokl. Akad. Nauk. USSR* **60** 207
——1957 *Nucl. Phys.* **3** 127
Landau L D and Lifshitz E M 1980 *Statistical Mechanics* part 1, 3rd edn (Oxford: Pergamon)
Langacker P (ed) 1995 *Precision Tests of the Standard Electroweak Model* (Singapore: World Scientific)
Larin S A and Vermaseren J A M 1991 *Phys. Lett.* B **259** 345
——1993 *Phys. Lett.* B **303** 334
Lautrup B 1967 *Kon. Dan. Vid. Selsk. Mat.-Fys. Med.* **35** 1
Lee B W *et al.* 1977a *Phys. Rev. Lett.* **38** 883
——1977b *Phys. Rev.* D **16** 1519
Lee T D and Nauenberg M 1964 *Phys. Rev.* B **133** 1549
Lee T D, Rosenbluth R and Yang C N 1949 *Phys. Rev.* **75** 9905
Lee T D and Yang C N 1956 *Phys. Rev.* **104** 254
——1957 *Phys. Rev.* **105** 1671
——1962 *Phys. Rev.* **128** 885
LEP 2003 (The LEP Working Group for Higgs Searches, ALEPH, DELPHI, L3 and OPAL Collaborations) *Phys. Lett.* B **565** 61
Lepage G P and Mackenzie P B 1993 *Phys. Rev.* **48** 2250
Leutwyler H 1996 *Phys. Lett.* B **378** 313
Lichtenberg D B 1970 *Unitary Symmetry and Elementary Particles* (New York: Academic)
Lipkin H J *et al.* 1991 *Phys. Rev.* D **44** 1454
Llewellyn Smith C H 1973 *Phys. Lett.* B **46** 233

Lobashev V *et al.* 2003 *Nucl. Phys.* A **719** 153c
London F 1950 *Superfluids Vol I, Macroscopic theory of Superconductivity* (New York: Wiley)
Lüscher M 1981 *Nucl. Phys.* B **180** 317
——1986 *Commun. Math. Phys.* **105** 153
——1991a *Nucl. Phys.* B **354** 531
——1991b *Nucl. Phys.* B **364** 237
——1998 *Phys. Lett.* B **428** 342
Lüscher M and Weisz P 1985 *Phys. Lett.* B **158** 250
Lüscher M *et al.* 1980 *Nucl. Phys.* B **173** 365
Majorana E 1937 *Nuovo Cimento* **5** 171
Maki Z, Nakagawa M and Sakata S 1962 *Prog. Theor. Phys.* **28** 870
Mandelstam S 1976 *Phys. Rep.* C **23** 245
Mandl F 1992 *Quantum Mechanics* (New York: Wiley)
Marciano W J and Sirlin A 1988 *Phys. Rev. Lett.* **61** 1815
Marshak R E *et al* 1969 *Theory of Weak Interactions in Particle Physics* (New York: Wiley)
Martin A D *et al.* 1994 *Phys. Rev.* D **50** 6734
——2002 *Eur. Phys. J.* C **23** 73
Maskawa T 2009 *Rev. Mod. Phys.* **81** 1027
Merzbacher E 1998 *Quantum Mechanics* 3rd edn (New York: Wiley)
Mikheev S P and Smirnov A Y 1985 *Sov. J. Nucl. Phys.* **42** 913
——1986 *Nuovo Cimento* **9** C 17
Minkowski P 1977 *Phys. Lett.* B **67** 421
Mohapatra R N *et al.* 1968 *Phys. Rev. Lett.* **20** 1081
Mohapatra R N and Senjanovic G 1980 *Phys. Rev. Lett.* **44** 912
——1981 *Phys. Rev.* D **23** 165
Montanet L *et al.* 1994 *Phys. Rev.* D **50** 1173
Montvay I and Munster G 1994 *Quantum Fields on a Lattice* (Cambridge: Cambridge University Press)
Morningstar C and Peardon M J 2004 *Phys. Rev.* D **69** 054501
