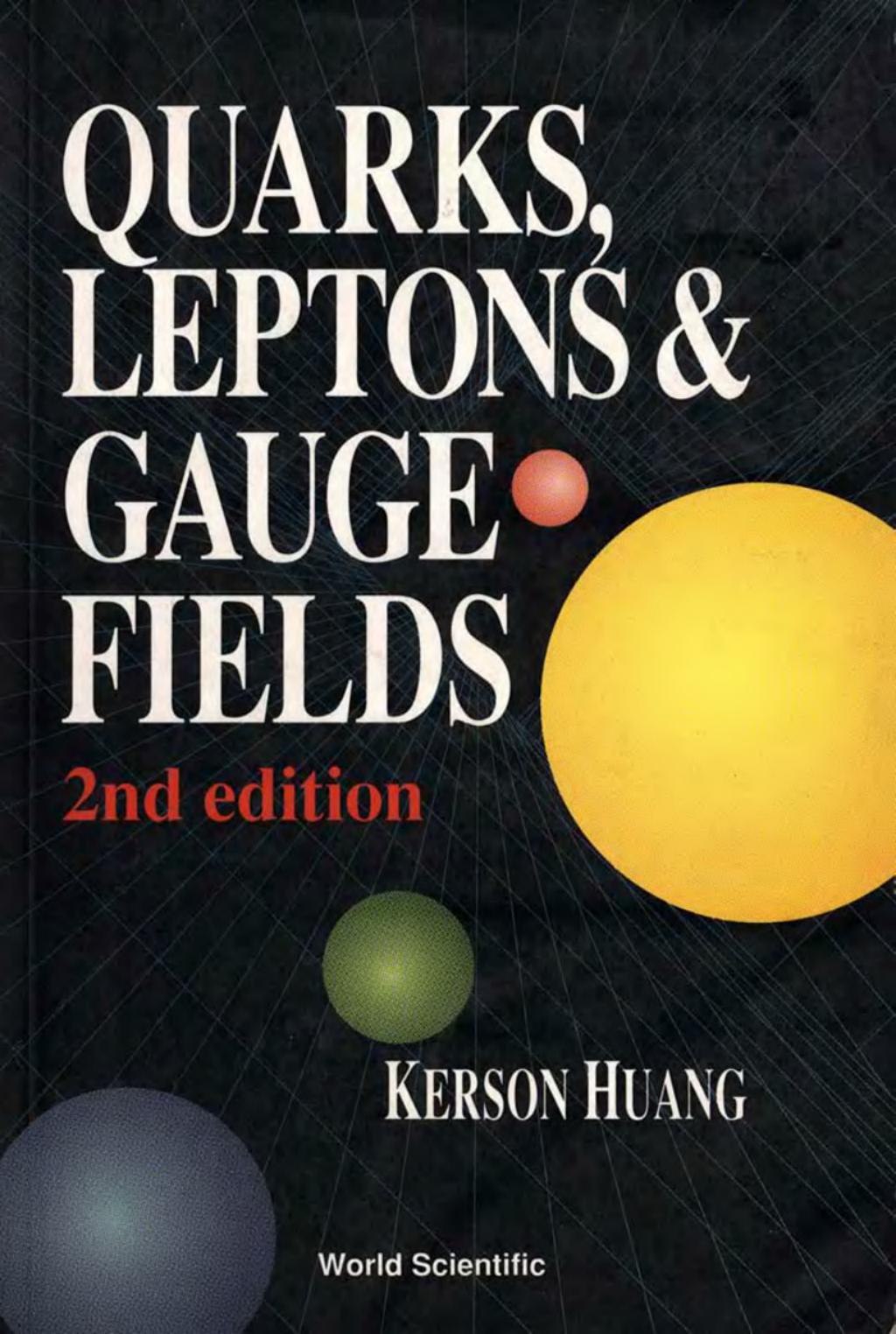


QUARKS, LEPTONS & GAUGE FIELDS

2nd edition

The background of the book cover features a dark, textured surface with several large, semi-transparent spheres of different colors and sizes. A large yellow sphere is positioned on the right side, while smaller green, red, and blue spheres are scattered across the lower half. The overall effect is a minimalist, scientific, or artistic representation of subatomic particles.

KERSON HUANG

World Scientific

QUARKS, LEPTONS & GAUGE FIELDS

2nd edition

QUARKS, LEPTONS & GAUGE FIELDS

2nd edition

Kerson Huang

*Professor of Physics
Massachusetts Institute of Technology*



World Scientific

Singapore • New Jersey • London • Hong Kong

Published by

World Scientific Publishing Co. Pte. Ltd.

P O Box 128, Farrer Road, Singapore 9128

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

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

Library of Congress Cataloging-in-Publication data is available.

QUARKS, LEPTONS AND GAUGE FIELDS (2nd Edition)

Copyright © 1992 by World Scientific Publishing Co. Pte. Ltd.

All rights reserved. This book, or parts thereof, may not be reproduced in any form or by any means, electronic or mechanical, including photocopying, recording or any information storage and retrieval system now known or to be invented, without written permission from the Publisher.

ISBN 981-02-0659-3

981-02-0660-7 (pbk)

Printed in Singapore by Singapore National Printers Ltd

For Kathryn Camille

CONTENTS

PREFACE

xiii

I. INTRODUCTION	1
1.1 Particles and Interactions	1
1.2 Gauge Theories of Interactions	6
1.3 Notations and Conventions	10
II. QUARKS	12
2.1 Internal Symmetries	12
1 Isospin	13
2 The gauge groups	14
3 More general internal symmetries: $SU(n)$	14
4 Unitary symmetry	15
2.2 Representation of $SU(3)$	17
1 The basic representation	17
2 Young's tableaux	18
3 Irreducible representations	20
2.3 The Quark Model	22
1 Quarks as basic triplets	22
2 Quarks as building blocks	25
3 Weight diagrams	25
4 The composition of hadrons	28
2.4 Color	28
1 Independent quark model	28
2 Color $SU(3)$ group	30
2.5 Electromagnetic and Weak Probes	33
1 Electromagnetic interactions	33
2 Parton model	35
3 Evidence for color	38
4 Weak interactions	40
2.6 Charm	43
1 The charmed quark	43
2 The J/ψ and its family	45
3 Correspondence between quarks and leptons	46
III. MAXWELL FIELD: $U(1)$ GAUGE THEORY	47
3.1 Global and Local Gauge Invariance	47
3.2 Spontaneous Breaking of Global Gauge Invariance: Goldstone Mode	50
3.3 Spontaneous Breaking of Local Gauge Invariance: Higgs Mode	53
3.4 Classical Finite-Energy Solutions	55

3.5 Magnetic Flux Quantization	56
3.6 Soliton Solutions: Vortex Lines	58
IV. YANG-MILLS FIELDS: NON-ABELIAN GAUGE THEORIES	61
4.1 Introductory Note	61
4.2 Lie Groups	61
1 Structure constants	61
2 Matrix representations	62
3 Topological properties	64
4 General remarks	66
4.3 The Yang-Mills Constructions	67
1 Global gauge invariance	67
2 Local gauge invariance	69
4.4 Properties of Yang-Mills Fields	72
1 Electric and magnetic fields	72
2 Dual tensor	73
3 Path representation of the gauge group	74
4.5 Canonical Formalism	77
1 Equations of motion	77
2 Hamiltonian	79
4.6 Spontaneous Symmetry Breaking	80
1 The little group	80
2 Higgs mechanism	83
V. TOPOLOGICAL SOLITONS	86
5.1 Solitons	86
5.2 The Instanton	88
1 Topological charge	88
2 Explicit solution	92
5.3 The Monopole	94
1 Topological stability	94
2 Flux quantization	95
3 Boundary conditions	98
4 Explicit solution	100
5 Physical fields	101
6 Spin from isospin	103
VI. WEINBERG-SALAM MODEL	105
6.1 The Matter Fields	105
6.2 The Gauge Fields	108
1 Gauging $SU(2) \times U(1)$	108
2 Determination of constants	111
3 Interactions	112
6.3 The General Theory	113
1 Mass terms	113
2 Cabibbo angle	117
3 Kobayashi-Maskawa matrix	118
4 Solitons	119

VII. METHOD OF PATH INTEGRALS	121
7.1 Non-Relativistic Quantum Mechanics	121
7.2 Quantum Field Theory	126
7.3 External Sources	128
7.4 Euclidean 4-Space	132
7.5 Calculation of Path Integrals	134
7.6 The Feynman Propagator	135
7.7 Feynman Graphs	137
7.8 Boson Loops and Fermion Loops	140
7.9 Fermion Fields	142
VIII. QUANTIZATION OF GAUGE FIELDS	147
8.1 Canonical Quantization	147
1 Free Maxwell field	147
2 Pure Yang-Mills fields	150
8.2 Path Integral Method in Hamiltonian Form	150
8.3 Feynman Path Integral: Fadeev-Popov Method	152
8.4 Free Maxwell Field	156
1 Lorentz gauge	156
2 Coulomb gauge	159
3 Temporal and axial gauges	160
8.5 Pure Yang-Mills Fields	162
1 Axial gauge	163
2 Lorentz gauge: Fadeev-Popov ghosts	164
8.6 The θ -World and the Instanton	165
1 Discovering the θ -world	165
2 Instanton as tunneling solution	168
3 The θ -action	171
8.7 Gribov Ambiguity	172
8.8 Projection Operator for Gauss' Law	174
IX. RENORMALIZATION	177
9.1 Charge Renormalization	177
9.2 Perturbative Renormalization in Quantum Electrodynamics	180
9.3 The Renormalization Group	183
1 Scale transformations	183
2 Scaling form	184
3 Fixed points	185
4 Callan-Symanzik equation	186
9.4 Scalar Fields	188
1 Renormalizability	188
2 ϕ^4 theory	189
3 "Triviality" and the Landau ghost	191
9.5 The Physics of Renormalization	192
1 Renormalization-group transformation	192
2 Real-space renormalization	195
3 Fixed points and relevancy	197
4 Renormalization and universality	199

Appendix to Chapter 9. Renormalization of QED	201
1 Vertex	201
2 Electron Propagator	201
3 Photon Propagator	202
4 Scaling Properties	205
5 Renormalization	206
6 Gauge Invariance and the Photon Mass	208
X. METHOD OF EFFECTIVE POTENTIAL	210
10.1 Spontaneous Symmetry Breaking	210
10.2 The Effective Action	210
10.3 The Effective Potential	212
10.4 The Loop Expansion	215
10.5 One-Loop Effective Potential	218
10.6 Renormalization	219
1 General scheme	219
2 Massive case	221
3 Massless case	221
10.7 Dimensional Transmutation	223
10.8 A Non-Relativistic Example	225
10.9 Application to Weinberg-Salam Model	227
XI. THE AXIAL ANOMALY	230
11.1 Origin of the Axial Anomaly	230
11.2 The Triangle Graph	231
11.3 Anomalous Divergence of the Chiral Current	237
11.4 Physical Explanation of the Axial Anomaly	238
11.5 Cancellation of Anomalies	242
11.6 't Hooft's Principle	247
XII. QUANTUM CHROMODYNAMICS	252
12.1 General Properties	252
1 Lagrangian density	252
2 Feynman rules	253
3 Quark-gluon interactions	255
4 Gluon self-interactions	256
12.2 The Color Gyromagnetic Ratio	260
12.3 Asymptotic Freedom	262
1 The running coupling constant	262
2 The vacuum as magnetic medium	265
3 The Nielsen-Hughes formula	268
12.4 The Pion as Goldstone Boson	269
1 The low-energy domain	269
2 Chiral symmetry: an idealized limit	269
3 PCAC	272
4 The decay $\pi^0 \rightarrow 2\gamma$	273
5 Extension to pion octet	275
12.5 The $U(1)$ Puzzle	276

12.6 θ -Worlds in QCD	278
1 Euclidean action	278
2 The axial anomaly and the index theorem	279
3 Chiral limit: Collapse of the θ -worlds	282
4 Quark mass matrix	283
5 Strong CP violation	286
XIII. LATTICE GAUGE THEORY	288
13.1 Wilson's Lattice Action	288
13.2 Transfer Matrix	291
13.3 Lattice Hamiltonian	293
13.4 Lattice Fermions	297
13.5 Wilson Loop and Confinement	299
13.6 Continuum Limit	303
13.7 Monte Carlo Methods	304
XIV. QUARK CONFINEMENT	308
14.1 Wilson Criterion and Electric Confinement	308
14.2 String Model of Hadrons	311
14.3 Superconductivity: Magnetic Confinement	312
1 Experimental manifestation	312
2 Theory	313
3 Mechanism for monopole confinement	314
14.4 Electric and Magnetic Order Parameters	316
14.5 Scenario for Quark Confinement	319
Appendix to Chapter 14. Symmetry and Confinement	322
1 Quark Propagator	322
2 Center Symmetry	324
3 Confinement as Symmetry	325
INDEX	327

PREFACE

According to the current view, the basic building blocks of matter are quarks and leptons, which interact with one another through the intermediaries of Yang-Mills gauge fields (gravity being ignored in this context). This means that the forms of the interactions are completely determined by the algebraic structure of certain internal symmetry groups. Thus, the strong interactions are associated with the group $SU(3)$, and is described by a gauge theory called quantum chromodynamics. The electro-weak interactions, as described by the now standard Weinberg-Salam model, is associated with the group $SU(2) \times U(1)$.

This book is a concise introduction to the physical motivation behind these ideas, and precise mathematical formulation thereof. The goal of the book is to explain why and how the mathematical formalism helps us to understand the relevant observed phenomena. The audience for which this book is written are graduate students in physics who have some knowledge of the experimental parts of particle physics, and an acquaintance with quantum field theory, including Feynman graphs and the notion of renormalization. This book might serve as a text for a one-semester course beyond quantum field theory. The first edition of this book, which came out in 1982, was based on a course I gave at M.I.T., and on lectures I gave in Santiago, Chile, in 1977, and in Beijing, China, in 1979. I am indebted to I. Saavedra for the opportunity to lecture in Chile, to Chang Wen-yu and S.C.C. Ting for the inducement to give the Beijing lecture, and to M. Jacob and K. K. Phua for the encouragement to bring out the first edition.

The main addition to the second edition are Wilson's approach to renormalization, lattice gauge theory, and quark confinement. I am grateful to the many readers who have pointed out errors in the first edition, which I hope have been corrected in this edition.

I owe special thanks to my colleagues at M.I.T., especially A. Guth, R. Jackiw, K. Johnson, and J. Polonyi, from whom I have learned much that is being passed along in this book.

Kerson Huang

*Marblehead, Mass.
February 1991*

CHAPTER I

INTRODUCTION

1.1 Particles and Interactions

一尺之棰 日取其半 万世不竭

*Take half from a foot-long stick each day;
you will not exhaust it in a million years.*

The thought experiment contemplated in this proposition by an ancient Chinese sophist¹ is an apt allegory for what physicists actually do in the laboratory, in their search for the ultimate constituents of matter.

During the three centuries since the birth of physics in the modern sense, we have done about 60 days' worth of "halving" (down to 10^{-16} cm). At around day 30 (at 10^{-8} cm), we encountered the first granular structure of matter—atoms, which appeared at first to be indivisible. As we know, they turned out to be divisible further into electrons and nuclei; and nuclei could in turn be split into nucleons. Now we are at the stage when constituents of the nucleon—quarks—can be confidently identified. Indications are that the subdividing process will continue. The ancient sophist seems to be right so far.

From an experimental point of view, particles are detectable packets of energy and momentum, be they billiard balls, photons, or lambda hyperons. At each stage of our understanding, we designate certain particles as "fundamental", in the sense that they are the most elementary interacting units in our theories. As our experimental knowledge expands, we have often been forced to revise our views. The necessity for such revisions rests with the stringent requirement we place upon our theories: they must, in principle, be able to predict the quantitative results of all possible experiments.

It is fortunate that, at any given stage, we were able to regard certain particles as provisionally fundamental, without jeopardizing the right to change our mind. The reason is that, according to quantum mechanics, it is a good approximation to ignore those quantum states of a system whose excitation energies lie far above the energy range being studied. For example, a nucleus could be treated phenomenologically as a point mass at energies far below 1 MeV. We have discovered many layers of substructure since the era of atomic physics; but it is a remarkable fact that the dynamical principles learned from that era, as synthesized by relativistic local quantum field theory², continues to work up to the present stage.

¹ Kungsun Lung (公孙龙), quoted in Chuang Chou, *Chuangtse* (ca. 300 B.C.), chapter 33. (莊子天下篇第三十三).

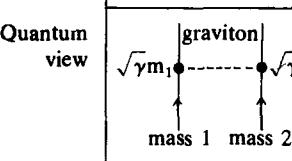
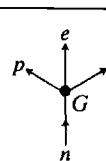
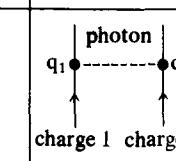
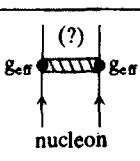
² J. D. Bjorken and S. D. Drell, *Relativistic Quantum Mechanics* (McGraw-Hill, New York, 1964); *Relativistic Quantum Fields* (McGraw-Hill, New York, 1965); C. Itzykson and J.-B. Zuber, *Quantum Field Theory* (McGraw-Hill, New York, 1980).

Interactions among experimentally observed particles fall into four types of markedly different strengths: gravitational, weak, electromagnetic, and strong interactions. These are briefly reviewed in Table 1.1.

In the current theoretical view, which has come to be known as the "standard model," the weak and electromagnetic interactions are low-energy manifestations of a single unified interaction, and the strong interactions originate in a hidden charge called "color", carried by quarks permanently confined in nucleons and other strongly interacting particles. All these interactions are supposed to be mediated via the exchange of vector mesons with "minimal" coupling, similar to the well-known situation in electrodynamics. We are even in a position to speculate that all the above interactions are really low-energy manifestations of a single "grand unified" interaction, whose simplicity will be directly revealed in experiments only at energies above 10^{17} GeV! Unfortunately, nothing reliable can be said about the microscopic aspects of the gravitational interaction, due to a total lack of experimental information. Important as it may be in an eventual grand synthesis of all the interactions, we will have nothing to say about gravity in this book.

Basic to the theoretical classification of particles is the assumption that physical laws are invariant under Poincaré transformations, i.e., Lorentz transformations and space-time translations. A particle, be it "fundamental" or composite, is defined as a state of a quantum field that transforms under elements of the Poincaré group according to a definite irreducible representation. This implies that a particle has definite mass and spin, and that to each particle is associated an antiparticle of the same mass and spin³. The assumption

Table 1.1 THE FOUR TYPES OF INTERACTIONS

Interaction	Gravitational	Weak	Electromagnetic	Strong
Manifestation	Celestial mechanics	β -radio-activity	Everyday world	Nuclear binding
Quantum view				
Static potential	$-\frac{\gamma m_1 m_2}{r}$ <p>r = distance between sources</p>	—	$\frac{q_1 q_2}{4\pi r}$	$-g_{\text{eff}} \frac{e^{-\mu r}}{4\pi r}$ $\frac{\hbar}{\mu c} \sim 10^{-13} \text{ cm}$
Coupling strength	$\frac{\gamma m_p^2}{\hbar c} = 5.76 \times 10^{-36}$ <p>m_p = proton mass</p>	$Gm_p^2 = 1.01 \times 10^{-5}$	$\frac{e^2}{4\pi\hbar c} = \frac{1}{137.036}$ <p>e = electron charge</p>	$\frac{g^2}{4\pi\hbar c} \approx 10$

³ E. P. Wigner, *Ann. Math.* **40**, 149 (1934).

of microcausality in local quantum field theory further implies a connection between spin and statistics: particles with integer spin are bosons, and those with half-integer spin are fermions⁴. The interactions among particles are required to be invariant under the Poincaré group; this imposes non-trivial conditions on possible local quantum field theories.⁵

In addition to Poincaré invariance, which is a space-time symmetry, there are also internal symmetries having to do with space-time-independent transformations of particle states. The invariance of interactions under internal symmetry groups gives rise to further quantum numbers that label particle states, such as electric charge, baryon number, isospin, etc.

A partial list of known particles, classified according to mass, spin, internal quantum numbers, and the types of interactions they have, is shown in Fig. 1.1.

“Hadrons” denote bosons and fermions having strong interactions, and “leptons” denote fermions without strong interactions^a. Among the hadrons, “mesons” are bosons with baryon number 0, and “baryons” are fermions with baryon number different from 0^b. Of all these particles (apart from the photon not shown in Fig. 1.1), only electrons and nucleons are relevant to our everyday experience. One might go a little further and include neutrinos as important catalysts for the generation of solar power, and μ mesons are free gifts from heaven^c. Everything else is created primarily in high-energy accelerators.

Two striking features should be mentioned. First, all the leptons appear to be point-like particles, the latest experimental upperbound on their “radii” being 10^{-16} cm.^d This is particularly remarkable for the τ , which is about twice as heavy as the proton. Secondly, there is a wild proliferation of hadrons. As noted by Hagedorn⁷, a plot of the density of hadronic states against mass suggests an exponential growth, as shown in Fig. 1.2.

If this trend continues to asymptotically large masses, there would exist an “ultimate temperature” of about 160 MeV (2×10^{13} K), beyond which no system could be heated⁸. If the growth were faster than exponential, the partition function of statistical mechanics would not exist. Thus, the density of hadronic states seems to be growing at the maximum rate consistent with thermodynamics.

Even if we had not detected experimentally a finite radius for the proton (which we have, at about 10^{-13} cm)⁹, the sheer number of the hadrons would make it absurd to suppose that they are all “fundamental”. A key to the inner

^a All observed bosons so far have strong interactions except the photon. Historically, leptons were so named because they were light; but this is no longer true with the discovery of the τ .

^b The reason that all baryons are fermions, while all mesons are bosons, comes from baryon conservation in relativistic field theory, i.e., fermion fields must occur bilinearly in the Lagrangian, but bosons can occur linearly.

^c “Who ordered them?” asked I. I. Rabi.

⁴ R. F. Streater and A. S. Wightman, *PCT, Spin and Statistics, and All That* (W. A. Benjamin, New York, 1964).

⁵ N. N. Bogolubov, G. G. Logunov, and I. T. Todorov, *Introduction to Axiomatic Quantum Field Theory* (W. A. Benjamin, Reading, Mass., 1975).

⁶ D. P. Barber *et al.*, *Phys. Rev. Lett.* **43**, 1915 (1979).

⁷ R. Hagedorn, *N. Cim.* **56A**, 1027 (1968).

⁸ K. Huang and S. Weinberg, *Phys. Rev. Lett.* **25**, 895 (1970).

⁹ R. Hofstadter and R. W. McAllister, *Phys. Rev.* **98**, 217 (1955).

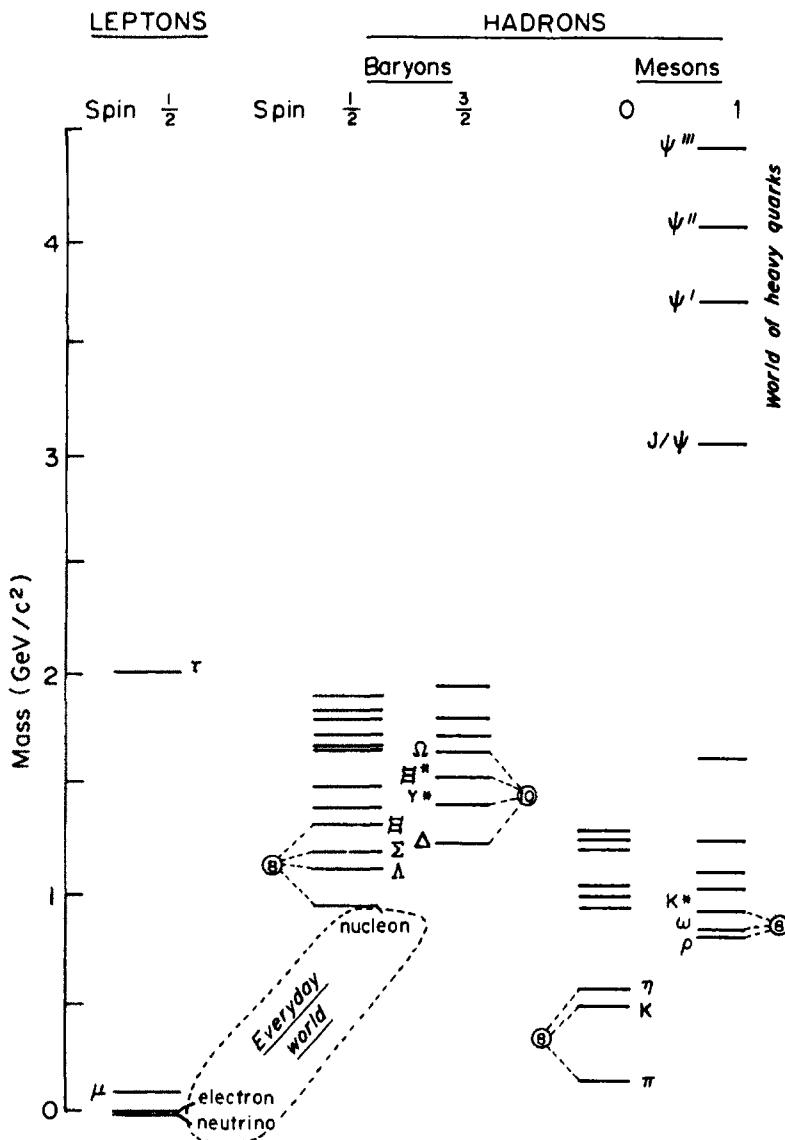


Fig. 1.1 Particle mass spectrum

structure of hadrons are the multiplet structures (e.g., **8** and **10** in Fig. 1.1) identifiable with irreducible representations of an internal symmetry group $SU(3)$. This is the first lead to the notion of quarks as hadronic constituents, namely, they form a fundamental representation of $SU(3)$. A more detailed discussion of the evidence for quarks and their interactions will be given in Chapter 2.

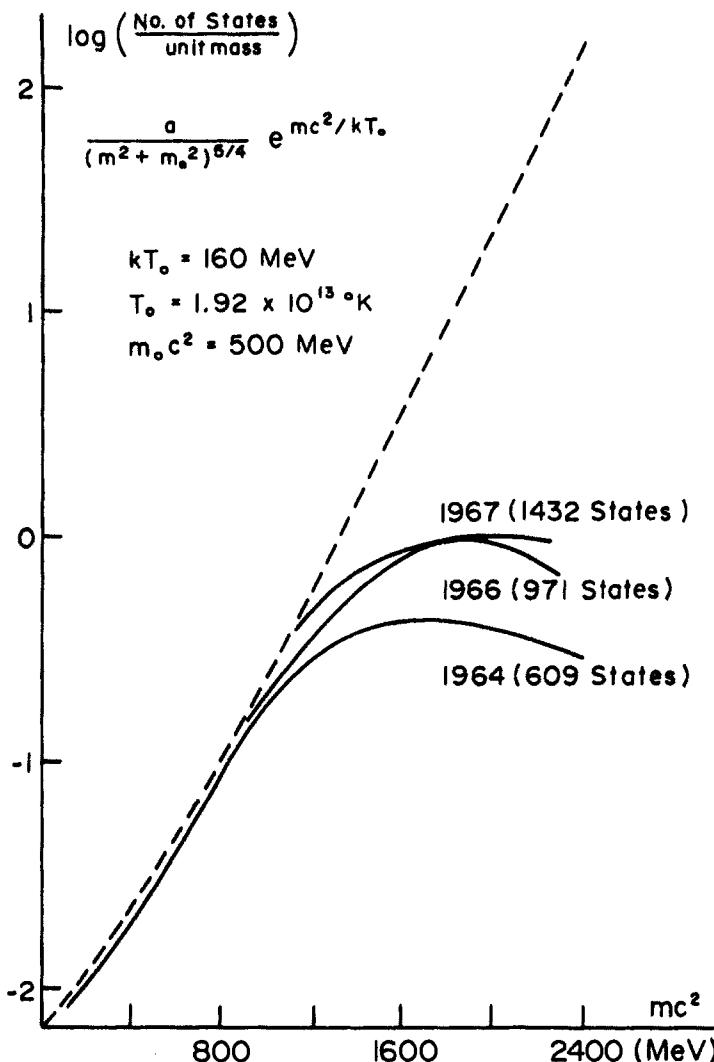


Fig. 1.2 Number of hadronic states as function of mass

1.2 Gauge Theories of Interactions

In the standard model, all the interactions are derived from a “gauge principle” similar to that in electromagnetism. We recall that the coupling of the electro-magnetic field A^μ to a charged matter field ψ can be derived through the following prescription: replace $\partial^\mu\psi$ in the matter Lagrangian by the covariant derivative $(\partial^\mu + ieA^\mu)\psi$, where e is the electric charge of ψ . Before we “turn on” the coupling (i.e., for $e = 0$), the matter Lagrangian must be invariant under constant phase changes of ψ , called “global gauge transformations”. What the prescription does is to enlarge this symmetry to a “local gauge invariance”, i.e., invariance under arbitrary space-time dependent phase changes of ψ (correlated with corresponding gauge transformations of A^μ)^d. The original global gauge invariance implies the existence of a conserved matter current j^μ , and the prescription leads to an interaction of the form $ej^\mu A_\mu$, in conformity with Maxwell’s theory. Under the usual assumptions of canonical field theory, the prescription is unique, and is called the “gauge principle”.

We may restate the gauge principle as follows. Consider a matter system originally invariant under a global $U(1)$ group of gauge transformations. We “gauge” this symmetry, i.e., enlarge it to a local $U(1)$ gauge invariance. This means that an independent $U(1)$ gauge group shall be associated with each space-time point. To do this it is necessary to introduce a vector gauge field, to which the matter field current becomes coupled. The coupling constant is the electric charge, the generator of $U(1)$. The original global symmetry can be gauged only if it is an exact symmetry.

We shall use a generalized gauge principle formulated by Yang and Mills¹⁰, which applies to a multicomponent matter field. Instead of $U(1)$, the gauge group is now a larger group of transformations that mix the different components of the matter field. There will now be more than one gauge field—the Yang-Mills fields. Their number is equal to the number of generators of the gauge group. The relevant group for the weak, electromagnetic and strong interactions is $SU(2) \times U(1) \times SU(3)$. To define this group, we must first describe the matter fields.

A well-known characteristic of the weak interactions is that they violate parity conservation to a maximal degree¹¹ by virtue of the V-A coupling¹². That is, only left-handed components of the leptons are coupled in the charge-changing sector; the right-handed components play a rather passive role—to provide mass. Similarly, hadronic weak interactions can be accounted for by assuming that quarks have the same kind of weak couplings. Thus, to the weak interactions, the elementary entities are states of definite chirality^e, which have

^d H. Weyl, *Ann. d. Physik*, **59**, 101 (1919), first introduced the term “gauge transformation” in an interesting but unsuccessful attempt to unify electromagnetism with gravity in a geometric theory, by extending the non-integrability of the direction of a vector in curved space-time to a non-integrability of its length (gauge) in an extended space called “gauge space”.

^e Chirality is defined as the eigenvalue of γ_5 , with $\gamma_5 = 1$ corresponding to right-handedness, and $\gamma_5 = -1$ to left-handedness.

¹⁰ C. N. Yang and R. L. Mills, *Phys. Rev.* **96**, 191 (1954).

¹¹ T. D. Lee and C. N. Yang, *Phys. Rev.* **104**, 254 (1956); C. S. Wu *et al.*, *Phys. Rev.*, **105**, 1413 (1957).

¹² R. P. Feynman and M. Gell-Mann, *Phys. Rev.* **109**, 193 (1958).

zero mass. (An eigenstate of finite mass is a superposition of left and right-handed states with equal weight). Glashow¹³ first proposed a unified gauge theory of electroweak interactions based on a gauge group $SU(2) \times U(1)$, which mixes different massless chiral states. However, the fact that physical particles have finite masses seems to violate this symmetry. The seeming impasse was overcome by Weinberg¹⁴ and Salam¹⁵ by appealing to the notion of “spontaneous symmetry breaking”. In the now-standard Weinberg-Salam model, “Higgs fields” are introduced to implement this idea, though they may be phenomenological parameters to be replaced by something more basic in a future theory. It is fair to say that at present we have no deep understanding of where masses come from.

The symmetries to be gauged refer to transformations among massless quarks and leptons of definite chirality. They come in at least six “flavors” (the sixth one being not yet experimentally confirmed). The lepton flavors are (e, ν) , (μ, ν') , (τ, ν'') , where the ν 's denote massless left-handed neutrinos. The quark flavors bear a one-to-one correspondence to the above: (u, d) , (s, c) , (t, b) . The parentheses group the particles into three families, which are indistinguishable copies as far as the weak interactions are concerned^f. In addition, each quark flavor comes in three (and only three) “colors”, while leptons have no color. Thus, the elementary particles are

$$\begin{aligned} \text{quarks: } q_{fn} & \left\{ \begin{array}{l} (f = 1, \dots, 6) \text{ (flavor index)} \\ (n = 1, 2, 3) \text{ (color index)} \end{array} \right. \\ \text{leptons: } l_f & \quad (f = 1, \dots, 6) \text{ (flavor index)} \end{aligned}$$

It is understood that, for example, q_{fn} denotes collectively $(q_R)_{fn}$ and $(q_L)_{fn}$, the right and left-handed components respectively, each regarded as an independent particle.

We list the quarks and leptons more explicitly in Table 1.2, and postulate the following internal symmetries:

(a) **Color $SU(3)$:** With respect to the color index, the three quarks of each flavor form a triplet representation of a “color group” $SU(3)$. The leptons are color singlets^g.

(b) **Weak isospin $SU(2)$:** In each family, the left-handed components of the upper and lower particles (e.g., ν_L and e_L) form a doublet representation of a “weak isospin group” $SU(2)$. All right-handed particles are $SU(2)$ singlets.

(c) **Weak hypercharge $U(1)$:** There is a $U(1)$ symmetry, called “weak hypercharge”, associated with simultaneous phase changes of each particle. The relative phases are fixed by definite “weak hypercharge” assignments.

The gauge group is then $SU(2) \times U(1) \times SU(3)$, a direct product of the three mutually commuting groups defined above. Gauging this group necessitates the

^f Rabi's question on p. 3 can be generalized, but remains unanswered.

^g This means that the theory is invariant under the group in question, and that the particles transform under the group according to the representations specified.

¹³ S. L. Glashow, *Nucl. Phys.* **22**, 579 (1961).

¹⁴ S. Weinberg, *Phys. Rev. Lett.* **19** 1264 (1967).

¹⁵ A. Salam, in *Elementary Particle Theory*, ed. N. Svartholm (Almqvist and Wiksell, Stockholm, 1968).

introduction of 12 vector gauge fields, one for each group generator, as listed in Table 1.3. The resulting interactions are described schematically by the Feynman vertices shown in Fig. 1.3.

The gauge fields generally have self-interactions because, unlike the photon, they generally carry "charge" by virtue of the non-Abelian nature of the group. It is to be noted that there are other exact symmetries of the theory, such as baryon number and lepton number, which are not gauged.

The theory so far has a serious defect, namely, all particles are massless. One cannot remedy this by simply including conventional mass terms in the Lagrangian, because such terms violate the $SU(2) \times U(1)$ symmetry, which we assume to be exact. Conventional vector boson mass terms also lead to non-renormalizable theories. A way out is to regard the masses as arising from "spontaneous breaking" of the $SU(2) \times U(1)$ symmetry, through couplings to scalar "Higgs fields". This will be fully explained in later chapters. It suffices to mention here that, by this method, one can obtain a renormalizable theory in which all particles can acquire arbitrary finite masses. The photon and the neutrinos can be arranged to remain massless in a natural way.

Since mass and chirality do not commute, physical particles are not necessarily members of $SU(2) \times U(1)$ multiplets. This leads to a mixing of flavors across families. For the same reason, the weak hypercharge $U(1)$ is not necessarily the electromagnetic $U(1)$. These points will be taken up in detail in Chapter 6.

Table 1.2 INTERNAL SYMMETRIES OF QUARKS AND LEPTONS

Family	Flavor f	Quarks q_{fn} Color: $n = 1, 2, 3$	Leptons l_f	
I	1	$u_1 \ u_2 \ u_3$	ν	$\Downarrow SU(2)$
	2	$d_1 \ d_2 \ d_3$	e	
II	3	$c_1 \ c_2 \ c_3$	ν'	$\Downarrow SU(2)$
	4	$s_1 \ s_2 \ s_3$	μ	
III	5	$t_1 \ t_2 \ t_3$	ν''	$\Downarrow SU(2)$
	6	$b_1 \ b_2 \ b_3$	τ	

\longleftrightarrow
 $SU(3)$

Table 1.3 THE GAUGE FIELDS

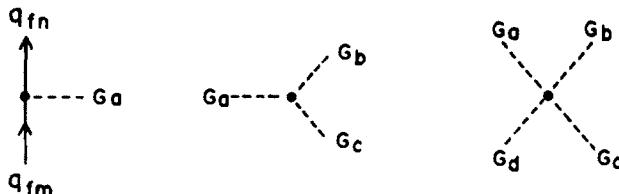
Gauge group	Number of Generators	Gauge Fields
Color $SU(3)$	8	$G_a^\mu \quad (a = 1, \dots, 8)$ (gluons)
Weak isospin $SU(2)$	3	$W_i^\mu \quad (i = 1, 2, 3)$
Weak hypercharge $U(1)$	1	W_0^μ

On the other hand, color multiplets are mass eigenstates, because right and left-handed quarks can have the same color. Experimental evidence indicates that color $SU(3)$ does not suffer spontaneous breakdown.

The electroweak interactions that result from gauging $SU(2) \times U(1)$ reproduce all known phenomena and predict new ones. Chief among these is the existence of "neutral currents" and the gauge vector bosons, which have all been triumphantly verified experimentally.

Due to the structure of color $SU(3)$, the quark-gluon coupling tends to vanish at large momenta (or small distances)—a phenomenon known as "asymptotic freedom". Thus, one should be able to detect quasi-free quarks inside a hadron by using probes that impart large momentum transfers to quarks. This has indeed been successfully demonstrated experimentally.

STRONG INTERACTION VERTICES

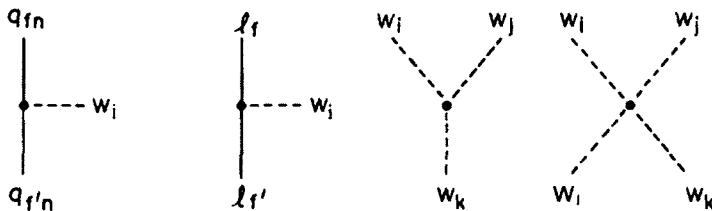


$n, m = 1, 2, 3$ (color index)

$a, b, c, d = 1, \dots, 8$ (gluon index)

Note: Flavor f does not change

ELECTROWEAK INTERACTION VERTICES



$f, f' = 1, \dots, 6$ (flavor index)

$i = 0, 1, 2, 3$ (vector boson index)

Note : Color n does not change

Fig. 1.3 Interaction vertices in gauge theory of strong and electroweak interactions

The quark-gluon coupling tends to grow as the momentum scale decreases, leading to very strong interactions at energies below ~ 0.5 GeV. It is believed that these interactions lead to “quark confinement.” That is, quarks (and gluons) cannot exist in isolation as physical states, but occur only as components of color-neutral bound states (hadrons.) Although a mathematical proof is lacking, computer studies have made this very plausible. We have also gained understanding of quark confinement through comparison with superconductivity, a dual phenomenon that can be characterized as “magnetic monopole confinement”.

The standard model has withstood all experimental tests in the decades since it took shape. All the gauge bosons corresponding to the gauge group $SU(3) \times SU(2) \times U(1)$ have been discovered, at the masses predicted by theory. It has opened our eyes to theoretical ideas like vacuum structures and topological excitations, which have found applications in cosmology and condensed matter physics.

An important open question concerns the existence of the Higgs boson and the t quark. A better theoretical understanding of the Higgs sector is also needed. Is the Higgs boson a composite of other particles, for example t and \bar{t} ?

From a theoretical point of view the greatest puzzle concerns the structure of the families. Is it an expression of a deeper symmetry, perhaps of constituents of the quarks and leptons? Attempts to answer such a question have spawned grand unified schemes, and superstring models, but it is too early to tell whether these theories contain any lasting ideas.

1.3 Notations and Conventions

We use units in which $\hbar = c = 1$ unless otherwise indicated. The diagonal metric tensor in Minkowski space-time has the diagonal elements

$$g^{00} = -g^{11} = -g^{22} = -g^{33} = 1.$$

The contravariant space-time 4-vector x^μ has components designated by

$$x^\mu = (x^0, x^1, x^2, x^3) \equiv (x^0, \mathbf{x}),$$

with the corresponding covariant vector given by

$$x_\mu = (x^0, -\mathbf{x}).$$

Some frequently used differential operators are

$$\partial_\mu \equiv \frac{\partial}{\partial x^\mu} = \left(\frac{\partial}{\partial x^0}, \boldsymbol{\nabla} \right),$$

$$\partial^\mu \equiv \frac{\partial}{\partial x_\mu} = \left(\frac{\partial}{\partial x^0}, -\boldsymbol{\nabla} \right),$$

$$\square^2 \equiv \partial_\mu \partial^\mu = \left(\frac{\partial}{\partial x^0} \right)^2 - \nabla^2,$$

$$\overleftrightarrow{A} \partial_\mu B \equiv (\partial_\mu A)B - A(\partial_\mu B).$$

The Dirac matrices γ^μ are chosen so that γ^0 is hermitian, while $\gamma^k (k = 1, 2, 3)$ are anti-hermitian:

$$\begin{aligned} (\gamma^0)^\dagger &= \gamma^0, & (\gamma^0)^2 &= 1, \\ (\gamma^k)^\dagger &= -\gamma^k, & (\gamma^k)^2 &= -1 \quad (k = 1, 2, 3). \end{aligned}$$

We define γ_5 as the hermitian matrix

$$\gamma_5 = -i\gamma^0\gamma^1\gamma^2\gamma^3, \quad (\gamma_5)^2 = 1.$$

In addition, we use the notation:

$$\begin{aligned} \alpha^k &= \gamma^0\gamma^k \quad (k = 1, 2, 3), & (\alpha^k)^\dagger &= \alpha^k, \\ \sigma^k &= -\epsilon^{klm}\gamma^l\gamma^m/2i, & (\sigma^k)^\dagger &= \sigma^k, \end{aligned}$$

from which follows

$$\alpha = \gamma_5\sigma.$$

A standard representation is the following:

$$\begin{aligned} \gamma^k &= \begin{pmatrix} 0 & -\underline{\sigma}^k \\ \underline{\sigma}^k & 0 \end{pmatrix}, & \gamma^0 &= \beta = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}. \\ \alpha^k &= \begin{pmatrix} 0 & \underline{\sigma}^k \\ \underline{\sigma}^k & 0 \end{pmatrix}, & \sigma^k &= \begin{pmatrix} \underline{\sigma}^k & 0 \\ 0 & \underline{\sigma}^k \end{pmatrix}, \\ \gamma_5 &= \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \end{aligned}$$

where underlined symbols denote 2×2 matrices:

$$\underline{\sigma}^1 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad \underline{\sigma}^2 = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}, \quad \underline{\sigma}^3 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}.$$

CHAPTER 2

QUARKS

2.1. Internal Symmetries

The hadrons we know all fall into multiplets that reflect underlying internal symmetries. To express this fact in a simple and concrete way, it was hypothesized that hadrons are composed of more elementary constituents with certain basic symmetries, called quarks.

The actual existence of quarks has been indirectly confirmed, in experiments that probe hadronic structure by means of electromagnetic and weak interactions, and with the discovery of heavy quark "atoms". The available evidence is consistent with the picture that hadrons participate in these interactions not as elementary entities, but through quarks. On the other hand, we have not completely understood why quarks have not been seen individually, although there are plausible explanations. This will be discussed in more detail in the chapter on quark confinement.

Internal symmetry refers to the fact that particles occur in families, called multiplets, that have degenerate or nearly degenerate masses. Each multiplet is looked upon as the realization of an irreducible representation of some internal symmetry group. One tries to identify such groups by the patterns of multiplets observed experimentally. If the masses in a multiplet are not exactly the same, one says that the associated symmetry is only an approximate one. Among hadrons, we have long recognized the internal symmetries I , S , B and Q , with varying degrees of exactness, as indicated in Table 2.1

We define the hypercharge Y by

$$Y \equiv B + S, \quad (2.1)$$

and state the empirical rule

$$Q = I_3 + \frac{1}{2}Y. \quad (2.2)$$

Table 2.1 INTERNAL SYMMETRIES OF HADRONS

Symbol	Quantum number	Symmetry group	Interactions conserving it	Interactions violating it
I	Isospin	$SU(2)$	strong	em, weak
S	Strangeness	$U(1)$	strong, em	weak
B	Baryon number	$U(1)$	all	none
Q	Charge	$U(1)$	all	none

1 Isospin

Let us review the familiar case of isospin. Experimental evidence suggests that the nucleons and the π -mesons may be grouped into the following multiplets:

$$N = \begin{pmatrix} p \\ n \end{pmatrix}, \quad \pi = \begin{pmatrix} \pi_1 \\ \pi_2 \\ \pi_3 \end{pmatrix}, \quad (2.3)$$

where π_1, π_2, π_3 are related to the observed pions by $\pi^\pm = 2^{-1/2} (\pi_1 \pm \pi_2)$, and $\pi^0 = \pi_3$. Members of each multiplet have nearly equal masses, and the small mass differences can be thought of as electromagnetic corrections.

Apart from electromagnetic corrections, systems of nucleons and pions are invariant under matrix transformations representing the isospin $SU(2)$ group:

$$\begin{aligned} N &\rightarrow N + \delta N, & \delta N &= -i\omega_\alpha I_\alpha N \\ \pi &\rightarrow \pi + \delta\pi, & \delta\pi &= -i\omega_\alpha I_\alpha \pi \end{aligned} \quad (2.4)$$

where ω_α are arbitrary infinitesimal real numbers, and the components of isospin I_α ($\alpha = 1, 2, 3$) are the generators of $SU(2)$, obeying the commutation relation

$$[I_\alpha, I_\beta] = i\epsilon_{\alpha\beta\gamma} I_\gamma. \quad (2.5)$$

The nucleon doublet forms a basis for a 2-dimensional representation^a, in which $2I_\alpha$ is represented by the 2×2 Pauli matrix τ_α :

$$\begin{aligned} 2: \quad I_\alpha &= \frac{1}{2}\tau_\alpha, \quad \tau_1 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad \tau_2 = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}, \quad \tau_3 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}. \\ I_\alpha N &= \frac{1}{2}\tau_\alpha N. \end{aligned} \quad (2.6)$$

$$\delta N_i = -\frac{i}{2} [\omega_1(\tau_1)_{ij} + \omega_2(\tau_2)_{ij} + \omega_3(\tau_3)_{ij}] N_j.$$

The π -meson triplet forms a basis for a 3-dimensional irreducible representation:

$$\begin{aligned} 3: \quad (I_\alpha)_{\beta\gamma} &= -i\epsilon_{\alpha\beta\gamma}. \\ (I_\alpha \pi)_\beta &= (I_\alpha)_{\beta\gamma} \pi_\gamma = -i\epsilon_{\alpha\beta\gamma} \pi_\gamma. \\ \delta \pi_\beta &= \epsilon_{\beta\alpha\gamma} \omega_\alpha \pi_\gamma, \\ \text{or, } \delta \pi &= \omega \times \pi. \end{aligned} \quad (2.7)$$

The representation 3 is special, in that its dimensionality equals the number of generators, and that the matrix representation for I_α is obtainable directly from the structure constants $\epsilon_{\alpha\beta\gamma}$ in (2.5). It is called the *adjoint representation*. Other irreducible representations of $SU(2)$ are familiar from the theory of angular momentum. Their possible dimensionalities are $2I + 1$ ($I = 0, 1/2, 1, 3/2, \dots$).

^a We use n to denote either an n -dimensional irreducible representation of a group, or the n -dimensional vector space on which the representation is realized.

The adjoint representation can also be represented in an alternative form as follows: if x_i ($i = 1, 2$) transforms as 2, i.e.,

$$2: \quad \delta x_i = -\frac{i}{2} \omega_\alpha (\tau_\alpha)_{ij} x_j, \quad (2.8)$$

then y_α ($\alpha = 1, 2, 3$), which transforms as 3, may be represented as

$$\begin{aligned} y_\alpha &= (x^\dagger I_\alpha x) = \frac{1}{2} x_i^* (\tau_\alpha)_{ij} x_j, \\ \text{or, } y &= \frac{1}{2} (x^\dagger \tau x). \end{aligned} \quad (2.9)$$

The equivalence between (2.9) and (2.7) can be shown by the following calculation:

$$\begin{aligned} \delta y_\alpha &= (\delta x^\dagger I_\alpha x) + (x^\dagger I_\alpha \delta x) \\ &= i \omega_\beta (x^\dagger [I_\beta, I_\alpha] x) \\ &= -\omega_\beta \epsilon_{\beta\alpha\gamma} (x^\dagger I_\gamma x) \\ &= -i \omega_\beta (-i \epsilon_{\beta\alpha\gamma}) y_\gamma. \end{aligned} \quad (2.10)$$

2 The Gauge Groups

The quantum numbers B, Q, S label one-dimensional representations of mutually commuting groups isomorphic to $U(1)$, the unitary group of dimension one. The group operation is multiplication of the particle state by a phase factor:

$$\begin{aligned} B: \quad \psi &\rightarrow e^{-i\alpha B} \psi, & B &= \text{baryon number of } \psi, \\ Q: \quad \psi &\rightarrow e^{-i\beta Q} \psi, & Q &= \text{charge of } \psi, \\ S: \quad \psi &\rightarrow e^{-i\gamma S} \psi, & S &= \text{strangeness of } \psi, \end{aligned} \quad (2.11)$$

where α, β, γ are arbitrary real numbers.

3 More General Internal Symmetries: $SU(n)$

More generally, particles may fall into multiplets forming representations of $SU(n)$, the group isomorphic to that of all $n \times n$ special unitary complex matrices U ($\det U = 1, U^\dagger U = 1$). The condition $\det U = 1$ singles out a connected subgroup of the group of matrices. The requirement $U^\dagger U = 1$ insures that the norms of particle states are preserved under the group transformations.

A general $n \times n$ complex matrix has $2n^2$ arbitrary real parameters. The requirement $U^\dagger U = 1$ imposes n^2 conditions, and $\det U = 1$ imposes one condition. Hence, there remains $n^2 - 1$ arbitrary parameters. Correspondingly, $SU(n)$ has $n^2 - 1$ generators L_α obeying

$$[L_\alpha, L_\beta] = i f_{\alpha\beta\gamma} L_\gamma. \quad (2.12)$$

An arbitrary infinitesimal element of the group is given by

$$U = 1 - i \omega_\alpha L_\alpha, \quad (2.13)$$

where ω_α are arbitrary infinitesimal real numbers. The generators L_α can be taken to be hermitian. The structure constants $f_{\alpha\beta\gamma}$ can be taken to be real and

completely antisymmetric with respect to α, β, γ . The smallest non-trivial irreducible representation (the fundamental representation) is of dimensionality n by definition. There always exists the adjoint representation, whose dimensionality equals the number of generators, with

$$(L_\alpha)_{\beta\gamma} = -if_{\alpha\beta\gamma}.$$

The possible dimensionalities of other irreducible representations depend on n .

4 Unitary Symmetry

The approximate symmetries corresponding to isospin and hypercharge conservation can be enlarged into the so-called “unitary symmetry”, associated with the group $SU(3)$. This symmetry, however, is approximate even with respect to the strong interactions. The motivation for the enlargement comes from the observation that hadrons can be grouped into larger multiplets containing isospin multiplets. For example, one can recognize an octet of spin 1/2 baryons (the N-octet), a decaplet of spin 3/2 baryons (the Δ^- decaplet), an octet of spin 0 mesons (the π -octet) and an octet of spin 1 mesons (the ρ -octet). These are indicated in Fig. 1.1 and displayed in Fig. 2.1 in the form of Y - I_3 plots.

Note that baryons and antibaryons form their own separate multiplets, whereas mesons and antimesons are in the same multiplet. The identification of $SU(3)$ as the relevant symmetry group is based on the fact that **8** and **10** are possible irreducible representations of that group, of which **8** is the adjoint representation.

The violation of unitary symmetry by the strong interactions is reflected in the large mass splittings within the multiplets. Gell-Mann and Okubo have proposed a mass formula¹, based on the assumption that the interaction that violates unitary symmetry transforms like Y under $SU(3)$:

$$\begin{aligned} M(m, n) = & a + bY - c[2I(I+1) - \frac{1}{2}Y^2 + \frac{4}{3}(n-m)Y] \\ & - \frac{1}{3}m(m+2) - \frac{1}{3}n(n+2) + \frac{1}{9}(m-n)^2, \end{aligned} \quad (2.14)$$

where a, b, c , are empirical constants, and (m, n) labels the irreducible representation: **8** corresponds to $(1, 1)$, and **10** corresponds to $(0, 3)$ (see Sec. 2.2). This formula accounts reasonably well for the observed mass splittings, and is historically important for establishing the case for unitary symmetry.

It is natural to surmise that the basic representation **3** might also be realized, and this leads to the quark hypothesis. The main motivation, of course, is to decrease the number of fundamental particles; but the quark hypothesis also gives a natural explanation for the difference between baryon and meson multiplets with respect to the inclusion of antiparticles, if quarks are assumed to be baryons (see Sec. 2.3). The mechanism for the violation of unitary symmetry will find a simple and concrete origin in the mass differences among different kinds of quarks.

¹ See M. Gell-Mann and Y. Ne'eman, *The Eightfold Way* (W. A. Benjamin, New York, 1964).

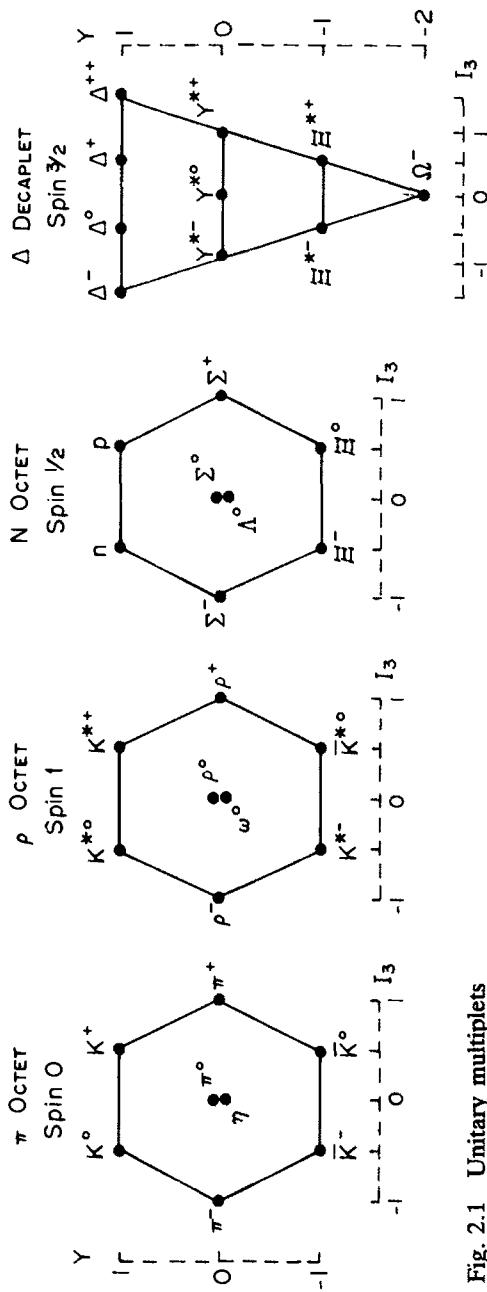


Fig. 2.1 Unitary multiplets

2.2 Representation of $SU(3)^2$

1 The basic representation

The group $SU(3)$ has 8 generators L_α ($\alpha = 1, \dots, 8$), satisfying

$$[L_\alpha, L_\beta] = if_{\alpha\beta\gamma}L_\gamma, \quad (2.15)$$

where $f_{\alpha\beta\gamma}$ are real constants completely antisymmetric in α, β, γ . The basic representation is 3, in which the generators are written in the form

$$L_\alpha = \frac{1}{2}\lambda_\alpha \quad (\alpha = 1, \dots, 8), \quad (2.16)$$

where λ_α are 3×3 hermitian matrices. They act on basis vectors, of the form

$$x = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}. \quad (2.17)$$

An infinitesimal element of the group is represented by the transformation

$$x' = Sx,$$

$$S = 1 - \frac{i}{2} \omega_\alpha \lambda_\alpha, \quad (2.18)$$

where ω_α ($\alpha = 1, \dots, 8$) are arbitrary infinitesimal real numbers. A set of matrices satisfying (2.15) is given in Table 2.2. By construction, the first three are respectively $2I_1, 2I_2, 2I_3$. The last one, which is diagonal and commutes with isospin, is identified with $\sqrt{3}Y$. These are called "Gell-Mann matrices", and are generalizations of the Pauli matrices.

The structure constants $f_{\alpha\beta\gamma}$ can be calculated by explicit commutation of matrices λ_α , and the results are listed in Table 2.3. They hold, of course, for any

Table 2.2 GELL-MANN MATRICES

$\lambda_1 = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$	$\lambda_4 = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}$	$\lambda_7 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & -i \\ 0 & i & 0 \end{pmatrix}$
$\lambda_2 = \begin{pmatrix} 0 & -i & 0 \\ i & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$	$\lambda_5 = \begin{pmatrix} 0 & 0 & -i \\ 0 & 0 & 0 \\ i & 0 & 0 \end{pmatrix}$	$\lambda_8 = \frac{1}{\sqrt{3}} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -2 \end{pmatrix}$
$\lambda_3 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 0 \end{pmatrix}$	$\lambda_6 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}$	

² For a general reference, see W. Miller, *Symmetry Groups and Their Applications* (Academic Press, New York, 1972).

representation of the group. However, other properties given in Table 2.3 are valid only for the fundamental representation.

We note that, in the representation given in Table 2.2, λ_4 and λ_5 form two members of another set of Pauli-like matrices (called *U*-spin). The same is true for λ_6 and λ_7 (called *V*-spin). We construct the missing member in each set by taking appropriate linear combinations of λ_3 and λ_8 . Thus, the original set of generators may be replaced by two other equivalent sets. These are listed in Table 2.4.

2 Young's tableaux

Taking the direct product of $\mathbf{3}$ with itself any number of times, we immediately obtain representations of higher dimensions $\mathbf{3} \times \mathbf{3}$, $\mathbf{3} \times \mathbf{3} \times \mathbf{3}$, etc. These representations, however, are reducible.

To decompose these reducible representations into irreducible ones, we decompose the corresponding product spaces into invariant irreducible subspaces^b. Then, the representations induced by the fundamental representation in these subspaces are irreducible representations.

Let $x_{i_1 \dots i_n}$ be a tensor that transforms like the product $x_{i_1} \dots x_{i_n}$. It can be decomposed into tensors of different symmetry classes with respect to a permutation of the indices i_1, \dots, i_n . By definition, a tensor belonging to a definite *symmetry class* is obtained from $x_{i_1 \dots i_n}$ through the following construction:

Table 2.3 PROPERTIES OF λ_α

$[\lambda_\alpha, \lambda_\beta] = 2if_{\alpha\beta\gamma}\lambda_\gamma$		$\{\lambda_\alpha, \lambda_\beta\} = \frac{4}{3}\delta_{\alpha\beta} \mathbf{1} + 2d_{\alpha\beta\gamma}\lambda_\gamma$	
$\alpha\beta\gamma$	$f_{\alpha\beta\gamma}$ (antisymmetric)	$\alpha\beta\gamma$	$d_{\alpha\beta\gamma}$ (symmetric)
123	1	118	$1/\sqrt{3}$
147	$1/2$	146	$1/2$
156	$-1/2$	157	$1/2$
246	$1/2$	228	$1/\sqrt{3}$
257	$1/2$	247	$-1/2$
345	$1/2$	256	$1/2$
367	$-1/2$	338	$1/\sqrt{3}$
458	$\sqrt{3}/2$	344	$1/2$
678	$\sqrt{3}/2$	355	$1/2$
		366	$-1/2$
		377	$-1/2$
$\text{Tr } \lambda_\alpha = 0$		448	$-1/(2\sqrt{3})$
$\text{Tr } \lambda_\alpha \lambda_\beta = 2\delta_{\alpha\beta}$		558	$-1/(2\sqrt{3})$
$\text{Tr } \lambda_\alpha [\lambda_\beta, \lambda_\gamma] = 4if_{\alpha\beta\gamma}$		668	$-1/(2\sqrt{3})$
$\text{Tr } \lambda_\alpha \{\lambda_\beta, \lambda_\gamma\} = 4id_{\alpha\beta\gamma}$		778	$-1/(2\sqrt{3})$
		888	$-1/\sqrt{3}$

^b "Invariant" means the space goes into itself under the group transformations. "Irreducible" means it does not contain a smaller invariant subspace.

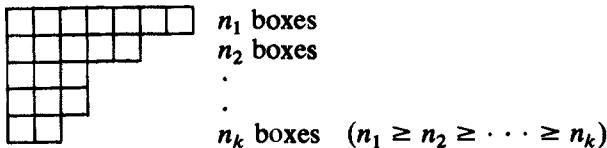
(a) First, pick n_1 of the indices, and symmetrize among them. Display this operation symbolically in the picture below:

 n_1 boxes (Fill in the chosen indices)

(b) Next, pick $n_2 \leq n_1$ of the remaining indices, and symmetrize among them:

 n_2 boxes (Fill in the chosen indices)

(c) Repeat the procedure until all indices have been used. Stack up the rows to form the following tableau, called a *Young's tableau*:



The set of integers $\{n_1, \dots, n_k\}$ is any possible partition of n :

$$n = n_1 + \dots + n_k, \quad (n_1 \geq n_2 \geq \dots \geq n_k). \quad (2.19)$$

Table 2.4 EQUIVALENT SETS OF $SU(3)$ GENERATORS

(1) $\underbrace{\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5, \lambda_6, \lambda_7, \lambda_8}_{2(I\text{-spin})} (= \sqrt{3}Y)$

(2) $\lambda_1, \lambda_2, \rho, \underbrace{\lambda_4, \lambda_5, \lambda_6, \lambda_7, \rho'}_{2(U\text{-spin})} (= \sqrt{3}Q)$

(3) $\lambda_1, \lambda_2, \lambda_4, \lambda_5, e, \underbrace{\lambda_6, \lambda_7, \epsilon'}_{2(V\text{-spin})}$

where

$$\rho = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & -1 \end{pmatrix} = \frac{1}{2}(\lambda_3 + \sqrt{3}\lambda_8) = I_3 + \frac{3}{2}Y$$

$$\rho' = \frac{1}{\sqrt{3}} \begin{pmatrix} 2 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -1 \end{pmatrix} = \frac{1}{2}(\sqrt{3}\lambda_3 + \lambda_8) = \sqrt{3}\left(I_3 + \frac{1}{2}Y\right) = \sqrt{3}Q$$

$$e = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -1 \end{pmatrix} = \frac{1}{2}(\sqrt{3}\lambda_8 - \lambda_3) = \frac{3}{2}Y - I_3$$

$$e' = \frac{1}{\sqrt{3}} \begin{pmatrix} -1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & -1 \end{pmatrix} = \frac{1}{2}(\lambda_8 - \sqrt{3}\lambda_3) = \sqrt{3}\left(\frac{1}{2}Y - I_3\right)$$

(d) Finally, antisymmetrize the indices in each column of the tableau separately and independently.

The geometrical pattern of a Young's tableau characterizes a symmetry class. The number of classes is therefore equal to the number of possible partitions of n . Tensors belonging to a given symmetry class can differ from one another only in the choice of indices in the various boxes of the Young's tableau. It is a matter of convention that the rows are permuted before the columns. In this convention, the indices appearing in each column of a Young's tableau are antisymmetric among themselves; but the indices appearing in each row are symmetric under a permutation if and only if the indices being permuted are not antisymmetrized with indices of a different row. For example:

These are not independent tensors:

$$\begin{array}{|c|c|c|} \hline 1 & 3 & 4 \\ \hline 2 & & \\ \hline \end{array} = - \begin{array}{|c|c|c|} \hline 2 & 3 & 4 \\ \hline 1 & & \\ \hline \end{array}, \quad (\text{where 1 means } i_1, \text{ etc.}).$$

These are not independent tensors:

$$\begin{array}{|c|c|c|} \hline 1 & 3 & 4 \\ \hline 2 & & \\ \hline \end{array} = \begin{array}{|c|c|c|} \hline 1 & 4 & 3 \\ \hline 2 & & \\ \hline \end{array}$$

These are independent tensors:

$$\begin{array}{|c|c|c|} \hline 1 & 3 & 4 \\ \hline 2 & & \\ \hline \end{array} + \begin{array}{|c|c|c|} \hline 3 & 1 & 4 \\ \hline 2 & & \\ \hline \end{array}$$

The central theorem, which we give without proof, is the following:

(a) Tensors in a given symmetry class form an invariant irreducible space. Hence the group representation induced in this space by the fundamental representation is irreducible.

(b) The irreducible representations generated through all symmetry classes are exhaustive.

3 Irreducible Representations

Since a tensor index takes on only three values $i = 1, 2, 3$, a column in a Young's tableau can have no more than three boxes. The tableau



represents a tensor of rank 0, and corresponds to the 1-dimensional trivial representation (i.e., it is invariant under the group). Such a tableau may be omitted, if it occurs as part of a larger tableau:

$$1 = \begin{array}{|c|c|} \hline \square & \square \\ \hline \square & \\ \hline \end{array} = \begin{array}{|c|c|c|} \hline \square & \square & \square \\ \hline \square & & \\ \hline \square & & \\ \hline \end{array} = \begin{array}{|c|c|c|c|} \hline \square & \square & \square & \square \\ \hline \square & & & \\ \hline \square & & & \\ \hline \square & & & \\ \hline \end{array} = \dots \quad (2.20)$$

Consequently, the most general Young's tableau for $SU(3)$ has at most two rows:

$$\begin{array}{|c|c|c|c|c|c|c|c|c|} \hline k_1 & k_2 & \cdots & k_m & i_1 & i_2 & \cdots & i_n \\ \hline l_1 & l_2 & \cdots & l_m & & & & \\ \hline \end{array} = \left(\begin{matrix} k_1 & \cdots & k_m \\ l_1 & \cdots & l_m \end{matrix} \middle| i_1 \cdots i_n \right) \quad (2.21)$$

An irreducible representation is therefore completely specified by two integers (m, n) .

It is convenient to replace the antisymmetric pair (k_λ, l_λ) by another index j_λ , in the following manner:

$$x_{i_1 \dots i_n}^{j_1 \dots j_m} = \begin{pmatrix} k_1 \dots k_m \\ l_1 \dots l_m \end{pmatrix}_{i_1 \dots i_n} \epsilon^{j_1 k_1 l_1} \dots \epsilon^{j_m k_m l_m}. \quad (2.22)$$

This tensor is symmetric in $\{j_1 \dots j_m\}$ and in $\{i_1 \dots i_n\}$. It is straightforward to show that

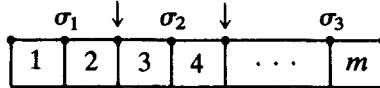
$$\sum_{j=1}^3 x_j^{j_1 j_2 \dots j_m} = 0. \quad (2.23)$$

Thus, components of the tensor $x_{i_1 \dots i_n}^{j_1 \dots j_m}$ are not all independent. The number of independent components is the dimension $D(m, n)$ of the irreducible representation (m, n) .

We now calculate $D(m, n)$. Suppose, among $\{j_1 \dots j_m\}$, 1 occurs σ_1 times, 2 occurs σ_2 times, and 3 occurs σ_3 times. Then the number of possible sets $\{j_1 \dots j_m\}$ is equal to

$$P_m = \text{No. of ordered sets } \{\sigma_1, \sigma_2, \sigma_3 | \sigma_1 + \sigma_2 + \sigma_3 = m\}. \quad (2.24)$$

Any possible set $\{\sigma_1, \sigma_2, \sigma_3\}$, with $\sigma_2 > 0$, can be chosen by choosing *two* of the dots in the picture below, [in $\binom{m+1}{2}$ ways]:



Any possible set $\{\sigma_1, \sigma_2, \sigma_3\}$ can be chosen by choosing only *one* dot, [in $\binom{m+1}{1}$ ways]. Hence,

$$P_m = \binom{m+1}{2} + \binom{m+1}{1} = \frac{1}{2}(m+1)(m+2). \quad (2.25)$$

The total number of ways of choosing $\{j_1 \dots j_m\}$ and $\{i_1 \dots i_n\}$ independently is $P_m P_n$. The requirement (2.23) imposes $P_{m-1} P_{n-1}$ conditions. Therefore $D(m, n) = P_m P_n - P_{m-1} P_{n-1}$, or

$$D(m, n) = \frac{1}{2}(n+1)(m+1)(n+m+2). \quad (2.26)$$

The representations (m, n) and (n, m) have the same dimensionality, and are said to be conjugate to each other:

$$\overline{(m, n)} = (n, m). \quad (2.27)$$

The representation (n, n) is self-conjugate, with dimensionality $D(n, n) = (n+1)^3$.

The matrices representing L_α in the irreducible representation (m, n) can be worked out from the transformation law,

$$x'^{j_1 \dots j_m}_{i_1 \dots i_n} = (S^*_{j_1 j_1'} \dots S^*_{j_m j_m'})(S_{i_1 i_1'} \dots S_{i_n i_n'}) x^{j_1' \dots j_m'}_{i_1' \dots i_n'}, \quad (2.28)$$

where S_{ij} is given in (2.18). This states that *an upper index transforms just like a lower index, except that λ_α is replaced by $-\lambda_\alpha^*$* .

The antisymmetric tensor ϵ^{ijk} can be used to raise or lower indices:

$$x^i = \epsilon^{ijk} A_{jk}, \quad (A_{jk} = -A_{kj}), \quad (2.29)$$

where A_{jk} transforms like $x_j x_k$.

In Table 2.5 we list some irreducible representations, in increasing order of dimensionality. Table 2.6 illustrates the rules for decomposing product representations into sums of irreducible representations.

2.3 The Quark Model³

1 Quarks as basic triplets

We think of the vector components x_i for the fundamental representation $\mathbf{3}$ as particle states called quarks. Those for the conjugate representation $\bar{\mathbf{3}}$ are called antiquarks^c. Then all higher representations can be regarded as composite states of quarks and/or antiquarks.

Actually, from the point of view of representing $SU(3)$, all we need is $\mathbf{3}$, because $\mathbf{3}$ can be generated by $\mathbf{3} \times \mathbf{3} = \bar{\mathbf{3}} + \mathbf{6}$, as shown in Table 2.6. However, we want to be able to distinguish between particle and antiparticle. To do this, we assign quarks and antiquarks to $\mathbf{3}$ and $\bar{\mathbf{3}}$ respectively, and assign them equal and opposite baryon numbers B , which is the generator of a $U(1)$ group that commutes with $SU(3)$. We assign to quarks $B = 1/3$, and to antiquarks $B = -1/3$. Then $\bar{\mathbf{3}}$ occurring in $\mathbf{3} \times \mathbf{3}$ is distinct from an antiquark, for it has $B = 2/3$. [Of course, under $SU(3)$ alone, it does transform like an antiquark].

If we take quarks seriously, we must eventually face the question of their interactions. Whether we regard quarks as real objects or mere mathematical constructions, the quantum numbers of their bound states are determined purely group-theoretically.

In accordance with current convention, we name the quarks u, d, s ; which stand respectively for “up”, “down” and “strange”. These are said to be the different “flavors” of a quark. Similarly, the antiquarks are named $\bar{u}, \bar{d}, \bar{s}$. There are three other known flavors, but we limit our discussion to these for simplicity. More explicitly, we write

$$x = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} u \\ d \\ s \end{pmatrix}, \quad \bar{x} = \begin{pmatrix} x^1 \\ x^2 \\ x^3 \end{pmatrix} = \begin{pmatrix} \bar{u} \\ \bar{d} \\ \bar{s} \end{pmatrix}. \quad (2.30)$$

^c Note that $\mathbf{3}$ and $\bar{\mathbf{3}}$ are distinct, unlike the situation in $SU(2)$, where $\mathbf{2}$ and $\bar{\mathbf{2}}$ are equivalent.