Muta T 2010 *Foundations of Quantum Chromodynamics* 3rd edtn (Singapore: World Scientific)
Nakamura K *et al.* 2010 (Particle Data Group) *J.Phys.* G **37** 075021
Nambu Y 1960 *Phys. Rev. Lett.* **4** 380
——1974 *Phys. Rev.* D **10** 4262
Nambu Y and Jona-Lasinio G 1961a *Phys. Rev.* **122** 345
——1961b *Phys. Rev.* **124** 246
Nambu Y and Lurie D 1962 *Phys. Rev.* **125** 1429
Nambu Y and Schrauner E 1962 *Phys. Rev.* **128** 862
Narayanan R and Neuberger H 1993a *Phys. Lett.* B **302** 62
——1993b *Phys. Rev. Lett.* **71** 3251
——1994 *Nucl. Phys.* B **412** 574
——1995 *Nucl. Phys.* B **443** 305
Nauenberg M 1999 *Phys. Lett.* B **447** 23
Ne'eman Y 1961 *Nucl. Phys.* **26** 222
Neuberger H 1998a *Phys. Lett.* B **417** 141
——1998b *Phys. Lett.* B **427** 353
Nielsen N K 1981 *Am. J. Phys.* **49** 1171

Nielsen H B and Ninomaya M 1981a *Nucl. Phys.* B **185** 20
——1981b *Nucl. Phys.* B **193** 173
——1981c *Nucl. Phys.* B **195** 541
Nir Y 1989 *Phys. Lett.* B **221** 184
Noaki J *et al.* 2008 *Phys. Rev. Lett.* **101** 202004
Noether E 1918 *Nachr. Ges. Wiss. Gottingen* 171
Oddone P 1989 *Ann. N.Y. Acad. Sci.* **578** 237
Okubo S 1962 *Prog. Theor. Phys.* **27** 949
Pais A 2000 *The Genius of Science* (Oxford: Oxford University Press)
Pak A and Czarnecki A 2008 *Phys. Rev. Lett.* **100** 241807
Parry W E 1973 *The Many Body Problem* (Oxford: Clarendon)
Pascoli S *et al.* 2007a *Phys. Rev.* D **75** 083511
——2007b *Nucl. Phys.* B **774** 1
Pauli W 1934 *Rapp. Septième Conseil Phys. Solvay, Brussels 1933* (Paris: Gautier-Villars), reprinted in Winter (2000) pp 7, 8
Peccei R D and Quinn H 1977a *Phys. Rev. Lett.* **38** 1440
——1977b *Phys. Rev.* D **16** 1791
Perkins D H 1975 in *Proc. Int. Symp. on Lepton and Photon Interactions at High Energies, Stanford, CA* p 571
Peskin M E 1997 in *1996 European School of High Energy Physics* ed N Ellis and M Neubert CERN 97-03 (Geneva) pp 49-142
Peskin M E and Schroeder D V 1995 *An Introduction to Quantum Field Theory* (Reading, MA: Addison-Wesley)
Politzer H D 1973 *Phys. Rev. Lett.* **30** 1346
Poluektov *et al.* 2010 (Belle Collaboration) *Phys. Rev.* D **81** 112002
Pontecorvo B 1946 *Chalk River Laboratory Report* PD-205
——1947 *Phys. Rev.* **72** 246
——1957 *Zh. Eksp. Theor. Phys.* **33** 549
——1958 *ibid.* **34** 247
——1967 *ibid.* **53** 1717 (Engl. transl. *Sov. Phys.JETP* **26** 984)
Prescott C Y *et al.* 1978 *Phys. Lett.* B **77** 347
Puppi G 1948 *Nuovo Cimento* **5** 505
Quigg C 1977 *Rev. Mod. Phys.* **49** 297
Rajaraman R 1982 *Solitons and Instantons* (Amsterdam: North-Holland)
Reines F and Cowan C 1956 *Nature* **178** 446
Reines F, Gurr H and Sobel H 1976 *Phys. Rev. Lett.* **37** 315
Renton P 1990 *Electroweak Interactions* (Cambridge: Cambridge University Press)