³ For a general reference, see J. J. J. Kokkedee, *The Quark Model* (W. A. Benjamin, New York, 1969).

Thus, a u -quark corresponds to $\begin{pmatrix} u \\ 0 \\ 0 \end{pmatrix}$, a d -quark corresponds to $\begin{pmatrix} 0 \\ d \\ 0 \end{pmatrix}$, etc. We define isospin and hypercharge Y by

$$\begin{aligned} I_\alpha &= \frac{1}{2}\lambda_\alpha \quad (\alpha = 1, 2, 3), \\ Y &= \lambda_8/\sqrt{3}. \end{aligned} \quad (2.31)$$

Table 2.5 IRREDUCIBLE REPRESENTATIONS OF $SU(3)$

(m, n)	$D(m, n)$	Tableau	Tensor
	1		1
(0, 1)	3		$x_i (i = 1, 2, 3)$
(1, 0)	$\bar{3}$		x^i
(0, 2)	6		x_{ij}
(2, 0)	$\bar{6}$		x^{ij}
(1, 1)	$8 = \bar{8}$		$x^i \left(\sum_{i=1}^3 x_i^i = 0 \right)$ Adjoint rep.
(0, 3)	10		x_{ijk}
(3, 0)	$\bar{10}$		x^{ijk}
(1, 2)	15		$x_{ik}^i \left(\sum_{i=1}^3 x_{ik}^i = 0 \right)$
(2, 1)	$\bar{15}$		$x_k^{ij} \left(\sum_{i=1}^3 x_k^{ij} = 0 \right)$
(0, 4)	15'		x_{ijkl}
(4, 0)	$\bar{15}'$		x^{ijkl}
(1, 3)	24		$x_{jkl}^i \left(\sum_{i=1}^3 x_{jkl}^i = 0 \right)$
(3, 1)	$\bar{24}$		$x_l^{ijk} \left(\sum_{i=1}^3 x_l^{ijk} = 0 \right)$
(2, 2)	$27 = \bar{27}$		$x_{kl}^{ij} \left(\sum_{i=1}^3 x_{kl}^{ij} = 0 \right)$

Two of these matrices, namely I_3 and Y , are diagonal in the quark basis:

$$I_3 = \begin{pmatrix} 1/2 & 0 & 0 \\ 0 & -1/2 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \quad Y = \begin{pmatrix} 1/3 & 0 & 0 \\ 0 & 1/3 & 0 \\ 0 & 0 & -2/3 \end{pmatrix}. \quad (2.32)$$

The square of isospin, as well as strangeness S and charge Q , are all diagonal:

$$I(I+1) \equiv I_1^2 + I_2^2 + I_3^2 = \begin{pmatrix} 3/4 & 0 & 0 \\ 0 & 3/4 & 0 \\ 0 & 0 & 0 \end{pmatrix},$$

$$S \equiv Y - B = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & -1 \end{pmatrix}, \quad (2.33)$$

$$Q \equiv I_3 + \frac{1}{2}Y = \begin{pmatrix} 2/3 & 0 & 0 \\ 0 & -1/3 & 0 \\ 0 & 0 & -1/3 \end{pmatrix}.$$

By (2.28), the I_3 and Y for antiquarks are the negatives of those for quarks.

Table 2.6 DECOMPOSITION OF PRODUCT REPRESENTATIONS

(a) $3 \times 3 = \square \times \square = \boxed{\square} + \boxed{\square\square} = \bar{3} + 6$

(b) $\bar{3} \times 3 = \boxed{\square} \times \square = \boxed{\square\square} + \boxed{\square\square\square} = 1 + 8$

(c) $6 \times 3 = \boxed{\square\square} \times \square = \boxed{\square\square\square} + \boxed{\square\square\square\square} = 8 + 10$

(d) $3 \times 3 \times 3 = (\bar{3} + 6) \times 3 = 1 + 8 + 8 + 10$

(e)* $8 \times 8 = \boxed{\begin{matrix} 1 & 2 \\ 3 & 3' \end{matrix}} \times \boxed{\begin{matrix} 1' & 2 \\ 3 & 3' \end{matrix}} = 1 + 8 + 8 + 10 + 10 + 27$

$\boxed{\begin{matrix} 1 & 1' & 2 & 2' \\ 3 & 3' \end{matrix}}$	$\boxed{\begin{matrix} 1 & 1' & 2 & 2' \\ 3 \\ 3' \end{matrix}}$	$\boxed{\begin{matrix} 1 & 3' & 2 \\ 3 & 1' & 2' \end{matrix}}$	(1 means i_1 etc)
27	10	$\bar{10}$	
$\boxed{\begin{matrix} 1' & 2' & 3 \\ 1 & 2 \\ 3' \end{matrix}}$	$\boxed{\begin{matrix} 1' & 2 & 3' \\ 1 & 2 \\ 3 \end{matrix}}$	$\boxed{\begin{matrix} 1 & 1' \\ 2 & 2' \\ 3 & 3' \end{matrix}}$	
8	8	1	

* Arrange the six labelled blocks in all possible combinations, but preserve antisymmetry in (1, 3) or (1', 3'), if both labels remain in the final tableau.

Consequently, they also have opposite signs for S and Q . However, $I(I + 1)$ is the same for quarks and antiquarks. These quantum numbers are tabulated in Table 2.7.

2 Quarks as building blocks

The irreducible representation (m, n) is realized in the space of the tensors $x_{i_1 \dots i_n}^{j_1 \dots j_m}$, whose independent components may be regarded as a multiplet of $D(m, n)$ particles, composed of quarks and/or antiquarks. We require each multiplet to have a definite baryon number that is the same for all particles in that multiplet. The quark and antiquark content of a multiplet is then uniquely determined.

In the tensor $x_{i_1 \dots i_n}^{j_1 \dots j_m}$, each lower index can be associated either with a quark, or an antisymmetric pair of antiquarks, according to (2.29). The same rule applies to an upper index, with quark and antiquark interchanged. The different choices are distinguished by baryon number. These possibilities are summarized in Table 2.8.

A straightforward application of (2.28) leads to the following theorem:

- (a) If λ is a generator that is diagonal in the fundamental representation 3, then it is diagonal in any irreducible representation.
- (b) If the eigenvalues of λ in the fundamental representation are c_i ($i = 1, 2, 3$), then in the irreducible representation (m, n) , where λ acts on $x_{i_1 \dots i_n}^{j_1 \dots j_m}$, the eigenvalues of λ are

$$(c_{i_1} + \dots + c_{i_n}) - (c_{j_1} + \dots + c_{j_m}).$$

In the quark model, the theorem merely states that in forming a composite state of quarks and antiquarks, their I_3 and Y are additive. It follows that charge and strangeness are also additive.

3 Weight diagrams

A convenient way to display the structure of a multiplet is to show all its components on a I_3 - Y plot, called the *weight diagram* of the irreducible

Table 2.7 QUANTUM NUMBERS OF QUARKS

	Quarks			Antiquarks		
	u	d	s	\bar{u}	\bar{d}	\bar{s}
I	1/2	1/2	0	1/2	1/2	0
I_3	1/2	-1/2	0	-1/2	1/2	0
Y	1/3	1/3	-2/3	-1/3	-1/3	2/3
Q	2/3	-1/3	-1/3	-2/3	1/3	1/3
B	1/3	1/3	1/3	-1/3	-1/3	-1/3
S	0	0	-1	0	0	1

representation. The weight diagrams for the quark $\mathbf{3}$ and the antiquark $\bar{\mathbf{3}}$ are shown in Fig. 2.2.

The weight diagram for (m, n) is constructed by identifying $D(m, n)$ lattice sites (not necessarily distinct), on the $Y-I_3$ plot, as follows. In Fig. 2.3, treat the vectors x_1, x_2, x_3 as lattice displacement vectors. Then $x_{i_1 \dots i_m}^{j_1 \dots j_n}$ is located at $(x_{i_1} + \dots + x_{i_m}) - (x_{j_1} + \dots + x_{j_n})$. The possible lattice sites form a diamond lattice.

Since different components may occupy the same site, each occupied site is characterized by a degeneracy. To obtain the correct degeneracy, it is important

Table 2.8 CORRESPONDENCE BETWEEN $SU(3)$ INDEX AND QUARK CONTENT

Each lower index i:

$$\text{Either } i = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix} \leftrightarrow \begin{pmatrix} u \\ d \\ s \end{pmatrix} \quad B = 1/3$$

$$\text{or* } i = \begin{bmatrix} (23) \\ (31) \\ (12) \end{bmatrix} \leftrightarrow \begin{bmatrix} (\bar{d}\bar{s}) \\ (\bar{s}u) \\ (\bar{u}\bar{d}) \end{bmatrix} \quad B = -2/3$$

Each upper index j:

$$\text{Either } j = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix} \leftrightarrow \begin{pmatrix} \bar{u} \\ \bar{d} \\ \bar{s} \end{pmatrix} \quad B = -1/3$$

$$\text{or* } j = \begin{bmatrix} (23) \\ (31) \\ (12) \end{bmatrix} \leftrightarrow \begin{bmatrix} (ds) \\ (su) \\ (ud) \end{bmatrix} \quad B = 2/3$$

* See Eq. 2.34 for definition of (ds) etc.

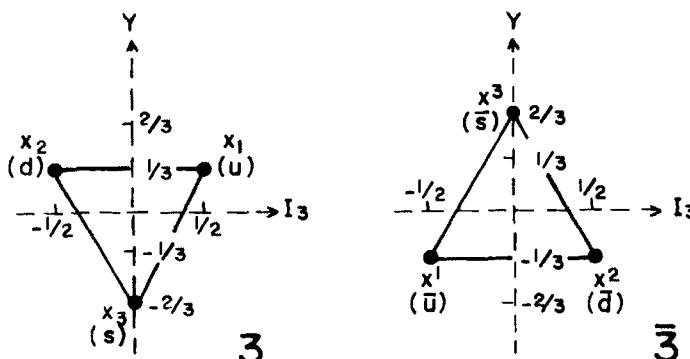


Fig. 2.2 Weight diagrams for $\mathbf{3}$ and $\bar{\mathbf{3}}$

to observe that a permutation of the upper or the lower indices among themselves does not lead to a distinct component. Furthermore, the condition (2.23) must be used to eliminate redundant components. Instead of giving general rules, we shall just work out some physically relevant examples.

The weight diagram, being a property of $SU(3)$ alone, knows nothing about baryon number. That is, multiplets of different baryon number have the same weight diagram, as long as they correspond to the same irreducible representation of $SU(3)$. The baryon number merely determines the quark content of the members of the multiplet.

Since no hadron of fractional baryon number has ever been observed, we shall consider as physical only multiplets with integer B , and assume that for some reason not understandable within this model, those with fractional B do not occur in nature. By this criterion, a multiplet is admissible only if it corresponds to a Young's tableau whose total number of boxes is divisible by 3 (since one box corresponds to one quark). From Table 2.5, we see that these are **8**, **$\bar{10}$** , **10**, **27**, . . . ; and this is consistent with experimental facts.

The weight diagram of **8**, **$\bar{10}$** and **10** are given in Fig. 2.4. For **10** and **$\bar{10}$** the lattice sites have no degeneracies. For **8**, the central site has a two-fold degeneracy (it is occupied by x_1^1, x_2^2, x_3^3 ; but only two are independent because $x_1^1 + x_2^2 + x_3^3 = 0$). The degenerate states are chosen to be $2^{-1/2} (x_1^1 + x_2^2)$ and $2^{-1/2} (x_1^1 - x_2^2)$. By noting that under $SU(3)$, they respectively transform like $\bar{u}u + \bar{d}d$ and $\bar{u}u - \bar{d}d$, it is clear that $2^{-1/2} (x_1^1 - x_2^2)$ is an isosinglet, while $2^{-1/2} (x_1^1 + x_2^2)$ forms an isotriplet with x_1^2 and x_2^1 .

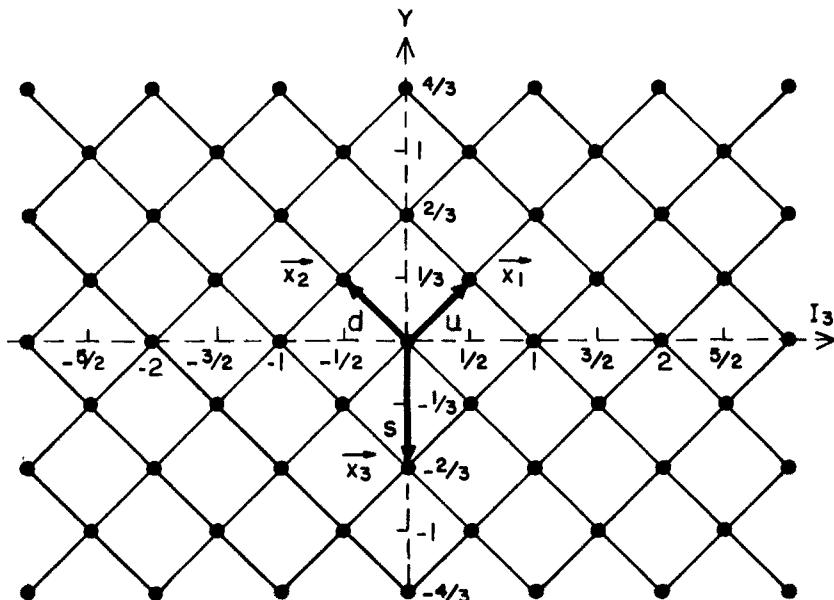


Fig. 2.3 Lattice on which weight diagrams are constructed

4 The composition of hadrons

We now examine the possible quark contents of the multiplets **8**, **$\bar{10}$** , and **10**. According to Table 2.8, **8** may have either $B = 0$ or $B = \pm 1$.

For $B = 0$, each member of the multiplet is composed of a quark and an antiquark. The quark contents are indicated in Fig. 2.5(a). We see that particles and antiparticles are included in the multiplet. Thus, the π -octet and ρ -octet can both be accounted for.

For $B = 1$, the quark contents of **8** are indicated in Fig. 2.5(b), where parentheses indicate antisymmetrization, for example,

$$(ud) \equiv u(1) d(2) - u(2) d(1), \quad (2.34)$$

where the labels 1 and 2 identify the two different quarks. That is, $\{u(1), d(1), s(1)\}$ and $\{u(2), d(2), s(2)\}$ are two independent vectors on which $SU(3)$ acts. The case $B = -1$ corresponds to the antiparticle multiplet. Thus we can account for the N -octet and the separate \bar{N} -octet.

The **10** with $B = 1$ is shown in Fig. 2.5(c). This accounts for the Δ -decaplet. Correspondingly, **$\bar{10}$** with $B = -1$ accounts for the separate $\bar{\Delta}$ -decaplet.

In principle, there could be **10**'s with $B = 0$ ($qq\bar{q}\bar{q}$), $B = -1$ ($q\bar{q}q\bar{q}\bar{q}$), and $B = -2$ ($\bar{q}\bar{q}\bar{q}\bar{q}\bar{q}\bar{q}$), where q stands for a quark. These states are called "exotic", and do not seem to occur among hadronic states of low masses.

2.4 Color

1 Independent quark model⁴

The quark model so far is purely algebraic. It is just a way of representing $SU(3)$, with the baryon number trivially tagged on. In order to say something

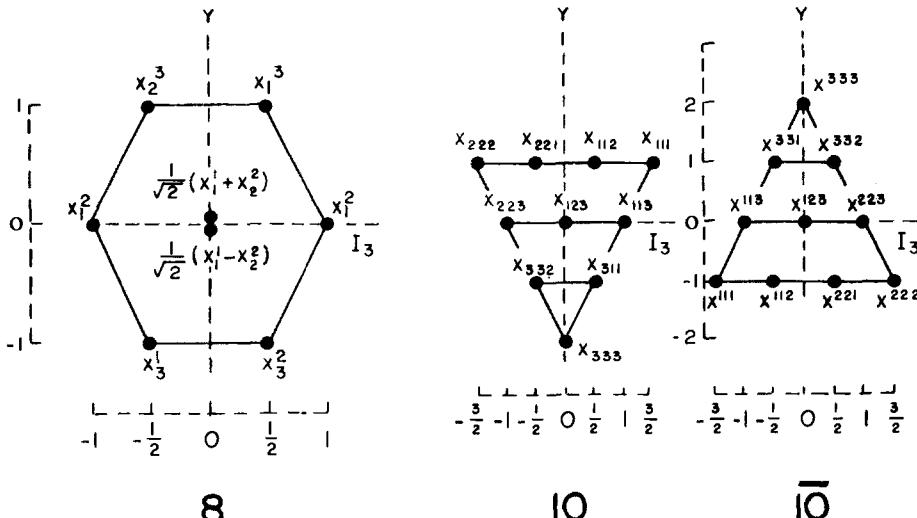


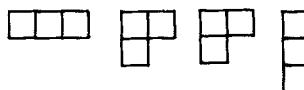
Fig. 2.4 Weight diagrams for **8**, **10** and **$\bar{10}$**

⁴ For a general reference, see J. J. J. Kokkedee, *op. cit.*

about the masses and the interactions of hadrons, we have to add dynamical content to the model, i.e., to construct a theory in space-time. At this point, it makes a difference whether or not we consider quarks to be real objects.

An extension of the model, still without dynamical content, is the $SU(6)$ quark model. One associates each quark with spin angular momentum, which generates the group $[SU(2)]_{\text{spin}}$ and embeds $[SU(3)]_{\text{flavor}} \times [SU(2)]_{\text{spin}}$ in a larger symmetry group $SU(6)$. The enlarged basis then consists of a sextet of quarks: $\{u(\uparrow), d(\uparrow), s(\uparrow), u(\downarrow), d(\downarrow), s(\downarrow)\}$. Among the irreducible representations of $SU(6)$, one finds a **35** with $B = 0$ (from $6 \times 6 = 1 + 35$). This can be identified with the union of the spin 0 π -octet and the spin 1 nonet consisting of the ρ -octet and the σ -singlet. One also finds a **56** with $B = 1$, identifiable with the union of the spin 1/2 N -octet and the spin 3/2 Δ -decaplet:

$$6 \times 6 \times 6 = \textbf{56} + \textbf{70} + \textbf{70} + \textbf{20}, \quad (2.35)$$



where corresponding Young's tableaux are indicated. As we can see, the **56** is completely symmetric under a permutation of the spin-flavors of the quarks.

The $SU(6)$ symmetry, however, distinguishes between spin and orbital angular momentum. This distinction can be made only in nonrelativistic models, and is inconsistent with Lorentz invariance. A straightforward relativistic extension of $SU(6)$ would be an embedding of $SU(6)$ in a larger group that contains the Poincaré group. Since internal symmetry would then be joined with space-time symmetries, such an extension might have dynamical content. However, all efforts along such directions have failed. In fact, there are theorems stating that a non-trivial imbedding is impossible. Nevertheless, the fact that $SU(6)$ correctly predicts the existence of the meson **35** and the baryon **56** may not be accidental. Perhaps we can view it as a sort of approximation to the non-relativistic limit of a more correct relativistic theory.

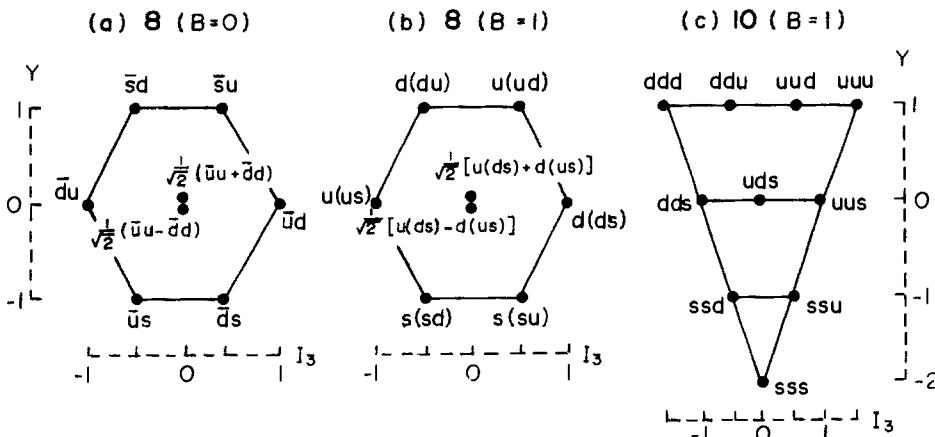


Fig. 2.5 Quark contents of **8** and **10**

From a phenomenological point of view, we can give the quark model some dynamical content by assuming that quarks are real particles moving in some effective potential, much like nucleons in the nuclear shell model. We call this an *independent quark model*, in which quarks occupy single-particle orbitals, which are described by spatial wave functions and have definite energies. Masses of hadrons can then be calculated in terms of the energies of occupied orbitals and the quark masses. The observed violation of unitary symmetry can be attributed to the fact that the s-quark has different mass from that of the *u* and *d*-quarks, and one can then understand the Gell-Mann-Okubo mass formula in a more concrete way.

Since quarks have spin 1/2, it is natural to assume that they are fermions and therefore obey the Pauli exclusion principle. That is, the total wave function of a hadron should be antisymmetric under the simultaneous interchange of flavor, spin, and orbital between two quarks. If we take the suggestion from (2.35) that the wave function of a baryon is symmetric under interchange of spin-flavor, then it must be antisymmetric under interchange of orbitals. Thus, three quarks in a baryon must be in different orbitals, only one of which can be the lowest *S* state (unless there are unsuspected degeneracies). On the other hand, magnetic moment calculations from the model agree with experiments only if there is no orbital contribution, suggesting that the three quarks are in the *S* state. This presents some sort of paradox.

The paradox is reinforced by considering the charge radii of hadrons. In a meson, the quark and antiquark can both be in the *S* state, and will be so, because that corresponds to the lowest energy. Since the charge radii of mesons are observed to be nearly the same as those for baryons, the three quarks in a baryon should all be in the *S* state.

One can always get out of the paradox by saying that $SU(6)$ is completely irrelevant, or that the independent quark model is totally wrong; but that would not lead us anywhere. Instead, one might resolve the paradox within the $SU(6)$ independent quark model by postulating that there are degeneracies which have so far been overlooked. The quantum number labelling this new degeneracy is called "color". If we assume that, for each flavor and spin, quarks come in different "colors", then all three quarks can be in the same state, as long as they have different colors. Clearly, this requires at least three colors. We now show the advantage of assuming *exactly* three colors.

2 Color $SU(3)$ group⁵

If there are only three colors, we can state a simple rule to insure that no isolated quark can be seen: *only "colorless" states can exist*. There is hope that this rule can be derived in a gauge theory based on color—quantum chromodynamics.

We assume, then, that there are exactly three colors, and that the world is invariant under color change. That is, color space is a representation space for a new $SU(3)$ symmetry group, denoted by $[SU(3)]_{\text{color}}$. A quark now transforms as

⁵ A review of the idea of color may be found in O. W. Greenberg and C. A. Nelson, *Physics Reports* 32C, 71 (1977).

q_{in} , with

$$\begin{aligned} \text{flavor index: } i &= 1, 2, 3 \quad (\text{or } u, d, s), \\ \text{color index: } n &= 1, 2, 3 \quad (\text{or red, yellow, green}). \end{aligned} \quad (2.36)$$

These indices transform respectively under $[SU(3)]_{\text{flavor}}$ and $[SU(3)]_{\text{color}}$. The antiquark \bar{q}_{in} transforms like q^{in} .

The empirical observation that there are no particles of fractional baryon number is guaranteed by the rule that *any physical state must be a color singlet*. This immediately implies that the number of quarks making up a baryon must be divisible by 3, as each quark corresponds to a square in a Young's tableau for $[SU(3)]_{\text{color}}$. (Note that this requirement does not depend on the number of flavors).

Since a color singlet corresponds to the Young's tableau in (2.20), three quarks in a baryon must be completely anti-symmetric in their color indices. Therefore, by the Pauli principle, they must be completely symmetric with respect to all other indices. Similarly, six quarks in a baryon (e.g., the deuteron) must consist of two color singlets. For a meson composed of quark and antiquark, its state must transform under $[SU(3)]_{\text{color}}$ as

$$\sum_{n=1}^3 q^{in} q_{jn} = \sum_{n=1}^3 \bar{q}_{in} q_{jn}. \quad (2.37)$$

As an illustration, let us write down the wave function for a proton in a non-relativistic independent quark model. We adopt a suitable notation for calculating matrix elements with respect to the wave function. The coordinates of a quark are collectively denoted by

$$z = \{n, i, s, \mathbf{r}\}, \quad (\text{coordinates}) \quad (2.38)$$

which are respectively the color, flavor, spin and position coordinates. The coordinates of the three quarks are denoted respectively by z_1, z_2, z_3 , with corresponding subscripts on n, i, s , and \mathbf{r} . The single-quark quantum numbers are denoted collectively by

$$\lambda = \{N, k, \sigma, l\}, \quad (\text{quantum numbers}) \quad (2.39)$$

which are respectively the color, flavor, spin and spatial quantum numbers. A single-quark wave function is written as

$$\psi_\lambda(z) = C_N(n) F_k(i) \chi_\sigma(s) R_l(\mathbf{r}). \quad (2.40)$$

Under internal symmetry operations, n and i transform like $SU(3)$ lower indices (i.e., like those on q_{in}), and s transforms like an $SU(2)$ index. The factors in (2.40) are chosen to be

$$\begin{aligned} C_N(n) &= \delta_{Nn}, \\ F_k(i) &= \delta_{ki}, \\ \chi_\sigma(s) &= \delta_{\sigma s}, \\ R_l(\mathbf{r}) &= R(r). \quad (S\text{-wave orbital}) \end{aligned} \quad (2.41)$$

It is convenient to write out all the components of C_N , F_k , χ_σ in the form of column vectors, and name these vectors

$$\begin{aligned} C_1 = r &= \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, & F_1 = u &= \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, & \chi_1 = \uparrow &= \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \\ C_2 = y &= \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, & F_2 = d &= \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, & \chi_2 = \downarrow &= \begin{pmatrix} 0 \\ 1 \end{pmatrix}. \\ C_3 = g &= \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}, & F_3 = s &= \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}. \end{aligned} \quad (2.42)$$

We can combine flavor and spin by writing

$$F_k \chi_\sigma = u_\uparrow, u_\downarrow, d_\uparrow, d_\downarrow, s_\uparrow, s_\downarrow. \quad (2.43)$$

Furthermore, as a shorthand, write

$$\begin{aligned} r(1) &\equiv r(n_1), \text{ etc.} \\ d_\uparrow(1) &\equiv d(i_1)[\uparrow(s_1)], \text{ etc.} \\ R(1) &\equiv R(r_1), \text{ etc.} \end{aligned} \quad (2.44)$$

The Pauli principle requires the total proton wave function $\Psi(1, 2, 3)$ to be completely antisymmetric under a permutation of 1, 2, 3. Therefore, $\Psi(1, 2, 3)$ is obtained by antisymmetrizing a linear combination of terms of the form $\psi_\lambda(1) \psi_\lambda(2) \psi_\lambda(3)$, the linear combination being dictated by unitary symmetry and spin. Requiring Ψ to be a color singlet, we write

$$\Psi(1, 2, 3) = \left[\sum_P \delta_P P_{123} r(1) y(2) g(3) \right] R(1) R(2) R(3) \Phi(1, 2, 3), \quad (2.45)$$

where P_{123} is a permutation of 1, 2, 3, and δ_P is its signature. The flavor-spin wave function $\Phi(1, 2, 3)$ must be completely symmetric in 1, 2, 3, and is chosen to be

$$\Phi(1, 2, 3) = \sum_P P_{123} u_\uparrow(1)[u_\uparrow(2)d_\downarrow(3) - d_\uparrow(2)u_\downarrow(3)]. \quad (2.46)$$

The choice is unique: it must be antisymmetric in one of the u-d pairs under $[SU(3)]_{\text{flavor}}$ in order that the proton be a member of **8**. The three spins must combine so as to give spin 1/2. The overall symmetrization then automatically makes the proton a member of **56** with respect to $SU(6)$. This, of course, is a consequence of the Pauli principle, and the assumption that the quarks are in the same spatial orbital. Writing out all the terms in (2.46), we obtain

$$\begin{aligned} \Phi(1, 2, 3) &= 2u_\uparrow u_\uparrow d_\downarrow + 2u_\uparrow d_\downarrow u_\uparrow + 2d_\downarrow u_\uparrow u_\uparrow \\ &\quad - u_\uparrow d_\downarrow u_\downarrow - u_\downarrow u_\downarrow d_\uparrow - u_\downarrow d_\uparrow u_\uparrow \\ &\quad - d_\uparrow u_\uparrow u_\downarrow - u_\downarrow u_\uparrow d_\uparrow - d_\uparrow u_\downarrow u_\uparrow, \end{aligned} \quad (2.47)$$

where it is understood that the coordinates in each term stand in the same order: 1, 2, 3.

As another example, the π^+ wave function in the same model is given by

$$\begin{aligned}\pi^+(1, 2) = & [\bar{r}(1)r(2) + \bar{y}(1)y(2) + \bar{g}(1)g(2)] \\ & \cdot [\bar{d}_\uparrow(1)u_\downarrow(2) - \bar{d}_\downarrow(1)u_\uparrow(2)].\end{aligned}\quad (2.48)$$

2.5 Electromagnetic and Weak Probes

If we assume that hadrons are made of quarks, then the electromagnetic and weak interactions of hadrons should be derived from those of quarks. That is, the basic electromagnetic and weak currents should be quark currents. Thus, electrons and neutrinos might “see” the quarks inside a hadron, through their electromagnetic and weak interactions respectively. In fact, experimental results from electron and neutrino scattering from nucleons are consistent with such an interpretation, and provide indirect evidence for quarks being dynamical objects. However, since we do not really know how to do dynamical calculations involving quarks, the interpretations of these experiments are necessarily based on intuitive models, and are plausible rather than conclusive.

1 Electromagnetic interactions

The quark picture assumes that the electromagnetic interactions of hadrons arise from those of the quarks. Thus the electromagnetic interaction Lagrangian density is written

$$\mathcal{L}_{em} = -eA^\mu(x)[J_\mu(x) + j_\mu(x)], \quad (2.49)$$

where $A^\mu(x)$ is the Maxwell field, e the magnitude of the electronic charge ($e^2/4\pi\hbar c = 1/137$), and $j_\mu(x)$ is the usual Dirac current of electrons and muons. The quark electromagnetic current is given by

$$\begin{aligned}J_\mu(x) = & \bar{q}_{in}(x)Q_{ij}\gamma_\mu q_{jn}(x) \\ = & \frac{2}{3}\bar{u}_n\gamma_\mu u_n - \frac{1}{3}\bar{d}_n\gamma_\mu d_n - \frac{1}{3}\bar{s}_n\gamma_\mu s_n,\end{aligned}\quad (2.50)$$

where the color index n is summed from 1 to 3, and Q_{ij} is the quark charge matrix given in (2.33). Since photons couple directly to the quarks, one might be able to “X-ray” a hadron to see them. Similar techniques employing neutrons as probes have been used to study the momentum distribution of atoms in liquid helium at very low temperatures⁶.

Consider the inclusive electron-proton scattering

$$e + p \rightarrow e + X, \quad (2.51)$$

where X is anything. The matrix element for this process is represented by the

⁶ K. Huang, in *Selected Topics in Physics, Astronomy and Biophysics*, A. DeLaredo and N. Jurisic, eds., (Reidel Publishing Co., Dordrecht, Holland, 1973), pp. 175–213.

Feynman graph in Fig. 2.6. The energy loss of the electron in the laboratory frame is denoted by ν . The 4-momentum transfer carried by a virtual photon is denoted by q_μ . The squared mass is negative: $q^2 < 0$. A Lorentz invariant expression for ν is

$$\nu = P \cdot q / M, \quad (2.52)$$

where P_μ is the proton 4-momentum, and M the proton mass. The interesting kinematic region for our purpose is the so-called deep-inelastic limit:

$$\begin{aligned} &\nu \rightarrow \infty, \\ &-q^2 \rightarrow \infty, \\ &x \equiv -q^2/2M\nu \quad (\text{fixed}), \\ &(0 \leq x \leq 1). \end{aligned} \quad (2.53)$$

The matrix element, as defined in Fig. 2.6, is

$$\mathcal{M} = [\bar{u}(\mathbf{k}')(-ie\gamma^\mu)u(\mathbf{k})] \frac{(-ie)}{q^2} \langle X | J_\mu | P \rangle, \quad (2.54)$$

where $J_\mu \equiv J_\mu(0)$, and the electron and proton states are covariantly normalized to E/m particles per unit volume. The laboratory differential cross section is

$$\frac{d\sigma}{dE' d\Omega} = \frac{4\alpha^2 M^2}{q^4} \frac{E'}{E} I^{\alpha\beta} W_{\alpha\beta}, \quad (2.55)$$

where $\alpha = e^2/4\pi\hbar c$, and $d\Omega$ is the solid angle in which the final electron

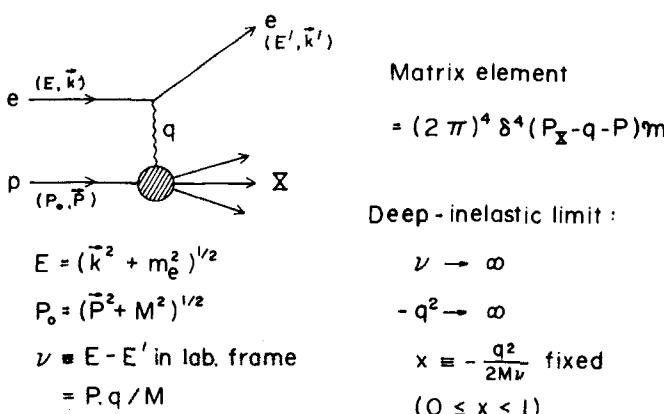


Fig. 2.6 Kinematics of electron-proton deep inelastic scattering

emerges. The tensor $I^{\alpha\beta}$ comes from electron spin averages:

$$\begin{aligned} I^{\alpha\beta} &\equiv \frac{1}{2} \text{Tr} \left[\gamma^\alpha \frac{\not{k}' + m_e}{2m_e} \gamma^\beta \frac{\not{k} + m_e}{2m_e} \right] \\ &= \frac{1}{2m_e^2} [k'^\alpha k^\beta + k^\alpha k'^\beta + g^{\alpha\beta}(m_e^2 - \not{k}' \cdot \not{k})]. \end{aligned} \quad (2.56)$$

The hadronic structure is entirely contained in the tensor

$$W_{\alpha\beta} \equiv (2\pi)^3 \sum_X \langle P | J_\alpha | X \rangle \langle X | J_\beta | P \rangle \delta^4(P_X - P - q). \quad (2.57)$$

Lorentz invariance and gauge invariance imply that $W_{\alpha\beta}$ can be expressed in terms of two Lorentz invariant form factors W_1 and W_2 :

$$W_{\alpha\beta} = -W_1 \left(g_{\alpha\beta} - \frac{q_\alpha q_\beta}{q^2} \right) + \frac{W_2}{M^2} \left(P_\alpha - q_\alpha \frac{q \cdot P}{q^2} \right) \left(P_\beta - q_\beta \frac{q \cdot P}{q^2} \right). \quad (2.58)$$

If $W_{\alpha\beta}$ is given, we can extract W_1 and W_2 through the following formulas:

$$\begin{aligned} W_1 &= \frac{1}{2} \left[C_2 - \left(1 - \frac{\nu^2}{q^2} \right) C_1 \right] \left(1 - \frac{\nu^2}{q^2} \right)^{-1}, \quad C_1 \equiv W^\alpha_\alpha, \\ W_2 &= \frac{1}{2} \left[3C_2 - \left(1 - \frac{\nu^2}{q^2} \right) C_1 \right] \left(1 - \frac{\nu^2}{q^2} \right)^{-2}, \quad C_2 \equiv \frac{P^\alpha P^\beta}{M^2} W_{\alpha\beta}. \end{aligned} \quad (2.59)$$

In the deep-inelastic limit,

$$\begin{aligned} W_1 &\approx -\frac{1}{2} C_1 + \frac{Mx}{\nu} C_2, \\ W_2 &\approx \frac{Mx}{\nu} \left(-C_1 + \frac{6Mx}{\nu} C_2 \right). \end{aligned} \quad (2.60)$$

After a certain amount of algebra, we can express the laboratory differential cross section (2.55) in the form

$$\frac{d\sigma}{dq^2 d\nu} = \frac{4\pi\alpha^2}{q^4} \frac{E'}{E} \left(W_2 \cos^2 \frac{\theta}{2} + 2W_1 \sin^2 \frac{\theta}{2} \right), \quad (2.61)$$

where θ is the laboratory scattering angle, and where the electron mass has been neglected. The main experimental results⁷ show that $W_1 \neq 0$, and that W_2 has the scaling property such that in the deep-inelastic limit, νW_2 is a function of x only. We shall argue that these results indicate that the proton is composed of spin 1/2 objects that are point-like with respect to photons.

2 Parton model

We have no way of calculating $W_{\alpha\beta}$ from first principles, because we have no theory for the states $|P\rangle$ and $|X\rangle$ in terms of the quark fields that appear in the

⁷ J. I. Friedman and H. Kendall, *Ann. Rev. Nucl. Sci.* **22**, 203 (1972).

current J_α . However, Feynman has suggested a simple intuitive model for this, the “parton” model, which we now describe.^{8,9}

Suppose the proton is made up of bound objects that appear point-like to the photon. Then the proton can exist in a transitory virtual state consisting of these free objects. The lifetime of the virtual state is inversely proportional to the difference between the virtual state energy and the proton energy. In a Lorentz frame in which the proton is moving arbitrarily fast, the relativistic time-dilation can make this lifetime arbitrarily long. We call such a frame an ∞ -momentum frame, and think of it as the limit of a sequence of Lorentz frames. In such a frame, a photon falling on the proton would see a collection of free point charges, which Feynman calls “partons”. (Since they are defined in an ∞ -momentum frame, partons do not necessarily have meaning in the proton rest frame). It is now imagined that the photon is absorbed by one of the partons. The absorption process lasts for a time of the order of the inverse photon energy, which we can control by selecting ν in the experiment. In the laboratory, we can in principle make ν as large as we please, so that the characteristic time ν^{-1} is as small as we please. However, this time also dilates when we pass to an ∞ -momentum frame, and it is not obvious whether a parton can live long enough as a free particle to absorb the photon. More detailed elementary considerations, which we shall not go into, show that the parton lifetime is indeed much longer than the absorption time, when the kinematic conditions (2.53) for deep-inelastic scattering is fulfilled, with $x > 0$.

Accepting the foregoing picture, we can calculate $W_{\alpha\beta}$ as follows. The matrix element $\langle X|J_\beta|P\rangle$ describes the absorption of a photon by a parton. We assume that the absorption takes such a relatively short time that both $|P\rangle$ and $|X\rangle$ can be described as free parton states. In the product $\langle P|J_\alpha|X\rangle\langle X|J_\beta|P\rangle$, the photon is absorbed, then re-emitted. We assume that both processes involve the same parton. (If two different partons were involved, their required momentum correlation would mean that one of them had very high momentum before interacting with the photon, and we deem this very unlikely). Accordingly, each parton contributes to $W_{\alpha\beta}$ singly and additively, as indicated schematically in Fig. 2.7.

In an ∞ -momentum frame, all masses can be neglected. Hence we take the partons to be massless Dirac particles, with wave functions covariantly normalized to $2p_0$ particles per unit volume. Then the one parton contribution to $W_{\alpha\beta}$ is

$$\begin{aligned} \tilde{W}_{\alpha\beta} &\equiv (2\pi)^3 \frac{1}{2} \sum_{\text{spins}} \sum_{\mathbf{p}'} \langle \mathbf{p}|J_\alpha|\mathbf{p}'\rangle \langle \mathbf{p}'|J_\beta|\mathbf{p}\rangle \delta^4(p' - p - q) \\ &= Q^2 \int \frac{d^3 p'}{2p'_0} \delta^4(p' - p - q) \frac{1}{2} \text{Tr}(\gamma_\alpha \not{p}' \gamma_\beta \not{p}) \\ &= Q^2 \delta\left(p \cdot q + \frac{1}{2}q^2\right) (2p_\alpha p_\beta + q_\alpha p_\beta + q_\beta p_\alpha - g_{\alpha\beta} q \cdot p), \end{aligned} \quad (2.62)$$

⁸ R. P. Feynman, *Phys. Rev. Lett.* **23**, 1415 (1969).

⁹ J. D. Bjorken and E. A. Paschos, *Phys. Rev.* **158**, 1975 (1969).

where p^α is the parton 4-momentum, and Q its charge in units of e . Now put

$$p^\alpha = y P^\alpha \quad (0 \leq y \leq 1). \quad (2.63)$$

That is, assume that all transverse momenta are negligible, and that no partons move oppositely to the proton. Then, using (2.60), we obtain the one-parton contributions to W_1 and W_2 :

$$\begin{aligned} \nu \tilde{W}_2 &= 2Q^2 Mx^2 \delta(y - x), \\ \tilde{W}_1 &= \nu \tilde{W}_2 / 2Mx. \end{aligned} \quad (2.64)$$

Suppose the proton state contains $f_i(y) dy$ parton states of the type i in the interval dy . Then

$$W_{1,2} = \sum_i \int_0^1 dy f_i(y) \tilde{W}_{1,2}. \quad (2.65)$$

The question is, how is $f_i(y)$ normalized? In our convention, a parton state has $2p_0$ partons per unit volume, while a proton state has P_0/M protons per unit volume. Therefore, in one proton, the number of partons of type i , in the interval dy , is $f_i(y)$ multiplied by $2p_0/(P_0/M) = 2My$:

$$n_i(y) dy = 2Myf_i(y) dy. \quad (2.66)$$

By (2.63), $n_i(y)$ must satisfy the condition

$$\sum_i \int_0^1 dy y n_i(y) = 1. \quad (2.67)$$

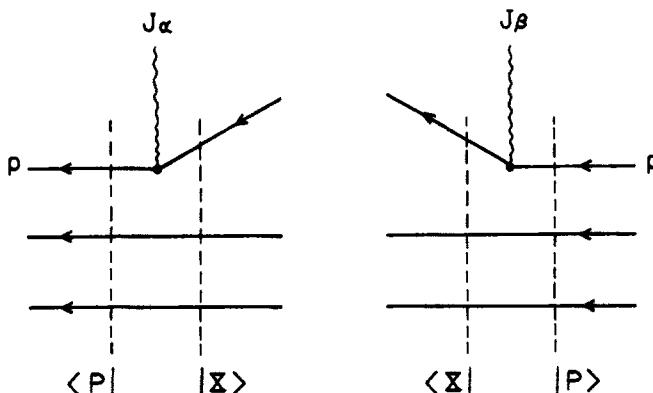


Fig. 2.7 The same parton absorbs and reemits the photon. Other partons are merely spectators.

In terms of this normalized parton distribution, (2.64) becomes

$$\nu W_2 = \sum_i Q_i^2 x n_i(x),$$

$$W_1 = \nu W_2 / 2Mx. \quad (2.68)$$

We see that νW_2 has the desired scaling property. If we had used spin 0 partons, we would have gotten $W_1 = 0$ (though W_2 would be the same). The model of spin 1/2 partons is therefore consistent with experiments, and we identify the partons with quarks or antiquarks. Experiments indicate that $n_i(x) \xrightarrow{x \rightarrow 0} x^{-1}$. Hence, the total number of partons is infinite. Thus, as seen by a high-energy photon, the proton is made of three quarks plus an infinite “sea” of quark-antiquark pairs.

3 Evidence for color

The most direct experimental evidence that quarks have color comes from measurements of the total cross section for the annihilation of electron-positron pairs in colliding beam experiments. The results are expressed in terms of the cross section ratio

$$R = \frac{e^+ e^- \rightarrow \text{Hadrons}}{e^+ e^- \rightarrow \mu^+ \mu^-}, \quad (2.69)$$

and the data up to c.m. energy 35.8 GeV is shown in Fig. 2.8.^{10,11} Note that heavy lepton final states such as $\tau^+ \tau^-$ are included, because they decay into hadrons. The final states $e^+ e^-$ and $\mu^+ \mu^-$ are excluded.

The gross features of the data can be understood as follows. The reactions occur predominantly through annihilation of the initial state into a single virtual photon, which then materializes into the final states, through production of lepton pairs and quark pairs (for these are the only particles coupled directly to the photon). The total cross section should be proportional to the sum of squared charges of all the leptons and quarks that can be energetically pair-produced, barring resonances and final-state interactions. Possible resonances are independently known, and can be subtracted if we wish. As long as the various thresholds for these productions are well-separated, there should be a plateau between thresholds, in which final-state interactions are relatively energy-independent. The sums of squares of charges at various plateaus are given in Table 2.9, for the case with color and that without color.

These plateau values, shown superimposed on the data in Fig. 2.8, indicate that the prediction with color is clearly favored. In the wide range from the $\bar{b}b$ threshold, at about 9 GeV, to the highest energy attained at 35.8 GeV, there are no new thresholds. This places a lower bound of 18 GeV/c² for masses of new leptons or quarks.

If the hadronic final states indeed come from a quark pair, then their angular distribution might retain a memory of the initial directions of the quark pair.

¹⁰ Particle Data Group, *Rev. Mod. Phys.* **52**, 556 (1980).

¹¹ Mark-J Collaboration, *The First Year of Mark-J*, MIT Lab for Nucl. Sci. Report 107 (April 1979).

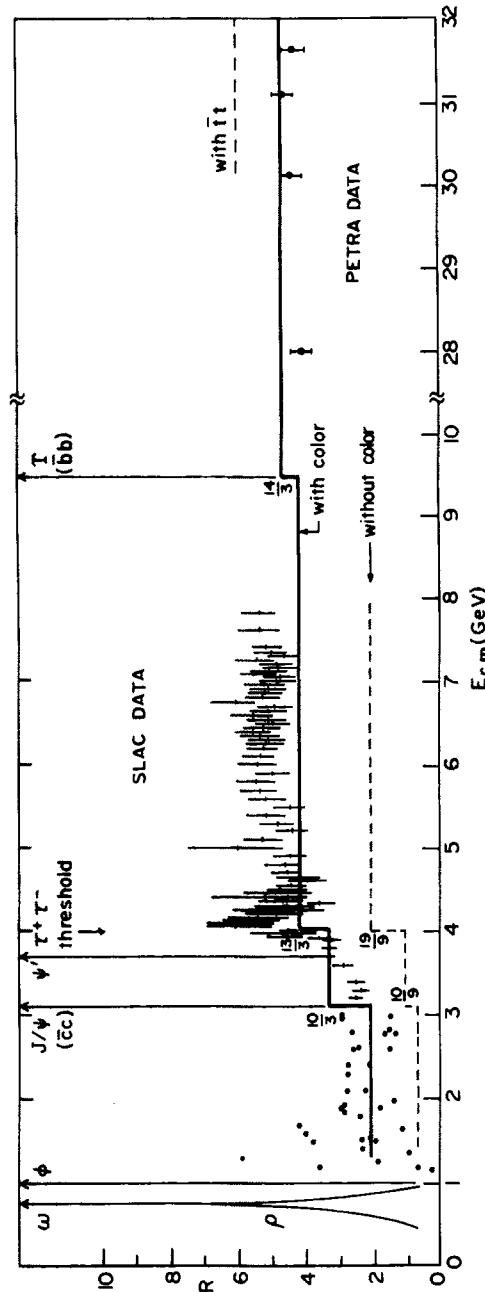


Fig. 2.8 The ratio $R = (e^+ e^- \rightarrow \text{Hadrons}) / (e^+ e^- \rightarrow \mu^+ \mu^-)$. The SLAC and PETRA data have different systematic errors, which are not indicated. Error bars for points below 3 GeV are large, and omitted for clarity. Points above 28 GeV are samplings of data.

This has been found to be the case, in the phenomena of "quark jets"¹². If, as we believe, color symmetry is a gauge symmetry that gives rise to the strong interactions, then the gauge bosons (gluons) can also be emitted by the quarks, resulting in three or more hadronic jets of a distinctive character (gluons have no electromagnetic interactions). This phenomena has also been observed.¹³

4 Weak interactions¹⁴

A concise summary of our present knowledge of the weak interactions is given by the interaction Lagrangian density

$$\mathcal{L}_{\text{wk}} = g W_\alpha(x) [J_{\text{wk}}^\alpha(x) + j_{\text{wk}}^\alpha(x)] + \text{h.c.}, \quad (2.70)$$

where $J_{\text{wk}}^\alpha(x)$ is the quark weak current, and $j_{\text{wk}}^\alpha(x)$ is the lepton weak current. From now on we drop the subscript "wk" for brevity. These weak currents are coupled to a massive charge vector field $W_\alpha(x)$.

The weak currents are given by

$$\begin{aligned} j^\alpha &= \bar{e}\gamma^\alpha(1 - \gamma_5)\nu + \bar{\mu}\gamma^\alpha(1 - \gamma_5)\nu', \\ J^\alpha &= [\bar{d}\gamma^\alpha(1 - \gamma_5)u] \cos \theta + [\bar{s}\gamma^\alpha(1 - \gamma_5)u] \sin \theta, \end{aligned} \quad (2.71)$$

where e and ν stand respectively for the Dirac field operators for the electron

Table 2.9 THE RATIO R

	u	d	s	c	t^+	b	(?)
Q^2	$\frac{4}{9}$	$\frac{1}{9}$	$\frac{1}{9}$	$\frac{4}{9}$		$\frac{1}{9}$	
	$\underbrace{\frac{2}{3}}$		$\underbrace{\frac{10}{9}}$				
WITHOUT COLOR					$\boxed{3x}$		
					$\underbrace{\frac{10}{3}}$		
WITH COLOR						$\frac{13}{3}$	

¹² G. Hansen *et al.*, *Phys. Rev. Lett.* **35**, 1609 (1975).

¹³ O. P. Barber *et al.*, *Phys. Rev. Lett.* **43**, 830 (1979).

¹⁴ For a general reference, see R. E. Marshak, Riazuddin, and C. P. Ryan, *Theory of Weak Interactions in Particle Physics* (Wiley-Interscience, New York, 1969).

and the electron-neutrino; and μ and ν' respectively for those of the negative muon and the muon-neutrino. The angle θ , called the Cabibbo angle, has an experimental value of $\theta \approx 1/4$.

The matrix $1 - \gamma_5$ acting on a Dirac spinor picks out the component with left-handed chirality. For a massless Dirac particle, or a massive Dirac particle of sufficiently high momentum, this means negative helicity (i.e., spin direction is opposite to momentum). We note that for any Dirac spinor ψ ,

$$\bar{\psi} \gamma^\alpha (1 - \gamma_5) \psi = \bar{\psi} (1 + \gamma_5) \gamma^\alpha \psi = [(1 - \gamma_5) \psi]^\dagger \gamma_0 \gamma^\alpha \psi. \quad (2.72)$$

Thus, only left-handed spinors enter into the weak currents. These currents give rise to 4 basic Feynman vertices, as shown in Fig. 2.9. The decay of the μ^- meson corresponds to the Feynman graph of Fig. 2.10.

Since in low-energy processes the W -meson carries small momentum, its propagator may be taken to be m_W^{-2} . We can then obtain the following relation between the coupling constant g and the experimental Fermi coupling

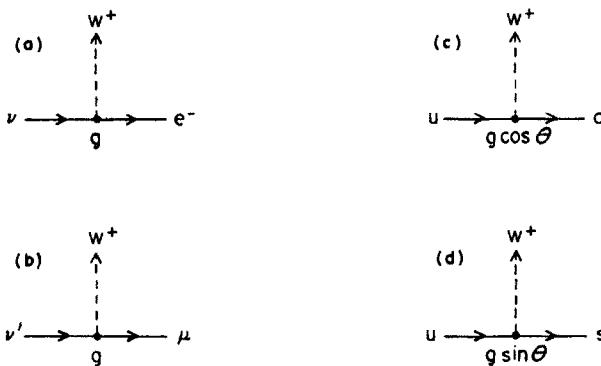


Fig. 2.9 Weak vertices

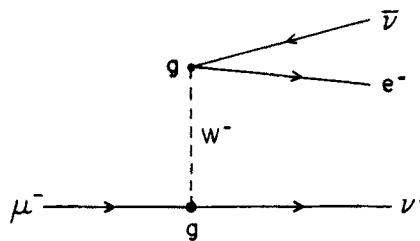


Fig. 2.10 μ^- decay

constant G :

$$\begin{aligned} g^2/m_W^2 &= G/\sqrt{2}, \\ G &= 1.01 \times 10^{-5} m_p^{-2}, \end{aligned} \quad (2.73)$$

where m_p is the proton mass.

Weak decays of hadrons occur through the quark vertices (c) and (d) of Fig. 2.9. Their structure immediately gives, to lowest order in $m_p^{-2}G$ ($\approx 10^{-5}$), the following selection-rules for hadronic decays which were established experimentally:

(a) The change in the hadron's strangeness does not exceed 1 in magnitude, i.e., $\Delta S = 0, \pm 1$.

(b) When there are leptons in the final state, and when $\Delta S = \pm 1$, the change in the hadron's charge is equal to the change in its strangeness, i.e., $\Delta Q = \Delta S = \pm 1$.

(c) When the hadron's strangeness does change, the change in the hadron's isospin is $\pm 1/2$, i.e., $|\Delta I| = 1/2$. (This rule, however, is violated by electromagnetic interactions, and in reality holds only to order $\alpha \sim 1\%$).

Some Feynman graphs for hadronic decays are shown in Fig. 2.11 in which the shaded blobs denote strong interactions.

We recall that the quarks (u, d) form an isodoublet, while s is an isosinglet. The vector part of the quark current can be written in the form

$$V^\alpha \equiv g \cos \theta (\bar{u} \quad d) \gamma^\alpha \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} u \\ d \end{pmatrix} = \sqrt{2} g \cos \theta (\bar{q} \gamma^\alpha I_+ + q), \quad (2.74)$$

where $I_+ = (I_1 + iI_2)/\sqrt{2}$ is the isospin raising operator. This is a succinct statement of the "conserved vector-current hypothesis" (CVC) of Feynman and Gell-Mann (in the form subsequently amended by the introduction of the

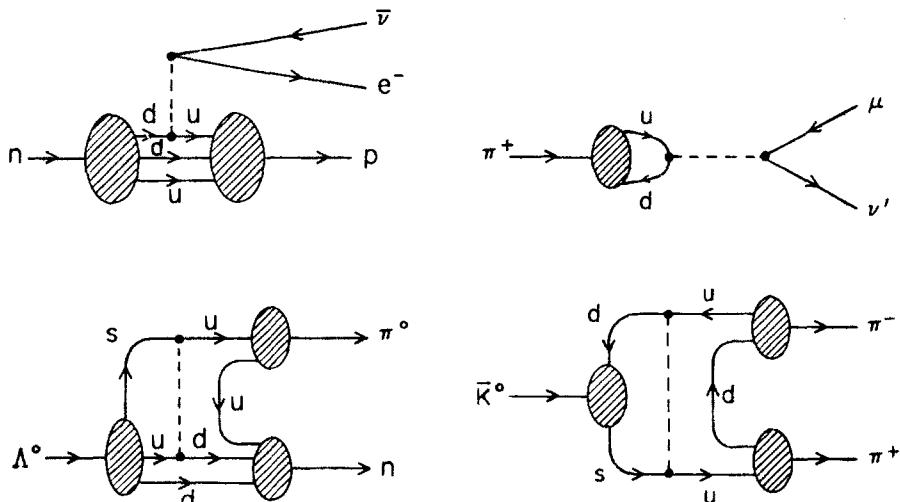


Fig. 2.11 Hadronic decays

Cabibbo angle, as required by experiments). Because strong interactions conserve isospin, the coupling constant g here is unaffected by strong renormalization effects, when V^α is sandwiched between hadronic states. On the other hand, the coupling constant in front of the axial vector part of the current is altered by such effects (experimentally, by a factor 1.25 in the nucleon state).

From the foregoing, we see that the quark hypothesis brings order into the structure of weak interactions. It neatly summarizes the selection rules, and disentangles weak effects from strong.

By analogy with the case of electromagnetic interactions, we can probe hadronic structure through deep-inelastic neutrino scattering from nucleons, and again try out the parton idea. The analysis in this case is more complicated than the electromagnetic case, because we no longer have gauge invariance, and because neutrino and antineutrino have definite and distinct helicities. We shall forego a discussion of this case. Available data are consistent with the predictions of a simple parton picture.¹⁵

2.6 Charm

1 The charmed quark

Glashow, Iliopoulos and Maiani¹⁶ (GIM) proposed that in addition to u , d , s , there should be another quark flavour which they named "charm" (c). The motivation is to rid the old theory of certain undesirable higher order weak effects, i.e., violations of the ΔS and $\Delta S = \Delta Q$ rules, and the occurrence of unobserved decays, such as $\bar{K}^0 \rightarrow \mu^+ + \mu^-$, through effective neutral currents. Their conjecture appears to have been independently and brilliantly confirmed, by the subsequent discovery of new families of hadrons beginning with the J/ψ . With their scheme, there emerges a remarkable one-to-one correspondence between quarks and leptons.

Experimentally, the decay mode $\bar{K}^0 \rightarrow \mu^+ + \mu^-$ is not observed; its branching ratio being less than 10^{-6} . In the three-flavoured quark theory, this decay mode occurs at an order of g^4 , through the Feynman graph of Fig. 2.12. The matrix element is divergent.

Although the theory can be made renormalizable by elaborating it with the addition of other fields (e.g., the Weinberg-Salam model), for qualitative purposes we keep the present model, but introduce a cutoff Λ . Taking the W -propagator to be

$$\Delta^{\alpha\beta}(k) = \frac{1}{k^2 - m_W^2} \left(g^{\alpha\beta} + \frac{k^\alpha k^\beta}{m_W^2} \right), \quad (2.75)$$

and assuming the integration over momentum k' in Fig. 2.12 converges by virtue of strong interactions, we see through simple power-counting that the degree of divergence is $g^4 \Lambda^2 \sim G^2 \Lambda^2$. To keep the branching ratio within the experimental

¹⁵ G. B. West, *Phys. Reports* **18C**, 263 (1975).

¹⁶ S. L. Glashow, J. Iliopoulos, and L. Maiani, *Phys. Rev. D* **2**, 1285 (1970).

upper bound, a cutoff $\Lambda \leq 3$ GeV is necessary, and this value seems unreasonably small. In any event, it seems more satisfactory, from a theoretical point of view, to suppress the decay through a dynamical mechanism, rather than through a cutoff, which merely relegates the explanation to things left out in the model.

The GIM proposal is to introduce a new quark c , with weak couplings chosen so as to cancel the Feynman graph of Fig. 2.12, in the limit of completely degenerate quark masses. The required couplings are indicated in Fig. 2.13.

In the real world, where unitary symmetry is inexact, the s quark should have a different mass from that of the u and d quarks. Presumably, the c quark will have a still different mass, and the $c-u$ mass difference will render the cancellation in Fig. 2.13 incomplete. To calculate the true rate of $\bar{K}^0 \rightarrow \mu^+ + \mu^-$, different ingredients would be involved. To give quark masses real meaning, one would have to have a model of quark binding. To calculate higher order weak processes in a meaningful way, one has to work with a renormalizable theory of weak interactions. As a rough guess, however, one might expect that the $c-u$ mass difference, defined in some effective way, takes the place of the cutoff Λ . The experimental bound cited earlier should then put this mass difference at ≤ 3 GeV/c².

The three flavors s, d, u were introduced to realize $[SU(3)]_{\text{flavor}}$. To add one more flavor means enlarging the group to $[SU(4)]_{\text{flavor}}$, which will be even more badly broken than $SU(3)$. Just as in going from $SU(2)$ to $SU(3)$ we added a new

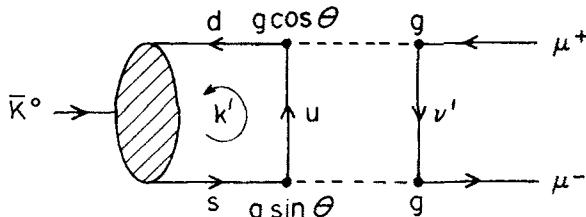


Fig. 2.12 Lowest-order matrix element for $\bar{K}^0 \rightarrow \mu^+ + \mu^-$, an unobserved decay mode

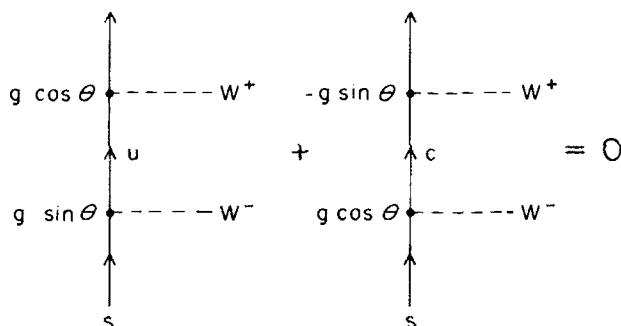


Fig. 2.13 Introducing the charmed quark c

hadronic quantum number hypercharge, we now add another hadronic quantum number charm (C). Like Y , it is assumed to be an additive quantum number conserved by the strong interactions. The charmed quark, by definition, is assigned $C = 1$, while u, d, s are assigned $C = 0$. The charmed quark is also supposed to come in the *same* three colors as the other quarks.

Other quantum numbers of the charmed quark c are assigned as follows, motivated by the structure of the vertices involving c in Fig. 2.13:

Spin:	$s = 1/2$	To conserve angular momentum.
Isospin:	$I = 0$	By definition.
Charge:	$Q = 2/3$	To conserve charge.
Baryon number:	$B = 1/3$	To conserve baryon number.
Strangeness:	$S = 0$	By definition.
Hypercharge:	$Y \equiv B + S + C = 4/3$.	

These assignments maintain the relation

$$Q = I_3 + \frac{1}{2} Y. \quad (2.76)$$

2 The J/ψ and its family

In 1974, Ting and Richter, with their respective collaborating teams, independently discovered the J/ψ .¹⁷ This particle is remarkable in that its mass ($3.1 \text{ GeV}/c^2$) is more than three times that of the proton, and yet its lifetime is a thousand times longer than the hadrons we had been familiar with (full width = 67 KeV). In Ting's words, "It's like stumbling upon a village inhabited by people who live to be ten thousand."¹⁸ Since then, other particles of the same family have been found: $\psi(3684)$, $\psi(3950)$, $\psi(4150)$, $\psi(4400)$. All are spin 1 mesons with $I = 0$, negative parity and G -parity.

It is now established that these particles are states of "charmonium", i.e., bound states of $\bar{c}c$. The J/ψ is the charmonium ground state. One can account for the J/ψ family by adopting a non-relativistic independent quark model, with a linear potential between c and \bar{c} , and taking the mass of c to be roughly half the J mass, i.e., $1.5 \text{ GeV}/c^2$ (thus making the non-relativistic assumption self-consistent). The higher ψ states are supposed to be radial excitations. According to such a model, there should also be orbital excitations, and these have been identified experimentally. Charmed mesons and baryons, i.e., bound states like (cd) and (cud) have also been found. The unusual narrowness of J/ψ has not been fully understood. It is thought to have a dynamical origin.

A new family of vector mesons, the Y family, starting at a mass of $9 \text{ GeV}/c^2$, has since been discovered, giving evidence of a new quark flavor b .¹⁹

There is thus a rich new heavy-quark spectroscopy, which forms a separate topic we cannot go into here²⁰. For our purpose, the clarification of this

¹⁷ Nobel lectures by Ting and Richter: S. C. C. Ting, *Rev. Mod. Phys.* **49**, 236 (1977); B. Richter, *Ibid.*, **49**, 251 (1977).

¹⁸ S. C. C. Ting, (private communications).

¹⁹ S. W. Herb *et al.*, *Phys. Rev. Lett.* **39**, 252 (1977).

²⁰ T. Appelquist, R. M. Barnett and K. D. Lane, *Ann Rev. Nucl. Part. Sci.* **28**, 387 (1978).

spectroscopy through the quark model constitutes strong evidence for the reality of quarks.

3 Correspondence between quarks and leptons

With the inclusion of the charmed quark, the quark weak current in (2.71) is amended to read

$$\begin{aligned} J^\alpha = & (\bar{d} \cos \theta + \bar{s} \sin \theta) \gamma^\alpha (1 - \gamma_5) u \\ & + (\bar{s} \cos \theta - \bar{d} \sin \theta) \gamma^\alpha (1 - \gamma_5) c. \end{aligned} \quad (2.77)$$

Apparently, the quark pair (d, s) , which was defined with respect to quantum numbers conserved by the strong interactions, partakes in the weak interactions through a linear recombination (d_θ, s_θ) , which is just a rotation through the Cabibbo angle:

$$\begin{pmatrix} d_\theta \\ s_\theta \end{pmatrix} = \begin{pmatrix} \cos \theta & \sin \theta \\ -\sin \theta & \cos \theta \end{pmatrix} \begin{pmatrix} d \\ s \end{pmatrix}. \quad (2.78)$$

In terms of these, the quark current and the lepton current have strikingly similar forms:

$$\begin{aligned} J^\alpha &= \bar{d}_\theta \gamma^\alpha (1 - \gamma_5) u + \bar{s}_\theta \gamma^\alpha (1 - \gamma_5) c, \\ j^\alpha &= \bar{e} \gamma^\alpha (1 - \gamma_5) \nu + \bar{\mu} \gamma^\alpha (1 - \gamma_5) \nu'. \end{aligned} \quad (2.79)$$

With the discovery of τ and b , the similarity persists, provided that one postulates a τ -neutrino, ν'' , and another quark flavor t . (These will be discussed in Chapter 6). We have no deeper understanding of this symmetry at present. It cannot help but suggest that the leptonic and quark flavors are manifestations of a deeper unity, and that these particles that seem elementary to us at this stage might yet be reducible to something simpler.

CHAPTER 3

MAXWELL FIELD: **$U(1)$** GAUGE THEORY

3.1 Global and Local Gauge Invariance

The Maxwell field, or electromagnetic field, is coupled solely to conserved currents of matter fields. This property may be derived from a “gauge principle”, namely, the electromagnetic coupling arises from an extension of “global gauge invariance” to “local gauge invariance”.

To illustrate the gauge principle, we consider a complex scalar field $\phi(x)$ whose classical Lagrangian density^a in the absence of electromagnetic coupling has the form

$$\mathcal{L}_0(\phi(x), \partial^\mu\phi(x)) = \partial_\mu\phi^*\partial^\mu\phi - V(\phi^*\phi), \quad (3.1)$$

which is obviously invariant under a constant phase change of $\phi(x)$:

$$\begin{aligned} \phi(x) &\rightarrow U\phi(x), \\ U &= e^{-i\alpha}, \end{aligned} \quad (3.2)$$

where α is an arbitrary real constant. This transformation is called a “global gauge transformation”, and the theory is said to have global gauge invariance under the group $U(1)$. By Noether’s theorem¹, there is a conserved current:

$$\begin{aligned} j^\mu &= \text{const. } \phi^*\tilde{\partial}^\mu\phi, \\ \partial_\mu j^\mu &= 0. \end{aligned} \quad (3.3)$$

Now we consider local gauge transformations:

$$\begin{aligned} \phi(x) &\rightarrow U(x)\phi(x), \\ U(x) &= e^{-i\alpha(x)}, \end{aligned} \quad (3.4)$$

where $\alpha(x)$ is an arbitrary real function. That is, let the gauge transformations at different points of space-time be independent of one another. Note that $\partial^\mu\phi(x)$ transforms in the same manner as $\phi(x)$ under a global gauge transformation, but acquires an extra term under a local gauge transformation:

$$\partial^\mu\phi(x) \rightarrow U(x)\partial^\mu\phi(x) + \phi(x)\partial^\mu U(x). \quad (3.5)$$

Therefore, $\mathcal{L}_0(\phi, \partial^\mu\phi)$ is not invariant under a local gauge transformation.

^a The theory has to be quantized. For expedience we work with the classical theory, and simply read off some simple properties in the quantized version whenever possible.

¹ See R. Jackiw in S. B. Treiman, R. Jackiw, B. Zumino, and E. Witten, *Current Algebra and Anomalies* (World Scientific, Singapore, 1985).

To make the theory locally gauge invariant, all we have to do is to replace $\partial^\mu\phi(x)$ by a suitable generalization that transforms in the same manner as $\phi(x)$. We define a vector field $A^\mu(x)$, called a “gauge field”, which transforms under the local gauge transformation (3.4) according to the rule

$$A^\mu(x) \rightarrow A^\mu(x) + \frac{1}{e} \partial^\mu \alpha(x), \quad (3.6)$$

where e is a given real number that fixes the scale of $A^\mu(x)$ relative to $\phi(x)$. Then, the “covariant derivative” defined by

$$D^\mu\phi(x) = [\partial^\mu + ieA^\mu(x)]\phi(x) \quad (3.7)$$

will transform in the same manner as $\phi(x)$:

$$\begin{aligned} D^\mu\phi(x) &\rightarrow U(x)D^\mu\phi(x), \\ [D^\mu\phi(x)]^* &\rightarrow U^*(x)[D^\mu\phi(x)]^*. \end{aligned} \quad (3.8)$$

By replacing $\partial^\mu\phi$ by $D^\mu\phi$, we obtain a new Lagrangian $\mathcal{L}(\phi, D^\mu\phi)$, which is obviously invariant under local gauge transformations. However, it contains the gauge field $A^\mu(x)$ as an external field. To define a closed dynamical system in the canonical sense, it is necessary to add a term involving $\partial^\nu A^\mu$ quadratically. The only gauge Lorentz scalar of this type is proportional to $F^{\mu\nu}F_{\mu\nu}$, where

$$F^{\mu\nu}(x) = \partial^\mu A^\nu(x) - \partial^\nu A^\mu(x) \quad (3.9)$$

is called the field strength tensor. Thus we arrive at a Lagrangian density for a closed dynamical system invariant under local gauge transformations:

$$\mathcal{L} = -\frac{1}{4}F^{\mu\nu}F_{\mu\nu} + \mathcal{L}_0(\phi, D^\mu\phi). \quad (3.10)$$

The factor $-1/4$ in the first term is purely conventional. The global $U(1)$ symmetry, which is now enlarged into a local symmetry, is said to have been “gauged”. Within the framework of canonical field theory, the procedure is unique. The classical equations of motion following from (3.10) are just Maxwell’s equations for scalar electrodynamics:

$$\begin{aligned} \partial_\mu F^{\mu\nu} &= j^\nu, \\ D_\mu D^\mu\phi &= -\partial V/\partial\phi^*, \\ (D_\mu D^\mu\phi)^* &= -\partial V/\partial\phi, \end{aligned} \quad (3.11)$$

where

$$\begin{aligned} j^\mu &= ie[\phi^*D^\mu\phi - (D^\mu\phi)^*\phi] = ie\phi^*\vec{\partial}^\mu\phi - 2e^2\phi^*\phi A^\mu, \\ \partial_\nu j^\nu &= 0, \end{aligned} \quad (3.12)$$

where the last statement is required by the equations of motion and the antisymmetry of $F^{\mu\nu}$. Thus we have “derived” electromagnetism.

Although the symmetry has been enlarged, no additional conserved currents arise. An application of Noether’s theorem will yield the obvious statement that the most general conserved current is $j^\mu(x)$ plus a term of the form $\partial_\nu[F^{\mu\nu}(x)f(x)]$, where $f(x)$ is an arbitrary function. The gauge principle merely determines the form of coupling between the matter field and the gauge field.

Summarizing the local gauge transformation (3.4) and (3.6) in the form

$$\begin{aligned}\phi(x) &\rightarrow e^{-ie\omega(x)}\phi(x), \\ A^\mu(x) &\rightarrow A^\mu(x) + \partial^\mu\omega(x),\end{aligned}\quad (3.13)$$

where $\omega(x)$ is an arbitrary real function, we see that the charge e (up to a constant factor) acts as the generator of the gauge group $U(1)$. In the original system with only global gauge invariance, the charge of a particle must be identified as e everywhere in space-time. However, in the enlarged system, the charge may be taken to be $\pm e$ independently at every space-time point, because the sign of $\omega(x)$ is free to change. Whatever the convention we choose, the physical content of the theory will be the same, for the gauge field maintains the correct bookkeeping.

In the most general case, we may begin with a set of matter fields $\{\phi_1(x), \dots, \phi_n(x)\}$ (which may be boson or fermion fields), that furnishes a representation (generally reducible) of the group $U(1)$. The generator of $U(1)$ is now represented by an $n \times n$ diagonal matrix, whose eigenvalues Q_1, \dots, Q_n are the respective charges of the matter fields. A global gauge transformation is an element of $U(1)$ represented by the transformation

$$\begin{pmatrix} \phi_1 \\ \phi_2 \\ \vdots \\ \phi_n \end{pmatrix} \rightarrow \begin{pmatrix} e^{-iQ_1\omega} & 0 & 0 & \dots \\ 0 & e^{-iQ_2\omega} & 0 & \dots \\ \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & e^{-iQ_n\omega} \end{pmatrix} \begin{pmatrix} \phi_1 \\ \phi_2 \\ \vdots \\ \phi_n \end{pmatrix}, \quad (3.14)$$

where ω is an arbitrary real constant. More compactly, we rewrite (3.14) in matrix form as

$$\begin{aligned}\phi(x) &\rightarrow U\phi(x), \\ U &= e^{-iQ\omega}.\end{aligned}\quad (3.15)$$

The Lagrangian density $\mathcal{L}_0(\phi, \partial^\mu\phi)$ is assumed to be invariant under (3.15). To enlarge the symmetry to a local one, we replace $\partial^\mu\phi$ by

$$D^\mu\phi(x) = [\partial^\mu + iQA^\mu(x)]\phi(x). \quad (3.16)$$

The Lagrangian density of the enlarged system has the same form as (3.10).

As a representation of $U(1)$, (3.15) places no constraint on the values of the charges Q_1, \dots, Q_n . It leaves unexplained the experimental fact that all observed charges in nature are multiples of a common unit (charge quantization). A possible scenario is that the electromagnetic $U(1)$ group is a subgroup of a *compact* symmetry group G . Then the matrix U in (3.15) will have to be a periodic function of ω , and charge will be quantized².

The matter Lagrangian density \mathcal{L}_0 may have other $U(1)$ symmetries that are not gauged, for example baryon number and lepton number. These will remain as global symmetries of \mathcal{L} , as long as they commute with the electromagnetic $U(1)$.

² C. N. Yang, *Phys. Rev. D* **1**, 2360 (1970).

It may seem that the theory predicts that the photon is massless, by the following argument. The equation of motion for the gauge field reads $\square^2 A^\mu = j^\mu$ in Lorentz gauge ($\partial_\mu A^\mu = 0$). Suppose $\phi = 0$ in the lowest state^b of the system, as gauge invariance would naively require (for ϕ is not gauge invariant, and naively one might expect the lowest state to be unique). Then the low-lying modes of the gauge field satisfy $\square^2 A^\mu = 0$, which leads to a plane-wave solution of wave vector k^μ , with $k^2 = 0$. These modes correspond to massless photons in quantum theory. Now, we can take j^μ into account by perturbation expansions in powers of e , resulting in a scheme that describes the interactions in terms of emissions and absorptions of massless photons. Quantum electrodynamics is such a scheme, and its predictions have been tested experimentally to extremely high precision. Therefore, it might seem that the masslessness of the photon is a consequence of gauge invariance.

However, the argument above breaks down if perturbation theory is invalid. This is obviously the case when $\phi \neq 0$ in the lowest state^c, a situation known as “spontaneous symmetry breaking”. In that case the photon does develop a mass, as we shall discuss in more detail in Sec. 3.3.

Even if $\phi = 0$ in the lowest state, the validity of perturbation theory cannot be proven mathematically; it remains a logical possibility that the photon may have mass³. In fact, the photon does acquire mass in the Schwinger model⁴—quantum electrodynamics in one spatial dimension. It can almost be proved⁵ that in quantum electrodynamics the photon is massless if the electron charge is smaller than a critical value, but acquires mass otherwise. The critical value is unknown, although one would like to believe that it is greater than the physical electronic charge. Thus, the masslessness of the photon in conventional quantum electrodynamics is not due to any known principle in the theory, but due to the assumption that conventional perturbation theory is valid.

3.2 Spontaneous Breaking of Global Gauge Invariance: Goldstone Mode

A symmetry of a system is said to be “spontaneously broken” if the lowest state of the system is not invariant under the operations of that symmetry. Far from being an esoteric situation, this is a common occurrence in the macroscopic world. For example, rotational symmetry is spontaneously broken in a ferromagnet, and translational symmetry is spontaneously broken in an infinite crystalline solid.

Here, we discuss the spontaneous breaking of global gauge invariance in a relativistic field theory. As a simple example, we turn again to the complex

^b Here we use the term “state” loosely, to denote either a classical solution or a state of the quantized theory.

^c In the quantized theory, the statement becomes $\langle \phi \rangle \neq 0$, where $\langle \phi \rangle$ is the vacuum expectation value of ϕ .

³ J. Schwinger, *Phys. Rev.* **125**, 397 (1962).

⁴ J. Schwinger, *Phys. Rev.* **128**, 2425 (1962).

⁵ K. Wilson, *Phys. Rev.* **D10**, 2445 (1974); A. Guth, *Phys. Rev.* **D21**, 2291 (1980).

scalar field defined by (3.1), with no electromagnetic coupling. The classical Hamiltonian of the system is

$$H = \int d^3x [\pi^* \pi + \nabla \phi^* \cdot \nabla \phi + V(\phi^* \phi)], \quad (3.17)$$

where $\pi = \partial \phi / \partial t$. A solution to the equations of motion with the lowest possible energy corresponds to a constant $\phi(x) = \phi_0$, such that $V(\phi_0^* \phi_0)$ is at its smallest possible value. This clearly minimizes H , and is a solution of the field equations, because for time-independent ϕ , H is proportional to the action. If $\phi_0 \neq 0$, the solution is clearly not invariant under a change of phase, hence global gauge invariance is spontaneously broken. The lowest state is then infinitely degenerate, corresponding to the fact that the phase of ϕ_0 is arbitrary.

To be definite let us choose

$$V(\phi^* \phi) = \mu^2 \phi^* \phi + \lambda (\phi^* \phi)^2 + \text{const.} \quad (3.18)$$

The reason for choosing a quartic polynomial form is that the corresponding quantum field theory is renormalizable, in a perturbation scheme based on expansions in powers of λ . The equation of motion now reads

$$(\square^2 + \mu^2) \phi = -2\lambda \phi^* \phi^2. \quad (3.19)$$

The free field case corresponds to $\lambda = 0$, $\mu^2 > 0$. In this case the classical modes above the lowest state are plane waves of wave vectors k^μ , with $k^2 = \mu^2$, and correspond to single-particle states of mass μ in the quantum theory.

If $\lambda < 0$, the theory does not exist because the Hamiltonian has no lower bound. If $\lambda < 0$ and $\mu^2 > 0$, then V has the form shown in Fig. 3.1(a). It may appear that the system can have a metastable state, with the field contained (in field space) within the local minimum of V . This would be true in particle quantum mechanics, but not in quantum field theory. In the latter case the decay rate of such an initial state in infinite space is infinite because it has finite value per unit volume.

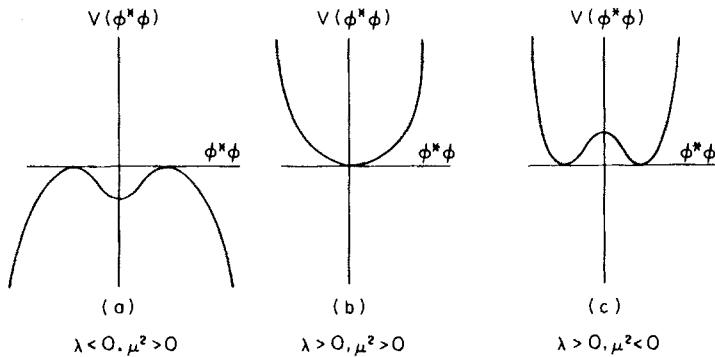


Fig. 3.1 The potential term for various choices of parameters

If $\lambda > 0$, we can have either $\mu^2 > 0$ or $\mu^2 < 0$, with the corresponding forms of V shown in Figs. 3.1(b) and 3.1(c). In the former case, the lowest solution is $\phi = 0$, and there is no spontaneous symmetry breaking. In the latter case, $\phi \neq 0$, which is what we want to study.

For convenience we rewrite (3.18) in the form

$$V(\phi^*\phi) = \lambda(\phi^*\phi - \phi_0^2)^2, \quad (\phi_0 \neq 0). \quad (3.20)$$

The lowest state corresponds to $\phi^*\phi = \phi_0^2$, or

$$\phi(x) = \phi_0 e^{i\alpha_0}, \quad (3.21)$$

where α_0 is an arbitrary real constant.

The low-lying states of the quantum theory can be deduced from the low-lying classical modes by inspection. Let us put

$$\phi(x) = [\phi_0 + \eta(x)]e^{i\alpha(x)}, \quad (3.22)$$

so that the complex fields $\phi^*(x)$, $\phi(x)$ are now replaced by the real fields $\eta(x)$ and $\alpha(x)$, in terms of which the Lagrangian density reads

$$\mathcal{L}_0 = \partial^\mu \eta \partial_\mu \eta - \lambda(2\phi_0 + \eta)^2 \eta^2 + (\phi_0 + \eta)^2 \partial^\mu \alpha \partial_\mu \alpha. \quad (3.23)$$

Assuming η to be small, and dropping terms higher than second order ones, we obtain

$$\mathcal{L}_0 \approx [\partial^\mu \eta \partial_\mu \eta - 4\lambda\phi_0^2 \eta^2] + \phi_0^2 \partial^\mu \alpha \partial_\mu \alpha + O(\eta^3). \quad (3.24)$$

The terms in the brackets describe free scalar particles of mass $2\phi_0\sqrt{\lambda}$. The next term describes free massless scalar particles. The neglected terms describe the interactions among the particles. These classical modes are illustrated in Fig. 3.2, where V is shown as a function of $\text{Re } \phi$ and $\text{Im } \phi$. We can see the direct connection between $\phi_0 \neq 0$ and the existence of a massless mode.

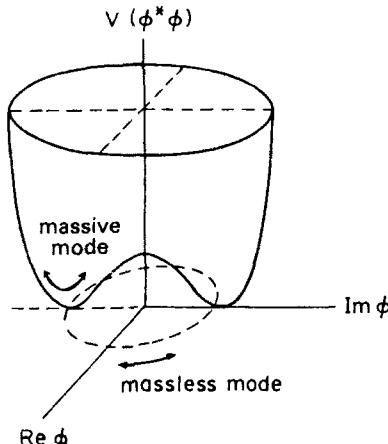


Fig. 3.2 Classical modes in the scalar field with symmetry-breaking potential

In the quantum theory, each value of α_0 in (3.21) gives a possible vacuum state. The transition amplitude between vacua with different values of α_0 vanishes for infinite spatial volume.

The statement that “spontaneous breaking of a continuous global symmetry implies the existence of a massless spin zero particle” is known as *Goldstone’s theorem*⁶, and the massless particle is called a *Goldstone boson*. The symmetry that is spontaneously broken is, of course, still a symmetry of the system. However, it is manifested not through the invariance of the lowest state, but in the “Goldstone mode”—through the existence of a Goldstone boson.

There are physical examples of Goldstone bosons in non-relativistic many-body systems: spin waves in a ferromagnet, phonons in a crystalline solid and in liquid helium.

3.3 Spontaneous Breaking of Local Gauge Invariance: Higgs Mode

When a *local* gauge symmetry is spontaneously broken, the symmetry is again manifested in a manner other than the invariance of the lowest state. In this case, however, no Goldstone boson occurs. Instead, the gauge field acquires mass. We say that the symmetry is manifested in the “Higgs mode”.

Let us consider scalar electrodynamics, and choose V to have the form (3.20):

$$\begin{aligned} \mathcal{L}(x) &= -\frac{1}{4}F^{\mu\nu}F_{\mu\nu} + (D^\mu\phi)^*(D_\mu\phi) - V(\phi^*\phi), \\ D^\mu\phi &= (\partial^\mu + ieA^\mu)\phi, \\ V(\phi^*\phi) &= \lambda(\phi^*\phi - \phi_0^2)^2, \quad (\phi_0 \neq 0). \end{aligned} \quad (3.25)$$

The Lagrangian is invariant under a local gauge transformation

$$\begin{aligned} A^\mu(x) &\rightarrow A^\mu(x) + \partial^\mu\omega(x), \\ \phi(x) &\rightarrow e^{-ie\omega(x)}\phi(x), \\ \phi^*(x) &\rightarrow e^{ie\omega(x)}\phi^*(x), \end{aligned} \quad (3.26)$$

where $\omega(x)$ is an arbitrary real function. The local gauge symmetry is spontaneously broken because $\phi_0 \neq 0$. The field $\phi(x)$, which causes the breakdown, is referred to as a “Higgs field”.

In canonical formalism the variables are as follows:

Field	Canonical conjugate
A^μ	$-F_{0\mu}$
ϕ	$\pi = (D^0\phi)^*$
ϕ^*	$\pi^* = D^0\phi$

Since the canonical conjugate to A^0 is identically zero, A^0 is not an independent variable but can be eliminated in terms of the others through the equations of

⁶ J. Goldstone, *N. Cim.* **19**, 154 (1961); J. Goldstone, A. Salam, and S. Weinberg, *Phys. Rev.* **127**, 965 (1962).

motion. Introduce the electric field \mathbf{E} and the magnetic field \mathbf{B} by

$$\begin{aligned} E^k &= F^{k0}, \\ B^k &= -\frac{1}{2}\epsilon^{ijk}F^{ij}, \end{aligned} \quad (3.28)$$

or

$$\begin{aligned} \mathbf{E} &= -\partial \mathbf{A}/\partial t - \nabla A^0, \\ \mathbf{B} &= \nabla \times \mathbf{A}. \end{aligned} \quad (3.29)$$

The Hamiltonian can be brought to the form

$$H = \int d^3x [\frac{1}{2}(\mathbf{B} \cdot \mathbf{B} + \mathbf{E} \cdot \mathbf{E}) + |\pi|^2 + |\mathbf{D}\phi|^2 + V]. \quad (3.30)$$

In deriving (3.30), the equations of motion have been used, and a surface integral $\int d^3x \nabla \cdot (\mathbf{E} A^0)$ has been dropped^d. The form (3.30) is manifestly gauge invariant, but it still contains A^0 . The final form, after A^0 has been eliminated, will depend on the particular gauge one chooses. The numerical value of H is of course gauge invariant.

It is clear from (3.30) that a lowest-energy solution is

$$\begin{aligned} A^\mu(x) &= 0, \\ \phi(x) &= \phi_0 e^{i\alpha_0}. \end{aligned} \quad (3.31)$$

To study the classical modes near this solution, it is convenient to go to “unitary gauge”, in which $\phi(x)$ is real, by transforming away its phase through a continuous local gauge transformation. This can always be done because ϕ satisfies a second order differential equation, and hence its phase must have continuous derivatives. Thus, we can always write

$$\phi(x) = \rho(x) \quad (\text{real}). \quad (3.32)$$

The equations of motion become

$$\begin{aligned} \partial_\mu F^{\mu\nu} &= -2e^2 \rho^2 A^\nu, \\ (\partial^\mu + ieA^\mu)(\partial_\mu + ieA_\mu)\rho &= 2\lambda\rho(\phi_0^2 - \rho^2). \end{aligned} \quad (3.33)$$

Since $\partial_\nu \partial_\mu F^{\mu\nu} \equiv 0$, we must have

$$\partial_\mu A^\mu(x) = 0 \quad \text{wherever } \rho(x) \neq 0. \quad (3.34)$$

Now we put

$$\rho(x) = \phi_0 + \eta(x), \quad (3.35)$$

and treat $\eta(x)$ and $A^\mu(x)$ as small quantities. The linearized equations of motion read

$$\begin{aligned} (\square^2 + 2e^2\phi_0^2)A^\mu &= 0 \quad (\partial_\mu A^\mu = 0), \\ (\square^2 + 4\lambda\phi_0^2)\eta &= 0. \end{aligned} \quad (3.36)$$

^d See the derivation of the more general expression (4.92) in chap. 4.

In the quantum theory, these lead to a spin 1 particle of mass $\sqrt{2}e\phi_0$, and a spin 0 particle of mass $2\sqrt{\lambda}\phi_0$. There is no massless particle. The original fields (in a fixed gauge) A^1, A^2, ϕ, ϕ^* are now replaced by A^1, A^2, A^3, η . (The subsidiary condition $\partial_\mu A^\mu = 0$ takes out the spin zero part of A). One could say that gauging the symmetry converts the “would-be Goldstone boson” into the longitudinal part of the resulting massive gauge field. If we had used some gauge other than the unitary gauge, the nature of the spectrum would be the same, but perhaps less obvious. The manner through which the photon mass comes about is called the “Higgs Mechanism”.⁷

A physical example in which the Higgs mechanism actually takes place is superconductivity. The Lagrangian of the theory is invariant under local phase changes of the electron field, but the ground state is not, owing to a condensation of Cooper pairs made up of two electrons. As a consequence, the photon becomes massive inside a superconducting body. In particular, an externally applied magnetic field can penetrate the body only to a finite depth equal to the inverse mass (Meissner effect). A phenomenological way to describe the condensation phenomenon is to introduce an “order parameter” $\phi(x)$ to describe the condensate, as done in the Landau-Ginzburg theory⁸. In such an approach, (3.25) serves as a phenomenology Lagrangian.

Spontaneous symmetry breaking without a fundamental Higgs field, such as in the case of superconductivity, is sometimes called “dynamical symmetry breaking”. This terminology carries the implication that Higgs fields are “normally” needed for symmetry breaking. This is a prejudice based not on physical fact, but solely on mathematical simplicity.

3.4 Classical Finite-Energy Solutions

We can see from (3.30) that a finite-energy solution must satisfy the requirements

$$\begin{aligned} \int d^3x (\mathbf{B} \cdot \mathbf{B} + \mathbf{E} \cdot \mathbf{E}) &< \infty, \\ \int d^3x V(\phi^* \phi) &< \infty. \end{aligned} \quad (3.37)$$

The first of these implies that asymptotically $F^{\mu\nu} \rightarrow O(x^{-2})$, and hence $A^\mu \rightarrow \partial^\mu \omega + O(x^{-1})$. That is,

$$A^\mu(x) \xrightarrow{x \rightarrow \infty} (\text{pure gauge form}) + O(x^{-1}). \quad (3.38)$$

The second condition in (3.37) implies that

$$\phi(x) \xrightarrow{x \rightarrow \infty} \phi_0 e^{i\alpha(x)} + O(x^{-4}). \quad (3.39)$$

We now impose these conditions on the equations of motion (3.11), which may be rewritten in the form

$$\partial_\mu F^{\mu\nu} = -2e \operatorname{Im}(\phi^* D^\nu \phi). \quad (3.40)$$

⁷ P. W. Higgs, *Phys. Rev. Lett.* **12**, 132, (1964); F. Englert and R. Brout, *ibid.* **13**, 321 (1964), G. S. Guralnik, C. R. Hagen, and T. W. B. Kibble, *ibid.* **13**, 585 (1964).

⁸ V. Ginzburg and L. Landau, *Zh. Theor. Fiz.* **47**, 2222 (1957). [*Sov. Phys. JETP* **5**, 1174 (1957)].

We know by (3.38) that $\partial_\mu F^{\mu\nu} \rightarrow O(x^{-3})$; the same must be true for the right hand side. From (3.38) and (3.39) we find that

$$\underset{x \rightarrow \infty}{D^\nu \phi} \longrightarrow i\phi \partial^\nu(e\omega + \alpha) + O(x^{-2}). \quad (3.41)$$

Hence,

$$\text{Im}(\phi^* D^\nu \phi) \underset{x \rightarrow \infty}{\longrightarrow} \phi_0^2 \partial^\nu(e\omega + \alpha) + O(x^{-2}). \quad (3.42)$$

For this to be $O(x^{-3})$, the first term must cancel the second, which requires $\partial^\nu(e\omega + \alpha) \rightarrow O(x^{-2})$, or

$$\underset{x \rightarrow \infty}{D^\nu \phi} \longrightarrow O(x^{-2}). \quad (3.43)$$

Along any infinitesimal segment dx on a circle of very large radius R , the length of the segment is $R d\theta$, where $d\theta$ is the angular displacement. Hence, by (3.43), $dx^\mu D_\mu \phi \sim R^{-1} d\theta$. That is, a classical finite-energy solution must obey the boundary condition

$$\underset{x \rightarrow \infty}{dx^\mu D_\mu \phi(x)} \longrightarrow 0. \quad (3.44)$$

Classical finite-energy solutions are interesting because they include “solitons” solutions that will be discussed later. This does not necessarily mean that classical infinite-energy solutions are irrelevant to physics. For example, classically, a plane wave has infinite energy, but leads to single particle states of finite energy in the quantum theory.

3.5 Magnetic Flux Quantization

Consider a static classical finite-energy solution. We can always choose $A^0 = 0$. At infinity, the boundary conditions (3.38) and (3.39) read

$$\begin{aligned} \mathbf{A}(\mathbf{x}) &\underset{x \rightarrow \infty}{\longrightarrow} \frac{1}{e} \nabla \alpha(\mathbf{x}), \\ \phi(\mathbf{x}) &\underset{x \rightarrow \infty}{\longrightarrow} \phi_0 e^{i\alpha(\mathbf{x})}. \end{aligned} \quad (3.45)$$

We denote $\alpha(\mathbf{x})$ and $\phi(\mathbf{x})$ on a fixed circle C of very large radius by $\alpha(\theta)$ and $\phi(\theta)$ respectively, for they depend only on the angular position $\theta(0 \leq \theta < 2\pi)$. Since $\phi(\theta)$ satisfies a differential equation, it must be continuous in space. Hence

$$\phi(2\pi) = \phi(0), \quad (3.46)$$

and this implies

$$\alpha(2\pi) - \alpha(0) = 2\pi n \quad (n = 0, \pm 1, \pm 2, \dots). \quad (3.47)$$

The net magnetic flux linking C is

$$\Phi = \iint \mathbf{dS} \cdot \mathbf{B} = \oint_C \mathbf{ds} \cdot \mathbf{A}, \quad (3.48)$$

where $d\mathbf{S}$ is a surface element of a surface spanning C , and ds is an element of arc along C . Using (3.45) and (3.47) we find

$$\Phi = \frac{2\pi n}{e} \quad (n = 0, \pm 1, \pm 2, \dots), \quad (3.49)$$

which is the statement of flux quantization, a necessary condition for a finite-energy solution.

In three spatial dimensions, we must choose $n = 0$; otherwise magnetic field lines would run off to infinity violating the condition (3.37) for finite energy. In two spatial dimensions, however, n can take on any integer value (with the magnetic field normal to the two-dimensional plane). The two-dimensional case may be looked upon as a three dimensional solution that is independent of the z coordinate, and has finite energy per unit length along the z axis.

Consider, then, a two dimensional problem in the x - y plane. Let us put $\alpha(\theta) = n\theta + \beta(\theta)$, where $\beta(2\pi) - \beta(0) = 0$. Thus, $\beta(0)$ is a continuous function, which can be transformed away through a continuous gauge transformation. Therefore there is a choice of gauge in which

$$\begin{aligned} \mathbf{A} &\xrightarrow[x \rightarrow \infty]{} \frac{1}{e} \nabla(n\theta) = \hat{\theta} \frac{n}{er} \quad (\text{pure gauge}), \\ \phi &\xrightarrow[x \rightarrow \infty]{} \phi_0 e^{in\theta}. \end{aligned} \quad (3.50)$$

Note that \mathbf{A} is multivalued, but $\oint ds \cdot \mathbf{A}$ is single-valued. For any n , the magnetic field approaches zero asymptotically, because \mathbf{A} approaches a pure-gauge form. The integer n labels different “gauge types”, which cannot be transformed into one another through continuous gauge transformations. This shows that magnetic flux confined to a finite portion of the x - y plane can make its presence known at infinity.

This fact underlies the Aharonov-Bohm effect⁹, namely, an electron is scattered in different manners by an impenetrable cylinder depending on the magnetic flux trapped entirely inside, even though the electron never enters the field region. The reason is that the phase of the electron wave function has an angular dependence correlated with the non-vanishing pure-gauge vector potential outside. The phase depends only on the total flux inside, and is therefore gauge invariant.

The flux quantization condition (3.49) arises from energetic considerations. One can always prepare an initial state that does not obey this condition; but then surface currents will be induced to adjust the flux to quantized values eventually. It should also be noted that (3.49) applies only to cylindrical geometry and does not hold, for example, for magnetic flux trapped inside a torus.¹⁰

In the classical electrodynamics of charged particles, a knowledge of $F^{\mu\nu}$ completely determines the properties of the system. A knowledge of A^μ is redundant there, because it is determined only up to gauge transformations,

⁹ Y. Aharonov and D. Bohm, *Phys. Rev.* **115**, 485 (1959).

¹⁰ K. Huang and R. Tipton, *Phys. Rev.* **23**, 3050 (1981), Appendix.

which do not affect $F^{\mu\nu}$. As we have seen, such is not the case in quantum theory, in which charged fields are coupled directly to A^μ ; a knowledge of $F^{\mu\nu}$ is not enough here. Although the continuous gauge transformations of A^μ remain physically irrelevant, discontinuous gauge transformations can generate different gauge types that give rise to different physical effects, the gauge invariance of $F^{\mu\nu}$ notwithstanding. The complete information that specifies the system, consists of $F^{\mu\nu}$ plus a specification of the gauge type.

3.6 Soliton Solutions: Vortices

A soliton solution is a classical finite-energy solution whose energy density remains non-vanishing in a finite region of space. In the three-dimensional case, the only possible solitons correspond to solutions in which magnetic flux is trapped inside a closed tube topologically equivalent to a torus¹¹. In the two-dimensional case (taken in the sense of a z-independent three-dimensional case), the solitons are “vortex lines”, configurations in which quantized magnetic flux is trapped inside a linear tube of finite radius, with the Higgs field assuming the normal value ϕ_0 outside the tube, but decreasing to zero towards the axis of the tube. Such vortex lines have been experimentally observed in superconducting bodies, with the ends of a vortex line attached to surfaces of the body.¹²

We set up cylindrical coordinates (r, θ) in the x - y plane, and seek static solutions to the equations of motion in Coulomb gauge ($\nabla \cdot A = 0$) subject to the boundary conditions (3.50). The static equations of motion read

$$\begin{aligned} (\nabla^2 - 2e^2\phi^*\phi)\mathbf{A} &= -ie\phi^*\vec{\nabla}\phi, \\ (\nabla - ie\mathbf{A})^2\phi &= 2\lambda\phi(\phi^*\phi - \phi_0^2), \end{aligned} \quad (3.51)$$

in which we have set $A^0 = 0$. To look for cylindrically symmetric solutions, put

$$\begin{aligned} \mathbf{A}(r, \theta) &= \hat{\theta}A(r), \\ \phi(r, \theta) &= \rho(r)e^{in\theta}. \end{aligned} \quad (3.52)$$

The magnetic field is given by

$$\mathbf{B} = \nabla \times \mathbf{A} = \hat{z} \frac{1}{r} \frac{d}{dr} [rA(r)]. \quad (3.53)$$

Now put

$$\begin{aligned} A(r) &= \frac{n}{er} [1 - F(r)], \\ B(r) &= -\frac{n}{er} F'(r), \end{aligned} \quad (3.54)$$

¹¹ K. Huang and R. Tipton, *op. cit.*

¹² K. Mendelsohn, *The Quest for Absolute Zero* (Taylor and Francis, London, 1977).

where a prime ('') denotes differentiation with respect to r . Then (3.51) can be rewritten in the form

$$\begin{aligned} F'' - \frac{F'}{r} - 2e^2 \rho^2 F &= 0, \\ \rho'' + \frac{\rho'}{r} - \frac{n^2 F^2}{r^2} \rho - 2\lambda \rho (\rho^2 - \phi_0^2) &= 0, \end{aligned} \quad (3.55)$$

with the boundary conditions

$$\begin{aligned} F &\xrightarrow[r \rightarrow \infty]{} 0, \\ \rho &\xrightarrow[r \rightarrow \infty]{} \phi_0. \end{aligned} \quad (3.56)$$

Flux quantization implies

$$2\pi \int_0^\infty dr r B(r) = \frac{2\pi n}{e}, \quad (3.57)$$

or

$$F(0) = 1 \quad (\text{for } n \neq 0).$$

Near $r = 0$, the solutions are

$$\begin{aligned} F &\xrightarrow[r \rightarrow 0]{} 1 - O(r^2), \\ B &\xrightarrow[r \rightarrow 0]{} \text{constant}, \\ \rho &\xrightarrow[r \rightarrow 0]{} ar^\mu. \end{aligned} \quad (3.58)$$

For $r \rightarrow \infty$, we may set $\rho = \phi_0$ in (3.55), and obtain the asymptotic forms

$$\begin{aligned} F(r) &\xrightarrow[r \rightarrow \infty]{} \text{const. } r^{1/2} \exp(-\sqrt{2}e\phi_0 r), \\ A(r) &\xrightarrow[r \rightarrow \infty]{} \frac{n}{er} + O(e^{-r}), \\ B(r) &\xrightarrow[r \rightarrow \infty]{} O(e^{-r}). \end{aligned} \quad (3.59)$$

That a solution exists can be shown by a variational principle. For static solutions, the action is proportional to the negative of the energy:

$$\mathcal{E} = \int_0^\infty dr r \left[\frac{1}{2} \left(\frac{n}{er} \right)^2 (F')^2 + \left(\frac{n}{r} \right)^2 F^2 \rho^2 + (\rho')^2 + V(\rho) \right], \quad (3.60)$$

with the conditions

$$\begin{aligned} F(0) &= 1, & F(\infty) &= 0, \\ \rho(0) &= 0, & \rho(\infty) &= \phi_0. \end{aligned} \quad (3.61)$$

Since every term in (3.60) is non-negative, \mathcal{E} has a minimum with respect to variations of F and ρ . The F and ρ that minimize \mathcal{E} are the solutions of lowest energy, with given n , and with the required boundary conditions. The qualitative nature of the solutions can be seen by inspection of (3.60). To minimize \mathcal{E} , $(F')^2$ wants to make F as smooth as possible, $(\rho')^2$ wants to make ρ as smooth as possible, and $F^2\rho^2$ wants to be as small as possible. The last condition means that where F is large, ρ wants to be small, and vice versa. Using these facts together with (3.58) and (3.59), we can make a rough sketch of F and ρ , as shown in Fig. 3.3., which shows the qualitative features of a vortex line.

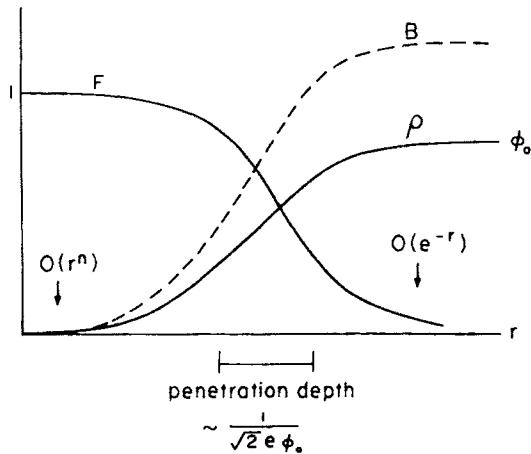


Fig. 3.3 Magnetic field B and Higgs field ρ in a vortex line, as function of normal distance r from axis of vortex.

CHAPTER 4

YANG-MILLS FIELDS: NON-ABELIAN GAUGE THEORIES

4.1 Introductory Note

The Maxwell field gives us the freedom to assign arbitrary signs to the charge of a particle at different space-time points. Yang-Mills fields play a similar role with respect to quantum numbers associated with higher symmetries. Historically, they were introduced with isospin in mind by Yang and Mills¹, who stated their motivation as follows:

The conservation of isotopic spin^a is identical with the requirement of invariance of all interactions under isotopic spin rotation. This means that . . . the orientation of the isotopic spin is of no physical significance. The differentiation between a neutron and a proton is then a purely arbitrary process. As usually conceived, however, once one chooses what to call a proton, what to call 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 . . .

We now believe that isospin is not a gauge symmetry, and cannot be, because it is not an exact symmetry in nature. But the idea of Yang and Mills, as applied to other internal symmetries, has led to the current gauge theories of interactions. The relevant internal symmetries are associated with Lie groups.

4.2 Lie Groups²

1 Structure Constants

For our purpose, a *Lie group* is a continuous group generated by a *Lie algebra*, which is a space whose basis consists of N generators L_a ($a = 1, \dots,$

^a Isotopic spin is the old name for isospin.

¹ C. N. Yang and R. L. Mills, *Phys. Rev.* **96**, 191 (1954).

² For general reference, see W. Miller, *Symmetry Groups and Their Applications* (Academic Press, New York, 1972); R. Gilmore, *Lie Groups, Lie Algebras, and Some of Their Applications* (Wiley-Interscience, New York, 1974); W. K. Tung, *Group Theory in Physics* (World Scientific, Singapore, 1985).

N) that are closed under commutations:

$$[L_a, L_b] = iC_{ab}^c L_c. \quad (4.1)$$

Here we use the summation convention for repeated indices. The constants C_{ab}^c are real numbers called *structure constants*, which completely characterize the Lie algebra. An element of G is of the form^b

$$U = e^{-i\omega_a L_a}, \quad (4.2)$$

where ω_a are arbitrary real numbers.^c

The structure constants are not unique, for they change under a linear transformation of the generators, which does not change the group. Some properties, however, are invariant. First, it is obvious from (4.1) that

$$C_{ab}^c = -C_{ba}^c. \quad (4.3)$$

Secondly, the *Jacobi identity* for commutators

$$[[L_a, L_b], L_c] + [[L_b, L_c], L_a] + [[L_c, L_a], L_b] = 0 \quad (4.4)$$

requires the identity

$$C_{ab}^n C_{nc}^d + C_{bc}^n C_{na}^d + C_{ca}^n C_{nb}^d = 0, \quad (4.5)$$

which we shall also refer to as the Jacobi identity.

If all the structure constants vanish, G would be an Abelian group—the direct product of N $U(1)$ groups. If not all the structure constants are zero, G is non-Abelian. The smallest non-Abelian Lie group is $SU(2)$, with $N = 3$.

If, after making a linear transformation on $\{L_a\}$ if necessary, the index a can be divided into two sets, such that $C_{ab}^c = 0$ whenever a belongs to one set and b to the other, then the Lie algebra breaks up into two commuting subalgebras. In this case G is a direct product of two independent Lie groups. A non-Abelian Lie group that cannot be so factorized is called *simple*. [Note that $U(1)$ is excluded]. It is characterized by the property that any two generators can be connected to each other through a chain of commutations. A direct product of simple Lie groups is called *semi-simple*.

2 Matrix Representations

We represent the generators by matrices, thereby inducing a matrix representation of G . To avoid needless complications, we consider only finite-dimensional matrices^d. If $\{L_a\}$ satisfies (4.1), so does the hermitian conjugate set $\{L_a^\dagger\}$. Hence, it is possible to represent $\{L_a\}$ by finite hermitian matrices. Group elements are then represented by finite unitary matrices.

^b That these form a group is a non-trivial fact, which depends on the Baker-Hausdorff-Campbell theorem (see Ref. 2).

^c We adopt a narrow definition of a Lie group that would suffice for later applications. To mathematicians, a Lie group is much more general, i.e., a continuous group that can be parametrized analytically.

^d This excludes the Lorentz group from our discussions, because its unitary representations are all infinite-dimensional. Its finite-dimensional representations are not unitary, and are thus excluded by (4.2).

The faithful representation of lowest dimensionality is called the *fundamental representation*. Other representations may be obtained by taking the repeated direct products of this representation with itself. An example of this procedure was discussed in Sec. 2.2 for the group $SU(3)$.

There is always one irreducible representation (not necessarily faithful) that is completely determined by the structure constants, namely, the *adjoint representation*. Its dimensionality is N (the number of generators), and the representative of L_a is given by

$$(L_a)_{bc} = -iC_{ab}^c. \quad (4.6)$$

This representation will play a central role in Yang-Mills gauge theory. To show that this is a representation, we calculate the commutator $[L_a, L_b]$:

$$\begin{aligned} [L_a, L_b]_{cd} &= (L_a)_{cn}(L_b)_{nd} - (L_b)_{cn}(L_a)_{nd} \\ &= -C_{ac}^n C_{bn}^d + C_{bc}^n C_{an}^d = C_{ca}^n C_{bn}^d + C_{bc}^n C_{an}^d \\ &= -C_{ab}^n C_{nc}^d \quad (\text{by the Jacobi identity}) \\ &= C_{ab}^n i(L_n)_{cd}. \blacksquare \end{aligned} \quad (4.7)$$

A simple Lie algebra can be made a metric space by defining the scalar product between L_a and L_b by

$$g_{ab} \equiv \text{Tr}(L_a L_b)_{\text{Adj. rep.}} = -C_{an}^m C_{bm}^n, \quad (4.8)$$

which is clearly a symmetric tensor. The norm of L_a is positive definite (i.e., $g_{aa} > 0$) because L_a is hermitian. Hence, by a linear transformation of the generators, g_{ab} can be diagonalized, and all the eigenvalues made unity:

$$g_{ab} = \delta_{ab}. \quad (4.9)$$

Defining new structure constants C_{abc} by

$$C_{abc} \equiv C_{ab}^n g_{nc}, \quad (4.10)$$

we can show that C_{abc} is completely antisymmetric in the indices a, b, c , by using the Jacobi identity. Using (4.9), we obtain

$$C_{ab}^c = C_{abc}. \quad (4.11)$$

Therefore, C_{ab}^c is completely antisymmetric in a, b, c . We shall write C_{abc} in place of C_{ab}^c from now on.

For any representation of a *simple* Lie group, the statements (4.8) and (4.9) can be generalized to

$$\text{Tr}(L_a L_b) = K \delta_{ab} \quad (\text{for simple Lie group}), \quad (4.12)$$

where K depends on the representation, but not on a . The proof depends on (4.11). First, note that the tensor $\text{Tr}(L_a L_b)$ can always be diagonalized by a suitable choice of $\{L_a\}$, so that

$$\text{Tr}(L_a L_b) = \begin{cases} 0 & \text{if } a \neq b \\ K_a & \text{if } a = b. \end{cases} \quad (4.13)$$

We only need to show that K_a is independent of a . To do this, define

$$d_{abc} \equiv \text{Tr}\{[L_a, L_b]L_c\} = \text{Tr}(L_aL_bL_c) - \text{Tr}(L_bL_aL_c). \quad (4.14)$$

Obviously, this is completely antisymmetric in a, b, c . Using (4.1) and (4.13) successively, we can write

$$d_{abc} = iC_{abn} \text{Tr}(L_nL_c) = iC_{abc}K_c \quad (\text{no sum on } c). \quad (4.15)$$

Interchanging the indices b and c , we have

$$d_{acb} = iC_{acb}K_b \quad (\text{no sum on } b). \quad (4.16)$$

Since $d_{abc} = -d_{acb}$, $C_{abc} = -C_{acb}$, on comparing (4.15) and (4.16), we conclude that

$$K_c = K_b, \quad (4.17)$$

whenever $[L_c, L_b] \neq 0$. Since the group is simple, any two generators are connected by a chain of commutations. Therefore all K_a 's are equal to one another, thus proving (4.11).

Finally, we note that the hermitian matrices $\{L_a\}$ can be replaced by real antisymmetric matrices $\{T_a\}$, defined by

$$T_a \equiv -iL_a. \quad (4.18)$$

The commutation relations read

$$[T_a, T_b] = -C_{abc}T_c, \quad (4.19)$$

and the adjoint representation is

$$(T_a)_{bc} = -C_{abc}. \quad (4.20)$$

Unitary matrices representing group elements take the form

$$U = e^{\omega_a T_a}, \quad U^{-1} = e^{-\omega_a T_a}. \quad (4.21)$$

3 Topological Properties

In a matrix representation, G is parametrized by $\{\omega_a\}$. In order that there be a one-to-one correspondence between $\{\omega_a\}$ and a group element, the possible values of $\{\omega_a\}$ must be suitably restricted. The space of the possible values of $\{\omega_a\}$ is called the *group manifold* of G . The group is called *compact* if the group manifold is a compact set. It is said to be *simply-connected* if every closed path in the group manifold can be continuously deformed to a point. Otherwise it is *multiply-connected*.

The correspondence between Lie group and Lie algebra is many-to-one. For example, $SU(2)$ and the rotation group $O(3)$ are different groups, but they share the same Lie algebra. The difference between the two groups is that $SU(2)$ is simply-connected, while $O(3)$ is doubly-connected.

Among all Lie groups sharing the same Lie algebra, only one is simply-connected, and this is called the *covering group*. Thus, $SU(2)$ is the covering group of $O(3)$.

To illustrate these ideas, we discuss $SU(2)$ and $O(3)$ in more detail. They are the only two groups sharing the Lie algebra characterized by $C_{abc} = \epsilon_{abc}$, and both are compact groups. It is well-known that $O(3)$ admits only integer angular momentum representations, whereas $SU(2)$ admits integer and half-integer representations. The difference arises from the topology of the respective group manifolds, as determined by the definitions of these groups.

We can associate a rotation with the tip of a 3-dimensional vector that points along the axis of rotation, with length equal to the angle of rotation. Thus, the manifold of $O(3)$ is the volume enclosed by a sphere of radius π , with diametrically opposite points on the surface identified as the same point. The last condition comes from the fact that rotations of $\pm\pi$ about the same axis are one and the same. There are two classes of closed paths in the group manifold: closed loops drawn within the sphere, and diameters of the sphere. The former can be continuously deformed to a point, while the latter cannot. Hence $O(3)$ is doubly-connected.

On the other hand, $SU(2)$ is defined as the group generated by the fundamental representation of the Lie algebra, i.e., by 2×2 matrices. Its most general element is of the form

$$\begin{aligned} U &= e^{i\theta \mathbf{\hat{b}} \cdot \boldsymbol{\sigma}/2} = b_0 + i\mathbf{b} \cdot \boldsymbol{\sigma}, \\ b_0 &\equiv \cos \theta/2, \\ \mathbf{b} &\equiv \hat{\mathbf{n}} \sin \theta/2. \end{aligned} \quad (4.22)$$

Hence, the group is parametrized by the four numbers b_0 , \mathbf{b} , with the condition

$$\sum_{i=1}^4 b_i^2 = 1. \quad (4.23)$$

The manifold of $SU(2)$ is therefore the surface of a unit sphere in 4-dimensional Euclidean space, and is obviously simply-connected.

The difference between $SU(2)$ and $O(3)$ can be illustrated by the following experiment. Hang a sign from the ceiling with rubber bands, and anchor it to the floor with rubber bands, as shown in Fig. 4.1(a). Rotating the sign about a vertical axis through angle 2π produces a twisting of the rubber bands that cannot be unravelled by deforming the rubber bands continuously [see Fig. 4.1(b)]. However, the twisting produced by a 4π rotation can be so unravelled [see Fig. 4.1(c)]. Fig. 4.2 shows how.

This experiment illustrates the fact that a 2π rotation changes the relationship between the rotated object and its surroundings, even though the object is returned to its original orientation. Only a 4π rotation truly changes nothing. The group $SU(2)$ distinguishes between a 2π rotation and no rotation, whereas $O(3)$ identifies them by ignoring the rubber bands. Stated more precisely, $SU(2)$ contains a center Z_2 , which is the subgroup (consisting of the two elements ± 1) that commutes with all group elements. The group $O(3)$ is the factor group with Z_2 taken to be the identity element, i.e., $O(3) = SU(2)/Z_2$.

As another example, consider $U(1)$. Since its elements have the form $e^{i\theta}$, its manifold is the unit circle, which is compact. However, the manifold is not simply connected. There are an infinite number of classes of closed paths labelled by the number of times the path winds around the circle, and paths with

different winding numbers cannot be deformed into each other continuously. The covering group is of course simply-connected; but it is not compact because its manifold is the real line.

4 General Remarks

We have confined our discussion to groups of finite unitary matrices. Some of the properties we mentioned are in fact more general. For example, the complete antisymmetry of C_{ab}^c , which follows from (4.11), is true for any compact semi-simple Lie group. The general definition of a semi-simple Lie

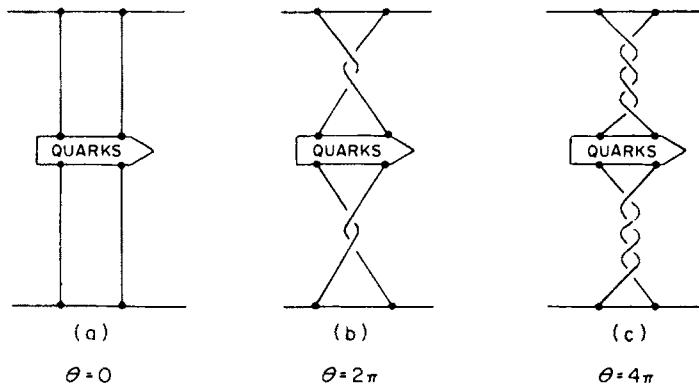


Fig. 4.1 (a) a sign held up by four rubber bands
 (b) The sign is rotated about a vertical axis through angle $\theta = 2\pi$. The twists in the rubber bands cannot be undone by continuous deformations of the rubber bands.
 (c) The sign is rotated through angle $\theta = 4\pi$. The situation is now topologically equivalent to (a), as shown in the next figure.

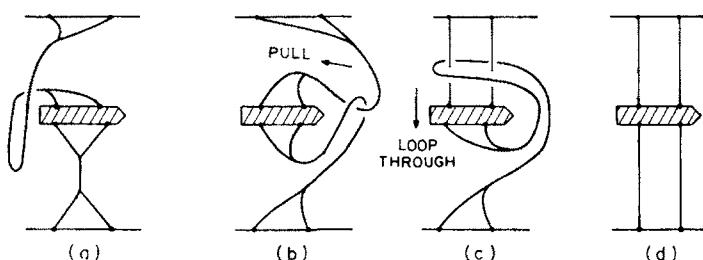


Fig. 4.2 How to undo the twists in the rubber bands in Fig. 4.1(c).

group is one whose Lie algebra does not contain an Abelian sub-algebra. Cartan's theorem states that a Lie group is semi-simple if and only if $\det \|g_{ab}\| \neq 0$, where g_{ab} is defined by (4.8). Hence, for a semi-simple Lie group, g_{ab} can be diagonalized, and all the eigenvalues can be made +1 or -1 by re-scaling L_a . If the group is compact, they can all be made +1, and (4.11) follows.

As we shall see in the next section, the necessary and sufficient conditions for the construction of a Yang-Mills gauge theory are the complete antisymmetry of C_{ab}^c , and the Jacobi identity. This means that the Lie group has to be a compact semi-simple group, and the Jacobi identity should hold. The Jacobi identity is automatic for a Lie algebra of finite matrices; but it can fail for infinite matrices.

4.3 The Yang-Mills Construction

1 Global Gauge Invariance

Yang-Mills fields are required when we enlarge global gauge symmetries to local ones. We begin with a description of globally gauge invariant systems.

A matrix representation of the Lie group element U is a generalization of the $U(1)$ element $e^{i\theta}$. We continue to refer to it as a gauge transformation (of the representational space on which it acts). If the matrix elements of U are space-time independent, U is called a global gauge transformation. Otherwise it is called a local gauge transformation.

A system of fields (called matter fields) that is globally gauge invariant generally contains multicomponent fields, which we denote collectively by $\Psi(x)$. The various components are grouped into multiplets that transform according to definite irreducible representations of the Lie group G . In general, these irreducible representations are different for different multiplets.

A multiplet may be boson or fermion; but a boson and a fermion field cannot be included in the same multiplet, because they respectively obey commutation and anticommutation rules in the quantum theory, and so cannot transform into each other.

Let the Lagrangian density be $\mathcal{L}_0(\Psi(x), \partial^\mu \Psi(x))$. Then, global gauge invariance states that

$$\mathcal{L}_0(U\Psi, \partial^\mu U\Psi) = \mathcal{L}_0(\Psi, \partial^\mu \Psi). \quad (4.24)$$

There are requirements arising from Lorentz invariance:

(a) Fermion fields ψ (of spin 1/2) are complex Dirac spinor fields. The canonical conjugate of ψ is $i\psi^\dagger$, i.e. ψ appears in L_0 in the form $i\psi^\dagger\psi$.

(b) Boson fields ϕ may be real or complex. The canonical conjugate to ϕ is ϕ^* , i.e., ϕ appears in L_0 in the form $\phi^*\phi$.

The two independent parts ϕ and ϕ^* of a complex boson field may be replaced by a pair of real fields, the real and imaginary parts of ϕ . Correspondingly, the representative of an element of G may be expressed either in complex form or real form. As an example, consider a complex boson field forming an n -dimensional irreducible representation of G . There are $2n$ independent fields,

which can be represented as two complex column vectors:

$$\phi = \begin{pmatrix} \phi_1 \\ \vdots \\ \phi_n \end{pmatrix}, \quad \phi^* = \begin{pmatrix} \phi_1^* \\ \vdots \\ \phi_n^* \end{pmatrix}. \quad (4.25)$$

An infinitesimal gauge transformation is of the form

$$\begin{aligned} \phi &\rightarrow \phi + \delta\phi, & \delta\phi &= -i\omega\phi, \\ \phi^* &\rightarrow \phi^* + \delta\phi^*, & \delta\phi^* &= i\omega^T\phi, \end{aligned} \quad (4.26)$$

where ω is an element of the Lie algebra:

$$\omega \equiv \omega_a L_a, \quad (4.27)$$

where ω_a are infinitesimal real numbers, and L_a is an $n \times n$ hermitian matrix representing a generator of G . ω^T denotes the transpose of ω .

Alternatively, we may consider the independent fields to be the real and imaginary parts of ϕ :

$$\phi = 2^{-1/2}(A + iB), \quad (4.28)$$

where A and B are both n -component real fields. From (4.26), we find easily that under an infinitesimal gauge transformation,

$$\begin{aligned} A &\rightarrow A + \delta A, & \delta A &= -\frac{i}{2}(\omega - \omega^T)A + \frac{1}{2}(\omega + \omega^T)B, \\ B &\rightarrow B + \delta B, & \delta B &= -\frac{1}{2}(\omega + \omega^T)A - \frac{i}{2}(\omega - \omega^T)B. \end{aligned} \quad (4.29)$$

These can be stated in terms of one real field of $2n$ components:

$$\begin{aligned} \Phi &= \begin{pmatrix} A \\ B \end{pmatrix} \\ \Phi &\rightarrow \Phi + \delta\Phi, & \delta\Phi &= \omega_a T_a \Phi, \\ T_a &= - \begin{pmatrix} \text{Im } L_a & \text{Re } L_a \\ -\text{Re } L_a & \text{Im } L_a \end{pmatrix} \quad (2n \times 2n \text{ matrix}). \end{aligned} \quad (4.30)$$

As an illustrative example of how the various fields appear in the Lagrangian density, let us take G to be the isospin group $SU(2)$, and choose the matter fields to consist of the following multiplets:

$$\Psi = \{\pi, K, N, \bar{N}, \Sigma, \bar{\Sigma}\}, \quad (4.31)$$

with transformation properties given in Table 4.1. A free Lagrangian density with global isospin invariance is

$$\begin{aligned} \mathcal{L}_0(\Psi, \partial^\mu \Psi) &= \frac{1}{2} \partial_\mu \pi_a \partial^\mu \pi_a + \partial_\mu K^* \partial^\mu K \\ &\quad + \bar{N}(i\gamma_\mu \partial^\mu - m)N + \bar{\Sigma}(i\gamma_\mu \partial^\mu - M)\Sigma. \end{aligned} \quad (4.32)$$

2 Local Gauge Invariance

Under a local gauge transformation,

$$\begin{aligned}\Psi(x) &\rightarrow U(x)\Psi(x), \\ \partial^\mu\Psi(x) &\rightarrow U(x)\partial^\mu\Psi(x) + [\partial^\mu U(x)]\Psi(x).\end{aligned}\quad (4.33)$$

Since $\partial^\mu U(x) \neq 0$, $\partial^\mu\Psi$ does not transform in the same manner as Ψ , and this spoils the invariance of \mathcal{L}_0 . To learn how to cancel the unwanted term $(\partial^\mu U)\Psi$, it suffices to consider infinitesimal local gauge transformations:

$$\begin{aligned}\delta\Psi(x) &= -i\omega(x)\Psi(x), \\ \delta[\partial^\mu\Psi(x)] &= -i\omega(x)\partial^\mu\Psi(x) - i[\partial^\mu\omega(x)]\Psi(x),\end{aligned}\quad (4.34)$$

where $\omega(x)$ is an element of the Lie algebra. We define the covariant derivative by

$$D^\mu\Psi(x) = [\partial^\mu + igA^\mu(x)]\Psi(x),\quad (4.35)$$

where $A^\mu(x)$ is an element of the Lie algebra:

$$A^\mu(x) = A_a^\mu(x)L_a.\quad (4.36)$$

Thus, we need N gauge fields $A_a^\mu(x)$ ($a = 1, \dots, N$) called Yang-Mills fields. Under an infinitesimal local gauge transformation,

$$D^\mu\Psi \rightarrow [\partial^\mu + ig(A^\mu + \delta A^\mu)](\Psi + \delta\Psi).\quad (4.37)$$

Note that $\delta\Psi = -i\omega\Psi$ does not commute with $A^\mu + \delta A^\mu$, because ω and

Table 4.1 EXAMPLES OF SU(2) MULTIPLETS

	Boson		Fermion	
Fields	$\pi = \begin{pmatrix} \pi_1 \\ \pi_2 \\ \pi_3 \end{pmatrix}$	$K = \begin{pmatrix} K^0 \\ \bar{K}^0 \end{pmatrix}$	$N = \begin{pmatrix} p \\ n \end{pmatrix}$	$\Sigma = \begin{pmatrix} \Sigma^+ \\ \Sigma^0 \\ \Sigma^- \end{pmatrix}$
			$\bar{N} = \begin{pmatrix} \bar{p} \\ \bar{n} \end{pmatrix}$	$\bar{\Sigma} = \begin{pmatrix} \bar{\Sigma}^+ \\ \bar{\Sigma}^0 \\ \bar{\Sigma}^- \end{pmatrix}$
Isospin	1	1/2	1/2	1
Generators L_a ($a = 1, 2, 3$)	$(L_a)_{bc} = -i\varepsilon_{abc}$ Adjoint representation	$L_a = \frac{1}{2}\tau_a$ Fundamental representation	$L_a = \frac{1}{2}\tau_a$	$(L_a)_{bc} = -i\varepsilon_{abc}$
Remarks	π_a ($a = 1, 2, 3$) are real numbers	K^0, \bar{K}^0 are complex numbers	Each entry in the column vector is a Dirac spinor with 4 complex components	

$A^\mu + \delta A^\mu$ are elements of the Lie algebra. The change in $D^\mu \Psi$ is given by

$$\delta(D^\mu \Psi) = -i\omega D^\mu \Psi - ig \left\{ \delta A^\mu - \frac{1}{g} \partial^\mu \omega + i[\omega, A^\mu] \right\}. \quad (4.38)$$

To make $D^\mu \Psi$ transform in the same manner as Ψ , we require that the last term vanish, namely,

$$\delta A^\mu(x) = \frac{1}{g} \partial^\mu \omega(x) - i[\omega(x), A^\mu(x)]. \quad (4.39)$$

Multiplying both sides by L_a , taking the trace, and using (4.12), we obtain

$$\delta A_a^\mu(x) = \frac{1}{g} \partial^\mu \omega_a(x) + C_{abc} \omega_b(x) A_c^\mu(x), \quad (4.40)$$

where we have made use of the antisymmetry of C_{abc} . This makes $\mathcal{L}_0(\Psi, D^\mu \Psi)$ invariant under local gauge transformations.

To make the fields A_a^μ dynamical objects, we still have to add to $\mathcal{L}_0(\Psi, D^\mu \Psi)$ their free Lagrangian density, which should be quadratic in space-time derivatives of A_a^μ , and should be both gauge invariant and Lorentz invariant. In analogy with the Maxwell case, we may try to define a field tensor $\partial^\mu A_a^\nu - \partial^\nu A_a^\mu$; but this transforms in a complicated way under (4.40). We seek a field tensor that transforms according to an irreducible representation of G that is solely determined by C_{abc} . Hence, the representation must be the adjoint representation. The following satisfies our requirements:

$$F_a^{\mu\nu}(x) \equiv \partial^\mu A_a^\nu(x) - \partial^\nu A_a^\mu(x) - g C_{abc} A_b^\mu(x) A_c^\nu(x). \quad (4.41)$$

The corresponding element of the Lie algebra, $F^{\mu\nu} \equiv F_a^{\mu\nu} L_a$, is given by

$$F^{\mu\nu}(x) = \partial^\mu A^\nu(x) - \partial^\nu A^\mu(x) + ig[A^\mu(x), A^\nu(x)]. \quad (4.42)$$

Under an infinitesimal local gauge transformation, it can be verified from (4.42), using the Jacobi identity, that

$$\begin{aligned} \delta F^{\mu\nu}(x) &= -i[\omega(x), F^{\mu\nu}(x)], \\ \delta F_a^{\mu\nu}(x) &= C_{abc} \omega_b(x) F_c^{\mu\nu}(x). \end{aligned} \quad (4.43)$$

From (4.6) and the antisymmetry of C_{abc} , we see that $C_{abc} = -i(L_b)_{ac}$. Hence

$$\delta F_a^{\mu\nu} = -i(\omega_b L_b)_{ac} F_c^{\mu\nu}. \quad (4.44)$$

This shows that $F_a^{\mu\nu}$ transforms according to the adjoint representation.

We have shown that the Jacobi identity and the complete antisymmetry of C_{abc} are necessary for (4.42) to lead to (4.43). Glashow and Gell-Mann³ have shown that the same conditions are also necessary for the converse, i.e., that (4.43) leads to (4.42). Therefore, $F_a^{\mu\nu}$ is uniquely given by (4.42).

³ S. L. Glashow and M. Gell-Mann, *Annals of Phys.* **15**, 437 (1961).

We now take the free Lagrangian density of the gauge field to be $-\frac{1}{4}F_a^{\mu\nu}F_{a\mu\nu}$, which is gauge invariant (see Table 4.2). The factor 1/4 is conventional.

The complete locally gauge invariant Lagrangian density is

$$\mathcal{L} = -\frac{1}{4}F_a^{\mu\nu}F_{a\mu\nu} + \mathcal{L}_0(\Psi, D^\mu\Psi). \quad (4.45)$$

This is the Yang-Mills Lagrangian density^c. The necessary and sufficient conditions for this construction to be possible are that.

- (a) the Jacobi identity holds,
- (b) C_{abc} is completely antisymmetric.

Condition (b) requires G to be a compact semi-simple Lie group (see Sec. 4.2). Condition (a) is automatic for finite matrices, but may not hold for infinite matrices.

In the definition (4.41), the only arbitrary parameters are g , and the relative normalization constants between subsets of C_{abc} corresponding to the various simple Lie groups contained in G , the arbitrariness of C_{abc} for a simple Lie group having been fixed by the convention (4.12). The parameter g is called a *gauge coupling constant*. It fixes the relative scale between the gauge fields and the

Table 4.2 LOCAL GAUGE TRANSFORMATIONS

Infinitesimal Transformations

$$\omega(x) = \omega_a(x)L_a$$

$$\delta\Psi_i(x) = -i\omega_a(x)(L_a)_{ij}\Psi_j(x),$$

$$\delta A_a^\mu(x) = \frac{1}{g}\partial^\mu\omega_a(x) + C_{abc}\omega_b(x)A_c^\mu(x)$$

$$\delta F_a^{\mu\nu}(x) = C_{abc}\omega_b(x)F_c^{\mu\nu}(x)$$

$$\delta\Psi(x) = -i\omega(x)\Psi(x)$$

$$\delta A^\mu(x) = \frac{1}{g}\partial^\mu\omega(x) - i[\omega(x), A^\mu(x)]$$

$$\delta F^{\mu\nu}(x) = -i[\omega(x), F^{\mu\nu}(x)]$$

Finite Transformations

$$U(x) = e^{-i\omega(x)}$$

$$\Psi(x) \rightarrow U(x)\Psi(x)$$

$$A^\mu(x) \rightarrow U(x)A^\mu(x)U^{-1}(x) - \frac{i}{g}U(x)\partial^\mu U^{-1}(x)$$

$$F^{\mu\nu}(x) \rightarrow U(x)F^{\mu\nu}(x)U^{-1}(x)$$

^c This does not include Einstein's theory of gravitation, which requires tensor gauge fields instead of vector gauge fields. The reason is that the Lorentz group is a group of space-time transformations, so that the index labelling a group generator is also a space-time index.

matter fields. For a simple G , it can be absorbed by redefining the gauge fields to be

$$\tilde{A}_a{}^\mu(x) = g A_a{}^\mu(x). \quad (4.46)$$

If G is not simple, there would be one independent gauge coupling constant for each simple sub-group, and each $U(1)$ sub-group. In general, not all of these can be absorbed simultaneously by re-scaling the gauge fields. From now on, we assume that G is simple [and not $U(1)$], for generalizations are straightforward.

The finite forms of the local gauge transformations can be obtained easily from the infinitesimal forms. We give the results in Table 4.2. Note $F_a{}^{\mu\nu}F_{a\mu\nu}$ is proportional to $\text{Tr}(F^{\mu\nu}F_{\mu\nu})$, which is obviously gauge invariant.

The pure-gauge form of $A^\mu(x)$ is

$$A^\mu(x) = -\frac{i}{g} U(x) \partial^\mu U^{-1}(x) \quad (\text{pure gauge}), \quad (4.47)$$

which is the generalization of $\partial^\mu \omega$ in the $U(1)$ case. It can be easily shown that (4.47) gives $F^{\mu\nu} = 0$. The matrix $U(x)$ is not necessarily obtainable from $U(x) = 1$ by continuous changes. This gives rise to different pure-gauge types.

To express (4.47) in terms of $\omega = i \ln U$, we note that an infinitesimal change in $\omega(x)$ leads to a change in $U(x)$ given by^f

$$dU \equiv e^{-i(\omega + d\omega)} - e^{-i\omega} = \int_0^1 dt e^{-(1-t)i\omega} (-i d\omega) e^{-it\omega}. \quad (4.48)$$

Using this in conjunction with (4.47) leads to

$$A^\mu(x) = -\frac{1}{g} \int_0^1 dt e^{-it\omega(x)} [\partial^\mu \omega(x)] e^{it\omega(x)} \quad (\text{pure gauge}). \quad (4.49)$$

4.4 Properties of Yang-Mills Fields

1 Electric and Magnetic Fields

We define electric and magnetic fields $\mathbf{E} = \mathbf{E}_a L_a$ and $\mathbf{B} = \mathbf{B}_a L_a$, which are elements of the Lie algebra, by designating the components of the antisymmetric tensor (or 6-vector) $F^{\mu\nu}$ as follows:

$$F^{\mu\nu} = \begin{pmatrix} 0 & -E^1 & -E^2 & -E^3 \\ E^1 & 0 & -B^3 & B^2 \\ E^2 & B^3 & 0 & -B^1 \\ E^3 & -B^2 & B^1 & 0 \end{pmatrix}, \quad (4.50)$$

^fOwing to the well-known formula

$$e^{A+B} = e^A + \int_0^1 dt e^{(1-t)A} B e^{tA} + \dots$$

or, (with $k = 1, 2, 3$),

$$\begin{aligned} E^k &= F^{k0}, \\ B^k &= -\frac{1}{2}\epsilon^{klj}F^{lj}, \quad F^{ij} = -\epsilon^{ijk}B^k. \end{aligned} \quad (4.51)$$

In terms of $A^\mu = (A^0, \mathbf{A})$:

$$\begin{aligned} E^k &= \partial^k A^0 - \partial^0 A^k + ig[A^k, A^0], \\ B^k &= (\partial^l A^m - \partial^m A^l) - ig[A^l, A^m] \quad (k, l, m \text{ cyclic}). \end{aligned} \quad (4.52)$$

These matrices are not gauge-invariant, but undergo unitary transformations under a group element U (see Table 4.2). We can also write

$$\mathbf{E}_a = -\nabla A_a^0 - \frac{\partial \mathbf{A}_a}{\partial t} - gC_{abc}\mathbf{A}_b A_c^0, \quad (4.53)$$

$$\mathbf{B}_a = \nabla \times \mathbf{A}_a + \frac{1}{2}gC_{abc}\mathbf{A}_b \times \mathbf{A}_c.$$

Note that

$$\nabla \cdot \mathbf{B}_a = \frac{1}{2}gC_{abc}\nabla \cdot (\mathbf{A}_b \times \mathbf{A}_c), \quad (4.54)$$

which shows that magnetic charge density can exist. The total magnetic charge is a surface integral at spatial infinity, and there are special solutions (monopole solutions) for which it does not vanish, as we shall discuss in Chapter 5.

2 Dual Tensor

Define the dual tensor by

$$\tilde{F}^{\mu\nu} = \frac{1}{2}\epsilon^{\mu\nu\alpha\beta}F_{\alpha\beta}, \quad (4.55)$$

which reads, as an antisymmetric tensor in μ, ν :

$$\tilde{F}^{\mu\nu} = \begin{pmatrix} 0 & -B^1 & -B^2 & -B^3 \\ B^1 & 0 & -E^3 & E^2 \\ B^2 & E^3 & 0 & -E^1 \\ B^3 & -E^2 & E^1 & 0 \end{pmatrix}. \quad (4.56)$$

It is obtainable from $F^{\mu\nu}$ by interchanging \mathbf{B} and \mathbf{E} . The following are Lorentz invariant quantities:

$$\begin{aligned} \frac{1}{4}F_a^{\mu\nu}F_{a\mu\nu} &= \frac{1}{2}(\mathbf{B}_a \cdot \mathbf{B}_a - \mathbf{E}_a \cdot \mathbf{E}_a) \quad (\text{scalar}), \\ \frac{1}{4}\tilde{F}_a^{\mu\nu}F_{a\mu\nu} &= -\mathbf{E}_a \cdot \mathbf{B}_a \quad (\text{pseudoscalar}). \end{aligned} \quad (4.57)$$

In the Abelian case, the absence of magnetic current is implied by the *identity* $\partial_\mu \tilde{F}^{\mu\nu} = 0$, which has no dynamical content, but is a direct consequence of the definition $F^{\mu\nu} = \partial^\mu A^\nu - \partial^\nu A^\mu$. In the non-Abelian case here, we can verify from (4.55), using the Jacobi identity, that

$$\begin{aligned} \partial^\mu \tilde{F}_{\mu\nu} + ig[A^\mu, \tilde{F}_{\mu\nu}] &= 0, \\ \partial^\mu \tilde{F}_{a\mu\nu} - gC_{abc}A_{b\mu} \tilde{F}_c^{\mu\nu} &= 0. \end{aligned} \quad (4.58)$$

We can also write

$$D^\mu \tilde{F}_{\mu\nu} \equiv 0, \quad (4.59)$$

where

$$D^\mu = \partial^\mu + ig[A^\mu, \quad] \quad (4.60)$$

is the covariant differentiation appropriate for the adjoint representation.

3 Path Representation of the Gauge Group

In the presence of Yang-Mills fields, each space-time point is associated with an independent choice of coordinate frames in the internal symmetry space of the matter fields. A change in the local frame between the space-time points x and $x + dx$ is correlated with a local gauge transformation of $A^\mu(x)$. The matter-field system does not care about the local frame, i.e., $\Psi(x)$ is physically indistinguishable from $U(x)\Psi(x)$. Similarly, local gauge transformations of $A^\mu(x)$ have no physical significance.

When there are no internal symmetries, $\Psi(x)$ is a single real field. The statement that $\Psi(x)$ does not change between x and $x + dx$ is then simply $dx_\mu \partial^\mu \Psi(x) = 0$. If $\Psi(x) = e^{-i\alpha(x)}|\Psi(x)|$ is a complex field coupled to a $U(1)$ gauge field, so that its phase has no physical significance, then for all physical purposes $\Psi(x)$ is constant over dx if $dx_\mu \partial^\mu |\Psi(x)| = 0$, or, $dx_\mu [\partial^\mu + i\partial^\mu \alpha(x)]\Psi(x) = 0$. In this case, we say that $\Psi(x)$ undergoes parallel displacement from x to $x + dx$, and we can restate that condition in terms of the $U(1)$ gauge field:

$$dx_\mu D^\mu \Psi(x) \equiv dx_\mu [\partial^\mu + igA^\mu(x)]\Psi(x) = 0. \quad (4.61)$$

In the Yang-Mills case, we take over this definition of *parallel displacement*, with $A^\mu(x) = A_a^\mu(x)L_a$. The local frame in internal space is specified by the matrix representing the Lie algebra element $A^\mu(x)$, i.e. by $A_a^\mu(x)$, when the matrix L_a is fixed.

Suppose $\Psi(x)$ undergoes parallel displacement along a path P , which is parametrized by $0 \leq s \leq 1$. Then at any point x on P ,

$$\frac{dx^\mu}{ds} [\partial^\mu + igA^\mu(s)]\Psi(s) = 0, \quad (4.62)$$

where $\Psi(x) \equiv \Psi(x(s))$, $A^\mu(s) \equiv A^\mu(x(s))$. The solution to the equation is

$$\Psi(s) = \mathcal{P} \left[\exp \left(-ig \int_0^s ds' (dx^\mu/ds') A_\mu(s') \right) \right] \Psi(0), \quad (4.63)$$

where \mathcal{P} is a “path-ordering operator” which instructs us to order the matrices $A_\mu(s')$ in increasing order of s' , in every term of the power series expansion of the exponential function. Thus, associated with every directed path P is the matrix

$$\Omega(P) \equiv \mathcal{P} \exp \left(-ig \int_P dx^\mu A_\mu(x) \right) \quad (4.64)$$

which is a path-dependent representation of an element of G , the particular representation being determined by the representation of L_a in the Lie algebra

element $A^\mu(x) = A_a^\mu(x)L_a$. The significance of $\Omega(P)$ is that it is the gauge transformation that takes $\Psi(x_0)$ to $\Psi(x_1)$ along P :

$$\Psi(x_1) = \Omega(P)\Psi(x_0), \quad (4.65)$$

where x_0 and x_1 are the endpoints of P : $x(0) = x_0$, $x(1) = x_1$.

Under a point-wise gauge transformation $U(x)$, $\Omega(P)$ transforms as follows:

$$\begin{aligned} \Omega(P) &\rightarrow \Omega'(P), \\ \Omega'(P) &= U(x_1)\Omega(P)U^{-1}(x_0). \end{aligned} \quad (4.66)$$

It follows that for a closed path C , $\text{Tr } \Omega(C)$ is gauge invariant.

Proof: Let $\Psi'(x) = U(x)\Psi(x)$. The definition of $\Omega'(P)$ is given by

$$\Psi'(x_1) = \Omega'(P)\Psi'(x_0) \quad (4.67)$$

or,

$$U(x_1)\Psi(x_1) = \Omega'(P)U(x_0)\Psi(x_0),$$

$$U(x_1)\Omega(P)\Psi(x_0) = \Omega'(P)U(x_0)\Psi(x_0),$$

$$U(x_1)\Omega(P) = \Omega'(P)U(x_0) \quad (\text{since } \Psi(x_0) \text{ is arbitrary}).$$

$$\therefore U(x_1)\Omega(P)U^{-1}(x_0) = \Omega'(P). \blacksquare$$

For an infinitesimal closed path at x , of area $dx^\mu dy^\nu$,

$$\Omega(dx dy) = 1 - ig dx^\mu dy^\nu F_{\mu\nu}(x). \quad (4.68)$$

Proof: Number the sides of the rectangle $n = 1, 2, 3, 4$ as shown in Fig. 4.3. Along any infinitesimal dx located at x , we have (to second order)

$$\Omega(dx) = 1 - ig dx_\mu A^\mu(x) - \frac{1}{2}g^2 dx_\mu dx_\nu A^\mu(x)A^\nu(x).$$

Let

$$\lambda_n = \int_{\text{side } n} dx^\mu A_\mu(x).$$

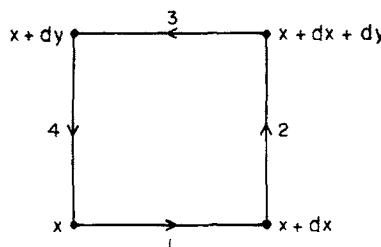


Fig. 4.3 An infinitesimal rectangle in space-time.