Richardson J L 1979 *Phys. Lett.* B **82** 272
Rodrigo G and Santamaria A 1993 *Phys. Lett.* B **313** 441
Ross D A and Veltman M 1975 *Nucl. Phys.* B **95** 135
Rubbia C *et al.* 1977 *Proc. Int. Neutrino Conf., Aachen, 1976* (Braunschweig: Vieweg) p 683
Ryder L H 1996 *Quantum Field Theory* 2nd edn (Cambridge: Cambridge University Press)
Sakurai J J 1958 *Nuovo Cimento* **7** 649
——1960 *Ann. Phys., NY* **11** 1
Salam A 1957 *Nuovo Cimento* **5** 299
——1968 *Elementary Particle Physics* ed N Svartholm (Stockholm: Almqvist and Wiksells)

Salam A and Ward J C 1964 *Phys. Lett.* **13** 168
Salam C P 2010 *Eur. Phys. J.* C **14** 47
Samuel M A and Surguladze L R 1991 *Phys. Rev. Lett.* **66** 560
Schael S *et al.* 2006 (ALEPH, DELPHI, L3, OPAL, SLD, LEP Electroweak Working Group, SLD Electroweak and Heavy Flavour Groups) *Phys. Reports* **427** 257
Schiff L I 1968 *Quantum Mechanics* 3rd edn (New York: McGraw-Hill)
Schrieffer J R 1964 *Theory of Superconductivity* (New York: Benjamin)
Schutz B F 1988 *A First Course in General Relativity* (Cambridge: Cambridge University Press)
Schwinger J 1957 *Ann. Phys., NY* **2** 407
——1962 *Phys. Rev.* **125** 397
Shaevitz M H *et al.* 1995 (CCFR Collaboration) *Nucl. Phys.* B *Proc. Suppl.* **38** 188
Sharpe S R 2006 *PoSLAT* 022 (hep-lat/0610094)
Shaw R 1995 The problem of particle types and other contributions to the theory of elementary particles *PhD Thesis* University of Cambridge
Sikivie P *et al.* 1980 *Nucl. Phys.* B **173** 189
Sirlin A 1980 *Phys. Rev.* D **22** 971
——1984 *Phys. Rev.* D **29** 89
Slavnov A A 1972 *Teor. Mat. Fiz.* **10** 153 (Engl. transl. *Theor. and Math. Phys.* **10** 99)
Snyder A E and Quinn H R 1993 *Phys. Rev.* D **48** 2139
Sommer R 1994 *Nucl. Phys.* B **411** 839
Spergel D *et al.* 2007 *Astrophys. J. Supp.* **170** 377
Staff of the CERN $\bar{\text{p}}$p project 1981 *Phys. Lett.* B **107** 306
Stange A *et al.* 1994a *Phys. Rev.* D **49** 1354
——1994b *Phys. Rev.* D **50** 4491
Staric M *et al.* 2007 (Belle Collaboration) *Phys. Rev. Lett.* **98** 211803
Steinberger J 1949 *Phys. Rev.* **76** 1180
Sterman G and Weinberg S 1977 *Phys. Rev. Lett.* **39** 1436
Stueckelberg E C G and Peterman A 1953 *Helv. Phys. Acta* **26** 499
Sudarshan E C G and Marshak R E 1958 *Phys. Rev.* **109** 1860
Susskind L 1977 *Phys. Rev.* D **16** 3031
——1979 *Phys. Rev.* D **19** 2619
Sutherland D G 1967 *Nucl. Phys.* B **2** 433
Symanzik K 1970 *Commun. Math. Phys.* **18** 227
——1983 *Nucl. Phys.* B **226** 187, 205
Tarasov O V *et al.* 1980 *Phys. Lett.* B **93** 429
Tavkhelidze A 1965 *Seminar on High Energy Physics and Elementary Particles* (Vienna: IAEA) p 763
Taylor J C 1971 *Nucl. Phys.* B **33** 436