By the group property of Ω ,

$$\begin{aligned}\Omega(dx dy) &= \Omega_1 \Omega_2 \Omega_3 \Omega_4 = 1 - ig(\lambda_1 + \lambda_2 + \lambda_3 + \lambda_4) \\ &\quad - g^2[\lambda_1 \lambda_2 + \lambda_1 \lambda_3 + \lambda_1 \lambda_4 + \lambda_2 \lambda_3 + \lambda_2 \lambda_4 + \lambda_3 \lambda_4 \\ &\quad + \frac{1}{2}(\lambda_1^2 + \lambda_2^2 + \lambda_3^2 + \lambda_4^2)].\end{aligned}$$

Sample side calculations are as follows:

$$\begin{aligned}\lambda_1 + \lambda_3 &= \left(\int_x^{x+dx} - \int_{x+dy}^{x+dx+dy} \right) dz^\mu A_\mu(z) \\ &= dy^\mu \frac{\partial}{\partial x^\mu} \int_x^{x+dx} dz^\mu A_\mu(z) = dy^\mu dx^2 \partial_\mu A_\nu(x).\end{aligned}$$

$$\lambda_1 + \lambda_2 + \lambda_3 + \lambda_4 = dx^\mu dy^\nu (\partial_\mu A_\nu - \partial_\nu A_\mu).$$

$$g^2[\lambda_1 \lambda_2 + \dots] = -g^2 dx^\mu dy^\nu [A_\mu(x), A_\nu(x)].$$

Hence,

$$\Omega(dx dy) = 1 - ig dx^\mu dy^\nu F_{\mu\nu}(x). \blacksquare$$

A knowledge of $\Omega(P)$ for all paths P determines $A_a^\mu(x)$, by differentiation with respect to an endpoint of P . The classical gauge theory can be formulated entirely in terms of $\Omega(P)$,⁴ but the canonical quantization is still based on the quantization of $A_a^\mu(x)$, which will be done in Chapter 8. A knowledge of $\Omega(C)$ for all closed loops C contains all the physical information about the classical gauge field without redundancy.⁵

To illustrate a practical use for $\Omega(P)$, we show that *any space-time component of $A_a^\mu(x)$ can be reduced to zero by means of a continuous local gauge transformation*. When a spatial component is zero, $A_a^\mu(x)$ is said to be in the *axial gauge*. When the time component is zero, it is said to be in the *temporal gauge*. Suppose we wish to make $A_a^0(x) = 0$. Let P_x be the space-time path directing linearly from $x_0 = (0, \mathbf{x})$ to $x = (t, \mathbf{x})$. Consider

$$\Omega(P_x) = \mathcal{P} \exp \left(-ig \int_0^t dt' A^0(t', \mathbf{x}) \right). \quad (4.69)$$

Clearly, $\Omega(P_x) = 1$ for $x = x_0$. If $A_a^0(x)$ is not already zero, make the continuous local gauge transformation

$$U(x) = [\Omega(P_x)]^{-1}, \quad U(x_0) = [\Omega(P_{x_0})]^{-1} = 1. \quad (4.70)$$

Then, according to (4.66),

$$\Omega(P_x) \rightarrow U(x) \Omega(P_x) U^{-1}(x_0) = \Omega^{-1}(P_x) \Omega(P_x) = 1, \quad (4.71)$$

which implies $A_a^0(x) = 0$. \blacksquare

⁴ C. N. Yang, *Phys. Rev. Lett.* **33**, 445 (1974).

⁵ T. T. Wu and C. N. Yang, *Phys. Rev. D* **12**, 3845 (1975).

4.5 Canonical Formalism

1 Equations of Motion

We consider the classical Lagrangian density

$$\begin{aligned}\mathcal{L} = & -\frac{1}{4}F_a^{\mu\nu}F_{a\mu\nu} + (D^\mu\phi)^*(D_\mu\phi) - V(\phi) \\ & + \bar{\psi}(i\gamma_\mu D^\mu - m)\psi,\end{aligned}\quad (4.72)$$

where ϕ is a set of complex boson fields, ψ is a set of fermion fields, and m is a constant mass matrix. The fields form sets of irreducible representations of the gauge group G . The boson self-interacting term is gauge-invariant:

$$V(\phi) = V(U\phi). \quad (4.73)$$

For renormalizability of the quantized theory, $V(\phi)$ must be a polynomial in ϕ and ϕ^* of degree no greater than 4. The same consideration rules out fermion self-interactions. Apart from $V(\phi)$, all interactions are mediated by the gauge fields, A_a^μ , which are coupled to ϕ and ψ solely through $D^\mu\phi$ and $D^\mu\psi$. We confine our discussions here to the classical theory, leaving quantization to later chapters.

The fields ϕ , ϕ^* can be replaced by the equivalent set $\text{Re } \phi$, $\text{Im } \phi$. The only case not covered by (4.72) is that of N real fields ϕ_a transforming according to the adjoint representation, whose free Lagrangian density $\frac{1}{2}D^\mu\phi_a D^\nu\phi_a$ can be added on if desired. We also leave out possible Yukawa couplings of the general form $\bar{\psi}_i\phi_j\psi_k$, where i, j, k are indices labelling field components. The allowed combinations of i, j, k depend on the group G , and on the representations to which the fields belong. These Yukawa couplings can contribute to the fermion mass, as we shall see in Chapter 6.

It should be noted that \mathcal{L} may be invariant under a global group larger than G , but that only G is gauged. For example, with $G = SU(2)$, take the boson fields to consist of a triplet $\{\phi_1, \phi_2, \phi_3\}$ and a doublet $\{K_1, K_2\}$, and take

$$V = a(\phi_1^2 + \phi_2^2 + \phi_3^2) + b(K_1^*K_1 + K_2^*K_2).$$

This is invariant under independent transformations of ϕ and K under $SU(2)$. Hence, the global symmetry group is $SU(2) \times SU(2)$; but the gauged symmetry is the $SU(2)$ group of simultaneous transformations of ϕ and K .

Table 4.3 FIELDS AND CANONICAL CONJUGATES

	<i>Field</i>	<i>Canonical Conjugate</i>
gauge fields	$A_a^\nu(x)$	$-F_a^{\nu 0}(x) = \begin{cases} -E_a^k(x) & (\nu = k = 1, 2, 3) \\ 0 & (\nu = 0) \end{cases}$
scalar fields	$\phi(x)$ $\phi^*(x)$	$[D^0\phi(x)]^* = \pi(x)$ $D^0\phi(x) = \pi^*(x)$
spinor fields	$\psi(x)$	$i\psi^\dagger(x)$

The independent fields and their canonical conjugates for our system are listed in Table 4.3. The equations of motion are

$$D_\mu D^\mu \phi = -\frac{\partial V}{\partial \phi^*}, \quad (D_\mu D^\mu \phi)^* = -\frac{\partial V}{\partial \phi}, \quad (4.74)$$

$$(i\gamma^\mu D_\mu - m)\psi = 0, \quad i(D_\mu \bar{\psi})\gamma^\mu + m\bar{\psi} = 0, \quad (4.75)$$

$$\partial_\mu F_a^{\mu\nu} - gC_{abc}A_{b\mu}F_c^{\mu\nu} = j_a^\nu, \quad (4.76)$$

where j_a^ν is the matter-field current:

$$\begin{aligned} j_a^\nu &= -ig[(D^\nu \phi)^* L_a \phi - \phi^* L_a (D^\nu \phi)] + (\bar{\psi} \gamma^\nu L_a \psi) \\ &= -ig(\phi^* \overset{\leftrightarrow}{\partial}^\nu L_a \phi) - g^2 A_b^\nu \phi^* \{L_a, L_b\} \phi + (\bar{\psi} \gamma^\nu L_a \psi). \end{aligned} \quad (4.77)$$

The current j_a^ν is gauge-covariant but not conserved. On the other hand, $j_a^\nu + gC_{abc}A_{b\mu}F_c^{\mu\nu}$ is conserved but not gauge-covariant.

We can rewrite (4.76) in the form

$$\partial_\mu F^{\mu\nu} + ig[A_\mu, F^{\mu\nu}] = j^\nu, \quad (4.78)$$

or

$$D_\mu F^{\mu\nu} = j^\nu, \quad (4.79)$$

where D^μ is defined in (4.60), and

$$j^\nu = j_a^\nu L_a. \quad (4.80)$$

We concentrate on the gauge fields, since the behavior of the matter fields is familiar, except for possible spontaneous symmetry breaking due to $V(\phi)$, which we discuss separately later.

The three-vector form of (4.76) is

$$\nabla \cdot \mathbf{E}_a + gC_{abc}\mathbf{A}_b \cdot \mathbf{E}_c = j_a^0, \quad (4.81)$$

$$\nabla \times \mathbf{B}_a - \frac{\partial \mathbf{E}_a}{\partial t} + gC_{abc}(A_b^0 \mathbf{E}_c + \mathbf{A}_b \times \mathbf{B}_c) = \mathbf{j}_a, \quad (4.82)$$

where \mathbf{E}_a and \mathbf{B}_a are defined in (4.53). These are generalizations of Maxwell's equations, but they cannot be expressed solely in terms of \mathbf{E}_a and \mathbf{B}_a .

Equation (4.81), which is the generalization of Gauss' Law, does not involve $\partial A_a^0 / \partial t$. Thus we can use it to eliminate A_a^0 in terms of the other fields. That this should be possible is indicated by the fact that the canonical conjugate of A_a^0 is identically zero (Table 4.3). It is also evident from the fact that A_a^0 can always be reduced to zero by means of a continuous gauge transformation (Sec. 4.4). If one imposes some initial gauge condition, then (4.81) is a constraint on the initial values of A_a^0 , so that the gauge condition is respected by the dynamics. In terms of A_a^0 , (4.81) can be rewritten in the form

$$\nabla^2 A_a^0 + M_{ab} A_b^0 + N_a = -\frac{\partial}{\partial t}(\nabla \cdot \mathbf{A}_a) + gC_{abc}(\nabla \cdot \mathbf{A}_b) A_c^0, \quad (4.83)$$

where

$$\begin{aligned} M_{ab} &= 2gC_{abc}\mathbf{A}_c \cdot \nabla + g^2C_{amn}C_{nlb}\mathbf{A}_m \cdot \mathbf{A}_l, \\ N_a &= gC_{abc}\mathbf{A}_b \cdot \frac{\partial \mathbf{A}_c}{\partial t} - j_a^0. \end{aligned} \quad (4.84)$$

If we choose the Coulomb gauge $\nabla \cdot \mathbf{A}_a = 0$ at a particular time, then the condition

$$\nabla^2 A_a^0 + M_{ab}A_b^0 + N_a = 0 \quad (4.85)$$

ensures $\partial(\nabla \cdot \mathbf{A}_a)/\partial t = 0$, so that the Coulomb gauge will be maintained at all times. In the $U(1)$ case, the analogue of (4.85) is $\nabla^2 A^0 = -j^0$, whose general solution leads to the Coulomb potential. In the present case, it is coupled to (4.82), because M_{ab} and N_a depend on \mathbf{A}_a . In the quantum theory, gauge-fixing presents problems, which we shall discuss in Chapter 8.

2 Hamiltonian

The Hamiltonian density can be obtained by the standard recipe $\mathcal{H} = p\dot{q} - \mathcal{L}$, where q is an independent field, and p its canonical conjugate:

$$\begin{aligned} \mathcal{H} &= \frac{1}{2}(\mathbf{B}_a \cdot \mathbf{B}_a - \mathbf{E}_a \cdot \mathbf{E}_a) - \mathbf{E}_a \cdot \dot{\mathbf{A}}_a + \pi\dot{\phi} + \pi^*\dot{\phi}^* \\ &\quad - (D^\mu\phi)^*(D_\mu\phi) + V(\phi) + \bar{\psi}\left(\frac{1}{i}\boldsymbol{\alpha} \cdot \mathbf{D} + m\beta\right)\psi. \end{aligned} \quad (4.86)$$

The second term can be re-expressed as follows:

$$\begin{aligned} -\mathbf{E}_a \cdot \dot{\mathbf{A}}_a &= \mathbf{E}_a \cdot (\mathbf{E}_a + \nabla A_a^0 + gC_{abc}\mathbf{A}_b A_c^0) \\ &= \mathbf{E}_a \cdot \mathbf{E}_a + \nabla \cdot (\mathbf{E}_a A_a^0) - (\nabla \cdot \mathbf{E}_a)A_a^0 \\ &\quad + gC_{abc}(\mathbf{E}_a \cdot \mathbf{A}_b)A_c^0. \end{aligned} \quad (4.87)$$

The next term can be re-expressed as follows:

$$\begin{aligned} \pi\dot{\phi} + \pi^*\dot{\phi}^* - (D^\mu\phi)^*(D_\mu\phi) &= \pi(\pi^* - igA_0\phi) \\ &\quad + \pi^*(\pi + igA_0\phi^*) - \pi^*\pi + (\mathbf{D}\phi)^* \cdot (\mathbf{D}\phi) \\ &= \pi^*\pi + (\mathbf{D}\phi)^* \cdot (\mathbf{D}\phi) + ig(\pi^*A_0\phi^* - \pi A_0\phi). \end{aligned} \quad (4.88)$$

Hence,

$$\begin{aligned} \mathcal{H} &= \frac{1}{2}(\mathbf{B}_a \cdot \mathbf{B}_a + \mathbf{E}_a \cdot \mathbf{E}_a) + \pi^*\pi + (\mathbf{D}\phi)^* \cdot (\mathbf{D}\phi) \\ &\quad + V(\phi) + \bar{\psi}\left(\frac{1}{i}\boldsymbol{\alpha} \cdot \mathbf{D} + m\beta\right)\psi + X, \end{aligned} \quad (4.89)$$

where

$$X = \nabla \cdot (\mathbf{E}_a A_a^0) - (\nabla \cdot \mathbf{E}_a)A_a^0 + gC_{abc}(\mathbf{E}_a \cdot \mathbf{A}_b)A_c^0 + ig(\pi^*A_0\phi^* - \pi A_0\phi). \quad (4.90)$$

Using (4.81) and (4.77), and choosing the gauge that reduces to $A_a^0 = 0$ in the

absence of matter fields, we find that all the terms after the first cancel one another. Therefore

$$X = \nabla \cdot (\mathbf{E}_a A_a^0), \quad (4.91)$$

which we can ignore upon integration over all space. Hence, the total Hamiltonian, which is the total energy, is given by

$$\begin{aligned} H = & \int d^3x \left[\frac{1}{2} (\mathbf{B}_a \cdot \mathbf{B}_a + \mathbf{E}_a \cdot \mathbf{E}_a) + \pi^* \pi + (\mathbf{D}\phi)^* \cdot (\mathbf{D}\phi) \right. \\ & \left. + V(\phi) + \bar{\psi} \left(\frac{1}{i} \boldsymbol{\alpha} \cdot \mathbf{D} + m\beta \right) \psi \right]. \end{aligned} \quad (4.92)$$

The independent gauge fields are \mathbf{A}_a , with canonical conjugates \mathbf{E}_a , and \mathbf{B}_a is a function of \mathbf{A}_a given by (4.53). All dependencies on A_a^0 have been eliminated. The Hamiltonian is clearly locally gauge invariant. Classically, one can fix the gauge completely so that for each a , only two of the three fields \mathbf{A}_a are independent. The form of H after gauge fixing will depend on the gauge choice. Gauge-fixing in the quantum theory will be discussed in Chapter 8.

4.6 Spontaneous Symmetry Breaking

1 The Little Group

Spontaneous symmetry breaking in the present model means that some spin 0 fields (Higgs fields) have non-vanishing vacuum values; how this comes about depends entirely on the form of $V(\phi)$. We assume that higher-spin fields have zero vacuum values, for otherwise Lorentz invariance would be spontaneously broken, in apparent contradiction to experimental evidence. As we mentioned in Chapter 3, the treatment of spontaneous symmetry in terms of a Higgs field provides us with a convenient mathematical description; but in reality the Higgs fields might turn out to be merely phenomenological devices. Here, we treat the problem classically, leaving a discussion of quantum corrections to Chapter 10.

Let the Higgs fields be denoted collectively by ϕ , and suppose that $V(\phi)$ has a lowest minimum at $\phi = \rho$, with the minimum value taken to be zero:

$$\begin{aligned} V(\rho) &= 0, \\ V'(\rho) &= 0, \\ V''(\rho) &> 0. \end{aligned} \quad (4.93)$$

Spontaneous symmetry breaking occurs if $\rho \neq 0$. In that event, a vacuum solution (lowest-energy solution) is

$$\begin{aligned} \phi(x) &= \rho, \\ A_a^\mu(x) &= 0, \end{aligned} \quad (4.94)$$

for it clearly satisfies the equations of motion, and has the lowest possible energy.

Just as in the $U(1)$ case, ρ is not unique, because $U(x)\rho$ is equivalent to ρ . But, in contrast to the $U(1)$ case, not all $U(x)\rho$ are independent of one another here. To emphasize this point, let us look at an example.

Take $G = SU(2)$, and $\phi = (\phi_1, \phi_2, \phi_3)$ transforming according to the adjoint representation. Take

$$V(\phi) = \lambda(\phi_1^2 + \phi_2^2 + \phi_3^2 - a^2)^2.$$

Then, $\rho = (\rho_1, \rho_2, \rho_3)$ is any 3-vector satisfying

$$\rho_1^2 + \rho_2^2 + \rho_3^2 = a^2.$$

That is, ρ is a 3-vector whose tip lies on a sphere of radius a . Therefore, $U\rho = \rho$, for any rotation U of the sphere about ρ . Choose the x_3 axis to lie along ρ . Then the most general rotation that leaves ρ invariant is $[\exp(-i\omega L_3)]\rho = \rho$, or $L_3\rho = 0$. Thus, the vacuum solution is invariant under the $U(1)$ subgroup of $SU(2)$ generated by L_3 . In this case, the $SU(2)$ symmetry is said to be spontaneously broken down to $U(1)$.

Returning to the general case, let us specify that ρ is constant (i.e., all components of ρ are independent of x). This is possible because $V(\rho)$ has no explicit x dependence. Still, ρ is not unique, because $U\rho$ will serve equally well as ρ , where U is a global gauge transformation. The set of all elements U of G that leaves ρ invariant forms a subgroup of G (obvious). Since $\rho \neq 0$ by assumption, this subgroup is not G itself, but a proper subgroup. The largest subgroup H that leaves ρ invariant is called the *little group* with respect to ρ . The Lie algebra of the little group consists of a subset $\{l_\alpha\}$ of the Lie algebra $\{L_\alpha\}$ of G , with

$$[l_\alpha, l_\beta] = iC_{\alpha\beta\gamma}l_\gamma. \quad (4.95)$$

Under an infinitesimal element of H ,

$$\delta\rho = -i\omega_\alpha l_\alpha \rho = 0. \quad (4.96)$$

Since $\{\omega_\alpha\}$ is arbitrary, we have

$$l_\alpha \rho = 0. \quad (4.97)$$

We say that the symmetry G is spontaneously broken down to H .

The elements of G fall into equivalence classes that are the distinct cosets of G with respect to H , namely H, U_1H, U_2H, \dots , where U_iH is the coset $\{U_iu \mid U_i \in G, u \in H\}$. The independent vectors among $U\rho$ are $\rho, C_1\rho, C_2\rho, \dots$, where $C_i \in U_iH$. The collection of all cosets, denoted by G/H , is in general not a group. It is a group if and only if H is a normal subgroup of G (i.e., $UH = HU$). In this case, G/H is called the factor group, with the multiplication rule $(U_1H)(U_2H) = (U_1U_2)H$.

The generators of G not in the set $\{l_\alpha\}$ cannot annihilate ρ , by definition. Thus, we can divide $\{L_\alpha\}$ into two disjoint subsets:

$$\begin{aligned} \{L_\alpha\} &= \{L_j, l_\alpha\}, \\ L_j\rho &\neq 0 \quad (j = 1, \dots, K), \\ l_\alpha\rho &= 0 \quad (\alpha = 1, \dots, N-K). \end{aligned} \quad (4.98)$$

The choice of $\{L_j\}$ and $\{l_\alpha\}$ depends on the particular gauge for ρ , and hence is non-unique; but the number of generators in each set is gauge-invariant, because

$$l_\alpha \rho = 0 \Rightarrow [U(x)l_\alpha U^{-1}(x)] U(x)\rho = 0. \quad (4.99)$$

The generators $\{L_j\}$ ($j = 1, \dots, k$) generate a group if and only if $\{l_\alpha\}$ is a normal subalgebra.

To proceed further, it will be convenient to represent the generators by real antisymmetric matrices in accordance with (4.18):

$$\begin{aligned} T_j &= -iL_j \quad (j = 1, \dots, K), \\ t_\alpha &= -il_\alpha \quad (\alpha = 1, \dots, N - K). \end{aligned} \quad (4.100)$$

We also take ρ to have R real components, so that the representational vector space is a real vector space of dimensionality R . Scalar products in this space are denoted by

$$(f, Og) = \sum_{n=1}^R \sum_{m=1}^R f_n O_{nm} g_m. \quad (4.101)$$

The real symmetric matrix $(T_i\rho, T_j\rho)$ has positive-definite eigenvalues. Therefore, $T_i\rho$ ($i = 1, \dots, K$) are independent vectors that span a K -dimensional subspace of the R -dimensional representational vector space. A necessary condition for the little group H to be non-empty is therefore

$$R - K > 0. \quad (4.102)$$

We call the K -dimensional space spanned by $T_i\rho$ ($i = 1, \dots, K$), the *Goldstone space*, and its complement, of dimensionality $R - K$, the *Higgs space*.

For any vector ϕ in the representational vector space, and for compact gauge group G , there is always a gauge transformation U_0 such that $U_0\phi$ is orthogonal to the Goldstone space:

$$(T_j\rho, U_0\phi) = 0 \quad (j = 1, \dots, K). \quad (4.103)$$

We say that $U_0\phi$ is in *unitary gauge*. That U_0 exists can be shown as follows.⁸

For given ρ and ϕ , consider a mapping of G to the real line defined by

$$f(U) = (\rho, U\phi).$$

Since G is compact, the values of $f(U)$ lie in a real compact set. Hence $f(U)$ has extrema. Let $f(U_0)$ be an extremum. A small variation of U near U_0 gives

$$\delta f = f(U_0 + \delta U) - f(U_0) = 0.$$

Now any change in a group element U can be written as a left multiplication by another group element. Hence we can write

$$\delta U = \omega_a T_a U_0,$$

where ω_a is arbitrary. Therefore,

$$0 = \delta f = (\rho, \omega_a T_a U_0 \phi) = \omega_j (\rho, T_j U_0 \phi) = -\omega_j (T_j \rho, U_0 \phi).$$

Since ω_j is arbitrary, we have $(T_j \rho, U_0 \phi) = 0$. ■

⁸ S. Weinberg, *Phys. Rev. D7*, 1068 (1973).

The above result holds at each space-time point. If $\phi(x)$ is a solution, it must be a continuous function of x . Therefore, there exists $U_0(x)$, continuous in x , such that $U_0(x)\phi(x)$ is in unitary gauge:

$$\phi(x) \rightarrow U_0(x)\phi(x),$$

$$U_0(x)\phi(x) = \begin{pmatrix} 0 \\ \tilde{\phi}(x) \end{pmatrix} \begin{array}{l} \text{Goldstone space, } K\text{-dimensional} \\ \text{Higgs space, } (R - K)\text{-dimensional.} \end{array} \quad (4.103)$$

2 Higgs Mechanism

In unitary gauge, the vacuum solution is

$$\begin{aligned} \phi(x) &= \rho = \begin{pmatrix} 0 \\ \tilde{\rho} \end{pmatrix}, \\ A_a^\mu(x) &= 0. \end{aligned} \quad (4.104)$$

Solutions with energy near the vacuum solution are of the form

$$\begin{aligned} \phi(x) &= \begin{pmatrix} 0 \\ \tilde{\rho} + \eta(x) \end{pmatrix}, \\ A_a^\mu(x) &\text{ small,} \end{aligned} \quad (4.105)$$

where $\eta(x)$ and $A_a^\mu(x)$ are both small because the energy is a continuous functional of η and A_a^μ .

To first order in these small quantities, we have

$$\begin{aligned} V(\phi) &= \frac{1}{2}(\eta, V''(\rho)\eta), \\ j_a^\mu &= -g^2(T_a\rho, T_b\rho)A_b^\mu. \end{aligned} \quad (4.106)$$

We define the following matrices, which will turn out to be mass matrices:

$$\begin{aligned} (\mu^2)_{rs} &= \begin{array}{|c|c|} \hline 0 & 0 \\ \hline 0 & V''(\rho) \\ \hline \end{array} & \begin{array}{l} \text{Goldstone space} \\ \text{Higgs space} \end{array} \\ (M^2)_{ab} &= g^2(T_a\rho, T_b\rho) = \begin{array}{|c|c|} \hline (M^2)_{ij} & 0 \\ \hline 0 & 0 \\ \hline \end{array} & \begin{array}{l} \text{Goldstone space} \\ \text{Higgs space} \end{array} \end{aligned} \quad (4.107)$$

Using these, the linearized equations can be written in the form

$$\begin{aligned} \square^2\eta_r + (\mu^2)_{rs}\eta_s &= 0 \quad (r = 1, \dots, R - K), \\ \square^2A_i^\nu + (M^2)_{ij}A_j^\nu &= 0 \quad (\partial_\mu A_i^\mu = 0) \quad (i = 1, \dots, K), \\ \square^2A_\alpha^\nu - \partial^\nu(\partial_\mu A_\alpha^\mu) &= 0 \quad (\alpha = 1, \dots, N - K). \end{aligned} \quad (4.108)$$

These respectively describe massive Higgs bosons, massive vector bosons, and massless vector bosons, as summarized in Table 4.4. The massless vector bosons are the gauge particles associated with the unbroken symmetry H .

The number of massive vector bosons is equal to the number of Goldstone bosons that would be present if there were no gauge coupling. As it is, there are

no Goldstone bosons; they have been “eaten up” by the massive vector bosons.

If the field ϕ is complex, and we choose not to put it in real form, then the vector meson mass matrix could be represented in the alternative form

$$(M^2)_{ab} = \frac{1}{2}g^2\rho^\dagger\{L_a, L_b\}\rho. \quad (4.109)$$

Examples

We illustrate the group-theoretic aspects of spontaneous symmetry breaking by two examples.

(i) First, consider $G = O(n)$, $\phi = \{\phi_1, \dots, \phi_n\}$ (fundamental representation), and

$$V(\phi) = \lambda[(\phi_1^2 + \dots + \phi_n^2) - a^2]^2.$$

We can choose $\rho = \{0, \dots, 0, a\}$. Clearly $H = O(n-1)$. The group $O(n)$ has $\frac{1}{2}n(n-1)$ generators, and $O(n-1)$ has $\frac{1}{2}(n-1)(n-2)$ generators. Therefore:

$$\text{no. of massless vector bosons} = \frac{1}{2}(n-1)(n-2),$$

$$\text{no. of massive vector bosons} = \frac{1}{2}n(n-1) - \frac{1}{2}(n-1)(n-2) = n-1,$$

$$\text{no. of Higgs bosons} = 1.$$

(ii) As a second example, consider $G = SU(2)$, with $N = 3$. Suppose the scalar field ϕ consists of a real triplet π and a complex doublet K :

$$\phi = \begin{pmatrix} \pi \\ K \end{pmatrix}, \quad \pi = \begin{pmatrix} \pi_1 \\ \pi_2 \\ \pi_3 \end{pmatrix}, \quad K = \begin{pmatrix} K_1 \\ K_2 \end{pmatrix}.$$

Choose

$$V = \lambda(\pi_1^2 + \pi_2^2 + \pi_3^2 - a^2)^2 + \lambda'(K_1^*K_1 + K_2^*K_2 - b^2)^2.$$

Table 4.4 HIGGS MECHANISM

G is spontaneously broken down to H

N = No. of generators of G

$N - K$ = No. of generators of H

R = Dimensionality of real representation of G

Field	No. of Fields	No. of indep. Components
η_r (Higgs, massive)	$R - K$	$R - K$
A_i^{μ} (gauge, massive)	K	$3K$
A_a^{μ} (gauge, massless)	$N - K$	$2(N - K)$
	Total	$N + R$

Then,

$$\rho = \begin{pmatrix} \phi_0 \\ K_0 \end{pmatrix}, \quad \phi_0 = \begin{pmatrix} 0 \\ 0 \\ a \end{pmatrix}, \quad K_0 = K_0^* = \begin{pmatrix} 0 \\ b \end{pmatrix}.$$

The generators are represented by

$$L_a = \begin{pmatrix} I_a & 0 \\ 0 & \frac{1}{2}\tau_a \end{pmatrix},$$

where I_a are 3×3 matrices for the adjoint representation, and τ_a are 2×2 Pauli matrices. Thus,

$$L_a \rho = \begin{pmatrix} I_a \phi_0 \\ \frac{1}{2}\tau_a K_0 \end{pmatrix}.$$

We distinguish the two cases $b \neq 0$, and $b = 0$.

If $b \neq 0$, then $K_0 \neq 0$. In this case no generator can annihilate ρ , because no τ_a can annihilate K_0 . The proof is as follows:

Assume the contrary, e.g., $\tau_3 K_0 = 0$. Then $-i[\tau_1, \tau_2]K_0 = 0$, or $\tau_1 \tau_2 K_0 = \tau_2 \tau_3 K_0$. But $\tau_1 \tau_2 = -\tau_2 \tau_3$. Hence $K_0 = 0$ (contradiction). Therefore, in this case the symmetry is completely broken, i.e., H is empty, and there are:

- no massless vector bosons,
- three massive vector bosons,
- three Higgs bosons (one real and one complex field).

If $b = 0$, then $H = U(1)$.

CHAPTER 5

TOPOLOGICAL SOLITONS

5.1 Solitons

As we have mentioned in Chapter 3, a classical soliton solution (soliton) is a solution of the classical equations of motion whose energy density is non-zero only in a finite region of space. Its total energy is therefore finite. Quantization leads to a corresponding quantum soliton. We discuss solitons in Yang-Mills field theory and deal only with the classical case here.

A Yang-Mills field can be looked upon as a map from space-time to the gauge group, and such maps may fall into topologically non-equivalent classes. Fields belonging to different classes are not deformable into each other through continuous changes. A topological soliton is a solution in a class different from that of the classical vacuum. There are also non-topological solitons, which owes their stability not to topology but to conservation laws¹; but we shall not discuss them here.

Solitons can be static or time-dependent. A static soliton is one for which $A^\mu(x)$ can be made time-independent through a continuous gauge transformation. This definition is equivalent to the statement that the time evolution of $A^\mu(x)$ is a continuous gauge transformation, i.e.,

$$A^\mu(x, t + dt) = A^\mu(x, t) + \frac{1}{g} \partial^\mu d\omega(x, t) - i[d\omega(x, t), A^\mu(x, t)], \quad (5.1)$$

or

$$\dot{A}^\mu(x) = \frac{1}{g} \partial^\mu \dot{\omega}(x) - i[\dot{\omega}(x), A^\mu(x)], \quad (5.2)$$

where $\omega(x)$ is an arbitrary continuous function. To show the equivalence, we note that the solution to (5.2) can be made time-independent through a gauge transformation $-\omega(x)$, with $\omega(x)$ satisfying

$$\partial^\mu \dot{\omega}(x) = ig[\dot{\omega}(x), A^\mu(x)]. \quad (5.3)$$

The solution is

$$\dot{\omega}(x) = \Omega(P)\dot{\omega}(x_0)\Omega^{-1}(P), \quad (5.4)$$

where $\Omega(P)$ is defined in (4.64), with the path P beginning at x_0 and ending at x . From (5.4) we can obtain $\omega(x)$ by integrating over time. Although $\omega(x)$ depends

¹ T. D. Lee, *Phys. Reports* 23, 254 (1976).

on the choice of the path P , its time-dependence is independent of P , because the endpoints of P are fixed. The freedom in the choice of P merely reflects the fact that a static $A^\mu(x)$ still has the freedom of time-independent gauge transformations. The gauge in which a static solution is time-independent will be referred to as the *static gauge*.

A static solution of finite energy is either the vacuum or a static soliton. For a pure Yang-Mills field (i.e., one without coupling to matter fields), we can show the following property:²

There are no static solitons in pure Yang-Mills theory except in 4 spatial dimensions.

Proof: In $(n + 1)$ -dimensional Minkowski space, let the space-time index be $\mu = 0, 1, \dots, n$, and the spatial index be $k = 1, \dots, n$. Consider static solutions of finite energy other than the vacuum. The canonical energy-momentum tensor is given by

$$\theta^{\mu\nu} = F_a^{\mu\lambda} F_{a\lambda}^\nu + \frac{1}{4} g^{\mu\nu} F_a^{\alpha\beta} F_{a\alpha\beta}, \quad (\partial_\mu \theta^{\mu\nu} = 0).$$

This is gauge-invariant, and hence independent of x^0 . The following quantities will be relevant:

$$\begin{aligned}\theta_{\mu}^{\mu} &= \frac{1}{4}(n - 3)F_a^{\alpha\beta}F_{a\alpha\beta}, \\ \theta^{00} &= \frac{1}{2}F_a^{k0}F_a^{k0} + \frac{1}{4}F_a^{ij}F_a^{ij}, \\ \theta^{0k} &= F_a^{j0}F_a^{jk}.\end{aligned}$$

The requirement of finite energy means

$$\int d^n x (\frac{1}{2}F_a^{k0}F_a^{k0} + \frac{1}{4}F_a^{ij}F_a^{ij}) < \infty,$$

which leads to the condition

$$F_a^{\mu\nu} \xrightarrow[r \rightarrow \infty]{} O(r^{-n/2-\varepsilon}), \quad (\varepsilon > 0),$$

$$r^2 = \sum_{k=1}^n x_k^2.$$

We can show that $F_a^{k0} = 0$ for a static soliton as follows. In the static gauge, $\partial^0 A^k = 0$. Hence, by (4.42)

$$F^{k0} = \partial^k A^0 + ig[A^k, A^0].$$

The relevant equation of motion in (4.78) reads

$$\partial_k F^{k0} + ig[A_k, F^{k0}] = 0.$$

Hence,

$$\begin{aligned}A_0 \{\partial_k F^{k0} + ig[A_k, F^{k0}]\} &= 0, \\ \partial_k (A_0 F^{k0}) - (\partial_k A^0) F^{k0} + ig(A_0 A_k F^{k0} - A_0 F^{k0} A_k) &= 0,\end{aligned}$$

² S. Deser, *Phys. Lett.* **64**, B463 (1976).

$$\text{Tr} \int d^n x \{\partial_k A^0 + ig[A_k, A_0]\} F^{k0} = 0,$$

$$\text{Tr} \int d^n x F_{k0} F^{k0} = 0,$$

$$\int d^n x F_a^{k0} F_a^{k0} = 0.$$

Since $F_a^{k0} F_a^{k0} = |\mathbf{E}_a|^2$ is non-negative, it must be zero. Hence, $E_a^k = F_a^{k0} = 0$ in the static gauge. Since $F^{k0} \rightarrow UF^{k0}U^{-1}$ under a gauge transformation, we conclude that $F_a^{0k} = 0$ for a static solution in any gauge.

Now consider

$$\partial_j(x^k \theta^{jk}) = \theta^k{}_k + x^k \partial_j \theta^{jk},$$

$$\partial_\mu \theta^{\mu k} = 0 \Rightarrow \partial_0 \theta^{0k} + \partial_j \theta^{jk} = 0.$$

$$\therefore \partial_j(x^k \theta^{jk}) = \theta^k{}_k - x^k \partial_0 \theta^{0k}.$$

The last term vanishes for static solutions. Hence,

$$0 = \int d^n x \partial_j(x^k \theta^{jk}) = \int d^n x \theta^k{}_k.$$

Using $\theta^k{}_k = \theta^\mu{}_\mu - \theta^{00}$, we obtain

$$\int d^n x [(2-n)F_a^{k0} F_a^{k0} - \frac{1}{2}(4-n)F_a^{jk} F_a^{jk}] = 0.$$

As shown earlier, $F_a^{k0} = 0$. Hence

$$(4-n) \int d^n x F_a^{ij} F_a^{ij} = 0.$$

Therefore $F_a^{ij} = 0$ unless $n = 4$. ■

For $n = 4$, a pure Yang-Mills static soliton can be constructed explicitly. An example is the instanton discussed in Sec. 5.2. For $n \neq 4$, there can be static solitons only if there are matter fields. An example is the monopole in 3 spatial dimensions, described in Sec. 5.3.

Quantization of a soliton solution requires appropriate handling of the translational motion of the soliton as a whole, and is quite involved³. We will not discuss it here.

5.2 The Instanton

1 Topological Charge

As we have shown, a pure Yang-Mills theory can have static soliton solutions only in 4 Euclidean dimensions. We now construct an example, the instanton solution, which is characterized by a “topological charge”. One might wonder why a phenomenon that exists only in Euclidean 4-space should concern us. The

³ J. Goldstone and R. Jackiw, *Phys. Rev.* D11, 1486 (1975); N. H. Christ and T. D. Lee, *Phys. Rev.* D12, 1606 (1975).

answer lies in the fact that a quantum field theory in Minkowski 4-space can be described in terms of the classical action in Euclidean 4-space, as we shall show in Chapter 7. The instanton can have physical applications, as will be discussed in Chapter 8.

In this section, we consider $G = SU(2)$. A vector in Euclidean 4-space is denoted by $x^\mu (\mu = 1, 2, 3, 4)$. There is no distinction between upper and lower indices. We may look upon x^μ either as the spatial components of a vector in Minkowski 5-space with metric $(1, -1, -1, -1, -1)$, or as a 4-vector in the usual Minkowski 4-space with the time-component continued to imaginary values: $x^0 \rightarrow -ix^4$. The topological charge is defined by

$$q = \frac{g^2}{16\pi^2} \int d^4x \operatorname{Tr} \tilde{F}^{\mu\nu} F_{\mu\nu}. \quad (5.5)$$

where the factor $16\pi^2$ is the group volume of $SU(2)$. First we show that the integrand is a total 4-divergence:

$$\begin{aligned} \frac{1}{4} \operatorname{Tr} \tilde{F}^{\mu\nu} F_{\mu\nu} &= \partial_\mu X^\mu, \\ X^\mu &= \epsilon^{\mu\alpha\beta\gamma} \operatorname{Tr} \left[\frac{1}{2} A_\alpha \partial_\beta A_\gamma + \frac{i}{3} g A_\alpha A_\beta A_\gamma \right]. \end{aligned} \quad (5.6)$$

Proof:

$$\begin{aligned} \frac{1}{4} \operatorname{Tr} \tilde{F}^{\mu\nu} F_{\mu\nu} &= \frac{1}{2} \epsilon^{\mu\nu\alpha\beta} \operatorname{Tr} [[(\partial_\alpha + igA_\alpha) A_\beta] [(\partial_\mu + igA_\mu) A_\nu]] \\ &= \frac{1}{2} \epsilon^{\mu\nu\alpha\beta} \operatorname{Tr} [(\partial_\alpha A_\beta) (\partial_\mu A_\nu) + 2ig(\partial_\alpha A_\beta) A_\mu A_\nu - g^2 A_\alpha A_\beta A_\mu A_\nu]. \end{aligned}$$

The last term will not contribute because it is symmetric in α, β . Side calculations:

$$\begin{aligned} (\partial_\alpha A_\beta)(\partial_\mu A_\nu) &= \partial_\alpha (A_\beta \partial_\mu A_\nu) - A_\beta (\partial_\alpha \partial_\mu A_\nu), \\ \epsilon^{\mu\nu\alpha\beta} \operatorname{Tr} [(\partial_\alpha A_\beta) A_\mu A_\nu] &= \epsilon^{\mu\nu\alpha\beta} \operatorname{Tr} [\partial_\alpha (A_\beta A_\mu A_\nu) - (\partial_\alpha A_\mu) A_\nu A_\beta - (\partial_\alpha A_\nu) A_\beta A_\mu]. \end{aligned}$$

The last two terms are the same, and equal to the left-hand side. Hence,

$$\epsilon^{\mu\nu\alpha\beta} \operatorname{Tr} [(\partial_\alpha A_\beta) A_\mu A_\nu] = \frac{1}{3} \epsilon^{\mu\nu\alpha\beta} \partial_\alpha \operatorname{Tr} (A_\beta A_\mu A_\nu).$$

Using the above, we obtain

$$\frac{1}{4} \operatorname{Tr} \tilde{F}^{\mu\nu} F_{\mu\nu} = \epsilon^{\mu\nu\alpha\beta} \partial_\alpha \operatorname{Tr} \left(\frac{1}{2} A_\beta \partial_\mu A_\nu + \frac{1}{3} ig A_\beta A_\mu A_\nu \right). \blacksquare$$

Note that the proof goes through for Euclidean as well as Minkowski 4-space. Using (5.6) we can write

$$q = \frac{g^2}{4\pi^2} \int d^4x \partial_\mu X^\mu = \frac{g^2}{4\pi^2} \int dS_\mu X^\mu, \quad (5.7)$$

where dS_μ is an element of a 3-dimensional spherical hyper-surface S^3 , with radius $R \rightarrow \infty$.

Now impose the condition of finite energy:

$$\begin{aligned} F^{\mu\nu} &\xrightarrow[x \rightarrow \infty]{} O(x^{-3}), \quad x^2 \equiv x_4^2 + |\mathbf{x}|^2, \\ A^\mu &\xrightarrow[x \rightarrow \infty]{} -\frac{i}{g} U \partial^\mu U^{-1} + O(x^{-2}), \end{aligned} \quad (5.8)$$

where $U \in SU(2)$. Let

$$\lambda^\mu \equiv U \partial^\mu U^{-1}. \quad (5.9)$$

Then $A^\mu \xrightarrow[x \rightarrow \infty]{} -(i/g)\lambda^\mu$. Using this form in (5.6) and (5.7) gives

$$\begin{aligned} X^\mu &= \frac{1}{6g^2} \epsilon^{\mu\alpha\beta\gamma} \text{Tr}(\lambda^\alpha \lambda^\beta \lambda^\gamma), \\ q[U] &= \frac{1}{24\pi^2} \int_{S^3} dS_\mu \epsilon^{\mu\alpha\beta\gamma} \text{Tr}(\lambda^\alpha \lambda^\beta \lambda^\gamma), \end{aligned} \quad (5.10)$$

where we regard $q[U]$ as a functional of $U(x)$.

The integrand in (5.10) depends on $U(x) \in SU(2)$, where $x \in S^3$, and hence represents a map $S^3 \rightarrow SU(2)$. Since $SU(2)$ has the topology of S^3 [cf. (4.22)], this is a map $S^3 \rightarrow S^3$, which falls into homotopy classes labeled by a winding number. We shall show that q is none other than the winding number of this map.

First we digress on group integration over $SU(2)$.⁴ An element of $SU(2)$ can always be parametrized by the Euler angles α, β, γ :

$$U = \exp\left(\frac{i\gamma\tau_3}{2}\right) \exp\left(\frac{i\beta\tau_1}{2}\right) \exp\left(\frac{i\alpha\tau_3}{2}\right),$$

$$0 \leq \alpha \leq 2\pi, \quad 0 \leq \beta \leq \pi, \quad 0 \leq \gamma \leq 4\pi,$$

where τ_i are the Pauli matrices. The volume element in parameter space is

$$d\mu(U) = \frac{1}{16\pi^2} \sin\beta \, d\alpha \, d\beta \, d\gamma,$$

which has been normalized to $\int d\mu = 1$. By explicit computation, it can be verified that

$$d\mu(U) = \frac{1}{4\pi^2} \text{Tr} \left(U \frac{\partial U^{-1}}{\partial \alpha} U \frac{\partial U^{-1}}{\partial \beta} U \frac{\partial U^{-1}}{\partial \gamma} \right) d\alpha d\beta d\gamma, \quad (5.12)$$

where the trace is antisymmetric in $\{\alpha, \beta, \gamma\}$. This is invariant under a change of parametrization, and has the property $\int d\mu(U) = \int d\mu(U_0 U)$, where the

⁴ See W. K. Tung, *Group Theory in Physics* (World Scientific, Singapore, 1985), Sec. 8.2.

integral extends over $U \in SU(2)$, and U_0 is any fixed element of $SU(2)$. Denoting the parameters more generally by ξ_1 , ξ_2 , ξ_3 , and noting the antisymmetry property mentioned, we can write

$$d\mu(U) = \frac{1}{24\pi^2} \epsilon^{ijk} \text{Tr} \left(U \frac{\partial U^{-1}}{\partial \xi_i} U \frac{\partial U^{-1}}{\partial \xi_j} U \frac{\partial U^{-1}}{\partial \xi_k} \right) d\xi_1 d\xi_2 d\xi_3, \quad (5.13)$$

Consider now the integrand in (5.10) at $x \in S^3$. We can set up local Cartesian coordinates x_1 , x_2 , x_3 and write

$$q[U] = \frac{1}{24\pi^2} \int dx_1 dx_2 dx_3 I(x_1, x_2, x_3),$$

$$I(x_1, x_2, x_3) = \epsilon^{ijk} \text{Tr} \left(U \frac{\partial U^{-1}}{\partial x_i} U \frac{\partial U^{-1}}{\partial x_j} U \frac{\partial U^{-1}}{\partial x_k} \right).$$

Changing variables to the group parameters ξ_1 , ξ_2 , ξ_3 , we have

$$I(x_1, x_2, x_3) = \text{Tr} \left(U \frac{\partial U^{-1}}{\partial \xi_a} U \frac{\partial U^{-1}}{\partial \xi_b} U \frac{\partial U^{-1}}{\partial \xi_c} \right) \epsilon^{ijk} \frac{\partial \xi_a}{\partial x_i} \frac{\partial \xi_b}{\partial x_j} \frac{\partial \xi_c}{\partial x_k}$$

$$= \text{Tr} \left(U \frac{\partial U^{-1}}{\partial \xi_a} U \frac{\partial U^{-1}}{\partial \xi_b} U \frac{\partial U^{-1}}{\partial \xi_c} \right) \frac{\partial(\xi_a \xi_b \xi_c)}{\partial(x_1 x_2 x_3)}$$

$$= 6 \text{Tr} \left(U \frac{\partial U^{-1}}{\partial \xi_1} U \frac{\partial U^{-1}}{\partial \xi_2} U \frac{\partial U^{-1}}{\partial \xi_3} \right) \frac{\partial(\xi_1 \xi_2 \xi_3)}{\partial(x_1 x_2 x_3)}.$$

Thus,

$$q[U] = \int_{x \in S^3} d\mu(U(x)), \quad (5.14)$$

where the integral extends over all $x \in S^3$. Therefore $q[U]$ is the net number of times the group manifold is covered, when x ranges over S^3 , i.e., the winding number of the map $S^3 \rightarrow SU(2)$.

An example of a map with $q = 0$ is obviously $U(x) = 1$. For a map with $q = 1$, we can take $U(x) = u(x)$, with

$$u(x) = \frac{1}{x} (x_4 + \mathbf{x} \cdot \boldsymbol{\tau}), \quad \left(x^2 = \sum_{i=1}^4 x_i^2 \right). \quad (5.15)$$

A map with $q = n$ is then given by

$$U(x) = [u(x)]^n \quad (n = 0, \pm 1, \pm 2, \dots). \quad (5.16)$$

To demonstrate this, let us parameterize $x \in S^3$ using polar coordinates:

$$\frac{x^\mu}{x} = (\cos \omega, \hat{n} \sin \omega), \quad \hat{n} = (\cos \theta, \sin \theta \cos \phi, \sin \theta \sin \phi) \\ (0 \leq \omega < \pi, 0 \leq \theta < \pi, 0 \leq \phi < 2\pi). \quad (5.17)$$

We can then write

$$u(x) = e^{-\omega \hat{n} \cdot \tau}, \quad (5.18)$$

which represents a rotation around \hat{n} through angle 2ω . When x ranges over S^3 once, a 2π rotation is made about each possible \hat{n} . The statement (5.16) now follows directly from the fact that $[u(x)]^n$ is obtainable from $u(x)$ through the replacement $\omega \rightarrow n\omega$.

2 Explicit Solution⁵

We now construct an explicit solution with $q = 1$. Take $L_a = \frac{1}{2}\tau_a$, and use the shorthand $F^2 = F^{\mu\nu}F_{\mu\nu}$, $\tilde{F} \cdot F = \tilde{F}^{\mu\nu}F_{\mu\nu}$. We note that $\tilde{F}^2 = F^2$. The Lagrangian density and the Euclidean action are given respectively by

$$\mathcal{L} = -\frac{1}{2}\text{Tr } F^2, \quad (5.19) \\ S = -\frac{1}{2} \int d^4x \text{Tr } F^2.$$

A solution corresponds to an extremum of S . Consider the inequality

$$\int d^4x \text{Tr}(F \pm \tilde{F})^2 \geq 0,$$

or

$$\int d^4x (F^2 \pm \tilde{F} \cdot F) \geq 0.$$

We choose the + sign if $\text{Tr } \tilde{F} \cdot F > 0$, and the - sign if $\text{Tr } \tilde{F} \cdot F < 0$. Hence

$$\int d^4x \text{Tr } F^2 \geq \left| \int d^4x \text{Tr } \tilde{F} \cdot F \right| \\ \text{or,} \\ S \leq -\frac{1}{2} \left| \int d^4x \text{Tr } \tilde{F} \cdot F \right|. \quad (5.20)$$

The equality holds if $F = \pm \tilde{F}$, and when this is true we say that F is self-dual (anti-self-dual). Since $F = \pm \tilde{F}$ corresponds to an extremum S , any self-dual or anti-self-dual F is a solution. An instanton solution is a self-dual F having $q = 1$. Its contribution to the (Euclidean) action is

$$S_{\text{instanton}} = -8\pi^2/g^2. \quad (5.21)$$

⁵ A. A. Belavin, A. M. Polyakov, A. S. Schwartz, and Yu. S. Tyupkin, *Phys. Lett.* **59**, B85 (1975); G. 't Hooft, *Phys. Rev. Lett.* **37**, 8 (1976).

Let

$$\begin{aligned}\tau_\mu &= (\tau, i) \quad (\mu = 1, 2, 3, 4), \\ \tau_\mu^\dagger &= (\tau, -i), \\ \tau_{\mu\nu} &= i(\tau_\mu \tau_\nu^\dagger - \delta_{\mu\nu}) \\ \tilde{\tau}_{\mu\nu} &= \frac{1}{2} \epsilon_{\mu\nu\alpha\beta} \tau_{\alpha\beta} = \tau_{\mu\nu}.\end{aligned}\tag{5.22}$$

Note that $\tau_{\mu\nu}$ is self-dual. Using the above notation we can write

$$\begin{aligned}(x \cdot \tau)(x \cdot \tau^\dagger) &= x^2, \\ u &= i \frac{x \cdot \tau^\dagger}{x}, \\ u^{-1} &= -i \frac{x \cdot \tau}{x},\end{aligned}\tag{5.23}$$

$$A_\mu^{\text{pure-gauge}} = \frac{i}{g} u \partial^\mu u^{-1} = \frac{1}{g} \frac{\tau_{\mu\nu} x_\nu}{x^2}.$$

For a finite-energy solution with $q = 1$, put

$$\begin{aligned}A_\mu &= \frac{1}{g} \frac{\tau_{\mu\nu} x_\nu}{x^2} f(x^2), \\ f(0) &= 0 \quad (\text{regularity at } x = 0), \\ f(\infty) &= 1 \quad (\text{finite energy, } q = 1).\end{aligned}\tag{5.24}$$

To make this a solution, we only need to choose f such that $F_{\mu\nu} = \tilde{F}_{\mu\nu}$. From (5.24) we obtain

$$F_{\mu\nu} = \frac{2}{g} \left\{ \frac{f(1-f)}{r^2} \tau_{\mu\nu} + \left[f' - \frac{f(1-f)}{r^2} \right] (\tau_{\mu\lambda} x_\lambda x_\nu - \tau_{\nu\lambda} x_\lambda x_\mu) \right\},\tag{5.25}$$

$$\tilde{F}_{\mu\nu} = \frac{2}{g} \left\{ \frac{f(1-f)}{r^2} \tau_{\mu\nu} + \left[f' - \frac{f(1-f)}{r^2} \right] \epsilon_{\mu\nu\alpha\beta} (\tau_{\alpha\lambda} x_\lambda x_\beta - \tau_{\beta\lambda} x_\lambda x_\alpha) \right\}.\tag{5.26}$$

Thus, $F_{\mu\nu}$ is self-dual if the second term vanishes;

$$f' - \frac{f(1-f)}{x^2} = 0.\tag{5.27}$$

The general solution satisfying the required boundary conditions is

$$f(r^2) = \frac{x^2}{\rho^2 + x^2},\tag{5.28}$$

where ρ is an arbitrary scale parameter. With this, we have

$$F_{\mu\nu} = \tilde{F}_{\mu\nu} = \frac{2}{g} \frac{\rho^2}{(x^2 + \rho^2)^2} \tau_{\mu\nu},\tag{5.29}$$

which is the field tensor for an instanton with $q = 1$. An instanton with $q = -1$ is called an anti-instanton, and may be obtained from (5.29) by replacing $\tau_{\mu\nu}$ by $\bar{\tau}_{\mu\nu}$, with

$$\begin{aligned}\bar{\tau}_{ij} &= \tau_{ij}, \\ \bar{\tau}_{i4} &= -\tau_{i4}.\end{aligned}\quad (5.30)$$

The physical relevance of the instanton will be discussed in Chap. 8.

5.3 The Monopole

1 Topological Stability

There are no solitons in pure Yang-Mills theory in 3 spatial dimensions; but they can exist when matter fields are present. We now discuss such an example. Consider any simple gauge group G , and introduce scalar matter fields $\phi(x) = \{\phi_1(x), \phi_2(x), \dots\}$, which form a (generally reducible) representation of G . Using a real representation, we write the total energy as

$$\begin{aligned}\mathcal{E} = \int d^3x [\frac{1}{2}(\mathbf{B}_a \cdot \mathbf{B}_a + \mathbf{E}_a \cdot \mathbf{E}_a + \frac{1}{2} D^0 \phi D^0 \phi \\ + \frac{1}{2} D^k \phi D^k \phi + V(\phi))].\end{aligned}\quad (5.31)$$

We assume that the lowest minimum of $V(\phi)$ occurs at $\phi(x) = \rho(x)$:

$$V(\rho) = 0, \quad V'(\rho) = 0, \quad V''(\rho) > 0. \quad (5.32)$$

The conditions for a finite-energy solution are

- (a) $F^{\mu\nu}(x) \xrightarrow[r \rightarrow \infty]{} O(r^{-2}), \quad r = |\mathbf{x}|,$
 $A^\mu(x) \xrightarrow[r \rightarrow \infty]{} -\frac{i}{g} U \partial^\mu U^{-1} + O(r^{-1}),$
- (b) $D^\mu \phi(x) \xrightarrow[r \rightarrow \infty]{} O(r^{-2}),$
- (c) $\phi(x) \xrightarrow[r \rightarrow \infty]{} \rho(x) + O(r^{-2})$

In (b), the first term of A^μ is of pure-gauge form, but not the $O(r^{-1})$ term. It is the latter that will be of interest.

We begin by considering the implications of (c). As noted in Sec. 4.6, $\rho(x)$ is not unique. The independent ρ 's are of the form $U(x)\rho_0$, where ρ_0 is a constant satisfying (5.32), and $U(x)$ belongs to a coset in G/H (H is the little group with respect to ρ , the largest subgroup of G that leaves ρ invariant). Conversely, a given $\rho(x)$ can always be written in the form $U(x)\rho_0$. Since $\rho(x)$ is the value of $\phi(x)$ at a point on S^2 (a 3-sphere of radius $r \rightarrow \infty$), the function $\rho(x)$ is a mapping of S^2 to G/H :

$$\rho(x): S^2 \rightarrow G/H. \quad (5.34)$$

The identity map corresponds to the physical vacuum, by definition. A topological soliton can exist only if there exists a non-trivial map. The topology is then different from that of the physical vacuum, making the soliton stable.

The mappings (5.34) fall into homotopy classes which form a group, namely, the second homotopy group of G/H , denoted by $\pi_2(G/H)$. In general, the n th homotopy group $\pi_n(X)$ is the group of inequivalent mappings $S^n \rightarrow X$. If $\pi_2(G/H) = 0$ (i.e. it contains only the identity map and its continuous deformations), then no topological solitons exist. Some general properties of $\pi_2(G/H)$ are as follows⁶:

- (a) For simply-connected G ,

$$\begin{aligned} \pi_2(G/H) &= \pi_1(H), \\ \pi_1(G/H) &= 0. \end{aligned} \quad (5.35)$$

(b) If $G = \tilde{G}/C$, where \tilde{G} is simply-connected with a finite center (i.e. a subgroup that commutes with all elements of \tilde{G}), and C is a subgroup of the center of \tilde{G} , then

$$\begin{aligned} \pi_2(G/H) &= \pi_2(\tilde{G}/H) = \pi_1(H), \\ \pi_1(G/H) &= C. \end{aligned} \quad (5.36)$$

We give some examples of these properties, assuming that H is non-empty:

Ex. 1. $G = SU(2)$.

- (a) For half-integer representations, $H = U(1)$, $\pi_1(H) = 0$.
- (b) For integer representations, $H = U(1)$, $\pi_1(H) = \mathbb{Z}$ (the set of all integers).

Ex. 2. $G = SU(3)$.

- (a) For the triplet representation, $H = SU(2)$, $\pi_1(H) = 0$.
- (b) For the octet representation, ρ can be represented as a 3×3 traceless matrix. If all eigenvalues are distinct, then $H = U(1) \times U(1)$, $\pi_1(H) = \mathbb{Z} + \mathbb{Z}$. If two of the eigenvalues are the same, then $H = U(2)$, $\pi_1(H) = \mathbb{Z}$.

Ex. 3. $G = SU(2) \times U(1)$. $H = U(1)$, $\pi_2(G/H) = 0$. This shows that the Weinberg-Salam model has no topological solitons. We shall discuss this in more detail in Chapter 6.

2 Flux Quantization⁷

Assume $\pi_2(G/H) \neq 0$. Then a topological soliton is possible, and condition (b) of (5.33) imposes boundary conditions on S^2 : to order $O(r^{-2})$,

$$D^\mu \rho(x) = \partial^\mu \rho(x) + ig A_a^\mu(x) L_a \rho(x) = 0. \quad (5.37)$$

The $O(r^{-1})$ terms in A_a^μ must be cancelled by $\partial^\mu \rho$. Thus, A_a^μ and ρ are related. In particular, if ρ is not constant, there will be a Coulombic potential.

⁶ M. I. Monastyrskii and A. M. Perelomov, *JETP Lett.* **21**, 43 (1975).

⁷ K. Huang and D. R. Stump, *Phys. Rev. D* **15**, 3660 (1977).

For static solutions, this cannot be an electric potential. Hence, there will be a magnetic potential corresponding to a magnetic monopole.

The condition (5.37) can always be fulfilled, if $V(\phi)$ is such that the lowest minima are all related by continuous gauge transformations. We assume this is true.

Consider three great circles $C_k (k = 1, 2, 3)$ on S^2 , whose normals are respectively $\hat{\mathbf{x}}_k$, as shown in Fig. 5.1. Let ΔC_k be a finite arc of C_k ; with endpoints y and z . According to (5.37), $\rho(x)$ undergoes parallel displacement from y and z to order r^{-2} :

$$\begin{aligned} \rho(z) &= \Omega(\Delta C_k) \rho(y), \\ \Omega(\Delta C_k) &= T \exp \left(ig \int_{\Delta C_k} \mathbf{ds} \cdot \mathbf{A}(x) \right), \end{aligned} \quad (5.38)$$

where \mathbf{ds} is an element of arc on S^2 . On the other hand, $\rho(z)$ can be obtained from $\rho(y)$ by a rotation through some angle $\Delta\theta$ about $\hat{\mathbf{x}}_k$, generated by a linear combination of the generator $\{L_a\}$ of G . Since these rotations form an Abelian group, there must be a choice of $\{L_a\}$ such that

$$[\Omega(\Delta C_k), \Omega(\Delta C'_k)] = 0, \quad (5.39)$$

where ΔC_k and $\Delta C'_k$ are two arcs of C_k . Therefore, the operation T in (5.38) can be dropped. Thus

$$\begin{aligned} \Delta\Omega(C_1) &= \exp \left(ig \int_{\Delta C_1} \mathbf{ds} \cdot \mathbf{A} \right) = \exp(-i(\Delta\theta_1)\mathcal{J}^1), \\ \Delta\Omega(C_2) &= \exp \left(ig \int_{\Delta C_2} \mathbf{ds} \cdot \mathbf{A} \right) = \exp(-i(\Delta\theta_2)\mathcal{J}^2), \\ \Delta\Omega(C_3) &= \exp \left(ig \int_{\Delta C_3} \mathbf{ds} \cdot \mathbf{A} \right) = \exp(-i(\Delta\theta_3)\mathcal{J}^3). \end{aligned} \quad (5.40)$$

It is possible to make all these rotations with the same choice of axes $\hat{\mathbf{x}}_1, \hat{\mathbf{x}}_2, \hat{\mathbf{x}}_3$; for example the closed circuit OPQ in Fig. 5.1. Therefore, with the same choice of $\{L_a\}$, \mathcal{J}^k can be constructed such that

$$[\mathcal{J}^1, \mathcal{J}^2] = i\mathcal{J}^3 \quad (1, 2, 3 \text{ cyclic}). \quad (5.41)$$

Since the eigenvalues of \mathcal{J}^k are integers or half-integers, the angles $\Delta\theta_k$ are defined only mod (4π) . This corresponds to the fact that a rotation through angle 4π leaves the system truly unchanged, as we pointed out in Sec. 4.2.

Now take ΔC_k to be a complete great circle, with $\theta_k = 2\pi$:

$$\Omega(C_k) = \exp \left(ig \oint_{C_k} \mathbf{ds} \cdot \mathbf{A} \right) = \exp(-2\pi i \mathcal{J}^k). \quad (5.42)$$

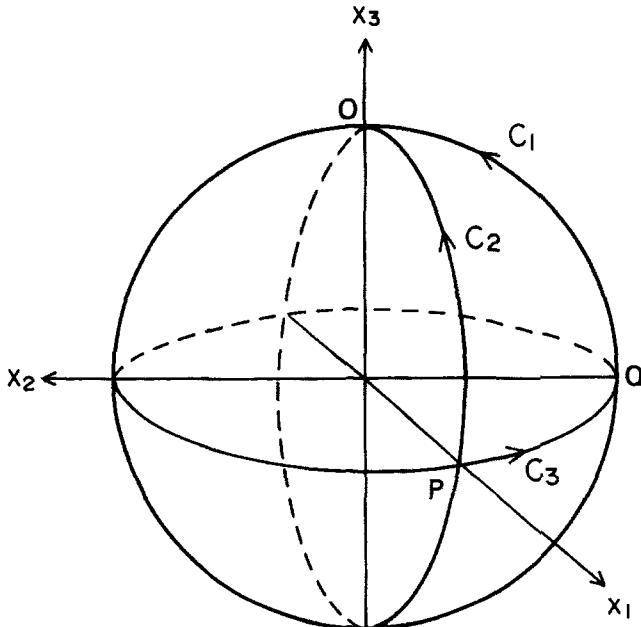


Fig. 5.1 Great circles on a sphere

Let the flux matrix Φ^k be defined by

$$\Phi^k = \oint_{C_k} \mathbf{ds} \cdot \mathbf{A}, \quad \Phi_k = -\Phi^k. \quad (5.43)$$

This refers to the flux of $\nabla \times \mathbf{A}$, which is divergenceless, and is in general not the flux of \mathbf{B} . Its eigenvalues are defined only mod $(4\pi/g)$. By (5.42) and (5.41) we have

$$\left[\frac{g}{2\pi} \Phi_1, \frac{g}{2\pi} \Phi_2 \right] = \frac{ig}{2\pi} \Phi_3 \quad (1, 2, 3 \text{ cyclic}). \quad (5.44)$$

This is the generalization of the flux quantization condition (3.49) in the $U(1)$ case. We have assumed that $\{L_a\}$ have been chosen in a special way. This is equivalent to saying that (5.42) is true in a gauge in which $\mathbf{A}(x)$ has no singularities on S^2 .

For a non-trivial solution to (5.44) to be possible, G must contain $SU(2)$. Choose $\{L_a\}$ such that the first three are generators of $SU(2)$:

$$\begin{aligned} \{L_a\} &= \{L_1, L_2, L_3, L_\alpha\} \quad (\alpha = 4, 5, \dots, N), \\ [L_1, L_2] &= iL_3 \quad (1, 2, 3 \text{ cyclic}). \end{aligned} \quad (5.45)$$

The choice, of course, depends on the choice of the coordinate system $\hat{x}_1, \hat{x}_2, \hat{x}_3$. Under rotations and reflections of the coordinate system both Φ_k and L_k ($k = 1, 2, 3$) must change like 3-vectors. Thus the following is a 3-vector equation:

$$\frac{g}{2\pi} \Phi_k = -\frac{g}{2\pi} \oint_{C_k} \mathbf{ds} \cdot \mathbf{A} = L_k, \quad (k = 1, 2, 3). \quad (5.46)$$

Note that the index k on L_k is an internal symmetry index that also serves as a 3-vector index. Thus, internal and spatial symmetries are coupled. The origin of this is the fact that parallel displacement of ρ is equivalent to rotation.

3 Boundary Conditions

Taking $L_k = \frac{1}{2}\tau_k$, we rewrite (5.46) as

$$-\frac{g}{2\pi} \Phi_a^k \tau_a \equiv -\frac{g}{2\pi} \oint_{C_k} \mathbf{ds} \cdot \mathbf{A}_a \tau_a = \tau_k \quad (k = 1, 2, 3), \quad (5.47)$$

which leads to

$$-\frac{g}{2\pi} \Phi_a^k \equiv -\frac{g}{2\pi} \oint_{C_k} \mathbf{ds} \cdot \mathbf{A}_a = \delta_{ak} \quad (k = 1, 2, 3; a = 1, 2, 3). \quad (5.48)$$

This requires A_a^k ($k = 1, 2, 3; a = 1, 2, 3$) to be a pseudotensor that falls off like r^{-1} asymptotically:

$$\begin{aligned} A_\alpha^k &= 0 \quad (\alpha = 4, 5, \dots, N), \\ A_a^k &= g' \epsilon^{akj} \frac{x^j}{r^2} + O(r^{-2}), \end{aligned} \quad (5.49)$$

where g' can be determined by substituting this into (5.48):

$$gg' = 1 \mod(2). \quad (5.50)$$

The case $gg' = \pm 2$ is gauge-equivalent to the trivial case $gg' = 0$, and $gg' = \pm 1$ are gauge-equivalent. Hence, the boundary conditions for $A_a^\mu(x)$ are⁸

$$A_a^k(x) \xrightarrow[r \rightarrow \infty]{} \frac{1}{g} \epsilon^{akj} \frac{x^j}{r^2} + O(r^{-2}) \quad (k = 1, 2, 3; a = 1, 2, 3). \quad (5.51)$$

This describes the only possible topological soliton with “spherical symmetry” (i.e., for which \mathbf{A} has no singularity on S^2).

Note that g is the gauge-coupling constant, or the “charge” of the gauge field. The charge of an irreducible multiplet ϕ is

$$e = \kappa g, \quad (5.52)$$

where κ is the smallest positive eigenvalue of the matrix representing a generator of G . For example, for fundamental representations, $e = g$ for $O(3)$, $e = g/2$ for $SU(2)$, and $e = g/3$ for $SU(3)$.

We now examine the boundary conditions for $\rho(x)$. Substituting (5.51) into (5.37), we obtain

⁸ T. T. Wu and C. N. Yang, in *Properties of Matter Under Unusual Conditions*, edited by H. Mark and S. Fernbach (Wiley-Interscience, New York, 1969), p. 349.

$$\mathbf{ds} \cdot \left(\nabla \rho + i \frac{\mathbf{r} \times \mathbf{L}}{r^2} \rho \right) = 0, \quad (5.53)$$

where $\mathbf{L} = (L_1, L_2, L_3)$ and \mathbf{ds} is an infinitesimal arc element on S^2 . Writing $\mathbf{ds} = \hat{n} \times \mathbf{r} d\theta$, with the symbols defined in Fig. 5.2, we obtain, after a little algebra,

$$\frac{\partial \rho}{\partial \theta} = i \hat{n} \cdot \mathbf{L} \rho. \quad (5.54)$$

In the representation for which $\hat{n} \cdot \mathbf{L}$ is diagonal with eigenvalue λ , the solution is $\rho = \rho_0 e^{i\lambda\theta}$. Since ρ must be a continuous function on S^2 , λ must be an integer. Therefore, a topological soliton can exist only if ρ belongs to an integer representation of the $SU(2)$ subgroup of G . Let us assume the simplest case of the adjoint representation. Then ρ is a triplet of fields ρ_a ($a = 1, 2, 3$), with $(L_a)_{bc} = -i\epsilon^{abc}$, and (5.53) can be written as

$$\begin{aligned} \partial^k \rho_b &= -\epsilon^{akj} \frac{x^j}{r^2} \epsilon^{abc} \rho_c \\ &= -\frac{1}{r^2} (\delta_{kb} x^c \rho_c - x^b \rho_k). \end{aligned} \quad (5.55)$$

The solution is

$$\rho_a(x) = \frac{x^a}{r} \rho_0, \quad (5.56)$$

where ρ_0 is any constant satisfying (5.32). In summary, the boundary conditions for a spherical symmetric topological soliton are

$$\begin{aligned} A_a{}^k(x) &\xrightarrow[r \rightarrow \infty]{1}{g} \epsilon^{akj} \frac{x^j}{r^2}, \\ \phi_a(x) &\xrightarrow[r \rightarrow \infty]{1}{r} \rho_0. \end{aligned} \quad (5.57)$$

The magnetic field can be shown to have the behavior

$$B_a{}^k(x) \xrightarrow[r \rightarrow \infty]{1}{g} \frac{1}{r^4} \frac{x^a x^k}{r^4}, \quad (5.58)$$

while the electric field vanishes to order r^{-2} .

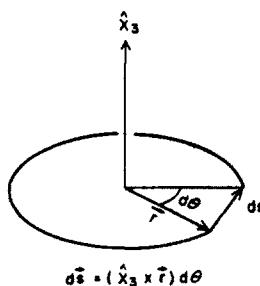


Fig. 5.2 Definition of symbols in Eq. (5.54).

4 Explicit Solution⁹

To show a solution actually exists, take $G = SU(2)$, with a triplet of Higgs fields $\phi_a(x)$. According to the Higgs mechanism (see Sec. 4.6), there should be one massless vector boson, two massive vector bosons, and one massive Higgs boson.

For a static solution, take $A_a^0 = 0$, and

$$\begin{aligned} A_a^k(x) &= \frac{1}{g} \epsilon^{akj} \frac{x^j}{r^2} F(r), \\ \phi_a(x) &= \frac{x^a}{r} \rho_0 \eta(r), \end{aligned} \quad (5.59)$$

with the boundary conditions

$$\begin{aligned} F(0) &= 0, & F(\infty) &= 1, \\ \eta(0) &= 0, & \eta(\infty) &= 1, \end{aligned} \quad (5.60)$$

where F and η are required to vanish at $r = 0$, in order to render the solution regular. The electric and magnetic fields are given by

$$\begin{aligned} E_a^k &= 0, \\ B_a^k &= \frac{1}{g} \frac{x^a x^k}{r^4} F(1 - F) + \frac{1}{g} \left(\delta_{ak} - \frac{x^a x^k}{r^2} \right) \frac{F'}{r}. \end{aligned} \quad (5.61)$$

The total energy is given by

$$\begin{aligned} \mathcal{E} = \int d^3x \left\{ \frac{1}{g^2} \left(\frac{F'}{r} \right)^2 + \frac{1}{2g^2} \left[\frac{F(1 - F)}{r^2} \right]^2 \right. \\ \left. + \rho_0^2 \left[\frac{\eta(1 - F)}{r} \right]^2 + \frac{\rho_0^2}{2} \left[\frac{\eta(1 - F)}{r} + \eta' F \right]^2 + V \right\}. \end{aligned} \quad (5.62)$$

All terms are positive-definite, and neither $F = 0$ nor $F = 1$ give the lowest minimum. Hence there is a solution, by the variational principle. Explicit numerical calculations with a quartic form for V show that the mass of the topological soliton is of the order of magnitude

$$M \sim \frac{m_V}{g^2/4\pi}, \quad (5.63)$$

where m_V is a vector boson mass.

We can see from (5.62) that for the pure Yang-Mills case ($\eta \equiv 0$), the F that minimizes \mathcal{E} would be $F = 1$, which leads to an unacceptable singularity in (5.59) at $r = 0$. This is why Higgs fields are needed to give a static soliton solution.

⁹ G. 't Hooft, *Nucl. Phys.* **B79**, 276 (1974); A. M. Polyakov, *JETP Lett.* **20**, 194 (1974).

5 Physical Fields

The boundary conditions (5.57) are valid only in a special gauge, the Coulomb gauge (we can verify that $\partial_k A_a^k = 0$). In this gauge most fields are not physical. It is difficult to see that there is only one massless vector field, and which combination of A_a^k it corresponds to. To display the physical fields, we go to unitary gauge, which is defined as the gauge in which $\phi(x)$ has only one non-vanishing component, as indicated in the following comparison:

$$\begin{aligned} \text{Coulomb gauge: } \phi(x) &\xrightarrow[r \rightarrow \infty]{} \rho_0 \begin{pmatrix} x/r \\ y/r \\ z/r \end{pmatrix}, \\ \text{Unitary gauge: } \phi(x) &\xrightarrow[r \rightarrow \infty]{} \rho_0 \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \end{aligned} \tag{5.64}$$

In the Coulomb gauge, the axis along which $\phi(x)$ points, in internal space (i.e. the 3-axis that defines the diagonal generator L_3), coincides with the radial vector r in ordinary space. In unitary gauge, the direction of $\phi(x)$ in internal space is independent of r . To transform from Coulomb gauge to unitary gauge, we make a gauge transformation to rotate \hat{r} into \hat{k}_3 at every point in space:

$$\begin{aligned} \phi &\rightarrow U\phi \\ U &= e^{-i\theta L_2} e^{-i\varphi L_3} \end{aligned} \tag{5.65}$$

Where θ, φ are the polar angles of r in a spherical coordinate system. It is clear that this transformation is ill-defined along the negative z -axis, because the downward vector \hat{r} has to be rotated to its opposite, and the way to do this is not unique. Therefore, in unitary gauge, A_a^k is not defined along the negative z -axis. The flux quantization condition (5.44) does not hold in this gauge.

The massless vector field tensor $B^{\mu\nu}$ must be the projection of $F_a^{\mu\nu}$ into the Higgs space (see Sec. 4.6.):

$$B^{\mu\nu} = F_a^{\mu\nu}\phi_a/\rho_0. \tag{5.66}$$

This is gauge-invariant, for under an infinitesimal gauge transformation,

$$\begin{aligned} \delta F_a^{\mu\nu} &= \epsilon_{abc}\omega_b F_c^{\mu\nu}, \\ \delta\phi_a &= \epsilon_{abc}\omega_b \phi_c, \end{aligned}$$

and hence

$$\delta(F_a^{\mu\nu}\phi_a) = \epsilon_{abc}\omega_b(F_c^{\mu\nu}\phi_a + F_a^{\mu\nu}\phi_c) = 0.$$

Therefore, $F_a^{\mu\nu}\phi_a$ is the same in unitary gauge as in the Coulomb gauge, and we can calculate it conveniently in the latter. We calculate the asymptotic form only:

$$B^k \xrightarrow[r \rightarrow \infty]{\frac{1}{g} \frac{x^a x^k}{r^4} \frac{x^a}{r}} = \frac{1}{g} \frac{x^k}{r^3}. \quad (5.67)$$

Note that this is well-defined all over S^2 , although A^k is not defined along the negative z -axis. To summarize, in unitary gauge we have

$$\begin{aligned} \mathbf{B} &\xrightarrow[r \rightarrow \infty]{\frac{1}{g} \frac{\hat{\mathbf{r}}}{r^2}} \text{(Massless gauge field),} \\ \phi &\xrightarrow[r \rightarrow \infty]{} \rho_0 \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \text{(Higgs field).} \end{aligned} \quad (5.68)$$

The other 2 vector fields are massive, and hence fall off exponentially as $r \rightarrow \infty$.

Far away from the origin, \mathbf{B} appears to be the field produced by a magnetic monopole of magnetic charge $1/g$ located at $\mathbf{r} = 0$, hence the name monopole solution. However, there is no singularity at $\mathbf{r} = 0$.

As a comparison, we give the vector potential and magnetic field for the Dirac monopole¹⁰, with a string along the negative z -axis:

$$\mathbf{A}_D = g' \frac{\hat{\mathbf{z}} \times \hat{\mathbf{r}}}{r + z}, \quad A_0 = 0, \quad (5.69)$$

where g' is such that $2g'e = \text{integer}$, e being the electronic charge. The components of \mathbf{A}_D are, in rectangular coordinates:

$$\begin{aligned} A^2 &= -\frac{g'}{r + z} \frac{y}{r}, \\ A^2 &= -\frac{g'}{r + z} \frac{x}{r}, \\ A^3 &= 0. \end{aligned} \quad (5.70)$$

and in polar coordinates:

$$\begin{aligned} A_r &= 0, \\ A_\theta &= 0, \end{aligned} \quad (5.71)$$

$$A_\varphi = \frac{g'}{r} \frac{1 - \cos \theta}{\sin \theta}.$$

The magnetic field is given by

$$\mathbf{B}_D = \frac{\hat{\mathbf{r}} g'}{r^2} + \hat{\mathbf{z}} b_0 \delta(x) \delta(y) \theta(1 - z). \quad (5.72)$$

This is the same as (5.68), except that this is supposed to hold everywhere, including $\mathbf{r} = 0$, except for the string. In the Dirac case, which is a construction not based on a complete theory, the string is a necessary artifact. In our case,

¹⁰ P. A. M. Dirac, *Proc. Roy. Soc. A* **133**, 60 (1931).

there is really no string because it does not appear in \mathbf{B} , and even in \mathbf{A} it can be transformed away by a gauge transformation.

So far, there is no experimental evidence for the magnetic monopole, possibly because even if it exists, the mass would have to be so large that it cannot be produced in present-day accelerators.

6 Spin from Isospin¹¹

A monopole can trap a boson of “isospin” 1/2 [i.e. a boson $SU(2)$ doublet], resulting in a topological soliton with intrinsic angular momentum $\hbar/2$, although Dirac fields are not present. The monopole converts isospin into spin. This phenomenon occurs only in the quantum theory.

The point is that the flux matrices $(g/2\pi)\Phi_k$ ($k = 1, 2, 3$) are generators of rotations in space, according to (5.44). Hence, they should be added to the total angular momentum of the system. Their eigenvalues are half-integers, if there are fields of isospin 1/2.

As an example, consider $G = SU(2)$, and take the matter fields to consist of a triplet of Higgs fields ϕ and a doublet K , with

$$\phi(x) \xrightarrow[r \rightarrow \infty]{} \rho \neq 0, \quad K(x) \xrightarrow[r \rightarrow \infty]{} 0. \quad (5.73)$$

The flux matrices can be represented by

$$\frac{g}{2\pi} \Phi_k = L_k = \begin{pmatrix} C_k & 0 \\ 0 & \frac{1}{2}\tau_k \end{pmatrix} \quad (k = 1, 2, 3), \quad (5.74)$$

where C_k are 3×3 matrices appropriate for the adjoint representation. We take this to be the extra angular momentum of the system, to be added to the “normal” total angular momentum, which has only integer eigenvalues. Note that the gauge field makes no contribution to it. In quantum theory, we put $\phi(x) = \rho + \eta(x)$, and represent $\eta(\mathbf{r})$ and $K(\mathbf{r})$ by quantum fields in the Schrödinger picture:

$$\begin{aligned} \eta(\mathbf{r}) &= \sum_{\mathbf{p}} \frac{1}{\sqrt{2\omega_p}} [e^{i\mathbf{p} \cdot \mathbf{r}} a(\mathbf{p}) + e^{-i\mathbf{p} \cdot \mathbf{r}} a^\dagger(\mathbf{p})], \\ K(\mathbf{r}) &= \sum_{\mathbf{p}} \frac{1}{\sqrt{2E_p}} [e^{-i\mathbf{p} \cdot \mathbf{r}} b(\mathbf{p}) + e^{-i\mathbf{p} \cdot \mathbf{r}} c^\dagger(\mathbf{p})], \end{aligned} \quad (5.75)$$

where $a(\mathbf{p})$ is an isovector, and $b(\mathbf{p}), c(\mathbf{p})$ are isospinors. The diagonal component of the extra angular momentum is

$$\mathcal{J}^3 = \sum_{\mathbf{p}} (a(\mathbf{p}), C_3 a(\mathbf{p})) + \sum_{\mathbf{p}} \left[b^\dagger(\mathbf{p}) \frac{\tau_3}{2} b(\mathbf{p}) - c^\dagger(\mathbf{p}) \frac{\tau_3}{2} c(\mathbf{p}) \right], \quad (5.76)$$

¹¹ R. Jackiw and C. Rebbi, *Phys. Rev. Lett.* **36**, 1116 (1976), P. Hasenfratz and G. 't Hooft, *Phys. Rev. Lett.* **36**, 1119 (1976).

where

$$C_3 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & -1 \end{pmatrix}, \quad \frac{\tau_3}{2} = \begin{pmatrix} \frac{1}{2} & 0 \\ 0 & -\frac{1}{2} \end{pmatrix}. \quad (5.77)$$

The second term in (5.76) has half-integer eigenvalues. The reasons why \mathcal{J}^3 must be included in the total angular momentum are Poincaré invariance, and the need for gauge-fixing in quantum theory.¹² Goldhaber¹³ has given an argument showing that monopoles with half-integer spin obey Fermi statistics.

A related phenomenon is that, in the presence of a monopole, a massless Dirac field possesses states of fermion number $1/2$.¹⁴ A simpler version of the phenomenon in one spatial dimension (where the topological soliton is known as a “kink”), can be applied to polymers, and predicts states of electron number $1/2$ in polyacetylene.¹⁵

¹² K. Huang and D. R. Stump, *op. cit.*

¹³ A. Goldhaber, *Phys. Rev. Lett.* **19**, 1122 (1976).

¹⁴ R. Jackiw and C. Rebbi, *Phys. Rev. D* **13**, 3398 (1976).

¹⁵ W. P. Su, J. R. Schrieffer, and A. J. Heeger, *Phys. Rev. B* **22**, 2099 (1980).

CHAPTER 6

WEINBERG-SALAM MODEL

6.1 The Matter Fields

Glashow advanced the idea that the electromagnetic and weak interactions may be unified in a gauge theory based on the group $SU(2) \times U(1)$. The problem of generating masses in a manner consistent with gauge invariance was solved later by Weinberg and Salam, using the idea of spontaneous symmetry breaking¹. The resulting theory, known as the Weinberg-Salam model, was shown by 't Hooft² to be a renormalizable quantum field theory. The inclusion of quarks in this theory was achieved by Glashow, Iliopoulos and Maiani³. It gained wide acceptance after experiments verified some of its predictions, chief among these being the structure of the neutral currents.

The basic idea of the Weinberg-Salam model has been discussed in Chapter 1, and basic experimental facts concerning the electromagnetic and weak interactions have been reviewed in Chapter 2. We reiterate some of these in a different form for emphasis.

Let us first review the definition of chirality, which is the eigenvalue of γ_5 , with $\gamma_5 = +1$ corresponding to right-handedness, and $\gamma_5 = -1$ to left-handedness:

$$\begin{aligned} \gamma_5 R &= R, & \bar{R} \gamma_5 &= -\bar{R}, \\ \gamma_5 L &= -L, & \bar{L} \gamma_5 &= \bar{L}, \end{aligned} \quad (6.1)$$

where R , L are Dirac spinors with only two independent components. They may be obtained from a 4-component Dirac spinor ψ by the following projections:

$$\begin{aligned} R &= \frac{1}{2}(1 + \gamma_5)\psi, & \bar{R} &= \frac{1}{2}\bar{\psi}(1 - \gamma_5), \\ L &= \frac{1}{2}(1 - \gamma_5)\psi, & \bar{L} &= \frac{1}{2}\bar{\psi}(1 + \gamma_5). \end{aligned} \quad (6.2)$$

For later applications, it is important to note that

$$\begin{aligned} \bar{\psi}\psi &= \bar{L}R + \bar{R}L, \\ \bar{\psi}\gamma^\mu\psi &= \bar{L}\gamma^\mu L + \bar{R}\gamma^\mu R. \end{aligned} \quad (6.3)$$

The Dirac equation for a massive particle of 4-momentum $p^\mu = (E, \mathbf{p})$ may be

¹ For historical accounts and references see the Nobel lectures of Glashow, Weinberg, and Salam: S. L. Glashow, *Rev. Mod. Phys.* **53**, 539 (1980); S. Weinberg, *ibid.* **52**, 515 (1980); A. Salam, *ibid.* **52**, 525 (1980).

² G. 't Hooft, *Nuclear Phys.* **B33**, 173 (1971); **B35**, 167 (1971).

³ S. L. Glashow, J. Iliopoulos, and L. Maiani, *Phys. Rev. D3*, 1043 (1981).

written as

$$(\alpha \cdot \mathbf{p} + \beta m)\psi = E\psi, \quad E = (p^2 + m^2)^{1/2}. \quad (6.4)$$

Using the identity $\alpha = \gamma_5 \sigma$, and the fact that γ_5 and σ commute, we can rewrite this in the form

$$\begin{aligned} \sigma \cdot \hat{\mathbf{p}} R &= \frac{E}{p} R - \frac{m}{p} \beta L, \quad p \equiv |\mathbf{p}|, \\ \sigma \cdot \hat{\mathbf{p}} L &= -\frac{E}{p} L + \frac{m}{p} \beta R. \end{aligned} \quad (6.5)$$

These equations become decoupled if $m = 0$:

$$\begin{aligned} \sigma \cdot \hat{\mathbf{p}} R &= R, \\ \sigma \cdot \hat{\mathbf{p}} L &= -L. \end{aligned} \quad (6.6)$$

Therefore, for massless Dirac particles, chirality is the same as helicity; for antiparticles, chirality is the opposite of helicity. [An antiparticle has the same chirality as the particle, by definition; but it has the opposite helicity due to a change in the sign of E in (6.4)].

The electromagnetic interaction Lagrangian density is given by

$$\mathcal{L}^{\text{em}} = e \bar{\psi} Q A \psi, \quad (6.7)$$

where Q is the charge matrix, and $A = A_\mu \gamma^\mu$. The charge-changing weak interaction Lagrangian density can be written in the form

$$\mathcal{L}^{\text{ch}} = \frac{g}{\sqrt{2}} \bar{L} (W_+ \tau_- + W_- \tau_+) L = \frac{g}{2} \bar{L} (W_1 \tau_1 + W_2 \tau_2) L, \quad (6.8)$$

where W_{\pm}^μ are the vector boson fields that mediate this interaction, with

$$\begin{aligned} W_{\pm}^\mu &= \frac{1}{\sqrt{2}} (W_1^\mu \pm i W_2^\mu), \\ \tau_{\pm} &= \frac{1}{2} (\tau_1 \pm i \tau_2). \end{aligned} \quad (6.9)$$

Adding (6.7) and (6.8), we have

$$\mathcal{L}^{\text{em}} + \mathcal{L}^{\text{ch}} = \bar{L} \left[g \left(W_1 \frac{\tau_1}{2} + W_2 \frac{\tau_2}{2} \right) + e A Q \right] L + \bar{R} e A Q R. \quad (6.10)$$

The form suggests that the basic spinor fermion fields are not 4-component Dirac spinors, but rather their left and right-handed projections, and that A^μ may be combined with W_1^μ and W_2^μ to form a multiplet of some kind.

To illustrate the multiplet structure of the matter fields, and the need for spontaneous symmetry breaking, we consider only one lepton doublet consisting of the electron and its neutrino:

$$L = \begin{pmatrix} \nu_L \\ e_L \end{pmatrix}, \quad R = e_R. \quad (6.11)$$

The neutrino is assumed to be massless, and hence ν_R is absent. The theory is assumed to be invariant under an $SU(2)$ group, under which L transforms as a doublet, and R transforms as a singlet.

A conventional mechanical mass term in the Lagrangian density cannot be invariant under $SU(2)$, because it is proportional to $\bar{\psi}\psi = \bar{L}R + \bar{R}L$. We cannot violate this symmetry (to however small a degree), because it is to be gauged. Therefore, in this theory the electron mass can arise only by virtue of a spontaneous breakdown of $SU(2)$. A convenient way to implement this is to introduce a doublet Higgs field

$$\phi = \begin{pmatrix} \phi_+ \\ \phi_0 \end{pmatrix}, \quad (6.12)$$

where the subscripts refer to the electric charge, and write the mass term as

$$\mathcal{L}^{\text{mass}} \propto \bar{L}\phi R + \bar{R}\phi^\dagger L, \quad (6.13)$$

where $\bar{L}\phi$ is an $SU(2)$ singlet, and a Dirac spinor. Thus, (6.13) is Lorentz invariant, and invariant under $SU(2)$. If ϕ has non-zero vacuum value, then for low excitations, (6.13) is indistinguishable from a conventional mass term. Writing out (6.13) in detail, we have

$$\begin{aligned} \mathcal{L}^{\text{mass}} &\propto (\bar{\nu}_L \quad \bar{e}_L) \begin{pmatrix} \phi_+ \\ \phi_0 \end{pmatrix} e_R + e_R (\phi_- \quad \phi_0) \begin{pmatrix} \nu_L \\ e_L \end{pmatrix} \\ &= (\bar{\nu}_L e_R) \phi_+ + (\bar{e}_R \nu_L) \phi_- + (\bar{e} e) \phi_0. \end{aligned} \quad (6.14)$$

The first two terms can be transformed away by a gauge transformation (unitary gauge). The last term gives mass to the electron if $\phi_0 \neq 0$ in the vacuum state. There is no mass term for the neutrino.

If one wishes to give the neutrino mass, one can make use of the conjugate Higgs doublet

$$\tilde{\phi} = \begin{pmatrix} \phi_0 \\ -\phi_- \end{pmatrix}, \quad \phi_- \equiv \phi_+^*, \quad (6.15)$$

which also transforms as an $SU(2)$ doublet, and add to (6.13) a term proportional to

$$\begin{aligned} \bar{L}\tilde{\phi}\nu_R + \bar{\nu}_R \tilde{\phi}^\dagger L &= (\bar{\nu}_L \bar{e}_L) \begin{pmatrix} \phi_0 \\ -\phi_- \end{pmatrix} \nu_R + \bar{\nu}_R (\phi_0 \quad -\phi_+) \begin{pmatrix} \nu_L \\ e_R \end{pmatrix} \\ &= (\bar{\nu}\nu) \phi_0 - (\bar{e}_L \nu_R) \phi_- - (\bar{\nu}_R e_R) \phi_+. \end{aligned}$$

The last two terms can be transformed away in unitary gauge, and the first term gives mass to the neutrino. However, we exclude neutrino mass for simplicity.

The Lagrangian density for the matter fields is taken to be

$$\begin{aligned} \mathcal{L}_0 &= \bar{L}i\not{D}L + \bar{R}i\not{D}R + (\partial\phi)^\dagger \cdot (\partial\phi) - V(\phi^\dagger\phi) - \frac{m}{\rho_0} (\bar{L}\phi R + \bar{R}\phi^\dagger L), \\ V(\phi^\dagger\phi) &= \lambda(\phi^\dagger\phi - \rho_0^2)^2, \end{aligned} \quad (6.17)$$

where ρ_0 , λ are real positive parameters, and m is the physical mass of the electron.

6.2 The Gauge Fields

1 Gauging $SU(2) \times U(1)$

The Lagrangian density (6.17) is globally invariant under $SU(2)$, whose generators will be denoted by t in general, and by $\tau/2$ in the fundamental representation. The global gauge transformations are

$$\begin{aligned} L &\rightarrow e^{-i\omega \cdot \tau/2}L, & \phi &\rightarrow e^{-i\omega \cdot \tau/2}\phi, \\ R &\rightarrow R. \end{aligned} \quad (6.18)$$

In addition to $SU(2)$, the Lagrangian density is also invariant under independent phase changes of L and R :

$$\begin{aligned} L &\rightarrow e^{-i\theta}L, \\ R &\rightarrow e^{-i\theta'}R, \\ \phi &\rightarrow e^{-i(\theta-\theta')}\phi. \end{aligned} \quad (6.19)$$

These transformations form a group $U(1) \times U(1)$. We can identify the generator of one of the $U(1)$ groups as lepton number N , and call the generator of the other $U(1)$ group, weak hypercharge t_0 :

$$\begin{aligned} L &\rightarrow e^{-i(\alpha t_0 + \beta N)}L \quad (t_0 = -\frac{1}{2}, N = 1), \\ R &\rightarrow e^{-i(\alpha t_0 + \beta N)}R \quad (t_0 = -1, N = 1), \\ \phi &\rightarrow e^{-i(\alpha t_0 + \beta N)}\phi \quad (t_0 = \frac{1}{2}, N = 0). \end{aligned} \quad (6.20)$$

The assignments for N are conventional, and t_0 satisfies the rule

$$Q = t_3 + t_0, \quad (6.21)$$

where Q is the electric charge, in units of the magnitude of the electronic charge. There is evidence that the $U(1)$ corresponding to lepton number is not a local gauge symmetry, because it is experimentally observed to be an unbroken symmetry; if it were gauged there would have to be a massless gauge field coupled to lepton number, which has not been observed.

The Weinberg-Salam model is obtained by gauging $SU(2) \times U(1)$, where the $U(1)$ is generated by weak hypercharge. To study the properties of the gauge fields, we continue to consider only one lepton doublet, since adding more doublets will not alter the gauge fields.

To gauge the symmetry $SU(2) \times U(1)$, we associate a gauge field with each of the generators, with notations given in Table 6.1. The locally gauge invariant

Table 6.1 GAUGE FIELDS OF THE WEINBERG-SALAM MODEL

Group	Generators	Gauge-Fields	Field Tensors
$SU(2)$	t	W^μ	$G^{\mu\nu} = \partial^\mu W^\nu - \partial^\nu W^\mu - g W^\mu \times W^\nu$
$U(1)$	t_0	W_0	$H^{\mu\nu} = \partial^\mu W_0^\nu - \partial^\nu W_0^\mu$

Lagrangian density is

$$\begin{aligned}\mathcal{L} = & -\frac{1}{4}(G^{\mu\nu}G_{\mu\nu} + H^{\mu\nu}H_{\mu\nu}) + \bar{L}i\cancel{D}L + \bar{R}i\cancel{D}R \\ & + (D\phi)^\dagger(D\phi) - V(\phi^\dagger\phi) - \frac{m}{\rho_0}(\bar{L}\phi R + \bar{R}\phi^\dagger L).\end{aligned}\quad (6.22)$$

The covariant derivative D^μ is defined by

$$D^\mu = \partial^\mu + ig\mathbf{W}^\mu \cdot \mathbf{t} + ig'W_0^\mu t_0, \quad (6.23)$$

where g and g' are two independent gauge coupling constants.

We require that there be only one massless neutral gauge field, the electromagnetic field A^μ , which is coupled to the charge eQ , with Q given by (6.21). Generally, it is a linear combination of W_3^μ and W_0^μ . Accordingly we put

$$\begin{aligned}W_3^\mu &= Z^\mu \cos \theta_W + A^\mu \sin \theta_W, \\ W_0^\mu &= -Z^\mu \sin \theta_W + A^\mu \cos \theta_W.\end{aligned}\quad (6.24)$$

Solving this for A^μ and Z^μ gives

$$\begin{aligned}A^\mu &= W_0^\mu \cos \theta_W + W_3^\mu \sin \theta_W, \\ Z^\mu &= -W_0^\mu \sin \theta_W + W_3^\mu \cos \theta_W,\end{aligned}\quad (6.25)$$

where θ_W is called the Weinberg angle, a free parameter to be determined by experiments.

Given θ_W , the requirement that A^μ be the electromagnetic field imposes relations between g and g' , as follows. Rewrite (6.23) as

$$\begin{aligned}D^\mu = & \partial^\mu + ig(W_1^\mu t_1 + W_2^\mu t_2) \\ & + i(gt_3 \sin \theta_W + g't_0 \cos \theta_W)A^\mu \\ & + i(gt_3 \cos \theta_W - g't_0 \sin \theta_W)Z^\mu\end{aligned}\quad (6.26)$$

Requiring the coefficient of A^μ be eQ , we obtain the condition

$$gt_3 \sin \theta_W + g't_0 \cos \theta_W = e(t_3 + t_0), \quad (6.27)$$

where $-e$ is the charge of the electron ($e^2/4\pi\hbar c \approx 1/137$). This leads to

$$e = g \sin \theta_W = g' \cos \theta_W, \quad (6.28)$$

or

$$\begin{aligned}g'/g &= \tan \theta_W, \\ e &= gg'/(g^2 + g'^2)^{1/2}.\end{aligned}\quad (6.29)$$

With these, the covariant derivative can be written in the form

$$D^\mu = \partial^\mu + ig(W_1^\mu t_1 + W_2^\mu t_2) + ieQA^\mu + ieQ'Z^\mu, \quad (6.30)$$

where the neutral charge matrix Q' is defined by

$$Q' = t_3 \cos \theta_W - t_0 \tan \theta_W. \quad (6.31)$$

To study the masses of the gauge fields, it is convenient to go to unitary gauge, in which

$$\phi = \begin{pmatrix} 0 \\ \rho \end{pmatrix}, \quad (6.32)$$

where ρ is a real field^a. We can write

$$\begin{aligned} D^\mu \phi &= \{\partial^\mu + ig[\tfrac{1}{2}(W_1^\mu - iW_2^\mu)\tau_+ + \tfrac{1}{2}(W_1^\mu + iW_2^\mu)\tau_- + \tfrac{1}{2}W_3^\mu\tau_3] \\ &\quad + ig'W_0^\mu t_0\} \begin{pmatrix} 0 \\ \rho \end{pmatrix} \\ &= \left(\partial^\mu \rho - \frac{ig}{2 \cos \theta_W} Z^\mu \rho \right). \end{aligned} \quad (6.33)$$

Hence, the kinetic term for the Higgs field in the Lagrangian density, which generates the masses, takes the form

$$(D^\mu \phi)^\dagger (D_\mu \phi) = \tfrac{1}{4}g^2 \rho^2 \left[(W_1^\mu W_{1\mu} + W_2^\mu W_{2\mu}) + \frac{Z^\mu Z_\mu}{\cos^2 \theta_W} \right] + \partial^\mu \rho \partial_\mu \rho. \quad (6.34)$$

Note that A^μ does not appear, because the charged component of ϕ has been transformed away in (6.32). Therefore, A^μ is a massless field as desired. There are, in general, 4 complex scalar fields (ϕ_+ , ϕ_0). In unitary gauge there is only one real scalar field. The 3 scalar fields that went away became the longitudinal components of W^\pm and Z .

In terms of fields in the unitary gauge, the Lagrangian density is

$$\begin{aligned} \mathcal{L} &= -\tfrac{1}{4}(\mathbf{G} \cdot \mathbf{G} + \mathbf{H} \cdot \mathbf{H}) + \tfrac{1}{4}g^2 \rho^2 \left(W_1^2 + W_2^2 + \frac{Z^2}{\cos^2 \theta_W} \right) \\ &\quad + \bar{\nu}_L i\cancel{D} \nu_L + \bar{e} \left(i\cancel{D} - \frac{\rho}{\rho_0} m \right) e \\ &\quad + \partial \rho \cdot \partial \rho - \lambda (\rho^2 - \rho_0^2)^2, \end{aligned} \quad (6.35)$$

in an abbreviated notation that should be obvious. The masses m_W , m_Z , m_H of the fields W_\pm^μ , Z^μ , and the Higgs fields $\eta = \rho - \rho_0$, are given by

$$\begin{aligned} m_W^2 &= \tfrac{1}{2}g^2 \rho_0^2, \\ m_Z/m_W &= 1/\cos \theta_W, \\ m_H &= 2\lambda^{1/2} \rho_0. \end{aligned} \quad (6.36)$$

^a It is more customary to write $\rho/\sqrt{2}$ in place of ρ in (6.32). We simply write ρ to avoid too many factors of 2.

The motivation for constructing a gauge theory, with all masses generated by spontaneous symmetry breaking, is to have a renormalizable quantum field theory.

2 Determination of Constants

The Fermi constant is given by

$$G = \frac{g^2}{2^{5/2} m_W^2} = 1.165 \times 10^{-5} (\text{GeV})^{-2}. \quad (6.37)$$

Substituting g^2 from (6.36) into (6.37), we have

$$m_W^2 = 2^{3/2} \rho_0^2 m_W^2 G. \quad (6.38)$$

Hence^b

$$\begin{aligned} \rho_0^2 &= 2^{-3/2} G^{-1} \\ \rho_0 &= 174 \text{ GeV}. \end{aligned} \quad (6.39)$$

Substituting (6.28) into (6.37), we obtain

$$\begin{aligned} m_W^2 &= \frac{e^2}{2^{5/2} G \sin^2 \theta_W}, \\ m_W &\approx \frac{40}{\sin \theta_W} \text{ GeV}/c^2. \end{aligned} \quad (6.40)$$

The Weinberg angle has been measured experimentally⁴:

$$\sin^2 \theta_W = 0.218 \pm 0.020. \quad (6.41)$$

Using this, we obtain

$$\begin{aligned} m_W &\approx 80 \text{ GeV}/c^2, \\ m_Z &\approx 90 \text{ GeV}/c^2. \end{aligned} \quad (6.42)$$

These are to be compared with the experimental values $m_w = 81 \pm 5 \text{ GeV}/c^2$, [G. Arnison *et al.*, *Phys. Lett.* **122B**, 103 (1983)], and $m_z = 95.2 \pm 2.5 \text{ GeV}/c^2$, [G. Arnison *et al.*, *Phys. Lett.* **126B**, 398 (1983)].

The Higgs-electron coupling is extremely weak:

$$\frac{m_e}{\rho_0} = 3.86 \times 10^{-6}. \quad (6.43)$$

The Higgs mass is not determined, and it depends on the unknown dimensionless parameter λ .

^b The vacuum expectation value of the Higgs field is often quoted as $\langle \phi \rangle = \sqrt{2}\rho_0 = 246 \text{ GeV}$.

⁴ I. Liede and M. Roos, *Nucl. Phys.* **B167**, 397 (1980).

3 Interactions

Rewrite (6.35) in unitary gauge as follows:

$$\mathcal{L} = \mathcal{L}_V + \mathcal{L}_F + \mathcal{L}_H + \mathcal{L}' + \mathcal{L}'', \quad (6.44)$$

where the first three terms denote the “free” Lagrangian density of the vector field, Fermi field, and Higgs field respectively:

$$\begin{aligned}\mathcal{L}_V &= -\frac{1}{4}(\mathbf{G} \cdot \mathbf{G} + \mathbf{H} \cdot \mathbf{H}) + \frac{1}{2}m_W^2(W_1^2 + W_2^2) + \frac{1}{2}m_Z^2Z^2, \\ \mathcal{L}_F &= \bar{L}i\cancel{\partial}L + \bar{R}i\cancel{\partial}R - m_e(\bar{L}R + \bar{R}L), \\ \mathcal{L}_H &= \partial_\mu\eta\partial^\mu\eta - m_H^2\eta\left(1 + \frac{\eta}{2\rho_0}\right)^2,\end{aligned}\quad (6.45)$$

where η is the real Higgs field in unitary gauge, in which $\phi(x)$ has the form (6.32), and

$$\rho(x) = \rho_0 + \eta(x). \quad (6.46)$$

The terms \mathcal{L}' and \mathcal{L}'' are interaction terms, with \mathcal{L}' containing the electromagnetic and weak currents, and \mathcal{L}'' containing the interactions between the Higgs field and other fields:

$$\mathcal{L}'' = [m_W^2(W_1^2 + W_2^2) + m_Z^2Z^2]\frac{\eta}{\rho_0}\left(1 + \frac{\eta}{2\rho_0}\right) + m(\bar{e}e)\frac{\eta}{\rho_0}, \quad (6.47)$$

$$\begin{aligned}\mathcal{L}' &= g\bar{\psi}(W_1t_1 + W_2t_2)\psi + e\bar{\psi}Q\cancel{A}\psi + e\bar{\psi}Q'\cancel{Z}\psi \\ &= \frac{g}{2}(W_1^\mu J_{1\mu}^{ch} + W_2^\mu J_{2\mu}^{ch}) + eA^\mu J_\mu^{em} + eZ^\mu J_\mu^{neut},\end{aligned}\quad (6.48)$$

where

$$\begin{aligned}J_{i\mu}^{ch} &= \bar{L}\tau_i\gamma_\mu L \quad (i = 1, 2), \\ J_\mu^{em} &= \bar{\psi}Q\gamma_\mu\psi, \\ J_\mu^{neut} &= \bar{\psi}Q'\gamma_\mu\psi.\end{aligned}\quad (6.49)$$

The “free” Lagrangian for the gauge fields is, again,

$$\mathcal{L}_V = -\frac{1}{4}(\mathbf{G} \cdot \mathbf{G} + \mathbf{H} \cdot \mathbf{H}) + \frac{1}{2}m_W^2(W_1^2 + W_2^2) + \frac{1}{2}m_Z^2Z^2. \quad (6.50)$$

To write this out more explicitly, we use the following notation:

$$\begin{aligned}(A \times B)^{\mu\nu} &\equiv A^\mu B^\nu - A^\nu B^\mu, \\ (A \times B)^2 &\equiv (A \times B)^{\mu\nu}(A \times B)_{\mu\nu}, \\ (A \times B) \cdot (C \times D) &\equiv (A \times B)^{\mu\nu}(C \times D)_{\mu\nu}.\end{aligned}\quad (6.51)$$

Then

$$\begin{aligned}\mathcal{L}_V &= -\frac{1}{4}[(\partial \times W_1)^2 + (\partial \times W_2)^2 + (\partial \times A)^2 + (\partial \times Z)^2] \\ &\quad + \frac{1}{2}m_W^2(W_1^2 + W_2^2) + \frac{1}{2}m_Z^2Z^2 \\ &\quad + \frac{1}{2}g[(\partial \times W_1) \cdot (W_2 \times W_3) + (\partial \times W_2) \cdot (W_3 \times W_1) \\ &\quad + (\partial \times W_3) \cdot (W_1 \times W_2)] \\ &\quad - \frac{1}{4}g^2[(W_1 \times W_2)^2 + (W_2 \times W_3)^2 + (W_3 \times W_1)^2],\end{aligned}\quad (6.52)$$

where W_3^μ is given by (6.24). The terms of orders g and g^2 describe interactions among the gauge fields, which arise from the non-Abelian structure of the gauge group. Note that Z^μ does not interact directly with A^μ , as we would expect of a neutral field.

The charged gauge fields W_1^μ , W_2^μ have electromagnetic interactions. The part of (6.52) containing these interactions can be written in the form

$$\begin{aligned}\mathcal{L}_W^{\text{em}} = & e W_1^\mu W_2^\nu F_{\mu\nu} + e[(\partial \times W_1)^{\mu\nu} W_{2\mu} - (\partial \times W_2)^{\mu\nu} W_{1\mu}] A_\nu \\ & + \frac{1}{2} e^2 [W_1^\mu W_2^\nu + W_2^\mu W_1^\nu - g^{\mu\nu} (W_1^2 + W_2^2)] A_\mu A_\nu,\end{aligned}\quad (6.53)$$

where $F^{\mu\nu} \equiv \partial^\mu A^\nu - \partial^\nu A^\mu$. In terms of the fields with definite charge:

$$\begin{aligned}W &= 2^{-1/2} (W_1 + iW_2), \\ W^* &= 2^{-1/2} (W_1 - iW_2),\end{aligned}\quad (6.54)$$

we can rewrite (6.53) as

$$\begin{aligned}\mathcal{L}_W^{\text{em}} = & -ie(W_\nu^* \overleftrightarrow{\partial}_\mu W^\nu) A^\mu - ie W^*{}^\mu W^\nu F_{\mu\nu} \\ & - ie(W^*{}^\mu W^\nu - \text{c.c.}) A_\nu \\ & + e^2 (W^*{}^\mu W^\nu - g^{\mu\nu} W^*{}^\lambda W_\lambda) A_\mu A_\nu.\end{aligned}\quad (6.55)$$

The second term gives rise to magnetic moments and electric quadrupole moments⁵. In general, a term of the form $\mathcal{L}_W^{\text{em}} = -ie\kappa W_\mu^* W_\nu F^{\mu\nu}$ gives rise to

$$\text{Magnetic dipole moment} = (1 + \kappa) \frac{e}{2m_W} \mathbf{s},$$

$$\text{Electric quadrupole moment} = \int d^3x \rho (3z^2 - r^2) = -\frac{e\kappa}{m_W^2}, \quad (6.56)$$

where \mathbf{s} is the spin vector, and ρ is the static electric charge density in the state $s_z = 1$. Since $\kappa = 1$, the g -factor of the W boson is equal to 2.

The complicated structure of (6.55) actually serves a simple purpose; when (6.55) is added to the kinetic term for the W field, we obtain the Lagrangian density for a charged vector theory:

$$\begin{aligned}& -\frac{1}{4}[(\partial \times W_1)^2 + (\partial \times W_2)^2] + \mathcal{L}_W^{\text{em}} \\ & = -\frac{1}{2}(D \times W)^* \cdot (D \times W) - ie W^*{}^\mu W^\nu F_{\mu\nu}, \\ D^\mu &= \partial^\mu - ie A^\mu.\end{aligned}$$

6.3 The General Theory

1 Mass Terms

To include all the quarks and leptons in the three families as described in Chapter 1, we make two modifications in (6.22).

First, the definitions of R and L are extended to include all quarks and leptons:

⁵ T. D. Lee and C. N. Yang, *Phys. Rev.* **128**, 885 (1962).

$$\begin{aligned} L &= \{l_{L\alpha}, q_{L\alpha}\} \quad (\alpha = 1, 2, 3), \\ R &= \{l_{R\alpha}, q_{R\alpha}\} \quad (\alpha = 1, 2, 3). \end{aligned} \quad (6.57)$$

where the symbols are defined in Table 6.2. Secondly, the mass terms are modified to enable us to assign arbitrary masses. The locally gauge invariant Lagrangian is

$$\begin{aligned} \mathcal{L} = & -\frac{1}{4}(\mathbf{G} \cdot \mathbf{G} + \mathbf{H} \cdot \mathbf{H}) + \bar{L}i\not{\partial}L + \bar{R}i\not{\partial}R + (D\phi)^\dagger(D\phi) \\ & - V(\phi^*\phi) + \mathcal{L}_{\text{mass}}. \end{aligned} \quad (6.58)$$

The mass term is further split into lepton and quark contributions:

$$\mathcal{L}_{\text{mass}} = \mathcal{L}_{\text{mass}}^{\text{lept}} + \mathcal{L}_{\text{mass}}^{\text{quark}}, \quad (6.59)$$

Table 6.2 QUANTUM NUMBERS

Subscripts L, R denote left- and right-handed fields, respectively. A tilde over quark symbols denote generalized Cabibbo mixed states.

The Weinberg angle is denoted by θ ($\sin^2\theta \cong 1/4$).

Electric charge = eQ , $Q = t_3 + t_0$

Neutral charge = eQ' , $Q' = t_3 \cot\theta - t_0 \tan\theta$

Particles			t	t_3	t_0	Q	$Q' \sin 2\theta$	
Leptons	ν_L	ν'_L	ν''_L	$\frac{1}{2}$	$\frac{1}{2}$	$-\frac{1}{2}$	0	1
$l_{L\alpha}$:	e_L	μ_L	τ_L	$\frac{1}{2}$	$-\frac{1}{2}$	$-\frac{1}{2}$	-1	$-1 + 2 \sin^2 \theta$
$l_{R\alpha}$:	e_R	μ_R	τ_R	0	0	-1	-1	$2 \sin^2 \theta$
Quarks	\tilde{u}_L	\tilde{c}_L	\tilde{t}_L	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{6}$	$\frac{2}{3}$	$1 - \frac{4}{3} \sin^2 \theta$
$q_{L\alpha}$:	d_L	s_L	b_L	$\frac{1}{2}$	$-\frac{1}{2}$	$\frac{1}{6}$	$-\frac{1}{3}$	$-1 + \frac{2}{3} \sin^2 \theta$
$\tilde{q}_{R\alpha}$:	u_R	c_R	t_R	0	0	$\frac{2}{3}$	$\frac{2}{3}$	$-\frac{4}{3} \sin^2 \theta$
$q_{R\alpha}$:	d_R	s_R	b_R	0	0	$-\frac{1}{3}$	$-\frac{1}{3}$	$\frac{2}{3} \sin^2 \theta$
Higgs	ϕ_+			$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	1	$1 - 2 \sin^2 \theta$
	ϕ_0			$\frac{1}{2}$	$-\frac{1}{2}$	$\frac{1}{2}$	0	-1

with the lepton contribution given by

$$\mathcal{L}_{\text{mass}}^{\text{lept}} = \sum_{\alpha, \beta=1}^3 \bar{l}_{L\alpha} \frac{\phi}{\rho_0} m_{\alpha\beta} l_{R\beta} + \text{c.c.}, \quad (6.60)$$

where $m_{\alpha\beta}$ is an arbitrary constant complex mass matrix. In unitary gauge, in which ϕ is given by (6.32), we have

$$\mathcal{L}_{\text{mass}}^{\text{lept}} = \frac{\rho}{\rho_0} (\bar{e}_L \quad \bar{\mu}_L \quad \bar{\tau}_L) m \begin{pmatrix} e_R \\ \mu_R \\ \tau_R \end{pmatrix} + \text{c.c.} \quad (6.61)$$

Now, an arbitrary complex matrix can be brought to diagonal form with real non-negative diagonal elements. That is, there exists non-singular matrices A and B , such that

$$AmB^{-1} = D, \quad (6.62)$$

where D is diagonal, with non-negative diagonal elements. To prove this, note that $m^\dagger m$ is hermitian, with real non-negative eigenvalues, and the same is true of mm^\dagger . The eigenvalues of $m^\dagger m$ are the same as those of mm^\dagger . Therefore, there exist non-singular A and B such that

$$\begin{aligned} Amm^\dagger A^{-1} &= D^2, \\ Bm^\dagger mB^{-1} &= D^2. \end{aligned} \quad (6.63)$$

The solutions to these equations are

$$\begin{aligned} m &= A^{-1}DB, \\ m^\dagger &= B^{-1}DA. \end{aligned} \quad (6.64)$$

Thus, we make independent linear transformations on right- and left-handed particles:

$$\begin{pmatrix} e_R \\ \mu_R \\ \tau_R \end{pmatrix} \rightarrow B^{-1} \begin{pmatrix} e_R \\ \mu_R \\ \tau_R \end{pmatrix}, \quad (6.65)$$

$$\begin{pmatrix} e_L \\ \mu_L \\ \tau_L \end{pmatrix} \rightarrow A^{-1} \begin{pmatrix} e_L \\ \mu_L \\ \tau_L \end{pmatrix}.$$

This makes the mass matrix diagonal:

$$\mathcal{L}_{\text{mass}}^{\text{lept}} = (\bar{e} \quad \bar{\mu} \quad \bar{\tau}) \begin{pmatrix} m_e & 0 & 0 \\ 0 & m_\mu & 0 \\ 0 & 0 & m_\tau \end{pmatrix} \begin{pmatrix} e \\ \mu \\ \tau \end{pmatrix} \frac{\rho}{\rho_0}, \quad (6.66)$$

where m_e, m_μ, m_τ are arbitrary non-negative real parameters. Note that the transformations (6.65) mix leptons of different families. The same transformations, of course, must also be made in the kinetic part of \mathcal{L} , i.e., in the terms $(\bar{L}i\cancel{D}L + \bar{R}i\cancel{D}R)$ in (6.58). The effect is to transform the charge-changing current in the following way:

$$(\bar{e}_L \quad \bar{\mu}_L \quad \bar{\tau}_L) \gamma^\mu \begin{pmatrix} \nu_L \\ \nu'_L \\ \nu''_L \end{pmatrix} \rightarrow (\bar{e}_L \quad \bar{\mu}_L \quad \bar{\tau}_L) \gamma^\mu A \begin{pmatrix} \nu_L \\ \nu'_L \\ \nu''_L \end{pmatrix}. \quad (6.67)$$

The non-charge-changing currents remain invariant, because they are of the form $\bar{L}\gamma^\mu L + \bar{R}\gamma^\mu R$.

The symbols e, μ, τ in (6.66) and in (6.67) denote mass eigenstates, because the mass matrix is diagonal in the corresponding basis. Therefore, the physical neutrinos are

$$\begin{pmatrix} \tilde{\nu} \\ \tilde{\nu}' \\ \tilde{\nu}'' \end{pmatrix}_L = A \begin{pmatrix} \nu \\ \nu' \\ \nu'' \end{pmatrix}_L. \quad (6.68)$$

But this makes no difference, since the neutrinos are massless. Hence we can drop the tildes over the ν 's. In the Weinberg-Salam model, there is no mixing of physical leptons of different families, when all neutrinos are assumed to be massless. For example, $\mu \rightarrow e + \gamma$ is forbidden.

We now turn to quark masses. The quarks denoted by q_R and \bar{q}_R in Table 6.2 are to be coupled to the Higgs doublet ϕ and the conjugation doublet $\tilde{\phi}$ respectively, to give a mass term invariant under weak hypercharge, as follows:

$$\mathcal{L}_{\text{mass}}^{\text{quark}} = \sum_{\alpha, \beta=1}^3 \left(\bar{q}_{L\alpha} \frac{\phi}{\rho_0} M_{\alpha\beta} q_{R\beta} + \bar{q}_{L\alpha} \frac{\tilde{\phi}}{\rho_0} \tilde{M}_{\alpha\beta} \tilde{q}_{R\beta} \right) + \text{c.c.}, \quad (6.69)$$

where $\tilde{\phi}$ is defined in (6.15), and M and \tilde{M} are arbitrary constant complex matrices. *A color sum over each flavor of quarks is understood.* In unitary gauge, this reads

$$\mathcal{L}_{\text{mass}}^{\text{quark}} = \frac{\rho}{\rho_0} \left[(d_L \quad \bar{s}_L \quad \bar{b}_L) M \begin{pmatrix} d_R \\ s_R \\ b_R \end{pmatrix} + (\bar{u}_L \quad \bar{c}_L \quad \bar{t}_L) \tilde{M} \begin{pmatrix} u_R \\ c_R \\ t_R \end{pmatrix} \right] + \text{c.c.} \quad (6.70)$$

We diagonalize M and \tilde{M} by making the transformations

$$\begin{aligned} \begin{pmatrix} d \\ s \\ b \end{pmatrix}_L &\rightarrow A_L^{-1} \begin{pmatrix} d \\ s \\ b \end{pmatrix}_L, & \begin{pmatrix} u \\ c \\ t \end{pmatrix}_L &\rightarrow B_L^{-1} \begin{pmatrix} u \\ c \\ t \end{pmatrix}_L, \\ \begin{pmatrix} d \\ s \\ b \end{pmatrix}_R &\rightarrow A_R^{-1} \begin{pmatrix} d \\ s \\ b \end{pmatrix}_R, & \begin{pmatrix} u \\ c \\ t \end{pmatrix}_R &\rightarrow B_R^{-1} \begin{pmatrix} u \\ c \\ t \end{pmatrix}_R, \end{aligned} \quad (6.71)$$

where A_L, A_R, B_L, B_R are such that

$$A_L M A_R^{-1} = \begin{pmatrix} m_d & 0 & 0 \\ 0 & m_s & 0 \\ 0 & 0 & m_b \end{pmatrix}, \quad (6.72)$$

$$B_L \tilde{M} B_R^{-1} = \begin{pmatrix} m_u & 0 & 0 \\ 0 & m_c & 0 \\ 0 & 0 & m_t \end{pmatrix},$$

where the diagonal elements are arbitrary real non-negative parameters. With this, we have

$$\mathcal{L}_{\text{mass}}^{\text{quark}} = \frac{\rho}{\rho_0} \sum_q m_q \bar{q} q, \quad (6.73)$$

where q denotes the 4-component Dirac spinor for a quark, m_q its mass parameter, and the sum extends over all six flavors u, d, s, c, b, t (plus a color sum for each flavor). These quarks are now mass eigenstates.

The transformation (6.71) in the kinetic part of the Lagrangian density leaves the electromagnetic and neutral currents invariant, but gives a charge-changing quark current of the form

$$J_\mu^{\text{ch}} \propto (\bar{d} \quad \bar{s} \quad \bar{b})_L \gamma_\mu C \begin{pmatrix} u \\ c \\ t \end{pmatrix}_L, \quad (6.74)$$

where the quark symbols refer to mass eigenstates, and C is the 3×3 matrix

$$C = A_L B_R^{-1}, \quad (6.75)$$

with the properties

$$C^\dagger C = 1, \quad |\det C|^2 = 1. \quad (6.76)$$

Since quarks are supposed to be confined by virtue of their strong interactions, the meaning of the quark mass is not obvious. We shall take this up when we discuss quantum chromodynamics, in Chapter 12.

2 Cabibbo Angle

To understand the effect of the matrix C in (6.74), let us first ignore the b and t quarks, so that

$$J_\mu^{\text{ch}} \propto (\bar{d} \quad \bar{s})_L \gamma_\mu C \begin{pmatrix} u \\ c \end{pmatrix}_L. \quad (6.77)$$

The most general form of the 2×2 matrix C is then

$$C = \begin{pmatrix} e^{i\theta} & 0 \\ 0 & 1 \end{pmatrix} \times [\text{SU}(2) \text{ matrix}]. \quad (6.78)$$

The first factor may be ignored, as it can be absorbed into a redefinition of the phases of u and c . The most general form of an $\text{SU}(2)$ matrix corresponds to a rotation through the Euler angles α, β, γ :

$$C = e^{i\gamma\tau_3/2} e^{i\beta\tau_2/2} e^{-i\alpha\tau_3/2}. \quad (6.79)$$

Again, the first and last factor can be absorbed into redefinitions of the phases of (d, s) and (u, c) respectively. Therefore, we may take $C = e^{-i\beta\tau_2/2}$. Putting $\theta = \beta/2$, we have

$$C = \frac{d}{s} \begin{pmatrix} u & c \\ \cos \theta & \sin \theta \\ -\sin \theta & \cos \theta \end{pmatrix}, \quad (6.80)$$

where θ is an arbitrary angle to be determined by experiments. This is the Cabibbo angle, as discussed earlier in Chapter 1, with the experimental value

$$\theta \approx \frac{1}{4}. \quad (6.81)$$

Since C is real, the Lagrangian density remains real, and is invariant under time-reversal. Through the CPT theorem, it is therefore invariant under CP.

3 Kobayashi-Maskawa Matrix

The most general form of C for the 3×3 case was first worked out by Kobayashi and Maskawa⁶. By (6.76), C can be reduced to an $SU(3)$ matrix, by redefining the phases of the quark fields. Hence it is a sum of products of the matrices $e^{i\theta_1\lambda_1}, \dots, e^{i\theta_8\lambda_8}$, where $\lambda_1, \dots, \lambda_8$ are the Gell-Mann matrices, of which λ_3 and λ_8 are chosen to be diagonal. Therefore, there are 8 free parameters $\theta_1, \dots, \theta_8$. We can write C in the form

$$C = e^{i\lambda_3\theta_3} e^{i\lambda_8\theta_8} U e^{i\lambda_3\theta_3'} e^{i\lambda_8\theta_8'}, \quad (6.82)$$

where U is an $SU(3)$ matrix containing only 4 parameters, and consequently can be constructed from any 4 generators, excluding λ_3 and λ_8 . For convenience we choose these to be $\lambda_1, \lambda_2, \lambda_5, \lambda_7$. We note that $\lambda_2, \lambda_5, \lambda_7$ generate a rotation group, because $[\lambda_2, \lambda_5] = i\lambda_7$. Therefore, U can be constructed from 3 orthogonal matrices with 3 arbitrary parameters, and one unitary matrix of determinant 1, with one arbitrary parameter:

$$C = R_2 \bar{U} R_1 R_3 \quad (6.83)$$

where

$$\bar{U} = \begin{pmatrix} e^{-i\delta/3} & 0 & 0 \\ 0 & e^{-i\delta/3} & 0 \\ 0 & 0 & e^{2i\delta/3} \end{pmatrix} = e^{-i\delta/3} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & e^{i\delta} \end{pmatrix}, \quad (6.84)$$

and

$$R_1 = \begin{pmatrix} c_1 & s_1 & 0 \\ -s_1 & c_1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad R_2 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & c_2 & s_2 \\ 0 & -s_2 & c_2 \end{pmatrix}, \quad R_3 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & c_3 & s_3 \\ 0 & -s_3 & c_3 \end{pmatrix}, \quad (6.85)$$

$$c_i = \cos \theta_i, \quad s_i = \sin \theta_i \quad (i = 1, 2, 3).$$

The overall phase $e^{-i\delta/3}$ in (6.84) may be absorbed into the quark fields.

⁶ M. Kobayashi and K. Maskawa, *Prog. Theo. Phys.* **49**, 652 (1975).

Therefore

$$\begin{aligned}
 C &= \begin{pmatrix} 1 & 0 & 0 \\ 0 & c_2 & s_2 \\ 0 & -s_2 & c_2 \end{pmatrix} \begin{pmatrix} c_1 & s_1 & 0 \\ -s_1 & c_1 & 0 \\ 0 & 0 & e^{i\delta} \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & c_3 & s_3 \\ 0 & -s_3 & c_3 \end{pmatrix} \\
 &\quad u \qquad \qquad \qquad c \qquad \qquad \qquad t \\
 &= s \begin{pmatrix} c_1 & s_1 c_3 & s_1 s_3 \\ -s_1 c_2 & c_1 c_2 c_3 - s_2 s_3 e^{i\delta} & c_1 c_2 c_3 + s_2 c_3 e^{i\delta} \\ s_1 s_2 & -c_1 s_2 c_3 - c_2 s_3 e^{i\delta} & -c_1 s_2 c_3 + c_2 c_3 e^{i\delta} \end{pmatrix}. \quad (6.86)
 \end{aligned}$$

In the literature, one may find (c_2, s_2) interchanged with $(c_3, \pm s_3)$, and $-\delta$ instead of δ . There are 3 arbitrary angles $\theta_1, \theta_2, \theta_3$, of which θ_1 is the Cabibbo angle. There is an arbitrary phase δ , which makes the Lagrangian density non-real. Therefore, a non-zero value of δ violates time-reversal invariance. By the CPT theorem, it violates CP invariance^c. The only parameters determined with precision from experimental data are:⁷

$$\begin{aligned}
 |c_1| &= 0.9737 + 0.0025, \\
 |s_1 c_2| &= 0.219 \pm 0.003. \quad (6.87)
 \end{aligned}$$

The parameters θ_3 and δ are essentially unknown, except for bounds obtained from a combination of experimental data and theoretical hypotheses.⁸

We summarize the Weinberg-Salam model by displaying the complete Lagrangian density in Table 6.3, which is self-explanatory.

4 Solitons

There are no static topological solitons in the Weinberg-Salam model. The gauge group is $G = SU(2) \times U(1)$, with elements of the form

$$g \in G, \quad g = e^{-i\omega \cdot t} e^{-i\nu t_0}.$$

The little group is $H = [U(1)]_{em}$, with elements

$$h \in H, \quad h = e^{-i\theta Q} = e^{-i\theta(t_3 + t_0)}.$$

It is clear that H is not a normal subgroup of G . A coset with respect to H is, for fixed ω and ν ,

$$gH = \{e^{-i\omega \cdot t} e^{-i\nu t_0} e^{-i\theta Q} \mid -\infty < \theta < \infty\}.$$

Parametrize $e^{-i\omega \cdot t}$ by Euler angles α, β, γ :

$$e^{-i\omega \cdot t} = e^{-i\alpha t_3} e^{-i\beta t_2} e^{-i\gamma t_3}.$$

Using $t_0 = Q - t_3$, we have

$$e^{-i\omega \cdot t} e^{-i\nu t_0} e^{-i\theta Q} = e^{-i\alpha t_3} e^{-i\beta t_2} e^{-i(\nu - \theta)t_3} e^{-i(\nu - \theta)Q}.$$

^c It is difficult to put a bound on δ by the magnitude of the observed CP violation in K^0 decay, because δ appears only in flavor-mixed components involving the heavy quarks c and t . In K^0 decay, it can contribute only through the virtual effects of the heavy quarks. Thus, the observed smallness of CP violation does not necessarily imply that δ is small.

⁷ R. E. Shrock and L. L. Wang, *Phys. Rev. Lett.* **41**, 1692 (1978).

⁸ R. E. Shrock, S. B. Treiman and L. L. Wang, *Phys. Rev. Lett.* **42**, 1589 (1979).

Noting that $\nu - \theta$ and $\nu + \theta$ are independent parameters, we can write the set of cosets as

$$gH = \{e^{-i\omega' \cdot t} e^{-i\theta' Q} | -\infty < \theta' < \infty\},$$

$$G/H = \{e^{-i\omega' \cdot t}\}.$$

Since G/H has the topology of $SU(2)$, $\pi_2(G/H) = 0$.

Semi-classical non-topological solitons do exist. A class of these consists of a torus in space in which the Higgs field is expelled by one quantum of the Z field. These have been called "vorticons", and their masses estimated to be of the order of $3000 \text{ GeV}/c^2$.⁹

Table 6.3 LAGRANGIAN DENSITY OF THE WEINBERG-SALAM MODEL

$\mathcal{L} = \mathcal{L}_V + \mathcal{L}_F + \mathcal{L}_H + \mathcal{L}_{int}$	[\mathcal{L}_V given in (6.50) and (6.52)]
$\mathcal{L}_F = \bar{\psi}(i\cancel{p} - m)\psi$	
$\mathcal{L}_H = \partial\eta \cdot \partial\eta - m_H^{-2}\eta^2$	
$\mathcal{L}_{int} = \mathcal{L}_{VV} + \mathcal{L}_{VF} + \mathcal{L}_{HV} + \mathcal{L}_{HF} + \mathcal{L}_{HH}$	
$\mathcal{L}_{VV} = \frac{1}{2}g[(\partial \times \mathbf{W}_1) \cdot (\mathbf{W}_2 \times \mathbf{W}_3) + (\partial \times \mathbf{W}_2) \cdot (\mathbf{W}_3 \times \mathbf{W}_1) + (\partial \times \mathbf{W}_3) \cdot (\mathbf{W}_1 \times \mathbf{W}_2)]$	
$- \frac{1}{4}g^2[(\mathbf{W}_1 \times \mathbf{W}_2)^2 + (\mathbf{W}_2 \times \mathbf{W}_3)^2 + (\mathbf{W}_3 \times \mathbf{W}_1)^2]$	
$\mathcal{L}_{VF} = \frac{1}{2}g(W_1 \cdot J_1^{ch} + W_2 \cdot J_2^{ch}) + eA \cdot J^{em} + \frac{g}{\cos \theta_W} Z \cdot J^{neut}$	
$J_\mu^{ch} = \bar{l}_L \tau_i \gamma_\mu l_L + \bar{q}_L \tau_i \gamma_\mu C q_L$	
$= \bar{\psi} \tau_i \gamma_\mu \frac{1 - \gamma_5}{2} C \psi$	
$J_\mu^{em} = \bar{\psi} Q \gamma_\mu \psi$	
$J_\mu^{neut} = \bar{\psi} Q' \gamma_\mu \psi$	
$\mathcal{L}_{HV} = [m_W^2(W_1^2 + W_2^2) + m_Z^2 Z^2] \frac{\eta}{\rho_0} \left(1 + \frac{\eta}{2\rho_0}\right)$	
$\mathcal{L}_{HF} = (\bar{\psi} M \psi) \frac{\eta}{\rho_0}$	
$\mathcal{L}_{HH} = -\frac{1}{6} \rho_0 \lambda \eta^3 - \frac{1}{24} \lambda \eta^4$	

⁹ K. Huang and R. Tipton, *Phys. Rev.* **23**, 3050 (1981).

CHAPTER 7

METHOD OF PATH INTEGRALS

7.1 Non-Relativistic Quantum Mechanics

Feynman's method of path integrals¹ is a method of formulating a quantum theory. It is equivalent to the more familiar method of canonical quantization, but much more convenient to use for gauge theories. We introduce the method by first considering non-relativistic quantum mechanics with one degree of freedom. Generalizations will then become obvious.

In the method of canonical quantization, we work with Hilbert-space operators q_{op} and p_{op} , representing respectively the coordinate and the conjugate momentum, as defined by the commutation relation^a

$$[p_{\text{op}}, q_{\text{op}}] = -i\hbar. \quad (7.1)$$

We denote the eigenstates of these operators by $|q\rangle$ and $|p\rangle$:

$$\begin{aligned} q_{\text{op}}|q'\rangle &= q'|q'\rangle, \\ p_{\text{op}}|p'\rangle &= p'|p'\rangle, \end{aligned} \quad (7.2)$$

and normalize them according to

$$\begin{aligned} \langle q''|q'\rangle &= \delta(q'' - q'), & \int_{-\infty}^{\infty} dq' |q'\rangle\langle q'| &= 1, \\ \langle p''|p'\rangle &= 2\pi\hbar\delta(p'' - p'), & \int_{-\infty}^{\infty} \frac{dp'}{2\pi\hbar} |p'\rangle\langle p'| &= 1, \\ \langle p'|q'\rangle &= e^{-ip'q'/\hbar}. \end{aligned} \quad (7.3)$$

The dynamics of the system is completely specified by the transition amplitude governing the time evolution of the system:

$$\langle q'', t''|q', t'\rangle = \left\langle q'' \left| \exp \left[-\frac{i}{\hbar} (t'' - t') H_{\text{op}} \right] \right| q' \right\rangle, \quad (7.4)$$

where H_{op} is the time-independent Hamiltonian operator. We may regard $|q', t'\rangle$ as the eigenstate of the Heisenberg operator $q_{\text{op}}(t)$ with the eigenvalue q' :

$$\begin{aligned} q_{\text{op}}(t) &= e^{itH_{\text{op}}/\hbar} q_{\text{op}} e^{-itH_{\text{op}}/\hbar}, \\ q_{\text{op}}(t')|q', t'\rangle &= q' |q', t'\rangle. \end{aligned} \quad (7.5)$$

^a In this chapter we do not set $\hbar = 1$.

¹ R. P. Feynman and A. R. Hibbs, *Quantum Mechanics and Path Integrals* (McGraw-Hill, New York, 1965).

The time evolution is then given by the equation

$$|q', t'\rangle = e^{it' H_{\text{op}}/\hbar} |q'\rangle. \quad (7.6)$$

The object of the method of path integrals is to express the transition amplitude entirely in terms of a classical Hamiltonian $H(p, q)$, without reference to operators and states in Hilbert space.

To proceed, we divide the time evolution interval $t'' - t'$ into N equal steps, and take the limit $N \rightarrow \infty$ later. Let

$$\begin{aligned} \Delta t &\equiv (t'' - t')/N, \\ \varepsilon &\equiv \Delta t/\hbar. \end{aligned} \quad (7.7)$$

We can write

$$\begin{aligned} \langle q'', t'' | q', t' \rangle &= \langle q'' | e^{-iH_{\text{op}}N\varepsilon} | q' \rangle = \langle q'' | (1 - i\varepsilon H_{\text{op}})^N | q' \rangle \\ &= \int dq_1 \cdots dq_{N-1} \langle q'' | (1 - i\varepsilon H_{\text{op}}) | q_{N-1} \rangle \cdots \langle q_1 | (1 - i\varepsilon H_{\text{op}}) | q' \rangle. \end{aligned} \quad (7.8)$$

Next we rewrite a typical factor in the integrand above as

$$\langle q_2 | (1 - i\varepsilon H_{\text{op}}) | q_1 \rangle = \int_{-\infty}^{\infty} \frac{dp_1}{2\pi\hbar} \langle q_2 | p_1 \rangle \langle p_1 | (1 - i\varepsilon H_{\text{op}}) | q_1 \rangle, \quad (7.9)$$

and define the classical Hamiltonian $H(p, q)$ by

$$\langle p | H_{\text{op}} | q \rangle \equiv \langle p | q \rangle H(p, q). \quad (7.10)$$

This definition gives the usual connection between the classical and quantum mechanical Hamiltonian, provided $H(p, q)$ does not contain cross-products of p and q . Otherwise, H_{op} must be normal ordered, such that all p_{op} 's appear to the left of all q_{op} 's. Using (7.10), we obtain

$$\begin{aligned} \langle q_2 | (1 - i\varepsilon H_{\text{op}}) | q_1 \rangle &= \int_{-\infty}^{\infty} \frac{dp_1}{2\pi\hbar} \langle q_2 | p_1 \rangle \langle p_1 | q_1 \rangle [1 - i\varepsilon H(p_1, q_1)] \\ &= \int_{-\infty}^{\infty} \frac{dp_1}{2\pi\hbar} e^{ip_1(q_2 - q_1)/\hbar} [1 - i\varepsilon H(p_1, q_1)]. \end{aligned} \quad (7.11)$$

Hence

$$\begin{aligned} \langle q'', t'' | q', t' \rangle &= \int \frac{dq_1 dp_1}{2\pi\hbar} \cdots \int \frac{dq_{N-1} dp_{N-1}}{2\pi\hbar} \frac{dp_0}{2\pi\hbar} \\ &\times \exp \left[\frac{i}{\hbar} \sum_{n=0}^{N-1} p_n (q_{n+1} - q_n) \right] \prod_{n=0}^{N-1} [1 - i\varepsilon H(p_n, q_n)], \end{aligned} \quad (7.12)$$

with the conditions

$$q_0 \equiv q', \quad q_N \equiv q''. \quad (7.13)$$

Now comes the key step in the development, we note that in (7.12) the factor $(1 - i\epsilon H)$ may be replaced by $\exp(-i\epsilon H)$. The reason is that, given N number z_1, \dots, z_n such that $X = \lim_{N \rightarrow \infty} N^{-1} \sum z_n$ exists, we have

$$\lim_{N \rightarrow \infty} \prod_{n=1}^N \left(1 + \frac{z_n}{N}\right) = \lim_{N \rightarrow \infty} \prod_{n=1}^N e^{z_n/N} = e^X, \quad (7.14)$$

which can be proved by power series expansion in z_n . The advantage of replacing $(1 - i\epsilon H)$ by $\exp(-i\epsilon H)$ is that we can now express the amplitude (7.12) as integrals over unitary amplitudes:

$$\begin{aligned} \langle q'', t'' | q', t' \rangle &= \int \frac{dq_1 dp_1}{2\pi\hbar} \dots \int \frac{dq_{N-1} dp_{N-1}}{2\pi\hbar} \frac{dp_0}{2\pi\hbar} \\ &\times \exp \left\{ \frac{i}{\hbar} \Delta t \sum_{n=0}^{N-1} \left[\frac{p_n(q_{n+1} - q_n)}{\Delta t} - H(p_n, q_n) \right] \right\} \end{aligned} \quad (7.15)$$

To approach the limit $N \rightarrow \infty$ (or $\Delta t \rightarrow 0$), we adopt the view that the set of values $\{q_1, p_1, \dots, q_{N-1}, p_{N-1}\}$ are successive values of certain functions $q(t)$, $p(t)$, which may be discontinuous functions. Accordingly, we use the notation

$$\begin{aligned} t_n &= t' + n \Delta t, \\ q_n &= q(t_n), \\ p_n &= p(t_n), \end{aligned} \quad (7.16)$$

and write

$$\begin{aligned} (q_{n+1} - q_n)/\Delta t &\xrightarrow[\Delta t \rightarrow 0]{} \dot{q}(t_n), \\ \sum_{n=0}^{N-1} f(t_n) \Delta t &\xrightarrow[\Delta t \rightarrow 0]{} \int_{t'}^{t''} dt f(t). \end{aligned} \quad (7.17)$$

Thus, we can rewrite the transition amplitude in the form

$$\begin{aligned} \langle q'', t'' | q', t' \rangle &= \int (Dq)(Dp) \exp \frac{i}{\hbar} \int_{t'}^{t''} dt [p\dot{q} - H(p, q)], \\ q(t') &= q', \quad q(t'') = q''. \end{aligned} \quad (7.18)$$

This represents an integral over all paths $p(t)$ in momentum space, and all paths $q(t)$ in coordinate space, between the times t' and t'' with fixed values of the coordinates at the endpoints. The volume elements in path space are denoted by

$$(Dq) = \prod_{n=1}^{N-1} dq(t_n), \quad (Dp) = \prod_{n=0}^{N-1} \frac{dp(t_n)}{2\pi\hbar}. \quad (7.19)$$

The generalization of (7.18) to more than one degree of freedom is immediate:

$$\begin{aligned} \langle q_1'', \dots, q_n''; t'' | q_1', \dots, q_n'; t' \rangle \\ = \int \prod_{\alpha} (Dq_{\alpha})(Dp_{\alpha}) \exp \frac{i}{\hbar} \int_{t'}^{t''} dt \left[\sum_{\alpha} p_{\alpha} \dot{q}_{\alpha} - H(p, q) \right]. \end{aligned} \quad (7.20)$$

Feynman's formula for the transition amplitude is derived from (7.18) by restricting the classical Hamiltonian to the special form

$$H(p, q) = \frac{p^2}{2m} + V(q). \quad (7.21)$$

One can then perform the momentum integrations explicitly to obtain

$$\int (Dp) \exp \frac{i}{\hbar} \int_{t'}^{t''} dt (p \dot{q} - H) = \left(\frac{m}{2\pi\hbar} \right)^{N/2} \exp \frac{i}{\hbar} \int_{t'}^{t''} dt L(q, \dot{q}), \quad (7.22)$$

where $L(q, \dot{q})$ is the classical Lagrangian:

$$L(q, \dot{q}) = \frac{1}{2} m \dot{q}^2 - V(q). \quad (7.23)$$

Substituting (7.22) into (7.18), we obtain

$$\langle q'', t'' | q', t' \rangle = \mathcal{N} \int (Dq) \exp \frac{i}{\hbar} \int_{t'}^{t''} dt L(q, \dot{q}), \quad (7.24)$$

which is the Feynman formula. Here, \mathcal{N} is a normalization constant which is usually infinite in the limit $N \rightarrow \infty$, but irrelevant to physical results. This is because it cancels in matrix elements of the form $\langle q'', t'' | O | q', t' \rangle / \langle q'', t'' | q', t' \rangle$. For this reason, one need only define (Dq) up to a multiplicative constant (possibly infinite).

Under the assumption (7.21) one can show

$$\begin{aligned} \langle q'', t'' | T[q_{\text{op}}(t_1) \cdots q_{\text{op}}(t_n)] | q', t' \rangle \\ = \mathcal{N} \int (Dq)[q(t_1) \cdots q(t_n)] \exp \frac{i}{\hbar} \int_{t'}^{t''} dt L(q, \dot{q}). \end{aligned} \quad (7.25)$$

We indicate the proof for $n = 2$, with $t_1 > t_2$:

$$\begin{aligned} \langle q'', t'' | q_{\text{op}}(t_1) q_{\text{op}}(t_2) | q', t' \rangle \\ = \int dq_1 dq_2 \langle q'', t'' | q_{\text{op}}(t_1) | q_1, t_1 \rangle \langle q_1, t_1 | q_{\text{op}}(t_2) | q_2, t_2 \rangle \langle q_2, t_2 | q', t' \rangle \\ = \int dq_1 dq_2 q_1 q_2 \langle q'', t'' | q_1, t_1 \rangle \langle q_1, t_1 | q_2, t_2 \rangle \langle q_2, t_2 | q', t' \rangle. \end{aligned}$$

From this point, we follow the steps beginning with (7.8) to obtain the final result.