——1976 *Gauge Theories of Weak Interactions* (Cambridge: Cambridge University Press)
't Hooft G 1971a *Nucl. Phys.* B **33** 173
——1971b *Nucl. Phys.* B **35** 167
——1971c *Phys. Lett.* B **37** 195
——1976a *Phys. Rev.* D **14** 3432
——1976b *High Energy Physics, Proc. European Physical Society Int. Conf.* ed A Zichichi (Bologna: Editrice Composition) p 1225

——1980 *Recent Developments in Gauge Theories, Cargese Summer Institute 1979* ed G 't Hooft *et al.* (New York: Plenum)
——1986 *Phys. Rep.* **142** 357
't Hooft G and Veltman M 1972 *Nucl. Phys.* B **44** 189 *Rev. Mod. Phys.* **21** 153
Tiomno J and Wheeler J A 1949 *Rev. Mod. Phys.* **21** 153
Ur A C *et al.* (GERDA Collaboration) 2011 *Nucl. Phys. Proc. Suppl.* **217** 38
Valatin J G 1958 *Nuovo Cimento* **7** 843
van der Bij J J 1984 *Nucl. Phys.* B **248** 141
van der Bij J J and Veltman M 1984 *Nucl. Phys.* B **231** 205
van der Neerven W L and Zijlstra E B 1992 *Nucl. Phys.* B **382** 11
van Ritbergen T and Stuart R G 1999 *Phys. Rev. Lett.* **82** 488
van Ritbergen *et al.* 1997 *Phys. Lett.* B **400** 379
Veltman M 1967 *Proc. R. Soc.* A **301** 107
——1968 *Nucl. Phys.* B **7** 637
——1970 *Nucl. Phys.* B **21** 288
——1977 *Acta Phys. Polon.* B **8** 475
von Weiszäcker C F 1934 *Z. Phys.* **88** 612
Wegner F 1972 *Phys. Rev.* B **5** 4529
Weinberg S 1966 *Phys. Rev. Lett.* **17** 616
——1967 *Phys. Rev. Lett.* **19** 1264
——1973 *Phys. Rev.* D **8** 605, especially footnote 8
——1975 *Phys. Rev.* D **11** 3583
——1978 *Phys. Rev. Lett.* **40** 223
——1979a *Physica* A **96** 327
——1979b *Phys. Rev.* D **19** 1277
——1996 *The Quantum Theory of Fields Vol II Modern Applications* (Cambridge: Cambridge University Press)
Weisberger W 1965 *Phys. Rev. Lett.* **14** 1047
Weisskopf V F and Wigner E P 1930a *Z. Phys.* **63** 54
——1930b *Z. Phys.* **65** 18
Wilczek F 1978 *Phys. Rev. Lett.* **40** 279
Williams E J 1934 *Phys. Rev.* **45** 729
Wilson K G 1969 *Phys. Rev.* **179** 1499
——1971a *Phys. Rev.* B **4** 3174
——1971b *Phys. Rev.* B **4** 3184
——1974 *Phys. Rev.* D **10** 2445
——1975 *New Phenomena in Subnuclear Physics, Proc. 1975 Int. School on Subnuclear Physics Ettore Majorana* ed A Zichichi (New York: Plenum)
Wilson K G and Kogut J 1974 *Phys. Rep.* **12C** 75
Winter K 2000 *Neutrino Physics* 2nd edn (Cambridge: Cambridge University Press)
Wolfenstein L 1978 *Phys. Rev.* D **17** 2369
——1983 *Phys. Rev. Lett.* **51** 1945
Wu C S *et al.* 1957 *Phys. Rev.* **105** 1413
Yanagida T 1979 *Proc. Workshop on Unified Theory and Baryon Number in the Universe* ed O Sawada and A Sugamoto (Tsukuba: KEK)
Yang C N 1950 *Phys. Rev.* **77** 242
Yang C N and Mills R L 1954 *Phys. Rev.* **96** 191
Yosida K 1958 *Phys. Rev.* **111** 1255

Index