The Feynman formulas (7.24) and (7.25) are in fact valid under conditions more general than (7.21). They hold for a classical Lagrangian of the general

form

$$L(q, \dot{q}) = \frac{1}{2} \sum_{\alpha, \beta} \dot{q}_\alpha A_{\alpha\beta} \dot{q}_\beta + \sum_\alpha b_\alpha(q) \dot{q}_\alpha - V(q), \quad (7.26)$$

where the matrix A is independent of q . (Here it is assumed that A is a real non-singular matrix). To show this, first define the canonical momenta:

$$p_\alpha = \frac{\partial L}{\partial \dot{q}_\alpha} = \sum_\beta A_{\alpha\beta} \dot{q}_\beta + b_\alpha. \quad (7.27)$$

Using the matrix notation

$$\begin{aligned} p &= A\dot{q} + b, \\ \dot{q} &= A^{-1}(p - b), \end{aligned} \quad (7.28)$$

we can write the classical Hamiltonian in the form

$$H(p, q) = \frac{1}{2}(\tilde{p}, A^{-1}\tilde{p}) + V, \quad \tilde{p} = p - b. \quad (7.29)$$

From (7.18) we have

$$\begin{aligned} \langle q'', t'' | q', t' \rangle &= N \int \prod_\alpha dq_\alpha dp_\alpha \exp \frac{i}{\hbar} \int_{t'}^{t''} dt \left(\sum_\alpha p_\alpha \dot{q}_\alpha - H \right). \\ \sum_\alpha p_\alpha \dot{q}_\alpha - H &= \sum_\alpha (\tilde{p}_\alpha + b_\alpha) \dot{q}_\alpha - \frac{1}{2} \sum_{\alpha, \beta} \tilde{p}_\alpha A_{\alpha\beta}^{-1} \tilde{p}_\beta - V \\ &= -\frac{1}{2}(\tilde{p}, A^{-1}\tilde{p}) + (\tilde{p}, \dot{q}) + (b, \dot{q}) - V. \end{aligned}$$

$$\begin{aligned} \int (Dp) \exp \frac{i}{\hbar} \int dt (p\dot{q} - H) &= \prod_t \int (Dp) \exp \left[\frac{i}{\hbar} (\Delta t) (p\dot{q} - H) \right] \quad (7.30) \\ &= \prod_t \int (D\tilde{p}) \exp \frac{i}{\hbar} (\Delta t) \left[-\frac{1}{2}(\tilde{p}, A^{-1}\tilde{p}) + (\tilde{p}, \dot{q}) + (b, \dot{q}) - V \right] \\ &= \text{const.} \prod_t \left\{ \exp \frac{i}{\hbar} (\Delta t) \left[\frac{1}{2}(\dot{q}, A\dot{q}) + (b, \dot{q}) - V \right] \right\} (\det A)^{-1/2} \\ &= \text{const.} \left[\exp \frac{i}{\hbar} \int_{t'}^{t''} dt L(q, \dot{q}) \right] \prod_t (\det A)^{-1/2}. \end{aligned}$$

If A is independent of q , then the last factor is a constant, which may be absorbed into an overall normalization factor. ■

If the matrix A in (7.26) does depend on q , then the Feynman formula requires modification. We rewrite

$$\prod_t (\det A)^{-1/2} = \exp \left[-\frac{1}{2} \sum_t \ln(\det A) \right]. \quad (7.31)$$

In the limit $\Delta t \rightarrow 0$,

$$\sum_t \rightarrow \delta(0) \int dt, \quad (7.32)$$

because a rectangle of unit area, with base dt , has height $\delta(0)$. Thus,

$$\prod_t (\det A)^{-1/2} \xrightarrow[\Delta t \rightarrow 0]{} \exp \left[-\frac{1}{2} \delta(0) \int_{t'}^{t''} dt \ln(\det A) \right]. \quad (7.33)$$

Therefore

$$\langle q'', t'' | q', t' \rangle = \mathcal{N} \int (Dq) \exp \frac{i}{\hbar} \int_{t'}^{t''} dt \left[L(q, \dot{q}) - \frac{\hbar}{2i} \delta(0) \ln \det A(q) \right]. \quad (7.34)$$

An example of the necessity for the singular term with the factor $\delta(0)$ is encountered in massive vector boson theory.²

The Feynman formula (7.24) does not apply when the matrix $A_{\alpha\beta}$ in (7.26) is singular. This is the case when there is a coordinate whose time derivative does not appear in the Lagrangian. In such a case, one has to return to the Hamiltonian form (7.18).

7.2 Quantum Field Theory

The extension of the Feynman formula to a quantum field theory is straightforward; one merely allows the number of coordinates to become non-denumerably infinite. We shall illustrate this extension for the case of one boson field. Generalization to more than one boson field is straightforward. The case of fermion fields will be discussed separately later.

In canonical quantization, the coordinates are the Schrödinger field operators $\phi_{op}(x)$, where the spatial variable x serves as a continuous label for the coordinates. We denote eigenstates of $\phi_{op}(x)$ by $|\phi\rangle$:

$$\phi_{op}(x)|\phi\rangle = \phi(x)|\phi\rangle, \quad (7.35)$$

where $\phi(x)$ is a c-number function. The transition amplitude is defined by

$$\langle \phi_2, t_2 | \phi_1, t_1 \rangle \equiv \langle \phi_2 | e^{-i(t_2-t_1)H_{op}/\hbar} | \phi_1 \rangle, \quad (7.36)$$

where H_{op} is the time-independent Hamiltonian operator of the system. Let the classical Lagrangian density be

$$\mathcal{L}(x) = \mathcal{L}(\phi(x), \partial^\mu \phi(x)), \quad (7.37)$$

which is assumed to be quadratic in $\partial_\mu \phi(x)$ with coefficients independent of $\phi(x)$. The Feynman formula reads:

$$\langle \phi_2, t_2 | \phi_1, t_1 \rangle = \mathcal{N} \int_{\phi_1}^{\phi_2} (D\phi) \exp \frac{i}{\hbar} \int_1^2 d^4x \mathcal{L}(x), \quad (7.38)$$

² T. D. Lee and C. N. Yang, *Phys. Rev.* **128**, 885 (1962), Appendix E.

where

$$\int_1^2 d^4x \equiv \int_{t_1}^{t_2} dx_0 \int_{\text{all space}} d^3x, \quad (7.39)$$

with the endpoint constraints

$$\phi(x, t_2) = \phi_2(x), \quad \phi(x, t_1) = \phi_1(x). \quad (7.40)$$

The functional integration $\int (D\phi)$ may be defined by first considering x to be a discrete variable, and then approaching the continuum limit in the final result of a physical calculation. Alternatively, we may enclose the system in a large but finite space-time volume, integrate independently over each of the discrete Fourier components of ϕ , and then approach the limit of infinite volume. We also have

$$\begin{aligned} & \langle \phi_2, t_2 | T\phi_{\text{op}}(x_1) \cdots \phi_{\text{op}}(x_n) | \phi_1, t_1 \rangle \\ &= N \int_{\phi_1}^{\phi_2} (D\phi) \phi(x_1) \cdots \phi(x_n) \exp \frac{i}{\hbar} \int_1^2 d^4x \mathcal{L}(x). \end{aligned} \quad (7.41)$$

The proof is similar to that for (7.25).

The transition amplitude may be continued analytically to complex times in a certain manner. To investigate this possibility, consider the eigenstates $|n\rangle$ of H_{op} , and assume that there is a unique vacuum state with zero energy:

$$\begin{aligned} H_{\text{op}}|n\rangle &= E_n|n\rangle, \quad (E_n \geq 0), \\ H_{\text{op}}|0\rangle &= 0, \quad \langle 0|0 \rangle = 1. \end{aligned} \quad (7.42)$$

Then we can write

$$\begin{aligned} \langle \phi_2, t_2 | \phi_1, t_1 \rangle &= \left\langle \phi_2 \left| \exp \left[-\frac{i}{\hbar} (t_2 - t_1) H_{\text{op}} \right] \right| \phi_1 \right\rangle \\ &= \sum_n \langle \phi_2 | n \rangle \langle n | \phi_1 \rangle \exp \left[-\frac{i}{\hbar} (t_2 - t_1) E_n \right] \\ &= \int_0^\infty dE \rho_{21}(E) \exp \left[-\frac{i}{\hbar} (t_2 - t_1) E \right], \end{aligned} \quad (7.43)$$

where

$$\rho_{21}(E) = \sum_n \langle \phi_2 | n \rangle \langle n | \phi_1 \rangle \delta(E - E_n). \quad (7.44)$$

It is clear that the transition amplitude can be continued into the lower half plane of $t_2 - t_1$, for example:

$$\langle \phi_2, -i\tau | \phi_1, 0 \rangle = \int_0^\infty dE \rho_{21}(E) e^{-\tau E/\hbar} \quad (\tau > 0). \quad (7.45)$$

After calculating the right-hand side, the expression for real times can be recovered by analytically continuing τ to the positive imaginary axis.

Taking the limit $\tau \rightarrow \infty$, we obtain

$$\langle \phi_2, -i\tau | \phi_1, 0 \rangle \xrightarrow{\tau \rightarrow \infty} \langle \phi_2 | 0 \rangle \langle 0 | \phi_1 \rangle = \Psi_0[\phi_2] \Psi_0^*[\phi_1], \quad (7.46)$$

where $\Psi_0[\phi] \equiv \langle \phi | 0 \rangle$ is the ground state wave function of the system. Combining (7.46) with (7.38), we obtain

$$\Psi_0[\phi_2] \Psi_0^*[\phi_1] = \mathcal{N} \int_{\phi_1}^{\phi_2} (\mathcal{D}\phi) \exp \frac{i}{\hbar} \int_{-\infty}^{\infty} d\tau \int d^3x \mathcal{L}(x, -i\tau), \quad (7.47)$$

with the endpoint constraints

$$\phi(x, t = i\infty) = \phi_2(x), \quad \phi(x, t = -i\infty) = \phi_1(x). \quad (7.48)$$

The Feynman formula (7.38) is not applicable as it stands, when there are fields whose conjugate momenta vanish identically. This is the case for gauge fields, which will be discussed separately in Chapter 8. It suffices to mention at this point that the Feynman formula still holds, provided $\mathcal{L}(x)$ is supplemented by a “gauge fixing” term. The general methods discussed in the rest of this chapter still apply.

7.3 External Sources

The system can be coupled to an external source by adding the term $\phi(x)J(x)$ to the classical Lagrangian density, where $J(x)$ is an arbitrary function. As we shall see, this is a useful mathematical device.

In the presence of an external source, the transition amplitude is denoted by

$$\langle \phi_2, t_2 | \phi_1, t_1 \rangle_J \equiv \mathcal{N} \int_{\phi_1}^{\phi_2} (\mathcal{D}\phi) \exp \frac{i}{\hbar} \int_1^2 d^4x [\mathcal{L}(x) + \phi(x)J(x)]. \quad (7.49)$$

Matrix elements of time-ordered products of fields can be obtained from the above by functional differentiations with respect to the external source:

$$\begin{aligned} & \frac{\delta \langle \phi_2, t_2 | \phi_1, t_1 \rangle_J}{\delta J(y_1) \cdots \delta J(y_n)} \\ &= \left(\frac{i}{\hbar} \right)^n \mathcal{N} \int_{\phi_1}^{\phi_2} (\mathcal{D}\phi) \phi(y_1) \cdots \phi(y_n) \exp \frac{i}{\hbar} \int_1^2 d^4x [\mathcal{L}(x) + \phi(x)J(x)]. \end{aligned} \quad (7.50)$$

The right-hand side reduces to that of (7.41) upon setting $J(x) \equiv 0$.

Suppose the Lagrangian density can be separated into an unperturbed term plus a perturbation term:

$$\mathcal{L}(x) = \mathcal{L}_0(x) + \mathcal{L}'(\phi(x)). \quad (7.51)$$

Note that \mathcal{L}' is considered a function of $\phi(x)$, even though it may involve

derivatives of $\phi(x)$. We can write

$$\begin{aligned} \langle \phi_2, t_2 | \phi_1, t_1 \rangle_J &= \left\{ \exp \frac{i}{\hbar} \int_1^2 d^4x \mathcal{L}' \left(\frac{\hbar}{i} \frac{\delta}{\delta J(x)} \right) \right\} \mathcal{N} \int_{\phi_1}^{\phi_2} (D\phi) \exp \frac{i}{\hbar} \int_1^2 d^4x [\mathcal{L}_0(x) + \phi(x)J(x)]. \end{aligned} \quad (7.52)$$

A perturbation series can be obtained by expanding the first exponential factor as a power series in \mathcal{L}' . We shall illustrate this technique in Section 7.7 by deriving Feynman rules for ϕ^4 theory.

An important application of the method of path integrals is the calculation of the vacuum-vacuum amplitude in the presence of an external source. Knowing this amplitude determines the complete dynamics of the system. In other words, *all dynamical information can be deduced from the response of the vacuum state to an arbitrary external source*.

We first define the vacuum-vacuum amplitude, and exhibit its relation to the Green's functions of the system. In the presence of an external source, the Hamiltonian of the system is changed from H_{op} to

$$\begin{aligned} H'_{\text{op}} &= H_{\text{op}} - H'_{\text{op}} \\ H'_{\text{op}} &= \int d^3x J(x) \phi_{\text{op}}(x), \end{aligned} \quad (7.53)$$

where $\phi_{\text{op}}(x)$ is a Heisenberg operator, and $J(x)$ is a c-number function, arbitrary except for the condition that it is to be turned off in the infinite past and the infinite future:

$$J(x) \xrightarrow{|x_0| \rightarrow \infty} 0. \quad (7.54)$$

This insures that the system can be in the vacuum state $|0\rangle$ of the Hamiltonian H_{op} in the infinite past and infinite future. The response of the vacuum state is described by the amplitude

$$\langle 0^+ | 0^- \rangle_J = \text{Probability amplitude that the system will be in state } |0\rangle \text{ at } x_0 = \infty, \text{ when it was known to be in state } |0\rangle \text{ at } x_0 = -\infty. \quad (7.55)$$

By unitarity, this amplitude can only be a phase factor:

$$\langle 0^+ | 0^- \rangle_J = \exp \frac{i}{\hbar} W[J]. \quad (7.56)$$

Now, go to the interaction picture with respect to H'_{op} . At $x_0 = 0$, the state of the system is given by

$$|0^-\rangle_J = T \left[\exp \frac{i}{\hbar} \int_{-\infty}^0 dt H'_{\text{op}}(t) \right] |0\rangle, \quad (7.57)$$

where the time evolution of $H'_{\text{op}}(t)$ is governed by H_{op} . That is, $H'_{\text{op}}(t)$ is a Heisenberg operator of the system without external sources. On the other hand,

the state of the system at $x_0 = 0$ that will evolve into $|0\rangle$ at $x_0 = \infty$ is given by

$$|0^+\rangle_J = T \left[\exp \frac{i}{\hbar} \int_{-\infty}^0 dt H'_{\text{op}}(t) \right] |0\rangle. \quad (7.58)$$

Therefore

$$\begin{aligned} \langle 0^+ | 0^- \rangle_J &= \left\langle 0 \left| T \left[\exp \frac{i}{\hbar} \int_{-\infty}^{\infty} dt H'_{\text{op}}(t) \right] \right| 0 \right\rangle \\ &= \left\langle 0 \left| T \left[\exp \frac{i}{\hbar} \int d^4x J(x) \phi_{\text{op}}(x) \right] \right| 0 \right\rangle. \end{aligned} \quad (7.59)$$

We see immediately that

$$\left[\frac{\delta \langle 0^+ | 0^- \rangle_J}{\delta J(x_1) \cdots \delta J(x_n)} \right]_{J=0} = \left(\frac{i}{\hbar} \right)^n g_n(x_1, \dots, x_n), \quad (7.60)$$

where

$$g_n(x_1, \dots, x_n) = \langle 0 | T \phi_{\text{op}}(x_1) \cdots \phi_{\text{op}}(x_n) | 0 \rangle \quad (n \geq 1) \quad (7.61)$$

are the Green's functions which completely describe the dynamics of the system in the absence of external sources. From (7.60), we obtain the expansion

$$\langle 0^+ | 0^- \rangle_J = \sum_{n=0}^{\infty} \left(\frac{i}{\hbar} \right)^n \frac{1}{n!} \int d^4x_1 \cdots d^4x_n g_n(x_1, \dots, x_n) J(x_1) \cdots J(x_n). \quad (7.62)$$

where $g_0 \equiv 1$. This shows that $\langle 0^+ | 0^- \rangle_J$ is the generating functional for the Green's functions.

The phase $W[J]$ defined in (7.56) is the generating functional for *connected* Green's functions:

$$\frac{i}{\hbar} W[J] = \sum_{n=0}^{\infty} \left(\frac{i}{\hbar} \right)^n \frac{1}{n!} \int d^4x_1 \cdots d^4x_n G_n(x_1, \dots, x_n) J(x_1) \cdots J(x_n), \quad (7.63)$$

where $G_n(x_1, \dots, x_n)$, ($n \geq 1$), is the connected n -point Green's function, i.e., the sum of all *connected* Feynman graphs with n external lines terminating at x_1, \dots, x_n . To show this, we note that g_n , ($n \geq 1$), is a sum of products of connected Green's functions G_m , ($m \geq 1$), which may be displayed as follows. Let $\{\sigma_1, \dots, \sigma_n\}$ be a partition of the integer n :

$$n = \sigma_1 + 2\sigma_2 + 3\sigma_3 + \cdots. \quad (7.64)$$

We can write

$$g_n(x_1, \dots, x_n) = \sum_{\substack{\{\sigma_i\} \\ \sum \sigma_i = n}} \sum_P P \underbrace{[G_1(\cdot) \cdots G_1(\cdot)]}_{\sigma_1 \text{ factors}} \underbrace{[G_2(\cdot) \cdots G_2(\cdot)]}_{\sigma_2 \text{ factors}} \cdots \quad (7.65)$$

The above parentheses contain a total of n dots which are to be put into one-to-one correspondence with x_1, \dots, x_n in some manner. The symbol P

denotes a distinct permutation of x_1, \dots, x_n . The number of such permutations is

$$\frac{n!}{(\sigma_1! \sigma_2! \dots)[(1!)^{\sigma_1}(2!)^{\sigma_2} \dots]}. \quad (7.66)$$

When (7.65) is substituted into (7.62), every term in the sum Σ_P gives the same contribution, upon integration over x_1, \dots, x_n . Hence

$$\langle 0^+ | 0^- \rangle_J = \sum_{n=0}^{\infty} \sum_{\substack{\{\sigma_l\} \\ \sum l \sigma_l = n}} \left(\frac{i}{\hbar} \right)^n \frac{[\int d^4x G_1(x) J(x)]^{\sigma_1}}{\sigma_1!} \frac{[\int d^4x d^4y G_2(x, y) J(x) J(y)]^{\sigma_2}}{\sigma_2! (2!)^{\sigma_2}} \dots \quad (7.67)$$

The double sum in (7.67) is the same as the sum over each of the integers σ_l independently. Therefore

$$\begin{aligned} \langle 0^+ | 0^- \rangle_J &= \sum_{\sigma_1=0}^{\infty} \frac{1}{\sigma_1!} \left[\frac{i}{\hbar} \int d^4x G_1(x) J(x) \right]^{\sigma_1} \\ &\times \sum_{\sigma_2=0}^{\infty} \frac{1}{\sigma_2!} \left[\left(\frac{i}{\hbar} \right)^2 \frac{1}{2!} \int d^4x d^4y G_2(x, y) J(x) J(y) \right]^{\sigma_2} \dots \\ &= \exp \sum_{n=1}^{\infty} \left(\frac{i}{\hbar} \right)^n \frac{1}{n!} \int d^4x_1 \dots d^4x_n G_n(x_1, \dots, x_n) J(x_1) \dots J(x_n). \blacksquare \end{aligned} \quad (7.68)$$

We now show that the vacuum-vacuum amplitude can be expressed as a path integral. For simplicity let us turn on the source only for a finite but large time interval:

$$J(x) = 0 \quad \text{for } |x_0| > T, \quad (T \rightarrow \infty). \quad (7.69)$$

Consider the transition amplitude from a time t_1 before the source was turned on, to a time t_2 after the source has been turned off:

$$\langle \phi_2, t_2 | \phi_1, t_1 \rangle = \int (D\phi)(D\phi') \langle \phi_2, t_2 | \phi, T \rangle \langle \phi, T | \phi', -T \rangle_J \langle \phi', -T | \phi_1, t_1 \rangle \quad (t_1 < -T, t_2 > T). \quad (7.70)$$

We can calculate the source-free amplitudes immediately. For example,

$$\begin{aligned} \langle \phi_2, t_2 | \phi, T \rangle &= \left\langle \phi_2 \left| \exp \left[-\frac{i}{\hbar} (t_2 - T) H_{\text{op}} \right] \right| \phi \right\rangle \\ &= \sum_n \langle \phi_2 | n \rangle \langle n | \phi \rangle \exp \left[-\frac{i}{\hbar} (t_2 - T) E_n \right] \end{aligned} \quad (7.71)$$

$$\xrightarrow[t_2 \rightarrow -i\infty]{T \rightarrow i\infty} \langle \phi_2 | 0 \rangle \langle 0 | \phi \rangle.$$

Thus, continuing the times:

$$t_1 \rightarrow -i\infty, \quad t_2 \rightarrow i\infty, \quad T \rightarrow i\infty, \quad (7.72)$$

we obtain

$$\begin{aligned} \frac{\langle \phi_2, t_2 | \phi_1, t_1 \rangle_J}{\langle \phi_2 | 0 \rangle \langle 0 | \phi_1 \rangle} &\rightarrow \int (D\phi)(D\phi') \langle \phi, T | \phi', -T \rangle_J \\ &= \left\langle 0 \left| \exp\left(-\frac{i}{\hbar} TH_{\text{op}}^J\right) \right| 0 \right\rangle = \langle 0^+ | 0^- \rangle_J, \end{aligned} \quad (7.73)$$

or,

$$\langle 0^+ | 0^- \rangle_J = \lim_{\substack{t_2 \rightarrow -i\infty \\ t_1 \rightarrow i\infty}} \frac{\langle \phi_2, t_2 | \phi_1, t_1 \rangle_J}{\langle \phi_2 | 0 \rangle \langle 0 | \phi_1 \rangle}. \quad (7.74)$$

Using (7.49) we can express the right-hand side as a path integral over fields defined in Euclidean 4-space. Note that the endpoint constraints are irrelevant because the right-hand side of (7.74) is independent of $\phi_2(x)$ and $\phi_1(x)$.

7.4 Euclidean 4-Space

Euclidean 4-space is obtained from Minkowski space by clockwise rotation of the real axis in the complex x_0 plane into the negative imaginary axis, as indicated in Fig. 7.1. A point in Euclidean 4-space is denoted by x_E , and is related to the corresponding point $x = (x_0, \mathbf{x})$ in Minkowski space by

$$\begin{aligned} x_E &= (\mathbf{x}, x_4), \\ x_4 &= ix_0 \quad (\text{real}), \\ d^4x &= -i d^4x_E, \\ x_E^2 &= x_1^2 + x_2^2 + x_3^2 + x_4^2 = -x^2. \end{aligned} \quad (7.75)$$

The corresponding Euclidean momentum space is defined so that $k_4 x_4 = k^0 x^0$. This convention is chosen so that in the propagation of a plane wave, a positive sense of x_4 corresponds to a positive sense of x^0 . Accordingly, we rotate the k^0 axis counter-clockwise into the positive imaginary axis, as indicated in Fig. 7.1. We have

$$\begin{aligned} k_E &= (\mathbf{k}, k_4), \\ k_4 &= -ik_0 \quad (\text{real}), \\ d^4k &= i d^4k_E, \\ k_E^2 &= k_1^2 + k_2^2 + k_3^2 + k_4^2 = -k^2. \end{aligned} \quad (7.76)$$

Note that $k \cdot x = k^0 x^0 - \mathbf{k} \cdot \mathbf{x}$ transforms into $k_4 x_4 - \mathbf{k} \cdot \mathbf{x}$, but in taking the Fourier transform of $f(k^2)$, we can replace $k \cdot x$ by $k_E \cdot x_E = k_4 x_4 + \mathbf{k} \cdot \mathbf{x}$.

Continuation to Euclidean space means that we consider the dynamical evolution of the system in imaginary time. That is, we must in principle solve the

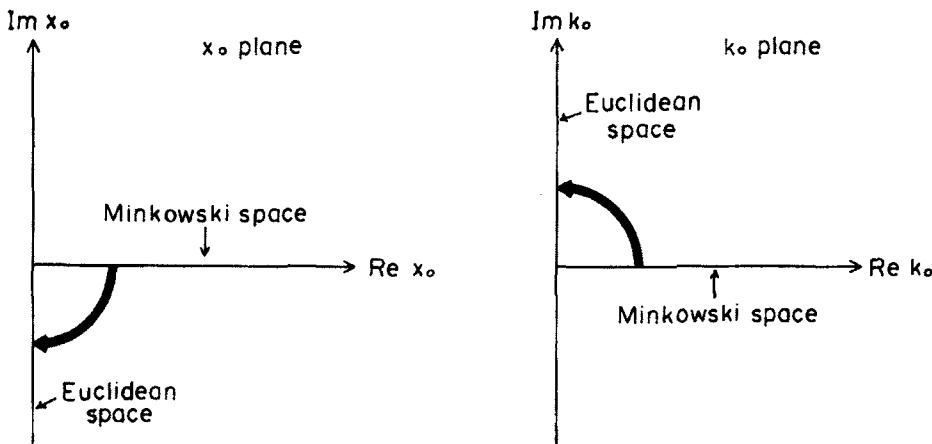


Fig. 7.1 How to continue Minkowski space into Euclidean space

equations of motion in which the time x^0 is replaced by $-ix_4$, where x_4 is a real parameter. The Lorentz invariance of the Lagrangian density is replaced by invariance with respect to $O(4)$ rotations in Euclidean space. The equations of motion will determine how the fields are to be continued into Euclidean space.

A real scalar field $\phi(x)$ defined in Minkowski space is replaced by a real scalar field $\phi(x_E)$ invariant under $O(4)$.

A massive vector field $A^\mu(x)$ with real components is replaced by a Euclidean vector field $A^\mu(x_E)$ with real components, according to the rule

$$\begin{aligned} A^k(x) &\rightarrow A^k(x_E) \quad (k = 1, 2, 3), \\ A^0(x) &\rightarrow iA_4(x_E). \end{aligned} \quad (7.77)$$

Note that A^0 continues to A_4 with sign opposite to that of x^0 because it should transform like $\partial/\partial x^0$. The subsidiary condition $\partial_\mu A^\mu = 0$ is replaced by

$$\nabla \cdot \mathbf{A} - \frac{\partial A_4}{\partial x_4} = 0.$$

For Euclidean vectors, there is no distinction between upper and lower indices.

For a gauge field, the answer depends on how we choose to handle the unphysical degrees of freedom associated with gauge invariance. These have no physical significance, and covariance arguments are not compelling. In a covariant gauge, (7.77) applies. In a non-covariant gauge, we could still use (7.77) as a formal device although it is not necessary.

The functional integral for the vacuum-vacuum amplitude in Euclidean form is as follows:

$$\langle 0^+ | 0^- \rangle_J = \exp \frac{i}{\hbar} W[J] = \mathcal{N} \int (\mathcal{D}\phi) \exp \left(-\frac{1}{\hbar} \{ S_E[\phi] - (J, \phi)_E \} \right), \quad (7.78)$$

where

$$\begin{aligned} S_E[\phi] &= \int d^4x_E \mathcal{L}(x_E) = -iS[\phi]. \\ (J, \phi)_E &= \int d^4x_E J(x_E)\phi(x_E). \end{aligned} \quad (7.79)$$

In this form, the functional integral is formally the partition function of a classical system in 4 dimensions, at temperature \hbar .

The Euclidean formulation is very practical when one wants to investigate which paths make important contributions to the transition amplitude. The path integral in Minkowski space assigns a phase factor to each path, the classical path being the one with stationary phase. The Euclidean formulation assigns a Boltzmann-like factor to each path, with the classical path corresponding to the one with least Euclidean action. The neighboring paths give contributions that are damped out, instead of oscillating ever more wildly.

In the Feynman graph expansion of the vacuum-vacuum amplitude, the Euclidean prescription merely supplies the correct *iε* in the propagators [see (7.99) and (7.100)].

7.5 Calculation of Path Integrals

Path integrals of the Gaussian type can be calculated by an extension of the formula for the ordinary Gaussian integral. Suppose $Q(x)$ is a quadratic form in one variable:

$$Q(x) = \frac{1}{2} ax^2 - bx = -\frac{b^2}{2a} + \frac{1}{2} a(x - x_0)^2, \quad (7.80)$$

where

$$x_0 = b/a. \quad (7.81)$$

It is well-known that

$$\int_{-\infty}^{\infty} dx e^{-Q(x)} = \left(\frac{2\pi}{a}\right)^{1/2} \exp\left(-\frac{b^2}{2a}\right). \quad (7.82)$$

This result can be immediately generalized to a quadratic form $Q(u)$ involving n variables $\{u_1, u_2, \dots, u_n\}$:

$$Q(u) = \frac{1}{2}(u, Au) - (b, u), \quad (7.83)$$

where b is a constant n -vector, A is a symmetric non-singular $n \times n$ matrix, and

$$(b, u) = \sum_{i=1}^n b_i u_i. \quad (7.84)$$

We can also write

$$\begin{aligned} Q(u) &= -\frac{1}{2}(b, A^{-1}b) + \frac{1}{2}((u - u_0), A(u - u_0)), \\ u_0 &= A^{-1}b. \end{aligned} \quad (7.85)$$

Then,

$$\begin{aligned} \int (\mathrm{D}u) e^{-Q(u)} &= (\det A)^{-1/2} \exp \frac{1}{2}(b, A^{-1}b), \\ (\mathrm{D}u) &\equiv (2\pi)^{-n/2} du_1 \cdots du_n, \\ \det A &= \prod_{i=1}^n a_i. \end{aligned} \quad (7.86)$$

For a real field $\phi(x)$, define a quadratic form $Q[\phi]$ by

$$Q[\phi] = \frac{1}{2}(\phi, A\phi) - (b, \phi), \quad (7.87)$$

where A is non-singular and self-adjoint, and

$$(\phi_1, \phi_2) = \int d^4x_E \phi_1(x_E) \phi_2(x_E). \quad (7.88)$$

We can rewrite $Q[\phi]$ as

$$\begin{aligned} Q[\phi] &= -\frac{1}{2}(b, A^{-1}b) + \frac{1}{2}((\phi - \phi_0), A(\phi - \phi_0)), \\ \phi_0 &= A^{-1}b. \end{aligned} \quad (7.89)$$

Then,

$$\int (\mathrm{D}\phi) e^{-Q[\phi]} = N(\det A)^{-1/2} \exp \frac{1}{2}(b, A^{-1}b). \quad (7.90)$$

The volume element $(\mathrm{D}\phi)$ may be defined up to multiplicative constant as $\prod_x d\phi(x)$, the ambiguous constant (possibly infinite) being absorbed into the normalization constant N .

7.6 The Feynman Propagator

We calculate the generating functional $W[J]$ for a free scalar field, to illustrate the fact that the Feynman propagator is the inverse of the “kinetic operator” in the Lagrangian density.

The Lagrangian density is

$$\mathcal{L}_0(x) = \frac{1}{2} \partial_\mu \phi(x) \partial^\mu \phi(x) - \frac{1}{2} m^2 \phi^2(x). \quad (7.91)$$

The classical action is

$$\begin{aligned} S_0[\phi] &= -\frac{1}{2} \int d^4x \phi(x) (\square^2 + m^2) \phi(x) \\ &= \frac{i}{2} \int d^4x_E \phi(x_E) (-\square_E^2 + m^2) \phi(x_E). \end{aligned} \quad (7.92)$$

The kinetic operator is $(-\square_E^2 + m^2)$. Using (7.78), we have

$$\begin{aligned}\exp \frac{i}{\hbar} W_0[J] &= \mathcal{N} \int (\mathbf{D}\phi) \exp \frac{i}{\hbar} \left\{ S_0[\phi] - i \int d^4x_E \phi(x_E) J(x_E) \right\} \\ &= \mathcal{N} \int (\mathbf{D}\phi) e^{-Q[\phi, J]},\end{aligned}\quad (7.93)$$

where

$$Q[\phi, J] = \frac{1}{2} (\phi, A\phi) - (b, \phi), \quad (7.94)$$

with

$$A = \frac{1}{\hbar} (-\square_E^2 + m^2), \quad (7.95)$$

$$b = \frac{1}{\hbar} J(x_E).$$

According to (7.90), therefore,

$$\exp \frac{i}{\hbar} W_0[J] = \mathcal{N} (\det A)^{-1/2} \exp \frac{1}{2}(b, A^{-1}b), \quad (7.96)$$

or,

$$\frac{i}{\hbar} W_0[J] = \frac{1}{2\hbar} (J, (-\square_E^2 + m^2)^{-1}J) + \ln[\mathcal{N} (\det A)^{-1/2}]. \quad (7.97)$$

The operator $(-\square_E^2 + m^2)^{-1}$ can be studied best in momentum space:

$$\begin{aligned}(-\square_E^2 + m^2)^{-1}f(x_E) &= (-\square_E^2 + m^2)^{-1} \int \frac{d^4k_E}{(2\pi)^4} e^{-ik_E \cdot x_E} \tilde{f}(k_E). \\ &= \int \frac{d^4k_E}{(2\pi)^4} \frac{e^{-ik_E \cdot x_E}}{k_E^2 + m^2} \tilde{f}(k_E) \\ &= i \int d^4y_E \Delta_F(x_E - y_E) f(y_E),\end{aligned}\quad (7.98)$$

where Δ_F is the Feynman propagator, defined by

$$\Delta_F(x_E) = -i \int \frac{d^4k_E}{(2\pi)^4} \frac{e^{-ik_E \cdot x_E}}{k_E^2 + m^2}. \quad (7.99)$$

When x_E is rotated into Minkowski space, this takes the familiar form

$$\Delta_F(x) = \int \frac{d^4k}{(2\pi)^4} \frac{e^{-ik \cdot x}}{k^2 - m^2 + i\varepsilon} \quad (\varepsilon \rightarrow 0^+). \quad (7.100)$$

We see that the Euclidean-space formulation amounts to specifying the contour of the k_0 integration in the usual manner.

The eigenvalues of $(-\square_E^2 + m^2)$ are $(k_E^2 + m^2)$. Hence

$$\det A = \prod_{k_E} \hbar^{-1}(k_E^2 + m^2), \quad (7.101)$$

which is a divergent quantity. However, since it is independent of $J(x)$, we can cancel it in (7.97) by choosing \mathcal{N} appropriately. Thus we obtain

$$\begin{aligned} W_0[J] &= \frac{1}{2} \int d^4x_E d^4y_E J(x_E) \Delta_F(x_E - y_E) J(y_E) \\ &= -\frac{1}{2} \int d^4x d^4y J(x) \Delta_F(x - y) J(y). \end{aligned} \quad (7.102)$$

Comparison with (7.63) shows that the only non-vanishing connected Green's function for the free field theory is the Feynman propagator:

$$G_2(x, y) = i\hbar \Delta_F(x - y). \quad (7.103)$$

7.7 Feynman Graphs

We illustrate Feynman graphs using ϕ^4 theory as an example. The Lagrangian density in the presence of an external source is

$$\mathcal{L}(x) = \mathcal{L}_0(x) + \mathcal{L}'(\phi(x)) + \phi(x)J(x), \quad (7.104)$$

where

$$\begin{aligned} \mathcal{L}_0(x) &= \frac{1}{2} \partial_\mu \phi(x) \partial^\mu \phi(x) - \frac{1}{2} m^2 \phi^2(x), \\ \mathcal{L}'(\phi(x)) &= \frac{\lambda}{4!} \phi^4(x). \end{aligned} \quad (7.105)$$

Using (7.56), (7.74), and (7.52), we can write

$$\exp \frac{i}{\hbar} W[J] = \left\{ \exp \frac{i}{\hbar} \int d^4x \mathcal{L}' \left(\frac{\hbar}{i} \frac{\delta}{\delta J(x)} \right) \right\} \cdot \exp \frac{i}{\hbar} W_0[J], \quad (7.106)$$

where $W_0[J]$ is given by (7.102). The usual Feynman graph expansion for the Green's functions is obtained by expanding the above in powers of \mathcal{L}' .

Without using the specific form of \mathcal{L}' , we can see that the classical limit (defined formally as the limit of the theory as $\hbar \rightarrow 0$), is obtained by retaining only the “tree graphs”. Consider any connected Feynman graph, and let

$$\begin{aligned} V &= \text{No. of vertices}, \\ I &= \text{No. of internal lines}, \\ E &= \text{No. of external lines}. \end{aligned} \quad (7.107)$$

Each internal line carries an internal 4-momentum that is integrated over. Each vertex imposes one condition of 4-momentum conservation. There is a

condition of total 4-momentum conservation for the external lines. Thus, the number of independent internal 4-momenta is

$$l = I - V + 1, \quad (7.108)$$

which is the number of closed loops in the graph. The graph is proportional to a certain power of \hbar , which comes from internal lines and vertices. Powers of \hbar coming from external lines are ignored, because the latter are to be replaced by wave functions in a S -matrix element.

According to (7.106), a vertex (which is associated with \mathcal{L}') comes with a factor \hbar^{-1} . Each internal line corresponds to a factor $\hbar \Delta_F$, which arises when we remove the two factors J/\hbar in the quantity $\hbar^{-1}(J, \Delta_F J)$ in $W_0[J]$, through differentiation by $\delta/\delta(J/\hbar)$. The contribution of an amputated connected graph to $(i/\hbar)W[J]$ is therefore proportional to \hbar^{I-V} , and the contribution to $W[J]$ is proportional to

$$\hbar^{I-V+1} = \hbar^l. \quad (7.109)$$

Since $W[J]$ is the generating functional for connected Green's functions, an l -loop graph of an amputated connected Green's function is proportional to \hbar^l . The tree graphs are those with $l = 0$, and are therefore the only surviving ones in the classical limit. Quantum corrections may be classified by the power of \hbar , and hence by the number of loops.

We now derive the Feynman rules, which depend on the explicit form of \mathcal{L}' . Eq. (7.106) can be written more explicitly as

$$\exp \frac{i}{\hbar} W[J] = \left\{ \exp \int d^4x \frac{i\lambda \hbar^3}{4!} \left[\frac{\delta}{\delta J(x)} \right]^4 \right\} \cdot \left\{ \exp \int d^4x d^4y J(x) \frac{\Delta_F(x-y)}{2i\hbar} J(y) \right\}. \quad (7.110)$$

Note that

$$\begin{aligned} \frac{\delta J(x)}{\delta J(z)} &= \delta^4(x-z), \\ \left[\frac{\delta}{\delta J(z)} \right]^4 \prod_{i=1}^4 J(x_i) &= 4! \prod_{i=1}^4 \delta^4(z-x_i). \end{aligned} \quad (7.111)$$

To work out (7.110), it is convenient to rewrite it using the following graphical short-hand:

Line: $x - y \equiv \Delta_F(x-y)/i\hbar$

Line End: $\circ_x \equiv J(x)$

Terminal: $(X)_z \equiv (\lambda \hbar^3/4!) [\delta/\delta J(z)]^4$

Vertex: $\bullet \equiv \begin{array}{c} 1 \circ \\ \diagup \quad \diagdown \\ 3 \circ \quad 4 \circ \\ \diagdown \quad \diagup \end{array} \circ^2 = i\lambda \hbar^3 \prod_{i=1}^4 \delta^4(z-x_i).$

As an example, we write

$$\underset{x}{\circlearrowleft} - \underset{y}{\circlearrowright} = J(x) \frac{\Delta_F(x - y)}{i\hbar} J(y). \quad (7.113)$$

With this notation, we have

$$\begin{aligned} \exp \frac{i}{\hbar} W[J] &= \left\{ \exp \int d^4 z (\overset{\mathbf{X}}{z}) \right\} \cdot \left\{ \exp \int d^4 x d^4 y \frac{1}{2} (\underset{x}{\circlearrowleft} - \underset{y}{\circlearrowright}) \right\} \\ &= \sum_{m=0}^{\infty} \sum_{n=0}^{\infty} \frac{1}{m! n!} \int (dx)(dy)(dz) \left[(\overset{\mathbf{X}}{z_1}) \cdots (\overset{\mathbf{X}}{z_m}) \right] \left[\frac{1}{2} (\underset{x_1}{\circlearrowleft} - \underset{y_1}{\circlearrowright}) \cdots \frac{1}{2} (\underset{x_n}{\circlearrowleft} - \underset{y_n}{\circlearrowright}) \right], \end{aligned} \quad (7.114)$$

where $\int (dx)(dy)(dz)$ denote integrations over all the x 's, y 's and z 's in the integrand. A non-vanishing term in the sum above has the following properties:

1. A terminal must seek out 4 line ends, and join them together to form a vertex.

2. A line end need not be joined to a terminal. [If it is not joined to a terminal, it is the free end of an external line and corresponds to a factor $J(x)$].

3. If both ends of a line are joined to the same terminal, the result is an infinite constant, which can be absorbed by mass renormalization. We can thus ignore this type of connection.

According to the above, each terminal must seek out 4 *different* lines and join them to form a vertex. A Feynman graph is a drawing showing the connectivity of lines and terminals corresponding to a non-vanishing term in (7.114). Since the coordinates attached to terminals and line ends are integrated over, two distinct terms in (7.114) may have the same connectivity, and hence, share the same Feynman graph. Thus, (7.114) is a weighted sum of Feynman graphs and $(i/\hbar)W[J]$ is a weighted sum of connected Feynman graphs.

A terminal can pick out one or the other end of a line, producing two distinct terms in (7.114) that are equal in numerical value. Hence the factor $\frac{1}{2}$ in $\frac{1}{2}(\underset{x}{\circlearrowleft} - \underset{y}{\circlearrowright})$ can be omitted, provided we do not distinguish the two ends of a line, for the purpose of drawing Feynman graphs. To incorporate this rule, we write symbolically

$$\exp \frac{i}{\hbar} W[J] = \sum_{m=0}^{\infty} \sum_{n=0}^{\infty} \frac{1}{m! n!} \int \left[(\overset{\mathbf{X}}{z_1}) \cdots (\overset{\mathbf{X}}{z_m}) \right] \left[(\underset{1}{\circlearrowleft} - \underset{1}{\circlearrowright}) \cdots (\underset{n}{\circlearrowleft} - \underset{n}{\circlearrowright}) \right]. \quad (7.115)$$

It is understood that attaching one end of a line to a terminal is not counted as distinct from attaching the other end. The integration is performed after all vertices have been formed, and extends over the coordinates of all vertices and all free ends of external lines.

From now on, we only consider connected graphs. Let N be the number of terms in (7.115) that have the same pattern of connectivity corresponding to a connected graph, having m vertices and n lines (internal or external). Its weight is given by $N/(m! n!)$. [By convention the factor $1/(m! n!)$ in (7.115) is not included in the definition of a graph]. We define the *symmetry number*³ s of a

³ T. T. Wu, *Phys. Rev.* **125**, 1436 (1962).

connected graph as

$$\frac{1}{s} = \frac{N}{m! n!} \quad (7.116)$$

Then

$$\frac{i}{\hbar} W[J] = \sum_G \frac{G}{s}, \quad (7.117)$$

where the sum extends over all connected graphs G .

Consider a connected non-vacuum graph (i.e., a graph with at least two external lines), with m vertices and n lines (external or internal). Every permutation of the m vertices corresponds to a distinct term in (7.115), because a vertex can be uniquely identified by its ordered position along some line that goes from one external end to another. Renumbering the vertices along this line amounts to a change in the assignment of the terminals in (7.115) to the vertices, and hence gives a new term. Therefore, N contains a factor $m!$. The interchange of two lines corresponds to a distinct term in (7.115), unless the two lines are internal lines that share the same vertices at both ends. Such internal lines are called “equivalent”. The number of distinct permutations of lines is therefore $n!/(k_1! k_2! \dots)$, where k_1, k_2, \dots are the numbers of internal lines in equivalent sets. Therefore, $N = n! m!/(k_1! k_2! \dots)$, and the symmetry number of a connected non-vacuum graph is

$$s = \prod_i (k_i!). \quad (7.118)$$

The symmetry number of a vacuum graph does not follow any simple general rule, and has to be worked out for each individual case^c. Some examples of symmetry numbers are given in Fig. 7.2.

7.8 Boson Loops and Fermion Loops

Consider a free complex scalar field coupled to an external source $\Omega(x)$ that creates particle-antiparticle pairs:

$$\begin{aligned} \mathcal{L}(x) &= \partial_\mu \phi^* \partial^\mu \phi - m^2 \phi^* \phi + \phi^* \phi \Omega \\ &= \tfrac{1}{2} [\partial_\mu \phi_1 \partial^\mu \phi_1 + (\Omega - m^2) \phi_1^2] + \tfrac{1}{2} [\partial_\mu \phi_2 \partial^\mu \phi_2 + (\Omega - m^2) \phi_2^2], \end{aligned} \quad (7.119)$$

where

$$\begin{aligned} \phi &= (\phi_1 + i\phi_2)/\sqrt{2}, \\ \phi^* &= (\phi_1 - i\phi_2)/\sqrt{2}. \end{aligned} \quad (7.120)$$

^c Consider, for example, the term

$$\frac{1}{2!} \left[\begin{smallmatrix} (\mathbf{X}) & (\mathbf{X}) \\ 1 & 2 \end{smallmatrix} \right] \frac{1}{4!} \left[\begin{smallmatrix} (\circ-\circ) & \cdots & (\circ-\circ) \\ 1 & & 4 \end{smallmatrix} \right].$$

There is only one way of connecting the lines to produce a vacuum graph, i.e., the fifth graph in Fig. 7.2. The symmetry number is therefore $2! 4! = 48$.

Any connected Feynman graph of this theory consists of a single loop, on which any number of vertices can be placed.

We can calculate the generating functional $W[\Omega]$ in closed form as follows: the classical action is

$$S[\phi_1, \phi_2, \Omega] = i \int d^4x_E \mathcal{L}(x_E) = i\hbar[(\phi_1, A\phi_1) + (\phi_2, A\phi_2)], \quad (7.121)$$

where

$$A = \hbar^{-1}(-\square_E^2 + m^2 - \Omega). \quad (7.122)$$

Hence

$$\begin{aligned} \exp \frac{i}{\hbar} W[\Omega] &= \mathcal{N} \int (D\phi_1)(D\phi_2) \exp \frac{i}{\hbar} S[\phi_1, \phi_2, \Omega] \\ &= \mathcal{N} \left[\int (D\phi_1) e^{-(\phi_1, A\phi_1)} \right]^2 = \mathcal{N} \det A^{-1}. \end{aligned} \quad (7.123)$$

The Feynman graph expansion is obtained by expanding the above in powers of Ω . To do this, let us introduce the notation

$$\begin{aligned} i\Delta_F &= (-\square_E^2 + m^2)^{-1}, \\ \langle x|\Delta_F|y\rangle &= \Delta_F(x - y), \\ \langle x|\Omega|y\rangle &= \delta^4(x - y)\Omega(x), \\ \text{Tr } f &= \int d^4x_E \langle x_E | f | x_E \rangle. \end{aligned} \quad (7.124)$$



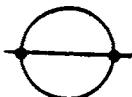
$S = 1$



$S = 2$



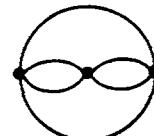
$S = 2$



$S = 6$



$S = 48$



$S = 8$

Fig. 7.2 Examples of symmetry number for Feynman graphs in ϕ^4 theory

We can then write

$$\begin{aligned} A &= \hbar^{-1}[(i\Delta_F)^{-1} - \Omega] \\ \det A &= \det(i\hbar\Delta_F)^{-1} \cdot \det(1 - i\Delta_F\Omega). \end{aligned} \quad (7.125)$$

Choosing $\mathcal{N} = \det(i\hbar\Delta_F)$, we have

$$\frac{i}{\hbar} W[\Omega] = -\ln \det(1 - i\Delta_F\Omega) = -\text{Tr} \ln(1 - i\Delta_F\Omega). \quad (7.126)$$

Expanding this in powers of Ω gives

$$\begin{aligned} \frac{i}{\hbar} W[\Omega] &= \sum_{n=1}^{\infty} \frac{1}{n} \text{Tr}(i\Delta_F\Omega)^n \\ &= \sum_{n=1}^{\infty} \frac{1}{n} \int d^4y_1 \cdots d^4y_n [i\Delta_F(y_1 - y_2)i\Delta_F(y_2 - y_3) \cdots i\Delta_F(y_n - y_1)] \cdot \Omega(y_1) \cdots \Omega(y_n), \end{aligned} \quad (7.127)$$

where all the y 's are Euclidean coordinates. This shows that there is only one n th order connected graph, consisting of one closed loop with n vertices on it, and the symmetry number is n .

If the field $\phi(x)$ obeyed Fermi statistics instead of Bose statistics, then every closed loop would be associated with an extra factor -1 , but the Feynman rules are otherwise unchanged. Since in this example all connected graphs are one-loop graphs, they would simply change sign, and so would $W[\Omega]$. To reproduce the Feynman rules, we must redefine the meaning of the path integral in a manner such that (7.123) turns into its reciprocal, i.e.,

$$\int (D\phi_1)(D\phi_2) \exp \frac{i}{\hbar} S[\phi_1; \phi_2, \Omega] = \det A. \quad (7.128)$$

In this particular example, the use of Fermi statistics is of course unphysical (making ϕ a "ghost field," which is used only as a formal device in the quantization of gauge fields). The main point of the example is to motivate a way for the treatment of a physical fermion field theory by the method of path integrals, which we shall discuss in the next section.

7.9 Fermion Fields

In canonical quantization, fermion fields $\psi_j(x)$, $\psi_k^\dagger(x)$ are defined by anticommutation rules:

$$\begin{aligned} \{\psi_j(x), \psi_k^\dagger(y)\}_{x_0=y_0} &= \hbar\delta^3(\mathbf{x} - \mathbf{y}), \\ \{\psi_j(x), \psi_k(y)\}_{x_0=y_0} &= 0. \end{aligned} \quad (7.129)$$

This suggests that in the formal classical limit $\hbar \rightarrow 0$, $\psi_j(x)$ and $\psi_k^\dagger(x)$ should be represented not by numbers, but by anticommuting c-numbers. To extend the method of path integrals to the fermion case, we would have to define functional integrals over anticommuting c-numbers.

For this purpose, we adopt the point of view that a quantum field theory is defined by its Feynman rules. Thus, fermion fields differ from boson fields *only* in that a closed fermion loop in a Feynman graph is associated with an extra factor -1 . We shall simply write down a path integral for the generating functional that reproduces this rule.

Consider a free spinor field coupled to an external source $\Omega(x)$ that creates fermion-antifermion pairs:

$$\mathcal{L}(x) = \bar{\psi}(i\gamma \cdot \partial - m)\psi + \bar{\psi}\psi\Omega, \quad (7.130)$$

with the action

$$\int d^4x \mathcal{L}(x) = i \int d^4x_E \mathcal{L}(x_E) = i\hbar[(\bar{\psi}, S_F^{-1}\psi) - (\bar{\psi}, \Omega\psi)] \quad (7.131)$$

where

$$(f, g) \equiv \int d^4x_E f(x_E)g(x_E), \\ S_F \equiv (i\gamma \cdot \partial_E + m)^{-1}.$$

All connected Feynman graphs are one-loop graphs, just as in the example in the last section. The Feynman rules would be correctly given if we could define the generating functional $W[\Omega]$ through a path integral formula such that

$$\exp \frac{i}{\hbar} W[\Omega] \equiv \mathcal{N} \int (D\psi)(D\bar{\psi}) \exp \left[\frac{i}{\hbar} (\bar{\psi}, (S_F^{-1} - \Omega)\psi) \right] \quad (7.133) \\ = \mathcal{N}' \det(S_F^{-1} - \Omega).$$

To see how this can be done, first consider two anticommuting c-numbers η_1 and η_2 , defined by

$$\{\eta_1, \eta_2\} = 0, \\ \eta_1^2 = \eta_2^2 = 0. \quad (7.134)$$

Any function of η_1 and η_2 can be written in the form

$$f(\eta_1, \eta_2) = C_0 + C_1\eta_1 + C_2\eta_2 + C_3\eta_1\eta_2. \quad (7.135)$$

where C_i is a complex number. The operation of “integration” is defined by the rules⁴

$$\int d\eta_1 = 0, \quad \int d\eta_2 = 0, \\ \int d\eta_1\eta_1 = 1, \quad \int d\eta_2\eta_2 = 1, \quad (7.136)$$

where $d\eta_1$ and $d\eta_2$ are supposed to anticommute with each other, and with η_1 and η_2 . We also define

⁴ F. A. Berezin, *Method of Second Quantization* (Academic Press, New York, 1966).

$$\int d\eta_1 \eta_2 \equiv -\eta_2 \int d\eta_1, \quad (7.137)$$

from which follows

$$\int d\eta f(\eta - \eta_0) = \int d\eta f(\eta). \quad (7.138)$$

Integrating (7.135) by these rules, we find

$$\begin{aligned} \int d\eta_1 f(\eta_1, \eta_2) &= C_1 + C_3 \eta_2, \\ \int d\eta_2 \int d\eta_1 f(\eta_1, \eta_2) &= C_3. \end{aligned} \quad (7.139)$$

In particular, for any number A ,

$$\int d\eta_2 \int d\eta_1 e^{A\eta_1 \eta_2} = \int d\eta_2 \int d\eta_1 (1 + A\eta_1 \eta_2) = A. \quad (7.140)$$

We can generalize the definitions above to any number of anticommuting c-numbers $\{\eta_i\}$:

$$\begin{aligned} \{\eta_i, \eta_j\} &= 0, \quad \eta_i^2 = 0, \\ \int d\eta_i &= 0, \quad \int d\eta_i \eta_i = 1 \quad (\text{no sum}). \end{aligned} \quad (7.141)$$

All $d\eta_i$ anticommute among themselves, and with all η_j . Let us divide $\{\eta_i\}$ into two disjoint sets $\{\eta_\alpha\}$ and $\{\eta_\alpha^*\}$. (The asterisk merely serves to distinguish between the two sets, and does not denote complex conjugation). Consider the quadratic form

$$X = \sum_{\alpha, \beta} \eta_\alpha^* A_{\alpha\beta} \eta_\beta \equiv \eta^* A \eta, \quad (7.142)$$

where A is a symmetric matrix whose elements are numbers. Through a linear transformation, this can be reduced to diagonal form:

$$X = \sum_{\alpha=1}^n A_\alpha \eta_\alpha^* \eta_\alpha, \quad (7.143)$$

where A_α are the eigenvalues of the matrix A . Because of (7.141),

$$e^X = 1 + X + \frac{1}{2!} X^2 + \dots + \frac{1}{n!} X^n, \quad (7.144)$$

$$\int (D\eta)(D\eta^*) e^X = \int (D\eta)(D\eta^*) \frac{X^n}{n!},$$

where

$$(D\eta) = \prod_{\alpha=1}^n d\eta_\alpha, \quad (D\eta^*) = \prod_{\alpha=1}^n d\eta_\alpha^*. \quad (7.145)$$

The signs of these quantities depend on the order in which the factors are arranged. To obtain one term in the expansion of X^n , we must choose one term from each of the n factors X and multiply them together. Since $\eta_\alpha^{*2} = \eta_\alpha^2 = 0$, all the terms we choose must be different from one another. Hence, the only possible result is $\prod_\alpha A_\alpha \eta_\alpha^* \eta_\alpha$. Since there are $n!$ ways of choosing, we have

$$X^n = n! \prod_\alpha A_\alpha \eta_\alpha^* \eta_\alpha. \quad (7.146)$$

Hence (up to a sign),

$$\int (D\eta)(D\eta^*) e^{\eta^* A \eta} = \prod_{\alpha=1}^n A_\alpha = \det A. \quad (7.147)$$

We may look upon this formula as a novel way of representing a determinant.

More generally, let

$$Q(\eta, \eta^*) = \eta^* A \eta - b^* \eta - b \eta^*, \quad (7.148)$$

where A is a non-singular symmetric matrix whose elements are numbers, and b^* and b are n -vectors whose components are anticommuting c-numbers. We rewrite

$$\begin{aligned} Q(\eta, \eta^*) &= (\eta^* - \eta_0^*) A (\eta - \eta_0) - b^* A^{-1} b, \\ \eta_0 &= A^{-1} b, \\ \eta_0^* &= b^* A^{-1}. \end{aligned} \quad (7.149)$$

Then, by (7.138) and (7.147), we have (up to a sign)

$$\int (D\eta)(D\eta^*) e^{Q(\eta, \eta^*)} = e^{-b^* A^{-1} b} \det A. \quad (7.150)$$

It is now obvious that to obtain (7.133), all we have to do is to represent the spinor fields $\psi(x)$ and $\bar{\psi}(x)$ by anticommuting c-numbers, regarding x as a discrete label first, and then pass to the continuum limit in the final result.

For a free spinor field coupled to external sources that create particles singly, we take

$$\mathcal{L}(x) = \bar{\psi}(i\gamma \cdot \partial - m)\psi + J\psi + J^*\bar{\psi} \quad (7.151)$$

where $J(x)$ and $J^*(x)$ are 4-component anticommuting c-numbers. The generating functional $W[J, J^*]$, which generates connected Feynman graphs with external lines, is then given by

$$\begin{aligned} \exp \frac{i}{\hbar} W[J, J^*] &= \mathcal{N} \int (D\psi)(D\bar{\psi}) \exp \left\{ \frac{i}{\hbar} [(\bar{\psi}, S_F^{-1}\psi) + (J, \psi) + (J^*, \bar{\psi})] \right\} \\ &= \mathcal{N}' e^{\hbar^{-1}(J^*, S_F J)} \det S_F^{-1}. \end{aligned} \quad (7.152)$$

To use the final result, it matters little whether J and J^* are anticommuting c-numbers or just ordinary numbers, for they occur paired. Indeed, the result may be derived easily by more conventional means; the use of anticommuting c-numbers here may seem to be a quaint way of doing it. The virtue of the

method, however, lies in the fact that it enables us to treat fermion fields and boson fields on equal footing by the method of path integrals. This is useful when we consider fermion and boson fields that interact with each other.

CHAPTER 8

QUANTIZATION OF GAUGE FIELDS

8.1 Canonical Quantization

1 Free Maxwell Field

The quantization of gauge fields presents special problems associated with the freedom of gauge transformations. To see how they arise, we first review the canonical quantization of the free Maxwell field, whose classical Lagrangian density is given by

$$\begin{aligned}\mathcal{L}(x) &= -\frac{1}{4}F^{\mu\nu}F_{\mu\nu} = \frac{1}{2}(\mathbf{E} \cdot \mathbf{E} - \mathbf{B} \cdot \mathbf{B}), \\ \mathbf{B} &\equiv \nabla \times \mathbf{A}, \\ \mathbf{E} &\equiv -\frac{\partial \mathbf{A}}{\partial t} - \nabla A^0.\end{aligned}\tag{8.1}$$

The independent variables are the field variables $A^\mu(x)$, regarded as coordinates, and their time derivatives $\dot{A}^\mu(x)$. Changing A^μ by a gauge transformation $A^\mu \rightarrow A^\mu + \partial^\mu\omega$ leaves $\mathcal{L}(x)$ invariant.

According to the rules of canonical quantization, we must first cast the theory in Hamiltonian form, and then impose canonical commutation relations between the coordinates and their corresponding conjugate momenta. Noting that $\mathcal{L}(x)$ is independent of $\dot{A}^0(x)$, we see that the momentum conjugate to $A^0(x)$ is identically zero. Therefore $A^0(x)$ is not an independent coordinate, and may be eliminated through the classical constraint equation $\nabla \cdot \mathbf{E} = 0$, or

$$\nabla^2 A^0 + \frac{\partial}{\partial t} \nabla \cdot \mathbf{A} = 0.$$

A common way to proceed is to set $A^0 = 0$, which corresponds to temporal gauge. Another way is to set $\nabla \cdot \mathbf{A} = 0$, which corresponds to Coulomb gauge. From the above equation we see that they are equivalent. We shall adopt temporal gauge here.

After setting $A^0 = 0$, $\mathcal{L}(x)$ is invariant under a more restrictive residual gauge transformation $\mathbf{A} \rightarrow \mathbf{A} - \nabla\omega$, where ω is a time-independent function. The classical Hamiltonian, which is also invariant under the residual gauge transformation, is given by

$$H = \frac{1}{2} \int d^3x [\mathbf{E} \cdot \mathbf{E} + (\nabla \times \mathbf{A})^2]. \quad (8.2)$$

The independent variables are the fields $\mathbf{A}(x)$, regarded as coordinates, and their conjugate momenta $-\mathbf{E}(x)$. The theory is quantized by imposing the equal-time commutation relations

$$[E^j(x), A^k(y)]_{x^0=y^0} = i\delta_{jk}\delta^3(\mathbf{x} - \mathbf{y}). \quad (8.3)$$

However, not all three components of \mathbf{A} represent dynamical degrees of freedom, owing to the residual gauge invariance. To complete the definition of the theory, we must explicitly eliminate the unphysical degrees of freedom.

The residual gauge invariance is expressed through the fact

$$[H, \nabla \cdot \mathbf{E}] = 0. \quad (8.4)$$

In other words, $\nabla \cdot \mathbf{E}$ is a constant of the motion. Consequently, the longitudinal part \mathbf{E}_{\parallel} is not a dynamical variable, but is completely determined by initial data and boundary conditions. The longitudinal part \mathbf{A}_{\parallel} is also completely determined, because

$$\frac{\partial \mathbf{A}}{\partial t} = -\mathbf{E}, \quad (8.5)$$

which follows directly from $\partial \mathbf{A} / \partial t = i[H, \mathbf{A}]$. Therefore, only the transverse parts \mathbf{E}_{\perp} and \mathbf{A}_{\perp} are truly dynamical variables. Their commutation relations are the transverse projections of (8.3):

$$[E_{\perp}^j(x), A_{\perp}^k(y)]_{x^0=y^0} = i\delta_{jk}^{\text{Tr}}(\mathbf{x} - \mathbf{y}), \quad (8.6)$$

where

$$\delta_{jk}^{\text{Tr}}(\mathbf{x} - \mathbf{y}) = \int \frac{d^3p}{(2\pi)^3} e^{i\mathbf{p} \cdot (\mathbf{x} - \mathbf{y})} \left[\delta_{jk} - \frac{p^j p^k}{|\mathbf{p}|^2} \right]. \quad (8.7)$$

The Hamiltonian can be rewritten in the form

$$H = \frac{1}{2} \int d^3x [\mathbf{E}_{\perp} \cdot \mathbf{E}_{\perp} + (\nabla \times \mathbf{A}_{\perp})^2] + \frac{1}{2} \int d^3x \mathbf{E}_{\parallel} \cdot \mathbf{E}_{\parallel}. \quad (8.8)$$

The last term is a c-number constant, which has no physical significance.

The residual gauge transformation $\mathbf{A} \rightarrow \mathbf{A} - \nabla\omega$ only affects \mathbf{A}_{\parallel} , and therefore merely changes the reference point of the total energy. Thus, we may choose $\mathbf{A}_{\parallel} = 0$ for simplicity. All states of the system will then satisfy the subsidiary condition $\nabla \cdot \mathbf{E} = 0$, which represents a restriction on the Hilbert space of the system.

To see this in more formal detail, let us represent a state in Hilbert space by the coordinate representative $\Psi[\mathbf{A}]$. Then the Schrödinger operator $\mathbf{E}(\mathbf{x})$ is represented by

$$E^k(\mathbf{x}) = i \frac{\delta}{\delta A^k(\mathbf{x})}. \quad (8.9)$$

By virtue of (8.4), H and $\nabla \cdot \mathbf{E}$ can be simultaneously diagonalized. Therefore, an eigenfunction $\Psi[\mathbf{A}]$ of H can be chosen to be an eigenfunction of $\nabla \cdot \mathbf{E}$:

$$\nabla \cdot \mathbf{E}(\mathbf{x})\Psi[\mathbf{A}] = \rho(\mathbf{x})\Psi[\mathbf{A}], \quad (8.10)$$

where $\rho(\mathbf{x})$ is an arbitrary function, the charge density of an external static charge distribution.

Under the residual gauge transformation, $\Psi[\mathbf{A}]$ changes to $\Psi[\mathbf{A} - \nabla\omega]$, which we can express in the form^a

$$\begin{aligned}\Psi[\mathbf{A} - \nabla\omega] &= \left\{ \exp \int d^3x [-\partial_i \omega(\mathbf{x})] \frac{\delta}{\delta A^i(\mathbf{x})} \right\} \Psi[\mathbf{A}] \\ &= \exp \left(i \int d^3x (\nabla\omega) \cdot \mathbf{E} \right) \Psi[\mathbf{A}] = \exp \left(-i \int d^3x \omega \nabla \cdot \mathbf{E} \right) \Psi[\mathbf{A}].\end{aligned} \quad (8.11)$$

Now, any $\mathbf{A}(\mathbf{x})$ can be written as $\mathbf{A}(\mathbf{x}) = \mathbf{A}_\perp(\mathbf{x}) - \nabla\omega(\mathbf{x})$. That is, the longitudinal part of \mathbf{A} is pure-gauge. In the absence of external magnetic flux, all pure-gauge fields can be transformed to zero continuously. Thus, by (8.10) and (8.11), we can always write, for any \mathbf{A} ,

$$\Psi[\mathbf{A}] = e^{-i\chi}\Psi[\mathbf{A}_\perp], \quad (8.12)$$

where

$$\begin{aligned}\chi &= \int d^3x \omega\rho = \int d^3x \omega \nabla^2 \left(\frac{1}{\nabla^2} \rho \right) = \int d^3x (\nabla^2 \omega) \frac{1}{\nabla^2} \rho \\ &= - \int d^3x (\nabla \cdot \mathbf{A}) \frac{1}{\nabla^2} \rho = \int d^3x d^3y \frac{[\nabla \cdot \mathbf{A}(\mathbf{x})]\rho(\mathbf{y})}{4\pi|\mathbf{x} - \mathbf{y}|}.\end{aligned} \quad (8.13)$$

The wave function $\Psi[\mathbf{A}_\perp]$ is an eigenfunction of the new Hamiltonian

$$e^{i\chi} H e^{-i\chi} = \frac{1}{2} \int d^3x \left[E_\perp^2 + (\nabla \times \mathbf{A}_\perp)^2 \right] + \frac{1}{2} \int d^3x d^3y \frac{\rho(\mathbf{x})\rho(\mathbf{y})}{4\pi|\mathbf{x} - \mathbf{y}|}, \quad (8.14)$$

which differs from H only by a constant, the energy of an arbitrary static charge distribution. In the physical Hilbert space spanned by $\Psi[\mathbf{A}_\perp]$, (8.10) becomes the subsidiary condition

$$\nabla \cdot \mathbf{E}(\mathbf{x})\Psi[\mathbf{A}_\perp] = 0, \quad (8.15)$$

which is the quantum version of Gauss' law.

In summary, the temporal gauge condition does not fix the gauge completely, but allows residual gauge transformations, which can be expressed in terms of an arbitrary external static charge density. Under a residual gauge transformation, the wave functional of the system changes only by a phase. When we fix the gauge completely by setting the external charge density to zero, the wave functional

^a We drop surface integrals, rather carelessly. The real justification is that in Maxwell theory there is no topological charge. (See the treatment of Yang-Mills fields in Sec. 8.6.).

becomes completely gauge invariant, and depends only on the transverse part of the vector potential.

The same results can be obtained in Coulomb gauge, in which the vector potential is required to be transverse from the beginning. The residual gauge freedom in this case corresponds to a gauge function that must be a solution of the Laplace equation, and is completely fixed by boundary conditions.

2 Pure Yang-Mills Fields

The quantization of Yang-Mills fields is full of problems, owing to the non-linear nature of the theory. For example, the Coulomb gauge turns out to be pathological, except in perturbation theory, due to the “Gribov ambiguity”, which will be discussed in a later section. A manifestation of the pathology is that the instanton solution in Coulomb gauge is discontinuous in time. This is a special case of a more general malaise. It has been proven, in fact, that no gauge fixing is possible in classical Yang-Mills theory. That is, given any condition that completely fixes the gauge, there does not always exist a continuous gauge transformation that would bring a continuous field configuration into the desired gauge.¹

The temporal gauge seems to be the only practical way to proceed. In this case, the residual gauge freedom has to be removed by imposing the non-Abelian version of Gauss’ law as a constraint on physical states:

$$(\nabla \cdot \mathbf{E}_a + gC_{abc}\mathbf{A}_b \cdot \mathbf{E}_c) \Psi[A] = 0.$$

These constraints are much more complicated than their counterparts in Maxwell theory, because the operators that act on Ψ do not commute with one another. Also, they do not naturally separate the fields into transverse and longitudinal parts. Consequently, the equivalence between temporal and Coulomb gauge, so simple in Maxwell theory, is completely lost here.

For pedagogical reasons, we postpone further discussions of canonical quantization in the temporal gauge to Sec. 8.6, in connection with the effect of topological charges, and Sec. 8.8, where projection operators will be introduced.

In the sections immediately following, we shall describe quantization via path integrals. This approach gives us a unified way to derive formal expressions for transition amplitudes. However, the above-mentioned problems are still present, in that the formal expressions may not exist.

It is interesting to note that, in contrast to the troubles we face here, quantization presents no problem whatsoever in lattice gauge theory, which we shall discuss in a later chapter. There, gauge fixing is not necessary, but trivial to carry out if desired.

8.2 Path Integral Method in Hamiltonian Form

The path integral in Hamiltonian form, given in the simplest instance by (7.18), was derived directly from canonical quantization and contains nothing new in principle. However, it enables us to eliminate the unphysical degrees of freedom in a much simpler fashion. We illustrate this with free Maxwell theory.

¹ I. M. Singer, *Commun. Math. Phys.* **60**, 7 (1978). See also T. P. Killingback, *Phys. Lett.* **138B**, 87 (1983).

The classical Hamiltonian is

$$H = \int d^3x \left\{ \frac{1}{2} [\mathbf{E} \cdot \mathbf{E} + (\nabla \times \mathbf{A})^2] + \mathbf{E} \cdot \nabla A^0 \right\}. \quad (8.17)$$

The independent coordinates are $\mathbf{A}(x)$ and $A^0(x)$, with $-\mathbf{E}(x)$ conjugate to $\mathbf{A}(x)$, while the momentum conjugate to $A^0(x)$ is absent. Thus $A^0(x)$ is not a dynamical variable. Let $|A, t\rangle$ be defined by

$$\mathbf{A}_{op}(\mathbf{x}, t)|A, t\rangle = \mathbf{A}(\mathbf{x})|A, t\rangle, \quad (8.18)$$

which is the analogue of (7.5). The transition amplitude is then given by

$$\langle \mathbf{A}_2, t_2 | \mathbf{A}_1, t_1 \rangle = \mathcal{N} \int (DA^0)(DA)(DE) \cdot \exp \frac{1}{i} \int_{t_1}^{t_2} d^4x \left[\frac{1}{2} E^2 + \frac{1}{2} (\nabla \times \mathbf{A})^2 + \mathbf{E} \cdot \left(\frac{\partial \mathbf{A}}{\partial t} + \nabla A^0 \right) \right], \quad (8.19)$$

where A^0 , \mathbf{A} , \mathbf{E} , are functions of \mathbf{x} and t , and the integral $\int_{t_1}^{t_2} d^4x$ extends over all space between the times t_1 and t_2 , with $\mathbf{A}(\mathbf{x}, t)$ constrained by

$$\begin{aligned} \mathbf{A}(\mathbf{x}, t_1) &= \mathbf{A}_1(\mathbf{x}), \\ \mathbf{A}(\mathbf{x}, t_2) &= \mathbf{A}_2(\mathbf{x}). \end{aligned} \quad (8.20)$$

We can immediately carry out the path integration over A^0 :

$$\int (DA^0) \exp i \int d^4x (\nabla \cdot \mathbf{E}) A^0 \propto \delta[\nabla \cdot \mathbf{E}]. \quad (8.21)$$

This requires the longitudinal part of \mathbf{E} to vanish. Thus, $\int (DE)$ may be replaced by $\int (DE_\perp)$, and $(E \cdot \partial \mathbf{A} / \partial t)$ by $(E_\perp \cdot \partial \mathbf{A}_\perp / \partial t)$, (since $\int d^3x E_\parallel \cdot \partial \mathbf{A}_\parallel / \partial t = 0$):

$$\begin{aligned} \langle \mathbf{A}_2, t_2 | \mathbf{A}_1, t_1 \rangle &= \mathcal{N} \int (DA_\parallel)(DA_\perp)(DE_\perp) \\ &\cdot \exp \frac{1}{i} \int_{t_1}^{t_2} d^4x \left[\frac{1}{2} E_\perp^2 + \frac{1}{2} (\nabla \times \mathbf{A}_\perp)^2 + \mathbf{E} \cdot \frac{\partial \mathbf{A}_\perp}{\partial t} \right] \\ &= \mathcal{N} \int (DA_\parallel)(DA_\perp) \exp \frac{i}{2} \int_{t_1}^{t_2} d^4x \left[\left(\frac{\partial \mathbf{A}_\perp}{\partial t} \right)^2 - (\nabla \times \mathbf{A}_\perp)^2 \right]. \end{aligned} \quad (8.22)$$

Since the integrand is independent of A_\parallel , $\int (DA_\parallel)$ may be absorbed into \mathcal{N} . This means that the longitudinal parts of \mathbf{A}_2 and \mathbf{A}_1 have no physical significance, for they only affect the normalization of the transition amplitude. This reflects the gauge invariance of the system, as \mathbf{A}_\parallel can be changed at will through a gauge transformation. Thus, we have recovered all the results of canonical quantization in Coulomb gauge in a much more streamlined fashion.

We note that (8.22) is in Feynman form:

$$\begin{aligned}\langle \mathbf{A}_{\perp 2}, t_2 | \mathbf{A}_{\perp 1}, t_1 \rangle &= \mathcal{N} \int (\mathrm{d}\mathbf{A}_{\perp}) \exp i \int_1^2 d^4x \mathcal{L}_{\perp}(x), \\ \mathcal{L}_{\perp}(x) &= \frac{1}{2} \left[\left(\frac{\partial \mathbf{A}_{\perp}}{\partial t} \right)^2 - (\nabla \times \mathbf{A}_{\perp})^2 \right].\end{aligned}\quad (8.23)$$

This result is obtained by eliminating the unphysical degrees of freedom in Coulomb gauge. Unphysical fields can be more generally defined as those that can be eliminated through a gauge transformation. Since physical results should be gauge invariant, we expect the Feynman form to be valid when any suitable gauge condition is used to eliminate the unphysical fields.

8.3 Feynman Path Integral: Fadeev-Popov Method

The previous example suggests that the Feynman path integral is valid provided we fix the gauge. Fadeev and Popov² showed how this could be done.

The failure of the unmodified Feynman formula and its possible remedy can be seen most clearly in Maxwell theory as follows. The action is given by

$$\begin{aligned}S[A] &= -\frac{1}{4} \int d^4x F^{\mu\nu} F_{\mu\nu} = -\frac{1}{2} \int d^4x (\partial^{\mu} A^{\nu} - \partial^{\nu} A^{\mu}) \partial_{\mu} A_{\nu} \\ &= \frac{1}{2} \int d^4x A_{\mu}(x) (g^{\mu\nu} \square^2 - \partial^{\mu} \partial^{\nu}) A_{\nu}(x) \\ &= -\frac{1}{2} \int \frac{d^4k}{(2\pi)^4} \tilde{A}_{\mu}(-k) (g^{\mu\nu} k^2 - k^{\mu} k^{\nu}) \tilde{A}_{\nu}(k),\end{aligned}\quad (8.24)$$

where $\tilde{A}^{\mu}(k)$ is the Fourier transform of $A^{\mu}(x)$. The kinetic operator $(g^{\mu\nu} k^2 - k^{\mu} k^{\nu})$ has no inverse, because it has eigenfunctions k^{ν} , with eigenvalue zero. The propagator therefore does not exist. One remedy is to restrict \tilde{A}^{μ} with the condition $k_{\mu} \tilde{A}^{\mu} = 0$ (Lorentz gauge), so that the eigenvectors k^{ν} are no longer physically relevant.

To see the problem from another point of view, note that $S[A]$ is gauge invariant, and therefore independent of any fields that can be gauge-transformed away. If we include such fields in the Feynman path integral, we would obtain an infinite volume factor. This can be absorbed into the normalization constant only if these fields are clearly identified, and hence gauge fixing is called for.

From yet another point of view, the necessity for gauge fixing comes from our knowledge that the Green's functions of the theory are gauge-dependent. Hence, the generating functional $W[J]$ is not uniquely defined unless the gauge is specified.

A gauge transformation

$$U(x) = e^{-i\omega_a(x)L_a}$$

² L. Fadeev and V. N. Popov, *Phys. Lett.* **25B**, 29 (1967).

is parametrized by the functions $\omega_a(x)$, denoted collectively by ω . The gauge fields $A^\mu(x) = A_a^\mu(x)L_a$ change according to

$$\begin{aligned} A_\mu &\rightarrow A_\mu \omega, \\ A_\mu \omega &= UA_\mu U^{-1} + \frac{i}{g} U\partial_\mu U^{-1}. \end{aligned} \quad (8.25)$$

Fixing the gauge means that we impose some condition on A_a^μ , called a gauge condition. Some commonly used gauge conditions are

$$\begin{aligned} \text{Lorentz gauge: } &\partial_\mu A_a^\mu = 0, \\ \text{Coulomb gauge: } &\nabla \cdot \mathbf{A}_a = 0, \\ \text{Axial gauge: } &A_a^3 = 0, \\ \text{Temporal gauge: } &A_a^0 = 0. \end{aligned} \quad (8.26)$$

A general gauge condition may be represented in the form

$$\hat{f}_a A = 0 \quad (a = 1, \dots, N), \quad (8.27)$$

where \hat{f}_a is a mapping of the function space of $A_a^\mu(x)$ into itself. The admissible A 's are those that are mapped into zero by each \hat{f}_a ($a = 1, \dots, N$).

As long as the gauge condition can eliminate all unphysical degrees of freedom, there will be no difficulty in passing from the Hamiltonian form of the path integral to the Feynman form. In that case, the transition amplitude in Feynman form is given by

$$\langle A_2, t_2 | A_1, t_1 \rangle = \mathcal{N} \int_{A_1}^{A_2} (\mathrm{D}A) e^{iS_{21}[A]} \mathcal{J}[A] \prod_{a=1}^N \delta[\hat{f}_a A], \quad (8.28)$$

where

$$S_{21}[A] = \int_1^2 d^4x \mathcal{L}(x),$$

and $\mathcal{J}[A]$ is a Jacobian defined by

$$\int (\mathrm{D}A) \mathcal{J}[A] \prod_{a=1}^N \delta[\hat{f}_a A] = 1. \quad (8.29)$$

In practice, one can tell whether a particular gauge condition works by seeing whether the kinetic operator has an inverse. In labelling the transition amplitude $\langle A_2, t_2 | A_1, t_1 \rangle$, we need not specify explicitly which components of A are the dynamical ones, as gauge-fixing does the job automatically; the unphysical ones affect only the normalization of the transition amplitude.

Faddeev and Popov showed a convenient way of calculating the Jacobian $\mathcal{J}[A]$. A side benefit of their method is that one can relax the gauge condition. Instead of demanding $\hat{f}_a A = 0$, one may allow $\hat{f}_a A$ to take on any functional form, but assign a definite weight to each functional form. This results in greater flexibility of the method. We now follow their development.

We assume that \hat{f}_a have the property that, given any A , there always exists a gauge transformation ω such that

$$\hat{f}_a A^\omega = 0. \quad (8.30)$$

Thus, given A , $\{\hat{f}_a\}$ determines a gauge transformation $U(x)$. As x varies, $U(x)$ traces out an orbit in the gauge-group manifold. The volume of this orbit is denoted by $1/\Delta_f[A]$:

$$\Delta_f[A] \int (D\omega) \delta[\hat{f}A^\omega] = 1, \quad (8.31)$$

where $(D\omega) = \prod_x \prod_a d\omega_a(x)$, and

$$\delta[\hat{f}A] = \prod_{a=1}^N \delta[\hat{f}_a A]. \quad (8.32)$$

Clearly $\Delta_f[A]$ is gauge invariant:

$$\Delta_f[A] = \Delta_f[A^\omega]. \quad (8.33)$$

By integrating both sides of (8.31) over A , one can manipulate it into the form

$$[f(D\omega)/f(DA)] \int (DA) \Delta_f[A] \delta[\hat{f}A] = 1, \quad (8.34)$$

which shows that $\Delta_f[A]$ is proportional to the Jacobian $J[A]$.

To calculate $\Delta_f[A]$, let $f_a(x)$ denote the numerical value of the function $\hat{f}_a A$ at x :

$$f_a(x) = \hat{f}_a A(x). \quad (8.35)$$

Then

$$\begin{aligned} \Delta_f^{-1}[A] &= \prod_x \prod_a \int_{-\infty}^{\infty} d\omega_a(x) \delta(f_a(x)) \\ &= \prod_x \prod_a \int_{-\infty}^{\infty} df_a(x) \delta(f_a(x)) \frac{\partial(\omega_1(x), \dots, \omega_N(x))}{\partial(f_1(x), \dots, f_N(x))} \\ &= \prod_x \det \left| \left| \frac{\partial \omega_a(x)}{\partial f_b(x)} \right| \right|_{f=0} = \det \left(\frac{\delta \omega}{\delta f} \right)_{f=0}. \end{aligned} \quad (8.36)$$

The last step defines the functional determinant of the continuous matrix $\delta \omega_a(x)/\delta f_b(y)$, with rows labelled by (a, x) and columns by (b, y) .

If A satisfies the gauge condition, then the condition $f = 0$ in (8.36) may be replaced by $\omega = 0$, and we can write

$$\Delta_f[A] = \det \left(\frac{\delta f}{\delta \omega} \right)_{\omega=0} \quad (\text{for } A \text{ satisfying } \hat{f}_a A = 0). \quad (8.37)$$

This in fact determines $\Delta_f[A']$ for arbitrary A' because, by assumption, A' can be gauge transformed to A , and by (8.33), $\Delta_f[A'] = \Delta_f[A]$. To calculate the

functional determinant in (8.37), it is only necessary to make infinitesimal gauge transformations from A , which is a great convenience.

The identity (8.31) can be generalized to

$$\Delta_f[A] \int (D\omega) \prod_a \delta(\hat{f}_a A^\omega - g_a) = 1, \quad (8.38)$$

where $\{g_a\}$ is a set of arbitrary functions. This merely redefines \hat{f}_a to make the gauge condition read $\hat{f}_a A = g_a$. Since g_a are independent of A , $\Delta_f[A]$ is unaffected. Now multiply both sides by an arbitrary functional $G[g] = G[g_1, \dots, g_N]$, and integrate over all g_a . The result is a more general identity:

$$\Delta_f[A] \frac{\int (D\omega) G[\hat{f}A]}{\int (Dg) G[g]} = 1. \quad (8.39)$$

Taking advantage of this, we can derive a more general representation for the transition amplitude than (8.28). The idea is to insert (8.39) into the unmodified Feynman path integral, and then try to factor out the infinite volume factor corresponding to integration over the unphysical fields.

We start by writing

$$\langle A_2, t_2 | A_1, t_1 \rangle = \frac{\mathcal{N}}{\int (Dg) G[g]} \int_{A_1}^{A_2} (DA) e^{iS_2[A]} \Delta_f[A] \int (D\omega) G[\hat{f}A^\omega]. \quad (8.40)$$

Now interchange the order of $\int (DA)$ and $\int (D\omega)$. Since (DA) , $S[A]$ and $\Delta_f[A]$ are all gauge invariant, they can be replaced respectively by (DA^ω) , $S[A^\omega]$ and $\Delta_f[A^\omega]$. The new integration variable A^ω can be renamed A . Thus we obtain

$$\langle A_2, t_2 | A_1, t_1 \rangle = \frac{\mathcal{N} \int (D\omega)}{\int (Dg) G[g]} \int_{A_1}^{A_2} (DA) e^{iS_2[A]} G[\hat{f}A] \Delta_f[A]. \quad (8.41)$$

Note that we have factored out $\int (D\omega)$, which is just the volume coming from integrating over fields that can be gauged away. Absorbing this constant together with $\int (Dg) G[g]$ into \mathcal{N} , we finally have

$$\langle A_2, t_2 | A_1, t_1 \rangle = \mathcal{N} \int_{A_1}^{A_2} (DA) e^{iS_2[A]} G[\hat{f}A] \Delta_f[A]. \quad (8.42)$$

The generating functional $W[J]$ for Green's functions in the gauge $\hat{f}_a A = 0$ is given by

$$e^{iW[J]} = \mathcal{N} \int (DA) e^{iS[A] + i(J, A)} G[\hat{f}A] \Delta_f[A], \quad (8.43)$$

where

$$S[A] = \int d^4x \mathcal{L}(x),$$

$$(J, A) = \int d^4x J_{a\mu}(x) A_a{}^\mu(x). \quad (8.44)$$

The integration $\int d^4x$ in $S[A]$ and (J, A) is defined by passage to Euclidean space, then continued back to Minkowski space. Since we know that this merely supplies the correct *i.e.* prescription in the propagators, we shall not always indicate it explicitly, but it should be understood as being done.

8.4 Free Maxwell Field

The action is

$$S[A] = -\frac{1}{2} \int \frac{d^4k}{(2\pi)^4} \tilde{A}_\mu (g^{\mu\nu} k^2 - k^\mu k^\nu) \tilde{A}_\nu. \quad (8.45)$$

It is convenient to decompose \tilde{A}^μ into transverse and longitudinal parts (in the 4-dimensional sense):

$$\tilde{A}^\mu = \tilde{A}_T{}^\mu + \tilde{A}_L{}^\mu, \quad (k \cdot \tilde{A}_T = 0). \quad (8.46)$$

The transverse and longitudinal projection operators are

$$\begin{aligned} P_T{}^{\mu\nu}(k) &= g^{\mu\nu} - \frac{k^\mu k^\nu}{k^2}, \\ P_L{}^{\mu\nu}(k) &= \frac{k^\mu k^\nu}{k^2}. \end{aligned} \quad (8.47)$$

As matrices in μ, ν with metric $g^{\mu\nu}$, they have the properties

$$\begin{aligned} P_T{}^2 &= P_T, & P_L{}^2 &= P_L, \\ P_T + P_L &= 1, & P_T \cdot P_L &= 0, \\ \tilde{A}_T &= P_T \tilde{A}, & \tilde{A}_L &= P_L \tilde{A}. \end{aligned} \quad (8.48)$$

We can then write, in matrix notation,

$$S[A] = -\frac{1}{2}(\tilde{A}, \tilde{K}\tilde{A}), \quad (8.49)$$

where

$$\tilde{K}^{\mu\nu} = k^2 g^{\mu\nu} - k^\mu k^\nu = k^2 P_T{}^{\mu\nu}(k). \quad (8.50)$$

1 Lorentz Gauge

Choose \hat{f} such that

$$\begin{aligned} \hat{f}A(x) &\equiv \partial_\mu A^\mu(x), \\ \hat{f}\tilde{A}(k) &\equiv ik \cdot \tilde{A}(k), \end{aligned} \quad (8.51)$$

where \tilde{A}^μ is the Fourier transform of A^μ . The weight assigned to $\hat{f}A$ is left open pending the choice of the functional G . First we calculate $\Delta_f[A]$. Let \tilde{A} be such that $k \cdot \tilde{A} = 0$. Under an infinitesimal gauge transformation, \tilde{A} and $\tilde{f} = \hat{f}A$ respectively change by

$$\begin{aligned}\delta\tilde{A}^\mu &= ik^\mu\delta\tilde{\omega}, \\ \delta\tilde{f} &= -k^2\delta\tilde{\omega}.\end{aligned}\quad (8.52)$$

Hence

$$\begin{aligned}\left(\frac{\delta\tilde{f}}{\delta\tilde{\omega}}\right)_{\omega=0} &= -k^2, \\ \Delta_f[A] &= \det\left(\frac{\delta\tilde{f}}{\delta\tilde{\omega}}\right)_{\omega=0} = -\prod_k k^2.\end{aligned}\quad (8.53)$$

We see that $\Delta_f[A]$ is divergent, but independent of A . Hence it can be absorbed into \mathcal{N} . Therefore

$$e^{iW[J]} = \mathcal{N} \int (DA) e^{iS[A]+i(J, A)} G[\hat{f}A]. \quad (8.54)$$

We still have the freedom to choose G .

First consider the choice $G[f] = \delta[f]$. This corresponds to the transverse Lorentz gauge, or Landau gauge. We have

$$\delta[\hat{f}A] = \prod_k \delta(k \cdot \tilde{A}) = \prod_k \delta(k \cdot \tilde{A}_L). \quad (8.55)$$

This requires $k \cdot \tilde{A}_L = 0$, which can be satisfied only by $\tilde{A}_L \equiv 0$. Hence

$$e^{iW[J]} = \mathcal{N} \int (DA_T) e^{iS[A_T]+i(J, A_T)}. \quad (8.56)$$

In fact, $S[A_T] = S[A]$. However, the restriction to the space of \tilde{A}_T renders \tilde{K} non-singular, for the only eigenfunctions of \tilde{K} with zero eigenvalue are longitudinal fields. It can be readily verified that

$$\tilde{K}^{-1} = \frac{1}{k^2} P_T(k), \quad (8.57)$$

in the sense

$$\tilde{K}\tilde{K}^{-1} = P_T(k). \quad (8.58)$$

Thus we can perform the path integral in (8.56), obtaining

$$\begin{aligned}e^{iW[J]} &= \mathcal{N} \int (DA_T) \exp\left(-\frac{i}{2} (A_T, \tilde{K}A_T) + i(J, \tilde{A}_T)\right) \\ &= \mathcal{N} (\det \tilde{K})^{-1/2} \exp\left(\frac{i}{2} (J_T, \tilde{K}^{-1}J_T)\right).\end{aligned}\quad (8.59)$$

Since \tilde{K}^{-1} is proportional to P_T , \tilde{J}_T may be replaced by J . Choosing the constant in front to be unity, we obtain the well-known result

$$\begin{aligned} W[J] &= \frac{1}{2} (\tilde{J}, \tilde{K}^{-1} \tilde{J}) = \frac{1}{2} \int d^4x J_\mu(x) D^{\mu\nu}(x-y) J_\nu(y), \\ D^{\mu\nu}(x) &= \int \frac{d^4k}{(2\pi)^4} \frac{e^{-ik \cdot x}}{k^2 + i\epsilon} \left(g^{\mu\nu} - \frac{k^\mu k^\nu}{k^2 + i\epsilon} \right). \end{aligned} \quad (8.60)$$

The $i\epsilon$ comes from the definition of $(\tilde{J}, \tilde{K}^{-1} \tilde{J})$ as an Euclidean integral.

Next, consider the choice

$$G[\hat{f}A] = \exp \frac{i}{2\lambda} \int d^4x (\hat{f}A)^2 = \exp \frac{i}{2\lambda} \int d^4x (\partial_\mu A^\mu)^2, \quad (8.61)$$

where λ is a real parameter. Common choices for λ are

$$\begin{aligned} \lambda &= 1 && \text{(Feynman gauge),} \\ \lambda &= 0 && \text{(Landau gauge).} \end{aligned} \quad (8.62)$$

The latter reduces to the previous case with $G[f] = \delta[f]$. We now have

$$e^{iW[J]} = \mathcal{N} \int (DA) \exp i[-\frac{1}{2}(A, KA) + (J, A)], \quad (8.63)$$

where

$$K^{\mu\nu} = - \left[g^{\mu\nu} \square^2 - \left(1 - \frac{1}{\lambda} \right) \partial^\mu \partial^\nu \right], \quad (8.64)$$

with Fourier transform

$$\begin{aligned} \tilde{K}^{\mu\nu} &= k^2 \left[g^{\mu\nu} k^2 - \left(1 - \frac{1}{\lambda} \right) \frac{k^\mu k^\nu}{k^2} \right] \\ &= k^2 \left[P_T^{\mu\nu}(k) + \frac{1}{\lambda} P_L^{\mu\nu}(k) \right]. \end{aligned} \quad (8.65)$$

It is easily verified that

$$\begin{aligned} (\tilde{K}^{-1})^{\mu\nu} &= \frac{1}{k^2} [P_T^{\mu\nu}(k) + \lambda P_L^{\mu\nu}(k)] \\ &= \frac{1}{k^2} \left[g^{\mu\nu} - (1 - \lambda) \frac{k^\mu k^\nu}{k^2} \right]. \end{aligned} \quad (8.66)$$

Thus

$$\begin{aligned} W[J] &= \frac{1}{2} (\tilde{J}, \tilde{K}^{-1} \tilde{J}) = \frac{1}{2} \int d^4x J_\mu(x) D^{\mu\nu}(x-y) J_\nu(y), \\ D^{\mu\nu}(x) &= \int \frac{d^4k}{(2\pi)^4} \frac{e^{-ik \cdot x}}{k^2 + i\epsilon} \left[g^{\mu\nu} - (1 - \lambda) \frac{k^\mu k^\nu}{k^2 + i\epsilon} \right]. \end{aligned} \quad (8.67)$$

As is well known, the term proportional to $k^\mu k^\nu$ does not contribute in Feynman graphs when the Maxwell field is coupled to a conserved current. Thus Lorentz gauges with different λ 's are physically equivalent.

The fact that $D^{\mu\nu}$ transforms under a Lorentz transformation as a tensor does not in itself make the theory Lorentz invariant, for $D^{\mu\nu}$ is not a physical quantity. To show Lorentz invariance, one must show that the S matrix is invariant. In the free field case, the matrix is $\epsilon_\mu D^{\mu\nu} \epsilon_\nu$, where ϵ^μ is the polarization vector of a free photon. Lorentz invariance depends on the fact that ϵ^μ is a 4-vector that can be brought to the form $\epsilon^\mu = (0, \epsilon_\perp)$ in some Lorentz frame. Thus, the manifest invariance of (8.67) is merely a formal nicety; the components J_{\parallel} and J_0 actually play no role in extracting physical results.

2 Coulomb Gauge

We choose \hat{f} and G as follows:

$$\begin{aligned}\hat{f}A(x) &\equiv \nabla \cdot \mathbf{A}(x), \\ G[\hat{f}A] &\equiv \exp \frac{i}{2\lambda} \int d^4x [\nabla \cdot \mathbf{A}(x)]^2.\end{aligned}\quad (8.68)$$

Then

$$\Delta_f[A] = -i \prod_k |\mathbf{k}|^2, \quad (8.69)$$

which is independent of A , and hence may be ignored. Thus

$$e^{iW[J]} = \mathcal{N} \int (DA) \exp i \left\{ S[A] + (J, A) + \frac{1}{2\lambda} \int d^4x (\nabla \cdot \mathbf{A})^2 \right\}. \quad (8.70)$$

To present this in mock-invariant form, we write

$$\begin{aligned}\int d^4x (\nabla \cdot \mathbf{A})^2 &= - \int d^4x A_i \partial^i \partial^j A_j = \int \frac{d^4k}{(2\pi)^4} \tilde{A}_i k^i k^j \tilde{A}_j \\ &= \int \frac{d^4k}{(2\pi)^4} \tilde{A}_\mu s^\mu s^\nu \tilde{A}_\nu,\end{aligned}\quad (8.71)$$

where

$$s^\mu \equiv (0, \mathbf{k}). \quad (8.72)$$

Then

$$\begin{aligned}e^{iW[J]} &= \mathcal{N} \int (DA) \exp i \left[-\frac{1}{2} (\tilde{A}, \tilde{K}_c \tilde{A}) + (J, \tilde{A}) \right] \\ &= \exp \frac{i}{2} (J, \tilde{K}_c^{-1} J),\end{aligned}\quad (8.73)$$

where

$$\tilde{K}_c^{\mu\nu} = k^2 g^{\mu\nu} - k^\mu k^\nu + \frac{1}{\lambda} s^\mu s^\nu. \quad (8.74)$$

To find the inverse of \tilde{K}_c , we write

$$(\tilde{K}_c^{-1})^{\mu\nu} = C_1 g^{\mu\nu} + C_2 k^\mu k^\nu + C_3 s^\mu s^\nu + C_4 k^\mu s^\nu + C_5 s^\mu k^\nu, \quad (8.75)$$

and determine the coefficients C_i by requiring $(\tilde{K}\tilde{K}^{-1})^{\mu\nu} = g^{\mu\nu}$. The result is

$$\begin{aligned} \tilde{D}_c^{\mu\nu} &= (\tilde{K}_c^{-1})^{\mu\nu} = \frac{1}{k^2 + i\epsilon} \left[g^{\mu\nu} + \left(1 + \frac{\lambda k^2}{s^2} \right) \frac{k^\mu k^\nu}{s^2} - \frac{1}{s^2} (k^\mu s^\nu + s^\mu k^\nu) \right] \\ &= \frac{1}{k^2 + i\epsilon} \left[g^{\mu\nu} - \frac{1}{|\mathbf{k}|^2} (k^\mu k^\nu - k^\mu s^\nu - s^\mu k^\nu) \right] + \frac{\lambda k^\mu k^\nu}{|\mathbf{k}|^4}, \end{aligned} \quad (8.76)$$

which is the photon propagator in Coulomb gauge. More explicitly,

$$\begin{aligned} \tilde{D}_c^{ij} &= -\frac{1}{k^2 + i\epsilon} \left(\delta_{ij} - \frac{k_i k_j}{|\mathbf{k}|^2} \right) + \frac{\lambda k_i k_j}{|\mathbf{k}|^4}, \\ \tilde{D}_c^{i0} &= \tilde{D}_c^{0i} = \frac{\lambda k^0 k^i}{|\mathbf{k}|^4}, \\ \tilde{D}_c^{00} &= -\frac{1}{|\mathbf{k}|^2} + \frac{\lambda k_0^2}{|\mathbf{k}|^4}. \end{aligned} \quad (8.77)$$

The final result is

$$W[J] = \frac{1}{2} (\tilde{J}, \tilde{D}_c \tilde{J}). \quad (8.78)$$

For the case $\lambda = 0$, which is equivalent to $G[f] = \delta[f]$,

$$W[J] = \frac{1}{2} \int \frac{d^4 k}{(2\pi)^4} \tilde{J}^i(k) \frac{1}{k^2 + i\epsilon} \left(-\delta_{ij} + \frac{k_i k_j}{|\mathbf{k}|^2} \right) \tilde{J}^j(k) - \int \frac{d^4 k}{(2\pi)^4} \frac{\tilde{J}_0^2(k)}{|\mathbf{k}|^2}, \quad (8.79)$$

or

$$W[J] = \frac{1}{2} \int d^4x J^i(x) D_c^{ij}(x - y) J^j(y) - \int dx_0 d^3x d^3y \frac{J_0(x) J_0(y)}{4\pi |\mathbf{x} - \mathbf{y}|}, \quad (8.80)$$

where D_c is the Fourier transform of \tilde{D}_c . The first term contains only \mathbf{J}_\perp , because $\tilde{D}_c(k)$ is proportional to $P_\perp(\mathbf{k})$. The last term is the analogue of the last term in (8.14). We see that the system is described in the most succinct physical terms: when one turns on an external current J^μ , only \mathbf{J}_\perp disturbs the system. \mathbf{J}_\parallel does not couple to the system at all, and J_0 merely shifts the reference point of the total energy.

3 Temporal and Axial Gauges

The temporal gauge $A^0 = 0$, and the axial gauge $A^3 = 0$, can be treated as one gauge, as $S[A]$ is really defined as an integral over Euclidean space, and the integrand is rotationally invariant. The distinction between the two gauges comes when we transform back to Minkowski space, in which physical photons are defined.

Denoting Euclidean quantities by the convention

$$\begin{aligned} p_\mu &= (\mathbf{k}, -ik_0), \\ A_\mu &= (\mathbf{A}, iA_0), \end{aligned} \quad (8.81)$$

we have

$$\begin{aligned} S[A] &= -\frac{1}{2} \int \frac{d^4k}{(2\pi)^4} [k^2 A^\mu A_\mu - (k^\mu A_\mu)^2] \\ &= -\frac{i}{2} \int \frac{d^4p}{(2\pi)^4} A_\mu (p^2 \delta_{\mu\nu} - p_\mu p_\nu) A_\nu, \end{aligned} \quad (8.82)$$

where the tildes denoting Fourier transforms have been omitted for simplicity. We set a component of A_μ to zero, say,

$$A_4 = 0. \quad (8.83)$$

The rest of the components are denoted by A_i ($i = 1, 2, 3$). Then

$$e^{iW[J]} = \mathcal{N} \int (DA) \exp[\frac{1}{2}(A, KA) + (J, A)], \quad (8.84)$$

where

$$\begin{aligned} (A, B) &\equiv \int \frac{d^4p}{(2\pi)^4} A_i(p) B_i(p), \\ K_{ij} &\equiv p^2 \delta_{ij} - p_i p_j. \end{aligned} \quad (8.85)$$

Calculating the path integral leads to

$$\begin{aligned} W[J] &= -\frac{i}{2} (J, K^{-1}J), \\ K_{ij}^{-1} &= \frac{1}{p^2} \left(\delta_{ij} + \frac{p_i p_j}{p_4^2} \right). \end{aligned} \quad (8.86)$$

Note that $p^2 = |\mathbf{p}|^2 + p_4^2$. The condition $A_4 = 0$ does not fix the gauge completely. This is reflected in the fact that K^{-1} diverges when $p_4 = 0$. However, this has no physical consequences, as we shall see.

Now we have to transform (8.86) back to Minkowski space, so that we can calculate the S -Matrix by replacing normal components of J (with respect to \mathbf{k}) by photon wave functions.

If we choose to regard p_4 as the imaginary energy, we would let

$$p_4 \rightarrow -ik_0 - i\epsilon, \quad (\epsilon \rightarrow 0^+),$$

and obtain $W[J]$ in the temporal gauge $A^0 = 0$:

$$W[J] = \frac{1}{2} \int \frac{d^4k}{(2\pi)^4} \frac{1}{k^2 + i\epsilon} J^i \left(\delta_{ij} - \frac{k_i k_j}{k_0^2 + i\epsilon} \right) J^j \quad (8.87)$$

Only the components of $\mathbf{J}_\perp(\mathbf{k})$, defined by $\mathbf{J}_\perp(\mathbf{k}) \cdot \mathbf{k} = 0$, are physically relevant.

If we choose to regard a component other than p_4 as the imaginary energy, say p_3 , we would take

$$\begin{aligned} p_3 &\rightarrow -ik_0 - i\epsilon, \quad (\epsilon \rightarrow 0^+), \\ p_1 &= k_1, \quad p_2 = k_2, \quad p_4 = k_3, \end{aligned}$$

and obtain $W[J]$ in the axial gauge $A^3 = 0$:

$$W[J] = \frac{1}{2} \int \frac{d^4k}{(2\pi)^4} \frac{1}{k^2 + i\epsilon} \left[\mathbf{J}_\perp^2 + \mathbf{J}_\parallel^2 - J_0^2 + \frac{(\mathbf{J}_\parallel \cdot \mathbf{k} - J_0 k_0)^2}{k_3^2} \right]. \quad (8.88)$$

Again, only the dependence on \mathbf{J}_\perp is physically relevant.

8.5 Pure Yang-Mills Fields

We describe in this section the Fadeev-Popov procedure in the non-Abelian case. In view of the problems in quantizing Yang-Mills fields discussed earlier, we should keep in mind that the procedure yields formal expression that may not exist.

It is plausible that the results are meaningful order by order in perturbation theory, because one is expanding about the Maxwell theory. Furthermore, problems such as the Gribov ambiguity are associated with the topological charge, which is always zero in perturbation theory.

On the other hand, non-perturbative phenomena are of physical interest too. These include the “ θ -worlds” to be discussed in a later section, and the physically important phenomena of quark confinement, to be covered in a separate chapter. In these cases, the path integral method gives us a conceptual framework to formulate these problems, but actual calculations, when possible, must be done with special care.

The complexity of Yang-Mills fields originates in the non-linear expression for the field tensor:

$$F_a^{\mu\nu} = \partial^\mu A_a^\nu - \partial^\nu A_a^\mu - g C_{abc} A_b^\mu A_c^\nu. \quad (8.89)$$

This makes the Lagrangian density more complicated:

$$\mathcal{L}_0(x) = -\frac{1}{4} F_a^{\mu\nu} F_{a\mu\nu} = \mathcal{L}_0(x) + \mathcal{L}_1(x), \quad (8.90)$$

where

$$\begin{aligned} \mathcal{L}_0(x) &= -\frac{1}{2} (\partial^\mu A_a^\nu - \partial^\nu A_a^\mu) \partial_\mu A_{a\nu}, \\ \mathcal{L}_1(x) &= -g C_{abc} A_a^\nu A_b^\mu \partial_\mu A_{c\nu} \\ &\quad - \frac{1}{4} g^2 C_{abc} C_{ab'c'} A_b^\mu A_{b'\mu} A_c^\nu A_{c'\nu}. \end{aligned} \quad (8.91)$$

The term $\mathcal{L}_0(x)$ contributes to the action a term of the same form as that of the Maxwell theory:

$$S_0[A] = \int d^4x \mathcal{L}_0(x) = \frac{1}{2} \int d^4x A_{a\mu} (g^{\mu\nu} \square^2 - \partial^\mu \partial^\nu) A_{a\nu}. \quad (8.92)$$

Unfortunately, $\mathcal{L}_1(x)$ makes the Feynman path integral non-Gaussian, and not calculable in closed form.

One way to proceed is to make use of (7.52) and write

$$\begin{aligned} e^{iW[J]} &= \mathcal{N} \int (DA) e^{iS[A]+i(J,A)} G[\hat{f}A] \Delta_f[A] \\ &= \left[\exp i \int d^4x \mathcal{L}_1 \left(\frac{1}{i} \frac{\delta}{\delta J(x)} \right) \right] e^{iW_0[J]}, \end{aligned} \quad (8.93)$$

where

$$e^{iW_0[J]} = \mathcal{N} \int (DA) e^{iS_0[A]+i(J,A)} G[\hat{f}A] \Delta_f[A]. \quad (8.94)$$

Gauge fixing is done in $W_0[J]$, from which $W[J]$ may be obtained by perturbation expansion in powers of g . This procedure is far from satisfactory for two reasons. First, we lose sight of the basic symmetry of the problem, because neither \mathcal{L}_0 nor \mathcal{L}_1 are separately gauge invariant. Secondly, non-perturbative effects, such as instantons and possibly quark confinement, would be difficult to discuss in this approach. Nevertheless, (8.93) is at least useful for perturbation theory, when one has reason to believe that to be valid.

In some problems for which standard perturbation theory is definitely unworkable, a better approach might be to expand $S[A]$ about the classical path, thus basing a new perturbation theory on the semi-classical approximation.³ But we shall not go into that.

Even though $S_0[A]$ has the same form as in Maxwell theory, the non-linear nature of the gauge transformation can make $\Delta_f[A]$, and therefore $W_0[J]$, quite different. In general, this gives rise to fictitious "ghost fields", as we shall see. In the rest of this section, we illustrate the calculation of $W_0[J]$, but forego the derivation of Feynman rules.

1 Axial Gauge

Choose $\hat{f}_a A = A_a$ ³. Under an infinitesimal gauge transformation ω about fields satisfying $\hat{f}_a A = 0$, we have

³ C. G. Callan, R. Dashen, and D. J. Gross, *Phys. Rev.* **D17**, 2717 (1978).

$$\hat{f}_a A^\omega = \frac{1}{g} \partial^3 \omega_a. \quad (8.95)$$

Hence, $\delta f / \delta \omega = g^{-1} \partial^3$, and $\Delta_f[A]$ are independent of A . Consequently $\Delta_f[A]$ can be absorbed into the normalization constant.

Note that this does not fix the gauge completely. The residual gauge freedom did not cause any problem in Maxwell theory, nor in perturbation theory here. In a non-perturbative context, it requires the inclusion of projection operators, which will be discussed later.

2 Lorentz Gauge: Faddeev-Popov Ghosts

Choose $\hat{f}_a A = \partial_\mu A_a^\mu$. For A such that $\partial_\mu A_a^\mu = 0$, an infinitesimal gauge transformation on A_a^μ gives

$$\begin{aligned} \hat{f}_a A_a^\omega &= \partial_\mu \left(A_a^\mu + \frac{1}{g} \partial^\mu \omega_a + C_{abc} \omega_b A_c^\mu \right) \\ &= \frac{1}{g} \square^2 \omega_a + C_{abc} (\partial_\mu \omega_b) A_c^\mu. \end{aligned} \quad (8.96)$$

Hence

$$\begin{aligned} \frac{\delta f_a}{\delta \omega_b} &= \frac{1}{g} \delta_{ab} \square^2 + C_{abc} A_c^\mu \partial_\mu, \\ \Delta_f[A] &= \det \left(\frac{1}{g} \delta_{ab} \square^2 + C_{abc} A_c^\mu \partial_\mu \right). \end{aligned} \quad (8.97)$$

Since $\Delta_f[A]$ depends on A , it cannot be absorbed. A convenient way of proceeding is to rewrite the determinant in terms of a path integral over anti-commuting c-number fields, as we did in Sec. 7.9:

$$\begin{aligned} \Delta_f[A] &= \int (D\eta^*) (D\eta) \exp \left[i \int d^4x \mathcal{L}_{\text{ghost}}(x) \right], \\ \mathcal{L}_{\text{ghost}}(x) &= \eta_a^*(x) [\delta_{ab} \square^2 + g C_{abc} A_c^\mu(x) \partial_\mu] \eta_b(x), \end{aligned} \quad (8.98)$$

where η^* and η are independent scalar fields obeying Fermi statistics, with anti-commutation and integration rules given in (7.141). They are called Faddeev-Popov ghosts.

Choosing

$$G[f] = \exp \left[\frac{i}{2\lambda} \int d^4x f^2 \right] = \exp \left[\frac{i}{2\lambda} \int d^4x (\partial_\mu A_a^\mu)^2 \right], \quad (8.99)$$

we obtain

$$e^{iW[J]} = \int (D\eta^*) (D\eta) (DA) \exp \left[i \int d^4x (\mathcal{L}_{\text{eff}} + J_{a\mu} A_a^\mu) \right], \quad (8.100)$$

where

$$\begin{aligned}\mathcal{L}_{\text{eff}} &= \frac{1}{2} A_{a\mu} \left[g^{\mu\nu} \square^2 - \left(1 - \frac{1}{\lambda}\right) \partial^\mu \partial^\nu \right] A_{a\nu} + \frac{1}{2} \eta_a^* \square^2 \eta_b + \mathcal{L}', \\ \mathcal{L}' &= \mathcal{L}_1 + \frac{1}{2} g C_{abc} (\eta_a^* \partial_\mu \eta_b) A_a{}^\mu,\end{aligned}\quad (8.101)$$

with \mathcal{L}_1 given in (8.91). Thus, the system is described as gauge fields with extra couplings to massless spinless ghost fields with derivative couplings. Since there is no physical need to introduce sources for the ghost fields, they only occur in closed loops in Feynman graphs.

As we mentioned before, the Coulomb gauge is plagued with the Gribov ambiguity, which also infects the Lorentz gauge. Fortunately it does not affect perturbation theory. In a general context, however, this pathology renders the Coulomb gauge and the Lorentz gauge meaningless.

8.6 The θ -World and the Instanton

1 Discovering the θ -World

To begin, let us set $A_a{}^0(x) = 0$. In the representation in which the Schrödinger operator $A_a(x)$ is diagonal, we denote a wave function of the system by $\Psi[A]$, and represent the conjugate momentum $-\mathbf{E}_a(x)$ by

$$E_a{}^k(x) = i \frac{\delta}{\delta A_a{}^k(x)}. \quad (8.102)$$

The Hamiltonian is given by

$$H = \frac{1}{2} \int d^3x (\mathbf{E}_a \cdot \mathbf{E}_a + \mathbf{B}_a \cdot \mathbf{B}_a), \quad (8.103)$$

where

$$\mathbf{B}_a = \nabla \times \mathbf{A}_a + \frac{1}{2} g C_{abc} \mathbf{A}_b \times \mathbf{A}_c. \quad (8.104)$$

Using $\partial \mathbf{A}_a / \partial t = i[H, \mathbf{A}_a]$, we find the operator equality

$$\frac{\partial \mathbf{A}_a}{\partial t} = -\mathbf{E}_a. \quad (8.105)$$

Classically, there is a residual gauge transformation $\mathbf{A} \rightarrow \mathbf{A}^\omega$, ($\mathbf{A} = \mathbf{A}_a L_a$), with

$$\mathbf{A}^\omega = U \mathbf{A} U^{-1} + \frac{i}{g} U \nabla U^{-1}, \quad (8.106)$$

$$U(\mathbf{x}) = \exp[-i\omega_a(\mathbf{x}) L_a],$$

where $\omega_a(\mathbf{x})$ is time-independent. For infinitesimal ω_a , the change in \mathbf{A}_a is given by

$$\delta\mathbf{A}_a = -\frac{1}{g} \nabla \omega_a + C_{abc} \omega_b \mathbf{A}_c. \quad (8.107)$$

Quantum-mechanically, the residual gauge invariance is expressed through the fact that

$$[\mathbf{D} \cdot \mathbf{E}_a, H] = 0, \quad (8.108)$$

$$\mathbf{D} \cdot \mathbf{E}_a \equiv \nabla \cdot \mathbf{E}_a + g C_{abc} \mathbf{A}_b \cdot \mathbf{E}_c.$$

Therefore, an eigenfunction $\Psi[\mathbf{A}]$ of H can be chosen to be a simultaneous eigenfunction of $\mathbf{D} \cdot \mathbf{E}_a$. Without further ado, we choose the eigenvalue of $\mathbf{D} \cdot \mathbf{E}_a$ to be zero, thus imposing on the Hilbert space of the system the subsidiary condition

$$\mathbf{D} \cdot \mathbf{E}_a(\mathbf{x}) \Psi[\mathbf{A}] = 0. \quad (8.109)$$

Consider continuous gauge transformations $U(\mathbf{x})$ that approach a constant at spatial infinity.^b

$$U(\mathbf{x}) \xrightarrow[|\mathbf{x}| \rightarrow \infty]{} \text{const.} \quad (8.110)$$

We can look upon $U(\mathbf{x})$ as a continuous mapping of ordinary 3-space into the gauge group G . The above condition identifies spatial infinity as one point, thus making the topology of space that of a 3-dimensional sphere S^3 . Hence, we are led to consider the continuous map $S^3 \rightarrow G$. A theorem due to Bott⁴ states:

Any continuous mapping of S^3 into a simple Lie group G can be continuously deformed into a mapping into an $SU(2)$ subgroup of G .

Therefore, for a Yang-Mills theory with simple gauge group G , it is only necessary to consider $S^3 \rightarrow SU(2)$. Since the manifold of $SU(2)$ has the topology of S^3 [See (4.22)], we consider continuous mappings $S^3 \rightarrow S^3$. As mentioned in Sec. 5.2, these mappings fall into homotopy classes characterized by winding numbers n , the number of times the spatial S^3 is covered by the group manifold S^3 . As a representative of the n th class, we choose

$$U_n(\mathbf{x}) = [v(\mathbf{x})]^n \quad (n = 0, \pm 1, \pm 2, \dots), \quad (8.111)$$

$$v(x) = \exp \frac{i \pi \mathbf{x} \cdot \boldsymbol{\pi}}{(r^2 + \rho^2)^{1/2}} \quad (r^2 \equiv |\mathbf{x}|^2), \quad (8.112)$$

^b We only consider $U \rightarrow \text{Const}$, because this is sufficient for our purpose. We can even rule out the possibility $U \not\rightarrow \text{Const}$. by placing the system in a large but finite box, and impose any definite boundary conditions on A . It is a common article of faith to assume that boundary conditions at large distances have no effect on local phenomena.

⁴ R. Bott, *Bull. Soc. Math. France* **84**, 251 (1956).

where ρ is an arbitrary number, and $\tau_a/2$ ($a = 1, 2, 3$), are the generators of a $SU(2)$ subgroup of G . A general member of class 0 is any gauge transformation that can be continuously deformed into the identity. A general member of class 1 is $v(x)$ times a member of class 0. A general member of class $n+m$ is a product of a member of class n with one of class m . Note that the map (8.112) is different from the instanton map (5.15).

Suppose \mathbf{A} undergoes an infinitesimal gauge transformation. Then $\Psi[\mathbf{A}]$ changes according to

$$\begin{aligned}\Psi[\mathbf{A} + \delta\mathbf{A}] - \Psi[\mathbf{A}] &= -\frac{i}{g} \int d^3x (-\nabla\omega_a + gC_{abc}\omega_b\mathbf{A}_c) \cdot \mathbf{E}_a \Psi[\mathbf{A}] \\ &= \frac{i}{g} \left[\int d\mathbf{S} \cdot \mathbf{E}_a \omega_a - \int d^3x \omega_a \mathbf{D} \cdot \mathbf{E}_a \right] \Psi[\mathbf{A}].\end{aligned}\quad (8.113)$$

For class 0 gauge transformations, the first term vanishes since ω_a approaches zero at spatial infinity. Thus, $\mathbf{D} \cdot \mathbf{E}_a$ is the generator of class 0 gauge transformations, under which $\Psi[\mathbf{A}]$ is invariant by (8.109). Under a gauge transformation of class $n \neq 0$, $\Psi[\mathbf{A}]$ is not necessarily invariant; but since the Hamiltonian is locally gauge invariant, all energy eigenfunctions can be chosen so that they change at most by a constant phase that is the same for all eigenfunctions:

$$\Psi[\mathbf{A}_n] = e^{in\theta} \Psi[\mathbf{A}] \quad (n = 0, \pm 1, \pm 2, \dots), \quad (8.114)$$

where \mathbf{A}_n is the transform of \mathbf{A} by a class n gauge transformation, a typical one being

$$\mathbf{A}_n = v^n \mathbf{A} v^{-n} + \frac{i}{g} v^n \nabla v^{-n}. \quad (8.115)$$

We shall denote the vacuum-state wave function by $\Psi_\theta[\mathbf{A}]$. The vacuum state characterized by θ is called the “ θ -vacuum”.

The Hilbert space of the system is divided into sectors labelled by the continuous parameter θ , each containing states built on the θ -vacuum. These “ θ -worlds” are isolated from one another by superselection rules, i.e., they are bridgeable only by non-gauge-invariant interactions. In each θ -world the vacuum state is unique.

If $\theta \neq 0$, then (8.114) requires the unique vacuum state to be complex, thus violating time-reversal invariance. By the CPT theorem, this implies CP violation.^c From (8.112), we see that under spatial reflection, $v(x) \rightarrow [v(x)]^{-1}$, and hence $n \rightarrow -n$. This leads to the conclusion that Ψ_θ is not an eigenstate of parity, if $\theta \neq 0$. The different θ -worlds are therefore physically inequivalent.

Let us call a gauge transformation of class 0 a “small” gauge transformation, and that of class ± 1 a “large” gauge transformation. Thus, Ψ_θ is invariant under a small gauge transformation, but changes by a phase factor $e^{\pm i\theta}$ under a large

^c CP violation in the θ -world does not directly translate into CP violation in the real world, because the situation changes dramatically when the gauge fields are coupled to fermions (See Sec. 12.6).

gauge transformation. To get a rough idea of what Ψ_θ might look like, let $\chi_n[A]$ be a functional of A that is “peaked” about A_n in the following sense: a small gauge transformation leaves χ_n invariant, while a large gauge transformation takes χ_n into χ_{n+1} . The χ_n ’s are like wave functions that peak about a particular potential minimum of a periodic potential in quantum mechanics. An intuitive representation of Ψ_θ might be

$$\Psi_\theta[A] \cong \sum_{n=-\infty}^{\infty} e^{in\theta} \chi_n[A], \quad (8.116)$$

for this realizes the property (8.114) in a concrete way. If we were to do a variational calculation, this might be a good guess for a trial wave function; but there is probably no χ_n that can make (8.116) an exact representation. This intuitive picture does suggest that quantum-mechanical tunnelling is responsible for the structure of Ψ_θ , and that the θ -vacua have different energies. This is indeed the case. The tunnelling mechanism is the instanton, as we shall see.

It might appear strange that θ is a free parameter, which cannot be determined from within our theory. After all, we can imagine obtaining the exact ground state wave function, which would then tell us precisely how it transforms under large gauge transformations. To do that without ambiguity, however, we must fix the gauge completely, and that proves to be impossible.¹

2 Instanton as Tunneling Solution

Let us enclose the system in a large box of volume V , and consider an evolution over a Euclidean time interval T , in which the initial and final configurations are pure gauges differing by a large gauge transformation. Let us denote a spatial 3-vector by \mathbf{r} , and Euclidean time by t , and choose the initial and final configurations to be of class 0 and class 1, respectively:

$$\begin{aligned} \mathbf{A}_{\text{initial}}(\mathbf{r}) &= 0, \\ \mathbf{A}_{\text{final}}(\mathbf{r}) &= \frac{i}{g} v(\mathbf{r}) \nabla v^{-1}(\mathbf{r}), \end{aligned} \quad (8.120)$$

where $v(\mathbf{r})$ is given by (8.112). We shall show that, for $T \rightarrow \infty$, a classical solution with these endpoints is the instanton. Let us recall the instanton solution (5.24) and (5.28), which can be written more explicitly in the form

$$\begin{aligned} \mathbf{A}(\mathbf{r}, t) &= \frac{1}{g} \frac{(\boldsymbol{\tau} \times \mathbf{r}) + \boldsymbol{\tau} t}{r^2 + t^2 + \rho^2}, \quad (r^2 = |\mathbf{r}|^2), \\ \mathbf{A}_4(\mathbf{r}, t) &= -\frac{1}{g} \frac{\boldsymbol{\tau} \cdot \mathbf{r}}{r^2 + t^2 + \rho^2}, \end{aligned} \quad (8.121)$$

where ρ is an arbitrary scale parameter. To conform to our gauge choice here, we must transform the above into temporal gauge, in which $A_4 = 0$. This can be achieved through the gauge transformation

$$\begin{aligned} \mathbf{U}(\mathbf{r}, t) &= \exp[i\tau_r F(r, t)], \quad (\tau_r = \boldsymbol{\tau} \cdot \mathbf{r}/r) \\ F(r, t) &= \frac{r}{\sqrt{r^2 + \rho^2}} \left[\frac{\pi}{2} + \tan^{-1} \frac{t}{\sqrt{r^2 + \rho^2}} \right]. \end{aligned} \quad (8.122)$$

Note that

$$F(r, t) \rightarrow \begin{cases} 0, & (t \rightarrow -\infty); \\ \pi r / \sqrt{r^2 + \rho^2}, & (t \rightarrow \infty). \end{cases} \quad (8.123)$$

Hence

$$U(\mathbf{r}, t) \rightarrow \begin{cases} 1 & (t \rightarrow -\infty); \\ v(\mathbf{r}), & (t \rightarrow \infty). \end{cases} \quad (8.124)$$

The instanton field in the temporal gauge is therefore

$$\mathbf{A}_{\text{tem}}(\mathbf{r}, t) = \frac{1}{g} U(\mathbf{r}, t) \left[\frac{(\boldsymbol{\tau} \times \mathbf{r}) + \boldsymbol{\tau} t}{r^2 + t^2 + \rho^2} \right] U^{-1}(\mathbf{r}, t) + \frac{i}{g} U(\mathbf{r}, t) \nabla U^{-1}(\mathbf{r}, t). \quad (8.125)$$

The first term vanishes for $|t| \rightarrow \infty$. Hence

$$\mathbf{A}_{\text{tem}}(\mathbf{r}, t) \rightarrow \begin{cases} 0, & (t \rightarrow -\infty); \\ \frac{i}{g} v(\mathbf{r}) \nabla v^{-1}(\mathbf{r}), & (t \rightarrow \infty). \end{cases} \quad (8.126)$$

This shows that the instanton interpolates, in Euclidean time, between two field configurations differing by a large gauge transformation. In this sense it is a “tunneling solution”.^{5,6}

The energy E_θ of the θ -vacuum is given by

$$\begin{aligned} e^{-E_\theta T} &= \langle \theta | e^{-HT} | \theta \rangle \\ &= \mathcal{N} \int (d\mathbf{A}_1)(d\mathbf{A}_2) \int_{\mathbf{A}_1}^{\mathbf{A}_2} (d\mathbf{A}) \Psi_\theta^*[\mathbf{A}_2] e^{-S_E[A]} \Psi_\theta[\mathbf{A}_1], \end{aligned} \quad (8.127)$$

where $\int (d\mathbf{A}_1)$ denotes integration over time-dependent functions $\mathbf{A}_1(\mathbf{x})$. E_θ can be calculated approximately by assuming that the path integral (8.127) is dominated by a dilute “instanton gas”.⁷

To study further the connection between topological charge and “large” gauge transformations, we refer to Fig. 8.1, which represents schematically the Euclidean space-time in which the tunneling process takes place. On the space-time boundary S the field is pure gauge:

$$A^\mu(x) = \frac{i}{g} U \partial^\mu U^{-1}, \quad (x \in S). \quad (8.128)$$

⁵ R. Jackiw and C. Rebbi, *Phys. Rev. Lett.* **37**, 172 (1976); C. G. Callan, R. F. Dashen and D. J. Gross, *Phys. Lett.* **63B**, 334 (1976).

⁶ K. M. Bitar and S.-J. Chang, *Phys. Rev.* **D17**, 486 (1978), give a description of tunneling in Minkowski space.

⁷ S. Coleman, in *The Whys of Subnuclear Physics*, Ed. A. Zichichi (Plenum, New York, 1980).

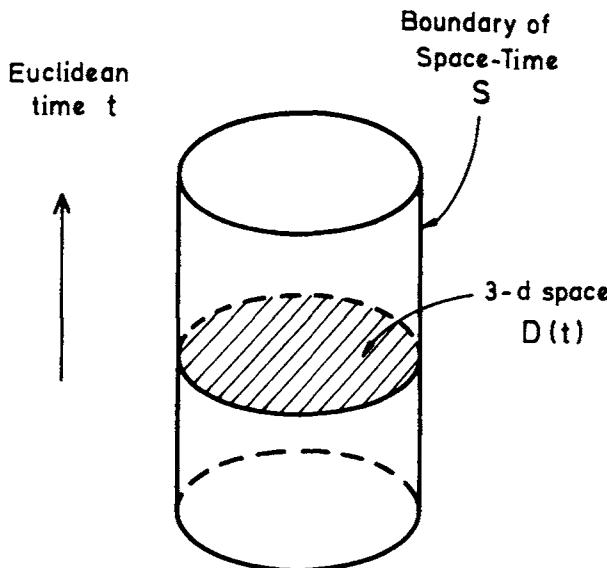


Fig. 8.1. Topological charge $q[A] = \text{winding number of } S \rightarrow SU(2)$. Index $n[A] = \text{winding number of } D(t) \rightarrow SU(2)$.

The topological charge $q[A]$ is just the winding number of $U(x)$, which is a map $S \rightarrow SU(2)$ [See (5.7), (5.14)]:

$$\begin{aligned} q[A] &= \frac{1}{4\pi^2} \int_S dS^\mu I_\mu(A), \\ I_\mu(A) &= (-ig)^3 \epsilon^{\mu\alpha\beta\gamma} \text{Tr}(A_\alpha A_\beta A_\gamma) \\ &= \epsilon^{\mu\alpha\beta\gamma} \text{Tr}[(U \partial^\alpha U^{-1})(U \partial^\beta U^{-1})(U \partial^\gamma U^{-1})]. \end{aligned} \quad (8.129)$$

At each Euclidean time slice t , the spatial volume is a “disk” $D(t)$, which also has the topology of S^3 , because spatial infinity is identified as one point by virtue of the condition (8.110). The winding number of the map $D(t) \rightarrow SU(2)$ is the index $n[A]$ that classifies “large” gauge transformations:

$$n[A] = \frac{1}{4\pi^2} \int_{x \in D(t)} d^3x I_4(A). \quad (8.130)$$

This expression is obtained by comparing I_4 from (8.129) with the $SU(2)$ group measure (5.12). In the temporal gauge $I_k(A) = 0$ ($k = 1, 2, 3$), because it contains a factor $A_4 \equiv 0$. Thus, the topological charge receives contributions only from the initial and final configurations:

$$\begin{aligned} q[A] &= \frac{1}{4\pi^2} \left[\int d^3r I_4(A) \right]_{t=-\infty}^{t=\infty} \\ &= n[\mathbf{A}_f] - n[\mathbf{A}_i]. \end{aligned} \quad (8.131)$$

3 The θ -Action

We can make a transformation in Hilbert space to take away the phase factor $e^{in\theta}$ in (8.114), thus making all wave functions completely gauge invariant, even under large gauge transformations. This is similar to what we did in Maxwell theory to make $\nabla \cdot E = 0$. The difference is that in the Maxwell case, the transformed Hamiltonian is the same as the original one, apart from being shifted by a constant. However, here the Hamiltonian will change in a substantial way. The easier way to make the transformation is to do it with path integrals. We shall then obtain a transformed action, which will be called the “ θ -action.”

Define new wave functions $\Phi[\mathbf{A}]$ by

$$\Phi[\mathbf{A}] \equiv e^{-i\theta N[\mathbf{A}]} \Psi[\mathbf{A}]. \quad (8.132)$$

It is clear that $\Phi[\mathbf{A}]$ is completely gauge invariant, even under large gauge transformations. The transformed θ -vacuum state is denoted by Φ_θ . The formula (8.127) giving E_θ can be rewritten in terms of $\Phi_\theta[\mathbf{A}]$ as

$$e^{-E_\theta T} = \mathcal{N} \int (d\mathbf{A}_2)(d\mathbf{A}_1) \int_{\mathbf{A}_1}^{\mathbf{A}_2} (D\mathbf{A}) \Phi_\theta^*[\mathbf{A}_2] e^{-S_E[\mathbf{A}, \theta]} \Phi_\theta[\mathbf{A}_1], \quad (8.133)$$

where the new action, called the “ θ -action,” is given by

$$S_E[\mathbf{A}, \theta] = S_E[\mathbf{A}] + i\theta q[\mathbf{A}]. \quad (8.134)$$

The proof is straightforward. In Minkowski space, the θ -action is given by^d

$$S[\mathbf{A}, \theta] = S[\mathbf{A}] - \theta q[\mathbf{A}]. \quad (8.135)$$

We note that the additional term does not change the equations of motion, because it is a total 4-divergence. In Maxwell theory, such a term is irrelevant because $q[\mathbf{A}] = 0$.

Let the “partition function” of the θ -world be denoted by Z_θ :

$$Z_\theta = \mathcal{N} \int (D\mathbf{A}) e^{-S_E[\mathbf{A}, \theta]} \quad (8.136)$$

The average topological charge of the θ -world is given by

$$\langle q \rangle = \frac{1}{i} \frac{\partial}{\partial \theta} \ln Z_\theta. \quad (8.137)$$

We may look upon θ as the Lagrange multiplier that fixes the θ -world average of the topological charge. An analogy may be made with the chemical potential in statistical mechanics, which is the Lagrange multiplier that fixes the ensemble average of the baryonic charge.

^d When transforming $q[\mathbf{A}]$ from Euclidean to Minkowski space, note that $\tilde{F}_{\mu\nu} F_{\mu\nu} \rightarrow \tilde{F}^{\mu\nu} F_{\mu\nu}$ (because $F_{\mu\nu} F_{\mu\nu} = -4\mathbf{B} \cdot \mathbf{E} = 4\mathbf{B} \cdot \partial\mathbf{A}/\partial x_4$), and the factor $-i$ cancels the i from $d^4x_E = id^4x$.

8.7 Gribov Ambiguity

Gribov⁸ discovered that, when one attempts to impose the Coulomb gauge in $SU(2)$ gauge theory, the field can become multivalued. As a consequence, the Hamiltonian in Coulomb gauge contains singular terms that lead to discontinuous time evolutions. This phenomenon has its origin in the existence of the topological charge. We shall demonstrate it by showing that the Euclidean time evolution of an instanton in Coulomb gauge is discontinuous. This phenomenon also occurs in Lorentz gauge, and renders the Coulomb and Lorentz gauges meaningless, except in perturbation theory. As mentioned earlier, these problems are illustrations of a general theorem¹ that, with definite boundary conditions, it is impossible to completely fix the gauge.

We continue to denote a spatial vector by \mathbf{r} , and Euclidean time by t . To demonstrate the Gribov phenomenon, it suffices to consider a pure-gauge field with “spherical symmetry”:

$$A^\mu = \frac{i}{g} U \partial^\mu U^{-1}, \quad (8.138)$$

$$U = \exp(i\phi\tau_r/2),$$

where $\tau_r = \boldsymbol{\tau} \cdot \hat{\mathbf{r}}$, $\hat{\mathbf{r}} = \mathbf{r}/r$. More explicitly we have

$$\begin{aligned} \mathbf{A} &= \frac{1}{2} \left[\hat{\mathbf{r}} \tau_r \frac{\partial \phi}{\partial r} + (\boldsymbol{\tau} \times \hat{\mathbf{r}}) \frac{1 - \cos \phi}{r} + (\boldsymbol{\tau} - \hat{\mathbf{r}} \tau_r) \frac{\sin \phi}{r} \right], \\ A_4 &= \frac{\tau_r}{2} \frac{\partial \phi}{\partial t}. \end{aligned} \quad (8.139)$$

It is straightforward to show that the transversality condition $\nabla \cdot \mathbf{A} = 0$ leads to the following differential equation for the gauge parameter ϕ :

$$\ddot{\phi} + \dot{\phi} - 2 \sin \phi = 0, \quad (8.140)$$

where a dot denotes partial derivative with respect to $s = \ln r$. If we think of s as time, then the above equation describes the motion of a damped pendulum with potential $V(\phi) = 2 \cos \phi$, which is depicted in Fig. 8.2. We can deduce the qualitative behavior of ϕ by inspection of the graph. At $r = 0$ ($s = -\infty$) we must have $\phi = 2\pi n$, in order to avoid a singularity in U . Let us take $\phi = \dot{\phi} = 0$ as initial condition (at $s = -\infty$). The pendulum starts at rest from point P in Fig. 8.2, at a potential maximum. As s increases, the pendulum either remains indefinitely in unstable equilibrium at point P , or slides down the potential, oscillates about, and eventually settles down at the point Q (where $\phi = \pi$). In the latter case the downward slide takes an infinitely long “time”.

The transversality condition does not fix the gauge completely, and we can further require

$$r \mathbf{A} \xrightarrow[r \rightarrow \infty]{} 0. \quad (8.141)$$

⁸ V. N. Gribov, *Nucl. Phys.* B139, 1 (1978).

$$V(\phi) = 2 \cos \phi$$

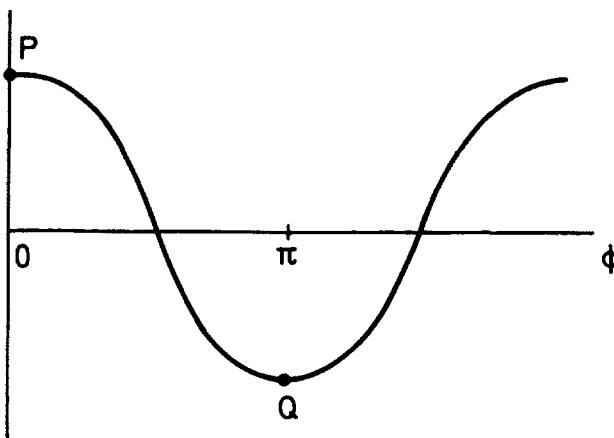


Fig. 8.2. The angle ϕ of a damped pendulum, initially at $\phi = 0$ with $\dot{\phi} = 0$, either remains at $\phi = 0$, or rolls down the potential to settle down at Q , after an infinite number of oscillations about Q .

With this condition, the only solution is $\phi = 0$. That is, the transverse pure gauge is uniquely $\mathbf{A} = 0$.

We have previously considered the instanton in temporal gauge, and calculated the topological charge in the geometry shown in Fig. 8.1. Let us transform that solution to Coulomb gauge with the boundary condition (8.141), by applying a gauge transformation. Recall that in temporal gauge the instanton was pure-gauge on the space-time boundary. Under a gauge transformation, the boundary values must remain pure-gauge, and therefore vanish in Coulomb gauge. Thus, the surface integral (8.129) gives the incorrect value zero for the topological charge. The trouble can be attributed to a failure of Gauss' theorem, which was used in (5.7) to convert a volume integral to a surface integral. Thus we are led to conclude that the time development in Coulomb gauge is discontinuous, so that Gauss' theorem gives the usual surface integral plus a contribution from the discontinuity.⁹

The real Gribov ambiguity is the following. Any Euclidean time evolution can be composed of a specific configuration, mixed in with an arbitrary time sequence of an equal number of instantons and anti-instantons. Thus, the overall configuration can have an arbitrary number of discontinuities in its time development, and it is not possible to specify and control the occurrence of such discontinuities. This pathology renders the Coulomb gauge meaningless in general. It is unambiguous only in perturbation theory, where instanton effects do not exist.

⁹ For graphs showing the actual discontinuity see R. Jackiw, I. Muzinich, and C. Rebbi, *Phys. Rev. D* 17, 1576 (1978).

8.8 Projection Operator for Gauss' Law

In the temporal gauge, the theory is quantized through the equal-time commutation relations

$$[E_a^j(\mathbf{x}), A_b^k(\mathbf{y})] = i\delta_{jk}\delta_{ab}\delta^3(\mathbf{x} - \mathbf{y}). \quad (8.142)$$

However, not all components of the fields are physically relevant, for the theory is still invariant under time-independent gauge transformations. This residual gauge freedom can be eliminated by imposing Gauss' law as a constraint on physical states:

$$G_a(\mathbf{x})\Psi[\mathbf{A}] = 0, \quad (8.143)$$

where

$$\begin{aligned} G_a(\mathbf{x}) &\equiv \mathbf{D} \cdot \mathbf{E}_a(\mathbf{x}) \\ &= \nabla \cdot \mathbf{E}_a(\mathbf{x}) + gC_{abc}\mathbf{A}_b(\mathbf{x}) \cdot \mathbf{E}_c(\mathbf{x}). \end{aligned} \quad (8.144)$$

It was shown in (8.113) that $G_a(\mathbf{x})$ is a generator of local gauge transformations. This can also be demonstrated by showing^c:

$$[G_a(\mathbf{x}), G_b(\mathbf{y})] = -iC_{abc}G_c(\mathbf{x})\delta^3(\mathbf{x} - \mathbf{y}). \quad (8.145)$$

This shows that $-G_a(\mathbf{x})$ are the generators of the local gauge group attached to the spatial point \mathbf{x} .

In Maxwell theory the constraint from Gauss' law was satisfied in a very simple way, namely, make $\Psi[\mathbf{A}]$ depend only on the transverse component of \mathbf{A} . Here an explicitly solution can be constructed, but it appears to be rather complicated and unwieldy.¹⁰ An alternative way to implement the constraint is to construct a projection operator onto the physical Hilbert space. We note that the constraint requires physical states to be annihilated by $G_a(\mathbf{x})$, for all a and \mathbf{x} . In other words, physical states are invariant under all local gauge transformations. The projection operator is therefore the integral over all group elements of the local gauge group¹¹:

$$P = \int [D\mu(\omega)] \exp \left[i \int d^3x \omega_a(\mathbf{x}) G_a(\mathbf{x}) \right], \quad \int [D\mu(\omega)] = \prod_{\mathbf{x}} \int d\mu(\omega(\mathbf{x})), \quad (8.146)$$

where $d\mu(\omega(\mathbf{x}))$ is the invariant group measure at the group element parameterized by $\omega_a(\mathbf{x})$, so normalized that $\int d\mu(\omega(\mathbf{x})) = 1$. The use of this measure

^c In showing this we use the Jacobi identity (4.5), and the relation

$$\frac{\partial}{\partial x} |\delta(x - y)f(y)| + \frac{\partial}{\partial y} |\delta(x - y)f(x)| = \delta(x - y) \frac{\partial f(x)}{\partial x}.$$

¹⁰ J. Goldstone and R. Jackiw, *Phys. Lett.* **74B**, 81 (1978).

¹¹ J. Polonyi, *Phys. Lett.* **213B**, 340 (1988); in *Quark Gluon Plasma*, Ed. R. C. Hwa (World Scientific, 1989).

ensures that the group manifold is properly covered in the integration.¹² The above formula is a special case of the well-known expression for the group projection operator onto a basis vector in a particular irreducible representation.¹³ In our case we project onto any basis vector of the trivial representation. By performing a partial integration and omitting surface terms, we can rewrite

$$P = \int [D\mu(\omega)] \exp \left[-i \int d^3x \mathbf{D}\omega_a(\mathbf{x}) \cdot \mathbf{E}_a(\mathbf{x}) \right],$$

$$\mathbf{D}\omega_a \equiv \nabla\omega_a + gC_{abc}\mathbf{A}_b\omega_c.$$

The omitted surface terms may give rise to a phase factor, which we can dispose of in the manner discussed in Sec. 8.6.

In the transition amplitude between basis states in the coordinate representation, the only practical way to enforce the constraint is to include the projection operator explicitly:

$$W_{fi} = \langle \mathbf{A}_f | e^{-itH} P | \mathbf{A}_i \rangle. \quad (8.148)$$

Note that $[H, P] = 0$, because $[H, \mathbf{D} \cdot \mathbf{E}_a] = 0$. We can therefore write

$$e^{-itH} P = e^{-it\tilde{H}},$$

$$W_{fi} = \int (DA) \exp \left[i \int_0^t dt' \int d^3x \mathcal{L}(\mathbf{x}, t') \right],$$

Comparing H with the Hamiltonian worked out in (4.89), we see that it is just the Hamiltonian with a static scalar potential

$$\mathbf{A}_a^0(\mathbf{x}) = t\omega_a^0(\mathbf{x}). \quad (8.149)$$

Thus, the effect of the projection operator is to restore the scalar potential, (which was set to zero when we adopted temporal gauge,) but we must integrate the transition amplitude over all static scalar potentials, using the group measure at the group element generated by tA^0 .

It is now straightforward to rewrite the transition amplitude as a Feynman path integral:

$$W_{fi} = \int (DA) \exp \left[i \int_0^t dt' \int d^3x \mathcal{L}(\mathbf{x}, t') \right], \quad (8.150)$$

$$\mathbf{A}(\mathbf{x}, 0) = \mathbf{A}_i(\mathbf{x}), \quad \mathbf{A}(\mathbf{x}, t) = \mathbf{A}_f(\mathbf{x}),$$

where the Lagrangian density $\mathcal{L}(\mathbf{x}, t')$ is that for a free Yang-Mills field, with static scalar potential $A_a^0(\mathbf{x})$. The integration extends over all space-time dependent vector potentials $\mathbf{A}(\mathbf{x}, t)$, and over all static scalar potentials $A^0(\mathbf{x})$, with the following integration measure:

$$(DA) \equiv \prod_{\mathbf{x}} [d\mu(tA^0(\mathbf{x}))] \prod_{\mathbf{x}, t'} [d\mathbf{A}(\mathbf{x}, t')]. \quad (8.151)$$

¹² D. J. Gross, R. D. Pisarski, and L. G. Yaffe, *Rev. Mod. Phys.* **53**, 43 (1981) introduced a similar projection operator, but neglected to include the invariant group measure.

¹³ W. K. Tung, *Group Theory in Physics* (World Scientific, Singapore, 1985), Sec. 4.2.

By going back to the derivation of the path integral, and inserting independent projection operators at different time slices, one can derive an alternative expression that involves a time-dependent scalar potential instead of a static one.¹⁴

The group measure introduces non-polynomial cutoff-dependent terms in the effective action, which makes the transition amplitude complicated, except in perturbation theory. Consider, for example, $SU(2)$ gauge theory. The group measure corresponding to the parameterization $\exp[i\omega_a \sigma_a/2]$ is¹⁵

$$d\mu(\omega_a) = \frac{1}{4\pi^2} \sin^2 \omega d\omega d\Omega = \frac{1}{4\pi^2} \left(\frac{\sin \omega}{\omega} \right)^2 d^3 \omega. \quad (8.152)$$

Thus, if we use the flat measure $[DA_a^0(x)]$ for the functional integral, then the action must be augmented by the term

$$S' = -\frac{i}{a^3} \int d^3x \ln \left[\frac{\sin(tA^0(x))}{tA^0(x)} \right]^2, \quad (8.153)$$

where a is a short-distance cutoff, necessary to provide the correct dimension for the spatial integral. In a power series expansion in A^0 , the A^0 -dependent terms become infinitely damped as $a \rightarrow 0$. Thus effectively $A^0 = 0$ in perturbation theory. In a path integral, the integration measure generally contain a cutoff implicitly; but when the measure is flat, the cutoff can be absorbed into an irrelevant normalization constant.

The explicit appearance of the cutoff in the effective action warns us that the theory may depend on how it is being regulated. A good way to regulate in a gauge-invariant manner is to put the theory on a lattice, which we shall discuss in the next chapter. We shall see that gauge-fixing is unnecessary on the lattice, but can be done very simply if desired. This indicates that the seemingly endless woes we face in the continuum stem from improper regularization. In the lattice theory, problems may resurface when we take the continuum limit; but those would be problems related to renormalization. Our conclusion is that the naive continuum formulation of the theory is useful only in perturbation theory.

¹⁴ J. Polonyi, *op. cit.*.

¹⁵ W. K. Tung, *ibid.*

CHAPTER 9

RENORMALIZATION

Renormalization was originally designed to deal with the divergences encountered in Feynman graphs. It furnishes a scheme for subtracting infinite contributions, and absorbing them into redefinitions of physical parameters such as coupling constants. From this point of view, renormalization seems to be no more than a recipe, devoid of physical significance. Its use, however, has yielded numerical results that agree with experiments to better than seven significant figures! One must wonder whether it has a physical basis, and indeed there is.

Through applications in critical phenomena, we have gained an understanding of the physics underlying renormalization. The central point is that the low-energy manifestations of a theory can be described by an effective Lagrangian, which has the same form as the original Lagrangian, except that it contains fewer degrees of freedom, and that the original coupling constants are replaced by renormalized ones that depend on the energy scale.

Although simple in conception, the renormalized coupling constants are difficult to calculate. In particular, the large number of degrees of freedoms involved in a field theory leads to divergences. Perturbative renormalization succeeded in taming them; but the preoccupation with the divergences tended to overshadow the underlying physics. In actuality, renormalization is necessary, even if there were no divergences. The techniques for subtracting and absorbing infinities, however, remain important, for that is what we use in practical calculations. For this reason, we shall introduce renormalization via the historical route.

9.1 Charge Renormalization

We begin with a calculation—charge renormalization to lowest order in quantum electrodynamics.

The full photon propagator to second order in the unrenormalized charge e_0 is represented by the Feynman graphs¹ in Fig. 9.1, which give the expression (in Feynman gauge)

$$iD'_{\mu\nu}(k) = \frac{g_{\mu\nu}}{ik^2} + \frac{g_{\mu\alpha}}{ik^2} i\Pi^{\alpha\beta}(k) \frac{g_{\beta\nu}}{ik^2}, \quad (9.1)$$

where $\Pi^{\mu\nu}(k)$ is the vacuum polarization tensor. Gauge invariance requires $k_\mu \Pi^{\mu\nu}(k) = 0$. Thus $\Pi^{\mu\nu}(k)$ must have the form

$$\Pi^{\mu\nu}(k) = e_0^{-2} (g^{\mu\nu} k^2 - k^\mu k^\nu) \Pi(k^2). \quad (9.2)$$

¹ For the Feynman rules, see J. D. Bjorken and S. D. Drell, *Relativistic Quantum Fields* (McGraw-Hill, New York, 1965), Appendix B.

$$iD'_{\mu\nu}(k) = \frac{\text{---}}{k} + \frac{\text{---}}{k} \circlearrowleft \frac{\text{---}}{k}$$

Fig. 9.1 Photon propagator to second order

From the lowest-order Feynman graph, one obtains

$$\Pi^{\mu\nu}(k) = ie_0^2 \int \frac{d^4 p}{(2\pi)^4} \text{Tr} \left(\gamma^\mu \frac{1}{p-m} \gamma^\nu \frac{1}{p-k-m} \right), \quad (9.3)$$

where m is the physical electron mass. This expression is quadratically divergent, and is not of the form (9.2). The recipe to remedy the situation, as explained in the appendix to this chapter, is to replace $\Pi^{\mu\nu}(k)$ by $\Pi^{\mu\nu}(k) - \Pi^{\mu\nu}(0)$, which satisfies (9.2). This procedure also reduces the quadratic divergence to a logarithmic one, and gives

$$\Pi(k^2) = -\frac{1}{12\pi^2} \ln \frac{\Lambda^2}{cm^2} + C(k^2), \quad (9.4)$$

where Λ is a cutoff momentum, c is a numerical constant that depends on the method of cutting off, and $C(k^2)$ is convergent:²

$$C(k^2) = \frac{1}{2\pi^2} \int_0^1 dx x(1-x) \ln \left[1 - x(1-x) \frac{k^2}{m^2} \right]. \quad (9.5)$$

This function is real for k^2 below the threshold of pair production, i.e., $k^2 < 4m^2$.

With (9.2), we have

$$D'_{\mu\nu}(k) = - \left(g_{\mu\nu} - \frac{k_\mu k_\nu}{k^2} \right) \frac{d'(k^2)}{k^2} + \begin{pmatrix} \text{gauge-dependent} \\ \text{terms} \end{pmatrix}, \quad (9.6)$$

$$d'(k^2) = 1 + e_0^2 \Pi(k^2).$$

Charge renormalization consists of subtracting off the logarithmic divergence in $d'(k^2)$, and absorbing it into a redefinition of the charge.

Define a finite function $\tilde{\Pi}$ by

$$\tilde{\Pi}(k^2, \mu^2) \equiv \Pi(k^2) - \Pi(\mu^2) = C(k^2) - C(\mu^2), \quad (9.7)$$

where μ^2 is an arbitrary number (the renormalization point). Then,

$$d'(k^2) = [1 + e_0^2 \Pi(\mu^2)] + e_0^2 \tilde{\Pi}(k^2, \mu^2). \quad (9.8)$$

Let

$$Z(\mu^2) \equiv 1 + e_0^2 \Pi(\mu^2). \quad (9.9)$$

² J. M. Jauch and F. Rohrlich, *The Theory of Photons and Electrons* (Addison-Wesley, Reading, Mass., 1955), p. 194, Eqs. (9-65); N. N. Bogolubov and D. V. Shirkov, *Introduction to the Theory of Quantized Fields* (Wiley-Interscience, New York, 1959), p. 296, Eq. (24.35).

To second-order accuracy, we can rewrite

$$d'(k^2) = Z(\mu^2)[1 + e_0^2 Z(\mu^2)\tilde{\Pi}(k^2, \mu^2)]. \quad (9.10)$$

Charge renormalization is carried out by identifying the physical charge as

$$e^2(\mu^2) = e_0^2 Z(\mu^2). \quad (9.11)$$

Using the notation $\alpha = e^2/4\pi$, we rewrite this as

$$\alpha(\mu^2) = \alpha_0 Z(\mu^2). \quad (9.12)$$

The physically relevant quantity $\alpha_0 d'(k^2)$ can now be written as

$$\alpha_0 d'(k^2) = \alpha(\mu^2)[1 + 4\pi\alpha(\mu^2)\tilde{\Pi}(k^2, \mu^2)]. \quad (9.13)$$

The standard renormalized propagator $d_c(k^2)$ is obtained by choosing $\mu = 0$:

$$d_c(k^2) = 1 + 4\pi\alpha\tilde{\Pi}(k^2, 0), \quad (9.14)$$

where α is the experimentally measured fine-structure constant:

$$\alpha \equiv \alpha(0) \cong 1/137. \quad (9.15)$$

The physical content of theory is of course independent of the choice of the renormalization point. The different $\alpha(\mu^2)$ merely correspond to different but equivalent definitions of the coupling constant. However, the function $\alpha(\mu^2)$, known as the “running coupling constant”, contains important physical information. To see this, note that the right side of (9.13) is actually independent of μ , and that $\tilde{\Pi}(k^2, k^2) = 0$ by (9.7). Thus, by choosing successively $\mu^2 = k^2$ and $\mu^2 = 0$, we obtain

$$\alpha(k^2) = \alpha d_c(k^2). \quad (9.16)$$

Thus, the effects of vacuum polarization may be viewed in two ways. We might say that it modifies the propagator of a virtual photon, or alternatively, that it makes the effective fine-structure constant momentum-dependent. It is important to keep in mind that $\alpha(k^2)$ and $\alpha d_c(k^2)$ are interchangeable concepts.

Our earlier calculations give, to lowest order in α ,

$$\frac{\alpha(k^2)}{\alpha} = 1 + \frac{2\alpha}{\pi} \int_0^1 dx x(1-x) \ln \left[1 - x(1-x) \frac{k^2}{m^2} \right]. \quad (9.17)$$

For large $|k^2/m^2|$, we can write

$$\alpha(k^2) = \alpha + \frac{\alpha^2}{3\pi} \ln \left| \frac{k^2}{m^2} \right|. \quad (9.18)$$

Since perturbation theory certainly becomes invalid when $\alpha \ln |k^2/m^2| \sim 1$, (9.18) should be used only for

$$1 \ll |k^2/m^2| \ll e^{137}. \quad (9.19)$$

The electrostatic potential energy between two unit test charges (in units of the electronic charge) is given by

$$\begin{aligned} V(r) &= e^2 \int \frac{d^3 k}{(2\pi)^3} e^{ik \cdot r} \left[\frac{d_c(k^2)}{-k^2} \right]_{k_0=0} \\ &= \int \frac{d^3 k}{(2\pi)^3} \frac{e^{ik \cdot r}}{\mathbf{k} \cdot \mathbf{k}} 4\pi\alpha(-\mathbf{k}^2). \end{aligned} \quad (9.20)$$

We can also write

$$V(r) = \int d^3 r' \frac{\rho(r')}{|\mathbf{r} - \mathbf{r}'|}, \quad (9.21)$$

where

$$\rho(r) = \int \frac{d^3 k}{(2\pi)^3} e^{ik \cdot r} 4\pi\alpha(-\mathbf{k}^2). \quad (9.22)$$

Thus, $4\pi\alpha(-\mathbf{k}^2)/e$ is the Fourier transform of the charge density surrounding a bare charge e_0 placed in the vacuum—the “charge form factor” associated with vacuum polarization.

Let us describe qualitatively the charge cloud induced by vacuum polarization. Far from the bare charge, one sees the renormalized charge, because

$$V(r) \xrightarrow[r \rightarrow \infty]{} \frac{\alpha}{r}, \quad (9.23)$$

which follows from the fact that $\alpha(0) = \alpha$. At distances comparable to the electron Compton wavelength $1/m$, one begins to see an effective charge *larger* than the renormalized charge. This can be deduced from the fact that $\alpha(-\mathbf{k}^2)$ *increases* as \mathbf{k}^2 becomes large. Thus, the effect of vacuum polarization is to screen the bare charge. As one penetrates deeper into the charge cloud, the screening becomes less effective, and one sees more of the bare charge. Our understanding stops at $r \sim e^{-137}/m$, because perturbation theory fails for shorter distances. From perturbation theory, we cannot tell whether the bare charge is infinite (as suggested by perturbation theory), or that it is really finite. The question probably has a mathematical answer within quantum electrodynamics; but it is not physically relevant, for long before we reach $r \sim e^{-137}/m$, other interactions not taken into account in quantum electrodynamics would surely become important.

In quantum chromodynamics the analog of $\alpha(-\mathbf{k}^2)$ is a *decreasing* function of \mathbf{k}^2 . Thus the effective coupling vanishes in the limit $\mathbf{k}^2 \rightarrow \infty$ —a phenomenon known as “asymptotic freedom”.

9.2 Perturbative Renormalization in Quantum Electrodynamics

Quantum electrodynamics can be shown to be renormalizable to all orders of perturbation theory. The result is that the relations (9.10), (9.12) are correct not just to second order, but to all orders. We describe the essence of the proof here. The actual proof is given in the appendix to this chapter.

The first step is to enumerate the types of divergent Feynman graphs. For any Feynman graph, let

- n = No. of vertices,
 E_i = No. of internal electron lines,
 E_e = No. of external electron lines,
 P_i = No. of internal photon lines,
 P_e = No. of external photon lines.

The following relations can be shown:

$$\begin{aligned} E_i &= n - \frac{1}{2} E_e, \\ P_i &= \frac{1}{2} (n - P_e). \end{aligned} \tag{9.24}$$

The number of independent internal 4-momenta is given by

$N = E_i + P_i - (n - 1)$, which, by (9.24), may be reexpressed as

$$N = \frac{1}{2}(n - E_e - P_e) + 1. \tag{9.25}$$

A graph may be represented schematically in the form

$$\text{Graph} \sim \int \frac{d^{4N}k}{(k^2)^{P_i} k^{E_i}}. \tag{9.26}$$

The integral above is generally divergent, and must be made finite by introducing a cut-off momentum Λ , which eventually tends to infinity. Renormalization is a procedure through which we subtract appropriate terms from the integral to render it finite in the limit $\Lambda \rightarrow \infty$, and absorb the subtracted terms into mass and charge renormalization.

We define a *primitively divergent graph* as a graph that is divergent, but becomes convergent when any internal line is cut (*i.e.*, when any integration variable is held fixed). Any divergent graph can be reduced to a primitively divergent one by cutting a sufficient number of internal lines. This is obvious because the graph becomes convergent when all internal lines are cut.

The superficial degree of divergence d of a primitively divergent graph may be obtained through naive power counting:^a

$$d = 4N - 2P_i - E_i. \tag{9.27}$$

The actual degree of divergence may be smaller than d . Using (9.24) and (9.25), we can rewrite

$$d = 4 - P_e - E_e. \tag{9.28}$$

Note that d is independent of n , and decreases as the number of external lines is increased. This is what makes the theory renormalizable.

^a The degree of divergence of a non-primitively divergent integral cannot be obtained by power counting. For example, the integral $\int dk dp k^{-1} p^{-2}$ is logarithmically divergent; but naive power counting would give $d = -1$.

There are only a finite number of types of primitively divergent graphs, and they can be classified according to P_e and E_e , as shown in Table 9.1. More detailed considerations, as indicated in Table 9.1, reduce these to only 3 types:

- proper electron self-energy (SE),
- proper photon self-energy,
- proper vertex.

The term “proper” is synonymous with “one-particle irreducible”. It denotes a connected graph that cannot be made disconnected by cutting only one internal line.

It is sufficient to consider only connected non-vacuum graphs. For any such graph, we define its *skeleton graph* as the graph obtained by removing all SE and vertex insertions inside the original graph. The skeleton graph may be convergent or divergent. If divergent, it must be primitively divergent.

To prove this last statement, assume the contrary. Then, by cutting a sufficient number of internal lines, the graph can be reduced to (possibly disconnected) components, of which one is primitively divergent. The latter must be either an SE or vertex graph, as indicated in Table 9.1. But these have been removed by definition. (Contradiction).

Table 9.1 PRIMITIVELY DIVERGENT GRAPHS

P_e	E_e	d	Example	Remarks
0	0	4		Vacuum graph. (May be ignored)
0	2	1		Electron SE. Superficially lin. div. Actually log. div.
1	2	0		Vertex. log. div
2	0	2		Photon SE. Superficially quad. div. Actually log. div. by gauge invar.
3	0	1		Cancelled by graph with electron arrow reversed, by Furry's theorem. (Ignore)
4	0	0		Sum of 4! graphs corresponding to permutation of ext. mom is convergent. Actually even more convergent than superficially indicated, by gauge invariance

For an arbitrary connected non-vacuum graph, we deal with its possible divergence as follows:

- Obtain the skeleton by removing all insertions.
- If the skeleton is convergent, no subtraction is needed. The graph is then renormalized by re-inserting renormalized SE and vertex insertions.
- If the skeleton is divergent, it must be either an SE or vertex graph. The problem is thus reduced to the renormalization of SE and vertex graphs.

We see from the above that, to renormalize a general graph of order n , it suffices to renormalize SE and vertex graphs to order less than or equal to n . Thus, the renormalization can be carried out order by order in perturbation theory.

The SE and vertex graphs are all logarithmically divergent. Therefore, subtracting the graphs at a fixed value of the external momenta renders them finite, and the divergences are isolated in the subtraction constants. There are only 2 independent subtraction constants, because the electron SE and the vertex graph are related by the Ward-Takahashi identity, which is described in the Appendix.

The final step is to show that the subtraction constants can be absorbed through redefinitions of the mass and the charge. The mass renormalization is trivial. The charge renormalization relies on certain scaling properties of Feynman graphs, which makes a subtraction equivalent to a multiplication, so to speak. Details will be given in the Appendix.

9.3 The Renormalization Group

1 Scale Transformations

Charge renormalization is expressed by the relation

$$d'(k^2) = Z(\mu^2)d(k^2). \quad (9.29)$$

The divergence in $d'(k^2)$ has been absorbed into the renormalization constant $Z(\mu^2)$, which depends on an arbitrary renormalization point μ . The renormalized quantity $d(k^2)$ is finite. A shift of the renormalization point is a scale transformation that induces a multiplicative transformation on $Z(\mu^2)$. The group of such scale transformations is called the renormalization group (RG).

To study the scaling properties of $Z(\mu^2)$, we must display all the arguments of the relevant functions. Note that unrenormalized quantities depend on $\alpha_0(\Lambda^2)$ and Λ , while renormalized quantities depends on $\alpha(\mu^2)$ and μ , where $\alpha(\mu^2)$ is the renormalized fine-structure constant at the scale μ . The relation (9.29) can be rewritten, with all arguments displayed, in the following form:

$$\alpha_0(\Lambda^2)d'\left(\frac{k^2}{\Lambda^2}, \frac{m^2}{\Lambda^2}, \alpha_0(\Lambda^2)\right) = \alpha(\mu^2)d\left(\frac{k^2}{\mu^2}, \frac{m^2}{\mu^2}, \alpha(\mu^2)\right), \quad (9.30)$$

where

$$\alpha(\mu^2) = \alpha_0(\Lambda^2)Z\left(\frac{m^2}{\Lambda^2}, \frac{\mu^2}{\Lambda^2}, \alpha_0(\Lambda^2)\right). \quad (9.31)$$

The scale parameter μ satisfies the normalization condition

$$d\left(1, \frac{m^2}{\mu^2}, \alpha(\mu^2)\right) = 1. \quad (9.32)$$

We note that the left side of (9.30) is independent of μ , and is therefore a renormalization-group invariant. Equating it to the expression obtained by setting $\mu^2 = k^2$, and making use of (9.32), we obtain

$$\alpha(k^2) = \alpha(\mu^2)d\left(\frac{k^2}{\mu^2}, \frac{m^2}{\mu^2}, \alpha(\mu^2)\right). \quad (9.33)$$

In principle, this relation gives us $\alpha(\mu^2)$, provided the function d is known. As a general functional relation, however, it does not restrict the form of $\alpha(\mu^2)$, for it can be easily shown that, given any function α , one can always find a function d that satisfies it.

2 Scaling Form

It is plausible that we can set $m = 0$ when all energy scales in the problem are much larger than m . Accordingly we assume that the following limit exists:

$$\lim_{m \rightarrow 0} \alpha(\mu^2)d\left(\frac{k^2}{\mu^2}, \frac{m^2}{\mu^2}, \alpha(\mu^2)\right) = f\left(\frac{k^2}{\mu^2}, \alpha(\mu^2)\right), \quad (9.34)$$

with the normalization condition $f(1, \alpha) = 1$. This statement has been verified to order α^2 in perturbation theory.³ With this, (9.33) becomes a functional equation that places a restriction on the form of the running coupling constant:

$$\begin{aligned} \alpha(k^2) &= \alpha(\mu^2)f\left(\frac{k^2}{\mu^2}, \alpha(\mu^2)\right), \\ k^2/m^2 &\gg 1, \quad \mu^2/m^2 \gg 1. \end{aligned} \quad (9.35)$$

This is a scaling form analogous to the ones in statistical mechanics, which are valid near a critical point, where correlation lengths (or inverse masses) go to infinity.

To find the restriction on the running coupling constant, rewrite the scaling form as follows:

$$\alpha(x) = \alpha(y)f\left(\frac{x}{y}, \alpha(y)\right). \quad (9.36)$$

Differentiating both sides with respect to x at fixed y , and then setting $y = x$, we obtain

$$x \frac{d\alpha(x)}{dx} = \beta(\alpha(x)), \quad (9.37)$$

³ R. Jost and J. M. Luttinger, *Helv. Phys. Acta* **23**, 20 (1950).

where

$$\beta(\alpha) \equiv \alpha \left[\frac{\partial f(x, \alpha)}{\partial x} \right]_{x=1} \quad (9.38)$$

This is usually referred to as the “ β -function”, or “Gell-Mann-Low” function, after the authors who first introduced it.⁴ Integrating (9.37), we have

$$\ln \frac{x}{y} = \int_{\alpha(y)}^{\alpha(x)} \frac{d\lambda}{\beta(\lambda)}. \quad (9.39)$$

For quantum electrodynamics, the β -function is given by

$$\beta(\alpha) = \frac{\alpha^2}{3\pi} \left(1 + \frac{3}{4\pi} \alpha + \dots \right), \quad (9.40)$$

the first term of which can be easily obtained from (9.18).

3 Fixed Points

Let us parametrize the energy scale by putting

$$\mu^2 = \mu_0^2 e^t \quad (-\infty < t < \infty), \quad (9.41)$$

where μ_0 is some fixed reference point. Then (9.39) can be rewritten as

$$t = \int_{\alpha(0)}^{\alpha(t)} \frac{d\lambda}{\beta(\lambda)}. \quad (9.42)$$

The integral on the right side must diverge as $|t| \rightarrow \infty$. This means that $\alpha(t)$ must either diverge, or approach a zero of $\beta(\alpha)$ as $|t| \rightarrow \infty$. We further analyze the latter possibility.

A zero of the β -function is called a fixed point, because $d\alpha(\mu^2)/d\mu^2 = 0$ at this point. In Fig. 9.2, we show some examples of what the β -function might look like. The arrows on the horizontal axis indicate the motion of $\alpha(t)$ as $t \rightarrow \infty$. It must approach a fixed point at which $\beta(\alpha)$ is a decreasing function, in order to make the left side of (9.42) positive. Such a fixed point is called an ultraviolet (UV) fixed point. Similarly, as $t \rightarrow -\infty$, $\alpha(t)$ must approach an infrared (UV) fixed point, at which $\beta(\alpha)$ is an increasing function.

In QED, $\alpha = 0$ is an IR fixed point. This is the region in which perturbation theory is valid, and that is why we can calculate low-energy ($t \rightarrow -\infty$) properties of QED. As t increases, $\alpha(t)$ tends to grow, and perturbation theory soon ceases to be applicable. The high-energy behavior of QED is unknown for this reason. An interesting possibility is that illustrated in Fig. 9.2(a), in which the β -function has a UV fixed point at some finite value α^* . In this case $\alpha(t) \rightarrow \alpha^*$ as $t \rightarrow \infty$, which means that α^* is the bare fine-structure constant. The finiteness of α^* would mean that QED is a finite theory.

An opposite behavior is illustrated in Fig. 9.2(b), where the UV fixed point is at the origin, as in QCD. In this case, perturbation theory gets better and better as

⁴ M. Gell-Mann and F. E. Low, *Phys. Rev.* **95**, 1300 (1954).

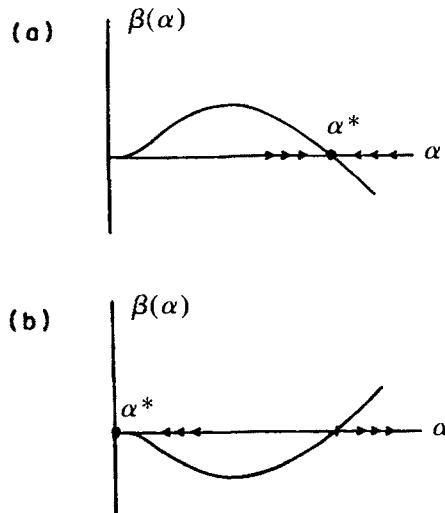


Fig. 9.2. Examples of the β -function (a) for a non-asymptotically free theory, and (b) for an asymptotically free theory. UV fixed points are marked by α^* .

$t \rightarrow \infty$, since $\alpha(t) \rightarrow 0$. This phenomenon is known as “asymptotic freedom”, and furnishes a basis for the parton picture discussed in Sec. 2.5. At low energies, however, the coupling grows and we can no longer calculate the β -functions of QCD. This is the regime in which quarks are believed to be confined.

4 Callan-Symanzik Equation

In a renormalizable field theory, a connected Green's function that is divergent can be expressed in the form

$$\begin{aligned} G'(p; \Lambda, g_0) &= Z(\Lambda/\mu, g_0)G(p; \mu, g), \\ g &= g(\Lambda/\mu, g_0), \end{aligned} \quad (9.43)$$

where Z is a dimensionless renormalization constant, p stands for all external momenta, Λ denotes the cutoff momentum, g_0 is a dimensionless coupling constant, and μ is the renormalization point. (Here we use μ , instead of μ^2 , as an independent variable.) All divergences are isolated in Z , and the renormalized Green's function G is a finite function of the renormalized coupling constant g . In this formula, we have set all masses equal to zero, which means that μ , and all external momenta, are assumed to be much greater than the masses. This is a generalization of a similar formula for the photon propagator, in which g was proportional to Z .

The relation (9.43) is usually the end result of a long and tedious proof; but the latter is merely a matter of technique, which might be improved upon in the future. The machinery for the treatment of divergences is really of secondary significance. Indeed, the significance of (9.43) is quite independent of the fact

that Z is divergent. We may thus look upon the scaling property of Green's functions, as expressed by (9.43), as the most significant result of renormalization theory. We shall examine its consequences here.

Note that the right side of (9.43) must be independent of μ , because the left side is. Therefore,

$$\frac{d}{d\mu} [Z(\Lambda/\mu, g_0)G(p; \mu, g)] = 0. \quad (9.44)$$

Carrying out the differentiation, we obtain

$$\mu \frac{\partial \ln Z}{\partial \mu} G + \mu \frac{\partial G}{\partial \mu} + \mu \frac{\partial g}{\partial \mu} \frac{\partial G}{\partial g} = 0, \quad (9.45)$$

where partial derivatives are carried out with all other arguments held fixed. Introducing two auxiliary dimensionless functions

$$\begin{aligned} \beta(g) &\equiv \mu \frac{\partial}{\partial \mu} g(\Lambda/\mu, g_0), \\ \gamma(\mu) &\equiv \mu \frac{\partial}{\partial \mu} \ln Z(\Lambda/\mu, g_0), \end{aligned} \quad (9.46)$$

we can rewrite (9.45) in the form

$$\left[\mu \frac{\partial}{\partial \mu} + \beta(g) \frac{\partial}{\partial g} + \gamma(\mu) \right] G(p; \mu, g) = 0. \quad (9.47)$$

This is the Callan-Symanzik equation.⁵ The function $\beta(g)$ is the β -function for the present case, and must be regarded as a function of the coupling constant. The function $\gamma(\mu)$ is known as the "anomalous dimension," whose significance will be explained later.

Let us parametrize the energy scale μ by writing

$$\begin{aligned} \mu(t) &= \mu_0 e^t, \\ g(t) &= g(\mu(t)), \\ \gamma(t) &= \gamma(\mu(t)). \end{aligned} \quad (9.48)$$

Then we can write

$$\frac{d}{dt} = \mu \frac{\partial}{\partial \mu} + \beta(g) \frac{\partial}{\partial g}. \quad (9.49)$$

Thus the Callan-Symanzik equation can be rewritten in the form

$$\left[\frac{d}{dt} + \gamma(t) \right] G(p; \mu(t), g(t)) = 0, \quad (9.50)$$

⁵ C. G. Callan, *Phys. Rev.* **D12**, 1541 (1970); K. Symanzik, *Commun. Math. Phys.* **18**, 227 (1970).

with solution

$$G(p; \mu(t), g(t)) = e^{-\Gamma(t)} G(p; \mu(0), g(0)),$$

$$\Gamma(t) = \int_0^t dt' \gamma(t'). \quad (9.51)$$

Now replace p by $e^t p$:

$$G(e^t p; \mu(t), g(t)) = e^{-\Gamma(t)} G(e^t p; \mu(0), g(0)). \quad (9.52)$$

The left side can be rewritten as

$$G(e^t p; e^t \mu(0), g(t)),$$

in which all arguments of dimension mass are multiplied by a common factor e^t . If G itself is of dimension (mass)^d, then

$$G(e^t p; e^t \mu(0), g(t)) = e^{td} G(p; \mu(0), g(t)). \quad (9.53)$$

Equating this to the right side of (9.52), we obtain

$$G(e^t p; \mu(0), g(0)) = e^{td + \Gamma(t)} G(p; \mu(0), g(t)). \quad (9.54)$$

This shows that a change in all the external momenta by a common scale factor is equivalent to changing the coupling constant.

The overall scale factor on the right side of (9.54) is not just e^{td} , as one might expect from naive dimensional analysis, but includes an extra factor $\exp[\int dt \gamma(t)]$ (hence the term ‘‘anomalous dimension’’). The latter arises from the fact that a scale change also changes the renormalization point, and G is not necessarily invariant under this operation.

9.4 Scalar Fields

1 Renormalizability

We consider a real scalar field ϕ in D -dimensional space-time, with a Lagrangian density of the form

$$\mathcal{L} = \frac{1}{2} \partial^\mu \phi \partial_\mu \phi - g_0 \phi^K, \quad (9.55)$$

where $\mu = (0, 1, \dots, D-1)$, and K is an integer. The purpose is to illustrate the fact that renormalizability depends on the interaction, as well as the dimension of space-time. It will be seen that terms not originally in the Lagrangian density can arise due to renormalization.

Feynman graphs of this theory contain vertices at which K lines meet. Each external line is emitted (or absorbed) by one vertex, while each internal line goes between two different vertices. Thus, an external line ‘‘uses up’’ $1/K$ vertices, while an internal line uses up $2/K$ vertices. For a graph with n vertices, L_e external lines, and L_i internal lines, we have the relation

$$\frac{L_e}{K} + \frac{2L_i}{K} = n. \quad (9.56)$$

The number of independent internal D -momenta is given by

$$N = L_i - (n - 1). \quad (9.57)$$

The naive degree of divergence d of a graph may be estimated by power counting:

$$\text{Graph} \sim \int \frac{(d^D k)^N}{(k^2)^{L_i}} \sim k^{ND - 2L_i}. \quad (9.58)$$

Thus the degree of divergence is $d = ND - 2L_i$, which, by virtue of (9.56) and (9.57), gives

$$d = D - \frac{1}{2}(D - 2)L_e + \frac{1}{2}[(D - 2)K - 2D]n. \quad (9.59)$$

For the theory to be renormalizable, this should be independent of n . Hence the coefficient of n must vanish, or

$$K = \frac{2D}{D - 2}. \quad (9.60)$$

Examples of renormalizable theories are ϕ^6 theory in 3D, ϕ^4 theory in 4D, and ϕ^3 theory in 6D.

We have used the term “renormalizable” in the conventional sense in perturbation theory, to mean that there are an infinite number of divergent graphs, but that they can be rendered finite by treating a finite number of primitive divergences. The number of divergent graphs becomes finite, if the coefficient of n is negative, which means

$$\frac{(K - 2)D}{K} < 2. \quad (9.61)$$

In this case the theory is called “super-renormalizable”.

For the ϕ^6 theory in 3D, we have $d = 3 - L_e/2$. Possible primitive divergences therefore correspond to $L_e = 2, 4, 6$. (There are no graphs with odd L_e .) Thus, we need 3 renormalization constants, one corresponding to mass renormalization ($L_e = 2$), and the others corresponding to the renormalization of 4- and 6-particle couplings. One way to subtract these divergences is to add “counter terms” to the Lagrangian density, of the form

$$\mathcal{L}_{\text{counter}} = \delta m \phi^2 + \delta g_4 \phi^4 + \delta g_6 \phi^6. \quad (9.62)$$

These will generate graphs to cancel the divergences coming from the “bare” interactions. We see that interactions can be generated through renormalization, even if the corresponding bare couplings happen to be zero.

2 ϕ^4 Theory

We illustrate coupling-constant renormalization in ϕ^4 theory in 4 dimensions.

Mass renormalization, which is trivial, will be ignored, and we set the renormalized mass to zero.

The renormalized coupling constant is defined by the 4-point Green's function, which to second order is given by the Feynman graphs in Fig. 9.3:

$$\begin{aligned} G'_4(p; \Lambda, g_0) = & -ig_0 + \frac{1}{2} (-ig_0)^2 [I(p_1 + p_2) \\ & + I(p_2 + p_3) + I(p_3 + p_1)], \end{aligned} \quad (9.61)$$

where the factor $\frac{1}{2}$ in the second term comes from the symmetry number, and

$$I(p) = \int \frac{d^4 k}{(2\pi)^4} \frac{1}{(k^2 + i\epsilon)[(k+p)^2 + i\epsilon]}. \quad (9.62)$$

The integral is logarithmically divergent, and is to be rotated into Euclidean momentum space and cut-off at momentum Λ . By this procedure we obtain

$$I(p) = \frac{i}{16\pi^2} \ln \frac{\Lambda^2}{-p^2 + i\epsilon}, \quad (9.63)$$

which has been rotated back into Minkowski space. Using this, we obtain

$$\begin{aligned} G'_4(p; \Lambda, g_0) = & -ig_0 + \frac{ig_0^2}{32\pi^2} \left(\ln \frac{\Lambda^2}{s} + \ln \frac{\Lambda^2}{t} + \ln \frac{\Lambda^2}{u} \right), \\ s = & -(p_1 + p_2)^2 + i\epsilon, \\ t = & -(p_2 + p_3)^2 + i\epsilon, \\ u = & -(p_3 + p_1)^2 + i\epsilon. \end{aligned} \quad (9.64)$$

The renormalized coupling constant α is defined by

$$g\left(\frac{\Lambda}{\mu}, g_0\right) \equiv iG'_4(\mu; \Lambda, g_0), \quad (9.65)$$

where the argument μ in G' indicates $p_1^2 = p_2^2 = p_3^2 = p_4^2 = -\mu^2$. (Note that we are renormalizing at a Euclidean point). Using (9.64), we obtain

$$g\left(\frac{\Lambda}{\mu}, g_0\right) = g_0 - \frac{3g_0^2}{32\pi^2} \ln \frac{\Lambda^2}{\mu^2}, \quad (9.66)$$

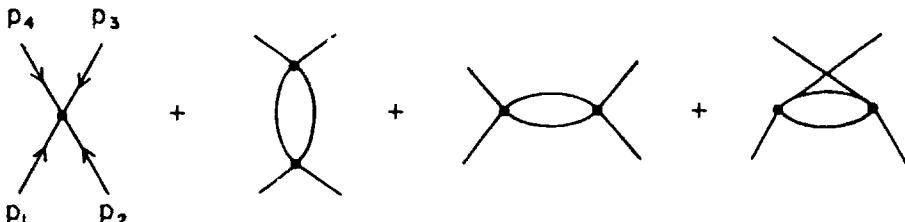


Fig. 9.3 Feynman graphs for 4-point function in ϕ^4 theory.

from which we obtain the β -function by using (9.46):

$$\beta(g) = \frac{3g^2}{16\pi^2}. \quad (9.67)$$

Since this is positive, the theory is not asymptotically free. We can obtain the running coupling constant $g(\mu)$ through

$$\int_{g(\mu_0)}^{g(\mu)} \frac{dg'}{\beta(g')} = \ln \frac{\mu}{\mu_0} \quad (9.68)$$

which gives

$$\frac{1}{g(\mu)} = \frac{1}{g(\mu_0)} - \frac{3}{32\pi^2} \ln \frac{\mu^2}{\mu_0^2}. \quad (9.69)$$

This result represents an effective summation of “leading logarithms”, i.e., graphs of order $(g \ln \Lambda)^n$ in the n -th order.

3 “Triviality” and the Landau Ghost

Let us rewrite (9.69) in the form

$$g(E) = \frac{g}{1 - (3g/32\pi^2) \ln (E/m)}, \quad (9.70)$$

where we regard $g(E)$ as a scattering amplitude at center-of-mass energy E , and $g \equiv g(m)$ is the renormalized coupling at some reference energy m . This amplitude has a pole at $E = M$, with negative residue:

$$g(E) \approx -\frac{32\pi^2 M}{3(E - M)},$$

$$M = m e^{32\pi^2/g}, \quad (9.71)$$

indicating the existence of a state whose wave function has negative squared modulus—a “ghost state”. To get rid of this state, we have to make $M \rightarrow \infty$ by making $g \rightarrow 0$, resulting in a “trivial” theory with no scattering.

Another way to arrive at the same conclusion is to rewrite (9.69) in the form

$$g_0 = \frac{g}{1 - (3g/32\pi^2) \ln (\Lambda/m)}, \quad (9.72)$$

where $g_0 = g(\Lambda)$ is the bare coupling. When we attempt to take the limit $\Lambda \rightarrow \infty$ while holding g fixed, we find that g_0 diverges at $\Lambda = M$, the location of the ghost pole. Thus, the cutoff cannot be lifted unless $g = 0$.

The triviality issue was first raised by Landau⁶ in the context of QED. A calculation of the running coupling constant via (9.42), using the lowest-order

⁶ L. D. Landau, in *Niels Bohr and the Development of Physics*, Ed. W. Pauli (McGraw-Hill, New York, 1955).

β -function, leads to the following expression for the photon propagator:

$$D(k) = \frac{1}{k^2[1 - (\alpha/3\pi) \ln(k^2/m^2)]}, \quad (9.73)$$

which has a ghost pole (the Landau ghost) at

$$k^2 = m^2 e^{3\pi/\alpha}. \quad (9.74)$$

Källen and Pauli⁷ have studied such ghost states in the exactly soluble Lee model,⁸ and found that they make the S -matrix non-unitary, and this pathology cannot be cured by redefining the Hilbert space to admit negative norms.

Our conclusions have been based on a partial sum of Feynman graphs, and therefore cannot be considered definitive. However, they are corroborated by Monte-Carlo simulations, and renormalization group analysis in $4-\epsilon$ dimensions. The ghost issue is significant only for a theoretical understanding of field theoretic interactions. It does not occur in asymptotically free theories, which is characterized by a dynamical suppression of short-distance interactions. Thus we can attribute the emergence of ghost states to the singular nature of unsuppressed point interactions.⁹

It should be emphasized that the phenomena discussed here have no practical implication, because we can keep the cutoff finite but large, and treat the renormalized coupling as a free parameter, as we usually do. We shall expand on this point at the end of this chapter.

9.5 The Physics of Renormalization

In this section we reformulate renormalization from Wilson's point of view,¹⁰ which is independent of perturbation theory, and brings out the physical meaning of renormalization.

1 Renormalization-Group Transformation

Consider a set of fields denoted collectively by $\phi(x)$, with a Lagrangian density of the form

$$\mathcal{L}(\phi(x)) = g_1 K_1(\phi(x)) + g_2 K_2(\phi(x)) + \dots, \quad (9.75)$$

where the g_λ are "bare" coupling constants, and the K_λ are functions of ϕ , referred to as "operators". A momentum cutoff Λ is assumed. For definiteness, we take K_1 to be the the free-field Lagrangian density of a scalar field:

$$K_1 = \frac{1}{2} (\partial \phi(x))^2. \quad (9.76)$$

⁷ G. Källen and W. Pauli, *K. Dan. Vidnsk. Selk. Mat. Fys. Medd.* **30**, No. 7 (1955).

⁸ T. D. Lee, *Phys. Rev.* **95**, 1329 (1954).

⁹ K. Huang, *Int. J. Mod. Phys.* **A4**, 1037 (1989).

¹⁰ K. G. Wilson, *Rev. Mod. Phys.* **55**, 583 (1983).

The scale of the field is fixed by choosing $g_1 = 1$. The other operators may be terms like $[\phi(x)]^m [\partial \phi(x)]^n$, for example.

The Euclidean action is denoted by

$$S[\phi] = \int d^Dx \left[\frac{1}{2} (\partial\phi(x))^2 + g_2 K_2(\phi(x)) + \dots \right]. \quad (9.77)$$

In momentum-space we write

$$S[\phi] = \int_{|k|<\Lambda} d^Dk \left[\frac{1}{2} k^2 \phi^2(k) + g_2 K_2(\phi(k)) + \dots \right], \quad (9.78)$$

where $\phi(k)$ is the Fourier transform of $\phi(x)$. We shall not bother to use a different symbol for the Fourier transform of K_λ . The partition function is given by

$$Q = \int \prod_{|k|<\Lambda} d\phi(k) e^{-S[\phi]}. \quad (9.79)$$

At an energy scale much smaller than Λ , we expect that physical processes are insensitive to the values of the individual Fourier components with large k . Renormalization is the process by which we expunge the irrelevant degrees of freedom, and readjust the coupling constants in compensation. In momentum space, the procedure is carried out through the following steps:

- A. Integrate out the Fourier components with $\Lambda/l < |k| < \Lambda$ to obtain a new action $\tilde{S}[\phi]$:

$$e^{-\tilde{S}[\phi]} = \int \prod_{\Lambda/l < |k| < \Lambda} d\phi(k) e^{-S[\phi]}. \quad (9.80)$$

The partition function is recovered by integrating over the remaining Fourier components:

$$Q = \int \prod_{|k|<\Lambda/l} d\phi(k) e^{-\tilde{S}[\phi]}.$$

There is thus no loss of information when we pass to the new action. The main point is that the new action can be written in terms of the original operators with new coupling constants, up to an additive constant:

$$\tilde{S}[\phi] = C + \int_{|k|<\Lambda/l} d^Dk \left[\frac{1}{2} \tilde{g}_1 k^2 \phi^2(k) + \tilde{g}_2 K_2(\phi(k)) + \dots \right]. \quad (9.82)$$

This is always possible if we allow for a sufficiently large number of operators in the beginning. That is, new operators that emerge can be regarded as present in the original Lagrangian density with zero coupling. The constant C has no physical relevance, but is generally needed mathematically.

- B. Rescale the cutoff to Λ , by changing the integration variable from k to

$$k' = lk. \quad (9.83)$$

Denoting the dimension of K_λ by d_λ , we can rewrite

$$\begin{aligned}\widetilde{S}[\phi] = C + \int_{|k'|<\Lambda} d^D k' \left[\frac{1}{2} \widetilde{g}_1 l^{-D-2} k'^2 \phi^2(k'/l) \right. \\ \left. + l^{-D-d_2} \widetilde{g}_2 K_2(\phi(k'/l)) + \dots \right].\end{aligned}\quad (9.84)$$

C. Renormalize the field to absorb \widetilde{g}_1 . Use as new field variable

$$\phi'(k) \equiv (l^{-D-2} \widetilde{g}_1)^{1/2} \phi(k/l). \quad (9.85)$$

The renormalized action is $S'[\phi'] \equiv \widetilde{S}[\phi]$, with the expansion

$$S'[\phi'] = C + \int_{|k|<\Lambda} d^D k \left[\frac{1}{2} k^2 \phi'^2(k) + g'_2 K_2(\phi'(k)) + \dots \right], \quad (9.86)$$

which defines the g'_λ . The net result of RG transformation is that the degrees of freedom of the system are thinned out, as illustrated in Fig. 9.4.

The partition function can be rewritten as

$$Q = \mathcal{N} \int \prod_{|k|<\Lambda} d\phi'(k) e^{-S'[\phi']}, \quad (9.87)$$

where \mathcal{N} is a constant. In all respects, the system behaves as if its Lagrangian density were

$$\mathcal{L}'(\phi'(x)) = \frac{1}{2} \partial \phi'(x) + g'_2 K_2(\phi'(x)) + \dots, \quad (9.88)$$

The transformation $\mathcal{L}(\phi) \rightarrow \mathcal{L}'(\phi')$, can also be described as a coupling-constant transformation:

$$\{g_2, g_3, \dots\} \rightarrow \{g'_2, g'_3, \dots\}. \quad (9.89)$$

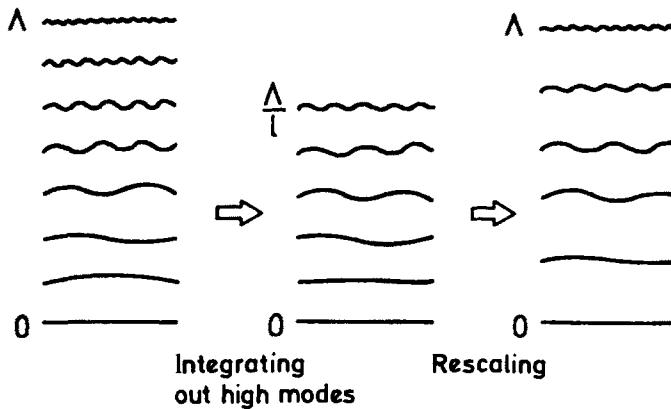


Fig. 9.4 RG transformation in momentum space. Λ is the cutoff momentum.

This is called a renormalization-group (RG) transformation, but is not exactly the same as the RG transformation in perturbative QED. The operations involved here do not even form a group, as they have no inverses. The relation between the present RG to the historical one will be pointed out later.

For many calculations, it is often advisable to first place the system in a large but finite space-time volume Ω , and take the infinite-volume limit later. Momentum space then becomes discrete, and the action reads

$$S[\phi] = \sum_{|k|<\Lambda} \left[\frac{1}{2} k^2 \phi_k + g_2 K_2(\phi_k) + \dots \right], \quad (9.90)$$

where the Fourier component ϕ_k is related to $\phi(k)$ through $\phi_k = \Omega^{-1/2} \phi(k)$. The analog of (9.85) reads

$$\phi'_k = (l^{-2} \tilde{g}_1)^{1/2} \phi_{k/l}. \quad (9.91)$$

Consider now a correlation function

$$G(k; g) = \langle \phi(k) \phi(-k) \rangle \equiv \frac{\int (D\phi) \phi(k) \phi(-k) e^{-S[\phi]}}{\int (D\phi) e^{-S[\phi]}}, \quad (9.92)$$

where $\phi(k)$ is the k th Fourier component of a scalar field, and g denotes a set of coupling constants. Since we are considering a single Fourier component, we first make the spectrum discrete by putting the system in a finite box, and re-express the above in terms of ϕ_k . For $k \ll \lambda$, an RG transformation integrates out momenta much larger than k , and the only effect it has on ϕ_k is the scaling transformation (9.91). It is important to note that in the coarse-graining and rescaling procedures in the RG transformation, the space-time volume must be held fixed. Thus, after restoring $\phi(k)$, we obtain the scaling form

$$G(k; g) = Z l^2 G(lk; g'), \quad Z = \tilde{g}_1. \quad (9.93)$$

This is valid in any number of dimensions, and is trivially satisfied by the free-field correlation function $(k^2 + m^2)^{-1}$. Being the analog of (9.54), it shows the equivalence between the RG transformation described here and the usual perturbative renormalization.

2 Real-Space Renormalization

An alternative way to introduce a cutoff is to place the system on a D -dimensional Euclidean lattice of spacing $a = \Lambda^{-1}$. The RG transformation then consists of averaging the fields within a block of unit cells, and regarding the block as a new unit cell. This is called ‘‘real-space renormalization’’. In this approach, the RG transformation is literally a coarse-graining operation.

Let us label the lattice sites by an integer i , and denote a position vector by

$$x_i = a n_i, \quad (9.94)$$

where the components of n_i are integers. The action has the form

$$S[\phi] = \sum_{\langle ij \rangle} [\phi(x_i) - \phi(x_j)]^2 + \sum_i g_2 K_2(\phi(x_i)) + \dots, \quad (9.95)$$

where $\langle ij \rangle$ denotes a pair of nearest-neighbor sites. The partition function is given by

$$Q = \int (D\phi) e^{-S[\phi]}, \quad \int (D\phi) = \int \prod_i d\phi(x_i). \quad (9.96)$$

Now group the sites into blocks, l sites on a side. [Figure 9.5 shows an example with $l = 2$.] We name the blocks B_α ($\alpha = 1, 2, \dots$), and denote the position of a block (say, its center) by

$$R_\alpha = lan_\alpha. \quad (9.97)$$

The average field in a block is given by

$$\tilde{\phi}(R_\alpha) = l^{-D} \sum_{i \in B_\alpha} \phi(x_i). \quad (9.98)$$

The steps for the real-space RG parallel those in momentum space:

- A. Integrate out the original field ϕ , while holding $\tilde{\phi}$ fixed, to obtain a new action $\tilde{S}[\tilde{\phi}]$:

$$e^{-\tilde{S}[\tilde{\phi}]} \equiv \int (D\phi) P(\tilde{\phi}, \phi) e^{-S[\phi]}, \quad (9.99)$$

where

$$P(\tilde{\phi}, \phi) \equiv \prod_\alpha \delta \left(\tilde{\phi}(R_\alpha) - l^{-D} \sum_{i \in B_\alpha} \phi(x_i) \right). \quad (9.100)$$

Since $\int (D\tilde{\phi}) P(\tilde{\phi}, \phi) = 1$, the partition function can be written as

$$Q = \int (D\tilde{\phi}) e^{-\tilde{S}[\tilde{\phi}]}, \quad \int (D\tilde{\phi}) = \int \prod_\alpha d\tilde{\phi}(R_\alpha). \quad (9.101)$$

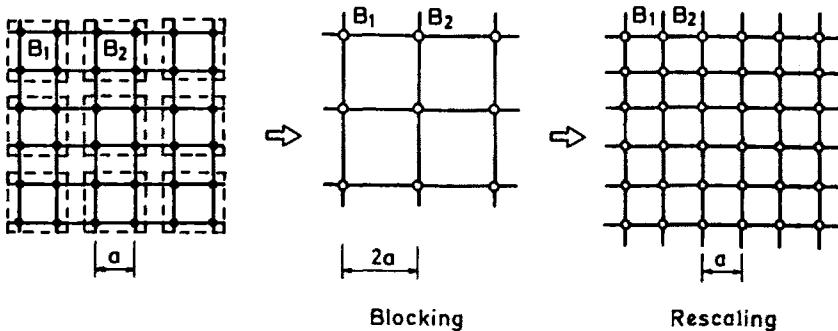


Fig. 9.5 RG transformation in real space.

The new action can be put in the form

$$\tilde{S}[\tilde{\phi}] = C + \tilde{g}_1 \sum_{\langle\alpha\beta\rangle} [\tilde{\phi}(R_\alpha) - \tilde{\phi}(R_\beta)]^2 + \sum_\alpha \tilde{g}_2 K_2(\tilde{\phi}(R_\alpha)) + \dots \quad (9.102)$$

B. Rescale the block lattice spacing to a . Use as position vectors

$$R'_\alpha = R_\alpha / l. \quad (9.103)$$

C. Renormalize the field to absorb \tilde{g}_1 . Use as new field

$$\phi'(R'_\alpha) \equiv \sqrt{\tilde{g}_1} \tilde{\phi}(R'_\alpha). \quad (9.104)$$

The renormalized action is defined as $S'[\phi'] \equiv \tilde{S}[\tilde{\phi}]$, with the expansion

$$S'[\phi'] = C + \sum_{\langle\alpha\beta\rangle} [\phi'(R'_\alpha) - \phi'(R'_\beta)]^2 + \sum_\alpha g'_2 K_2(\phi'(R'_\alpha)) + \dots \quad (9.105)$$

which defines the g'_λ . The partition function can be rewritten

$$Q = \mathcal{N} \int (\mathcal{D}\phi') e^{-S'[\phi']} \quad (9.106)$$

The real-space RG transformation is illustrated in Fig. 9.5.

3. Fixed Points and Relevancy

Let us put $l = e^t$, so that the rescaled cutoff is

$$\Lambda/l = \Lambda e^{-t} \quad (0 \leq t < \infty). \quad (9.107)$$

An RG transformation corresponds to an increase of t . By making successive infinitesimal RG transformations, we can find the g_λ as functions of t . Assuming that they change smoothly with t , we can describe their behavior by differential equations called RG equations:

$$\frac{dg_\lambda(t)}{dt} = \beta_\lambda(g_2, g_3, \dots), \quad (9.108)$$

where β_λ are the generalized β -functions. They do not depend on t explicitly, because the effective action does not depend on t explicitly.

In coupling-constant space, the point $g = \{g_2, g_3, \dots\}$ traces out a trajectory with t as parameter, referred to as an ‘‘RG trajectory’’. The dimensionality of the coupling-constant space should be as large as is necessary to contain all the RG trajectories corresponding to different initial conditions. The historical RG of perturbative QED is just the group of motions along a trajectory, as described by (9.108).

A fixed point g^* in coupling-constant space is a root of the β -functions:

$$\beta_\lambda(g^*) = 0 \quad (\text{all } \lambda). \quad (9.109)$$

At such a point the system is invariant under coarse-graining. Defining the correlation length ξ through the asymptotic form of the correlation function

$G(R) \sim \exp(-R/\xi)$, we see that under coarse-graining $\xi \rightarrow \xi/l$. Thus, at a fixed point the correlation length must be 0 or ∞ .

The β -functions generally depend on all the g_λ . Therefore the operators generally mix with one another under the RG transformation. If we are only interested in the behavior near a fixed point g^* , however, we can shift the origin to g^* , and diagonalize the RG equations in its neighborhood. We then have a set of "principle axes", along which the new coupling constant v_λ do not mix. This is illustrated in Fig. 9.6, in which the arrows indicate the coarse-graining direction. The Lagrangian density near g^* would have the form

$$\mathcal{L} = \mathcal{L}^* + v_1 O_1 + v_2 O_2, \quad (9.110)$$

where \mathcal{L}^* is the fixed-point Lagrangian density, which is invariant under the RG. It is clear from Fig. 9.6 that $v_2 \rightarrow 0$ as we approach the fixed point. The associated operator O_2 is called an "irrelevant" operator, because it can be ignored at the fixed point. On the other hand, v_1 generally tends to grow, unless the system happens to be on a trajectory that goes into the fixed point. The associated operator O_1 is said to be "relevant".

The definition of relevant and irrelevant operators can be given more formally in terms of a matrix M , which describes the behavior of g near g^* :

$$\frac{d(g - g^*)}{dt} = M(g - g^*), \quad (9.111)$$

where g is considered a column vector, and only linear terms are kept on the right side. The principle axes are those defined by the eigenvectors of M . A positive eigenvalue corresponds to a relevant operator, and a negative one corresponds to an irrelevant operator.

It should be emphasized that relevancy or irrelevancy are properties associated with a particular fixed point. Operators that are irrelevant near one fixed point may be relevant in the neighborhood of another, and vice versa. In fact, two different fixed points connected by a trajectory cannot have the same set of

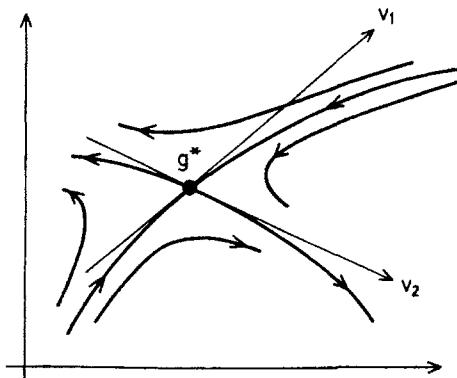


Fig. 9.6. "Principal axes" v_1 and v_2 at a fixed point g^* .

relevant operators. For this reason, the character of a system changes drastically when we go from one fixed point to another. Ignoring the irrelevant operators simplifies the problem in a particular neighborhood; but in so doing, we lose the very information necessary to make a crossover to another fixed point.

It should be noted that the RG is not unique. The one defined via momentum space is not the same as the one defined in real space. Even in the momentum-space or real-space method, the coarse-graining procedure is not unique, and one can have different RG transformations. Thus, the β -functions are not unique. It is believed, on the assumption that the theory makes physical sense, that the number and nature of the fixed points are common to all schemes, though not necessarily their locations. In terms of the matrix M mentioned earlier, its eigenvalues should be independent of the details of the RG scheme, though not its matrix elements.

4 Renormalization and Universality

Let us summarize the basic features of RG transformations by looking at a representative trajectory, as shown in Fig. 9.7. The bare system, defined at a scale set by the cutoff Λ , is represented by the point B . Successive coarse-graining will bring it to the renormalized system at point R , at a lower energy scale Λe^{-t} . By integrating the RG equations along the trajectory, we obtain the renormalized couplings $g(t, \Lambda)$. The theory is considered renormalizable if the number of relevant operators is finite.

In Fig. 9.7 we show a number of fixed points. The local properties of a trajectory is strongly affected by an adjacent fixed point. In particular, the fixed point determines which operators are relevant and which irrelevant. If we extend the trajectories in both directions, we might find that it came from a fixed point (or nearly so), or that it is going into a fixed point (or nearly so). The former is called a UV fixed point, and the latter an IR fixed point. In its flow towards the IR, the trajectory may experience near collisions with many fixed points. Each

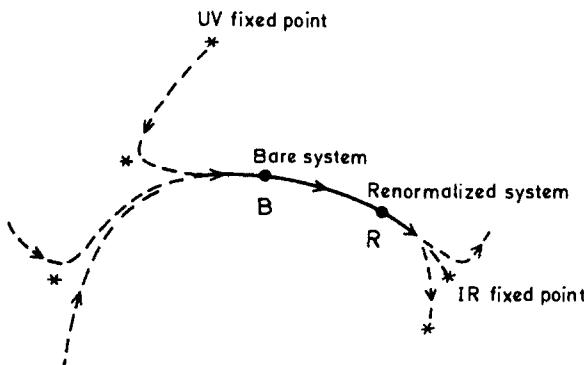


Fig. 9.7 A representative RG trajectory. B marks the "bare" system, which flows to the renormalized system R under successive coarse-graining. The trajectory is influenced by fixed points marked with asterisks.

time it skirts a fixed point, the character of the system changes, since new relevant operators emerge, and old ones die out. In physical situations, the point R lies in a low-energy domain, very far from B. Thus the relevant operators at R are likely to be determined by IR fixed points, if any.

At fixed points, the correlation length ξ is either 0 or ∞ . Those with $\xi = \infty$ are called critical points. Since ξ decreases under coarse-graining, a trajectory coming out of a fixed point can have non-zero ξ only if the fixed point is a critical point. A trajectory that goes into a critical point must have $\xi = \infty$ along the entire trajectory. Systems lying on such a trajectory are called “critical”, in that they are scale invariant.

In actual practice, we treat the renormalized couplings as free parameters, to be fitted to experiments. It is usually said that the cutoff is being sent to infinity. Actually, this is never literally carried out, nor is it necessary. In fact, when we adjust $g(t, \Lambda)$ at fixed t to fit experiments, we are allowing Λ to vary. This implicit variation of Λ is what makes the renormalization prescription work, for it makes allowance for possible modifications at high energies that do not alter the low-energy content of the theory. This renormalization procedure is useful not only in particle physics, where Λ lies in the unknown, but also in condensed matter physics, where Λ is a physical cutoff provided by atomicity.

When we seriously consider the limit $\Lambda \rightarrow \infty$, we are going beyond renormalization. To carry out this limiting process, we push the bare system further and further back along the trajectory (against the direction of the RG flow). The limit exists if the trajectory eventually backs into a fixed point (the UV fixed point), which should be a critical point. When this is so, the renormalized coupling $g_\lambda(t, \Lambda)$ approaches a limit $g_\lambda(t, \infty)$. This means that all properties of the system are calculable from “first principles”. Different trajectories that came from the same critical point describe systems in the same universality class. As we have noted, a trajectory may go past many different fixed points. If it went sufficiently near a fixed point, it behaves further downstream as if it came from that fixed point. Thus, in searching for an UV fixed point, the best we can hope for is to find the nearest one, which could serve as a staging area for looking further upstream.

The idea of universality first came up in condensed matter physics. Here, one looks not for UV fixed points, but IR fixed points, which define universality classes with characteristic critical exponents. This is feasible, because the relevant IR fixed points are located in the same region where the physical phenomena occur.

Let us now turn to some specific cases. In QCD we do have a UV fixed point, which corresponds to a free theory. The fact that the fixed point exists means that massless QCD is a theory with no free parameters. Unfortunately, knowing that things are calculable does not help us to calculate them. So far we are unable to calculate the low-energy properties, which includes quark confinement.

In QED, renormalization has given us the most accurately tested theory we have; but the trajectories in the UV remains unknown. The Landau ghost discussed earlier suggests that the theory may not be self-consistent; but that either means that the approximations leading to the Landau ghost is invalid, or that we just need a wee bit of irrelevant operator to fix things up at high energies.

The search for the nearest UV fixed point is at best a treacherous pursuit. But to find a likely nearest fixed point is synonymous with finding a nearly correct theory. And that, as Einstein said, is like “hunting ducks where there are few ducks, on a very dark night”.

Appendix to Chapter 9. Renormalization of QED

We outline perturbative renormalization in QED to all orders, following the method of Dyson and Ward.¹¹ We use the following notation:

$S'(p)$ = Full electron propagator,

$D'_{\mu\nu}$ = Full photon propagator,

$\Gamma_\mu(p, p')$ = Full vertex.

1 Vertex

We put

$$\Gamma_\mu(p, p') = \gamma_\mu + \Lambda_\mu(p, p'). \quad (9.A1)$$

The graphs for Λ_μ are shown in Fig. 9.A1 (a), and the skeleton expansion is shown in Fig. 9.A1 (b). The sum of all skeleton graphs, denoted by Λ_μ^* , is logarithmically divergent.

To obtain Λ_μ from Λ_μ^* , we replace all free propagators and bare vertices in the graphs of Λ_μ^* by the corresponding full propagators and full vertices. This may be expressed as follows:

$$\Lambda_\mu(p, p') = \Lambda_\mu^*[S', D'_{\alpha\beta}, \Gamma_\nu; e_0^2, p, p'], \quad (9.A2)$$

where Λ_μ^* is regarded as a functional of the propagator functions and the vertex function, with the bare squared charge e_0^2 and the momenta p, p' appearing as parameters of the functional. In this sense, the graphical series in Fig. 9.A2 (b) corresponds to $\Lambda_\mu^*[S, D_{\alpha\beta}, \gamma_\nu; e_0^2, p, p']$, where S and $D_{\alpha\beta}$ are free propagators.

The functional Λ_μ^* is logarithmically divergent, *i.e.*, it depends on the cut-off Λ as a parameter, and diverges like $\ln\Lambda$ as $\Lambda \rightarrow \infty$. The full vertex function Λ_μ is much more divergent, due to divergences coming from the insertions.

2 Electron Propagator

There is an ambiguity in the way insertions should be made in the skeletons of SE graphs, as illustrated in Fig. 9.A2. These are known as “overlapping divergences”. For the electron propagator, we can avoid the problem by making use of the Ward-Takahashi identity,¹² to express it in terms of the vertex:

$$[S'(p)]^{-1} = [S'(p_0)]^{-1} + (p - p_0)^\mu \Gamma_\mu(p, p_0). \quad (9.A3)$$

¹¹ F. J. Dyson, *Phys. Rev.* **75**, 486, 1736 (1949); J. C. Ward, *Proc. Phys. Soc. (London)* **A64**, 54 (1951).

¹² J. C. Ward, *Phys. Rev.* **78**, 1824 (1950); Y. Takahashi, *N. Cimento* **6**, 370 (1957).

$$(a) \quad \Lambda_\mu(p, p') = \frac{i}{-ie_0} \left\{ \begin{array}{c} \text{Diagram with } p' - p \text{ above vertex} \\ \text{Diagram with } p \text{ below vertex} \\ \text{Diagram with } p' \text{ below vertex} \end{array} \right\} + \begin{array}{c} \text{Diagram with } p' - p \text{ above vertex} \\ \text{Diagram with } p \text{ below vertex} \end{array} + \begin{array}{c} \text{Diagram with } p' - p \text{ above vertex} \\ \text{Diagram with } p \text{ below vertex} \end{array} + \dots + O(e_0^6)$$

$$(b) \quad \Lambda_\mu^*(p, p') = \frac{i}{-ie_0} \left\{ \begin{array}{c} \text{Diagram with } p' - p \text{ above vertex} \\ \text{Diagram with } p \text{ below vertex} \end{array} \right\} + \dots + O(e_0^6)$$

Fig. 9.A1 (a) Proper vertex graphs. (b) Skeleton vertex graphs.

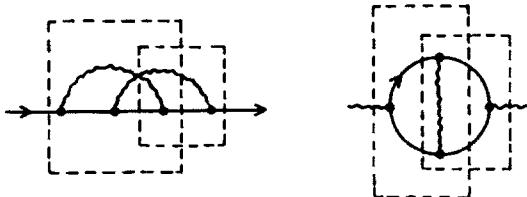


Fig. 9.A2 Overlapping divergences.

The right-hand side is actually independent of p_0 , but for the sake of definiteness we take p_0 to be the momentum of an electron on the mass shell. "Mass renormalization" consists in asserting that

$$[S'(p_0)]^{-1} = C(p_0 - m), \quad (9.A4)$$

where m is the physical mass of the electron, and C is a cutoff-dependent constant.

3 Photon Propagator

The free photon propagator is

$$D_{\mu\nu}(k) = - \left[g_{\mu\nu} - (1 - \lambda) \frac{k_\mu k_\nu}{k^2} \right] \frac{1}{k^2}, \quad (9.A5)$$

where λ is the gauge parameter. Regarding $D_{\mu\nu}$ as a matrix with Minkowski metric, we can rewrite (9.A5) in matrix form:

$$\mathbb{D}(k) = - \left[\mathbb{P}_T(k) + \lambda \mathbb{P}_L(k) \right] \frac{1}{k^2}, \quad (9.A6)$$

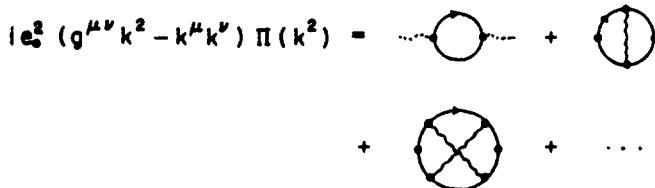


Fig. 9.A3 Vacuum polarization tensor.

where \mathbb{P}_T and \mathbb{P}_L are the transverse and longitudinal projection operators defined in (8.48). The full photon propagator is given in matrix notation by

$$\begin{aligned} iD'(k) &= iD(k) + iD(k)i\Pi(k)iD(k) + \dots \\ &= iD(k)[1 - i\Pi(k)iD(k)]^{-1}, \end{aligned} \quad (9.A7)$$

where $i\Pi_{\mu\nu}$ is the vacuum polarization tensor—the sum of all proper photon SE graphs, with external lines omitted. The graphical expansion for $\Pi_{\mu\nu}$ is shown in Fig. 9.A3.

Gauge invariance (or current conservation) requires that

$$k^\mu \Pi_{\mu\nu}(k) = 0. \quad (9.A8)$$

Therefore $\Pi_{\mu\nu}$ must have the form^b

$$\Pi_{\mu\nu}(k) = (g_{\mu\nu}k^2 - k_\mu k_\nu)e_0^2\Pi(k^2). \quad (9.A9)$$

Since, by power counting, $\Pi_{\mu\nu}(k)$ has quadratic skeletal divergences, $\Pi(k^2)$ has only logarithmic skeletal divergences, because two powers of the momentum have been factored out. In matrix notation, (9.A9) reads

$$\Pi(k) = \mathbb{P}_T(k)e_0^2k^2\Pi(k^2). \quad (9.A10)$$

Substituting this into (9.A7), we obtain

$$iD'(k) = iD(k)\left[\frac{1}{1 - e_0^2\Pi(k^2)}\mathbb{P}_T(k) + \lambda\mathbb{P}_L(k)\right]. \quad (9.A11)$$

Restoring the indices, we write

$$iD'_{\mu\nu}(k) = i\left(g_{\mu\nu} - \frac{k_\mu k_\nu}{k^2}\right)D'(k^2) + \frac{\lambda k_\mu k_\nu}{ik^2}, \quad (9.A12)$$

where

$$[iD'(k^2)]^{-1} = ik^2[1 - e_0^2\Pi(k^2)]. \quad (9.A13)$$

This is a gauge-invariant function. The second term in (9.A12) depends on the gauge, and is physically irrelevant.

^b Methods to insure gauge invariance in the calculation of $\Pi_{\mu\nu}$ are discussed at the end of this appendix.

To deal with overlapping divergences, define an auxiliary function $W_\mu(k)$ by

$$W_\mu(k) \equiv \frac{\partial}{\partial k^\mu} [iD'(k^2)]^{-1}. \quad (9.A14)$$

Using (9.A13), we write

$$W_\mu(k) = 2ik_\mu - ik_\mu T(k^2), \quad (9.A15)$$

where

$$T(k^2) = \frac{e_0^2}{k^2} k^\mu \frac{\partial}{\partial k^\mu} [k^2 \Pi(k^2)]. \quad (9.A16)$$

To recover D' from W_μ , use the formula

$$[iD'(k^2)]^{-1} = \int_0^1 dx k^\mu W_\mu(xk). \quad (9.A17)$$

The graphical expansion for W_μ is shown in Fig. 9.A4 (a). There are no overlapping divergences. A skeleton expansion for T is obtained by removing from the graphs not only all SE and vertex insertions, but also insertions of W_μ . The expansion, denoted by T^* , is shown in Fig. 9.A4 (b). An example illustrat-

$$(a) \quad k_\mu T_\mu(k^2) = e_0^2 k_\mu \frac{\partial}{\partial k^\mu} [k^2 \Pi(k^2)]$$

can be obtained from the graphs:

$$\begin{array}{c} \text{---} \\ \text{---} \end{array} + \begin{array}{c} \text{---} \\ \text{---} \end{array} + O(e_0^2)$$

$$(b) \quad k_\mu T_\mu^*(k^2) \text{ can be obtained from the graphs:}$$

$$\begin{array}{c} \text{---} \\ \text{---} \end{array} + \begin{array}{c} \text{---} \\ \text{---} \end{array} + O(e_0^2)$$

Fig. 9.A4 (a) Graphs for the auxiliary function $T(k^2)$. A cross indicates differentiation with respect to the external momentum k , which is by convention routed exclusively along the upper half of the closed loops. (b) Skeleton expansion for $T(k^2)$, denoted by $T^*(k^2)$.

ing the removal of W_μ is shown in Fig. 9.A5.^c We can now write

$$T(k^2) = T^*[S', D'_{\alpha\beta}, \Gamma_\nu, W_\mu; e_0^2, k^2]. \quad (9.A18)$$

The functional T^* is logarithmically divergent.

4 Scaling Properties

The reason that divergences in perturbation theory can be absorbed into renormalized coupling constants is that, in a certain sense, a subtraction is equivalent to a multiplication. This property depends on the scaling properties of the insertions, i.e., their behavior under a scale change of their arguments. The relations derived below therefore lie at the heart of renormalizability.

Graphs defining Λ_μ^* are of even order, and a graph of order $2n$ contains factors of e_0^2 , S , $D_{\alpha\beta}$ and γ_ν to various powers, as indicated schematically below:

$$\Lambda_{2n}^* \sim e_0^{2n} S^{2n} (D_{\alpha\beta})^n (\gamma_\nu)^{2n+1}. \quad (9.A19)$$

Under the transformation

$$\begin{aligned} \gamma_\nu &\rightarrow a\gamma_\nu, \\ D_{\alpha\beta} &\rightarrow bD_{\alpha\beta}, \\ S &\rightarrow a^{-1}S, \end{aligned} \quad (9.A20)$$

where a and b are arbitrary numbers,

$$\Lambda_{2n}^* \rightarrow ab^n \Lambda_{2n}^*. \quad (9.A21)$$

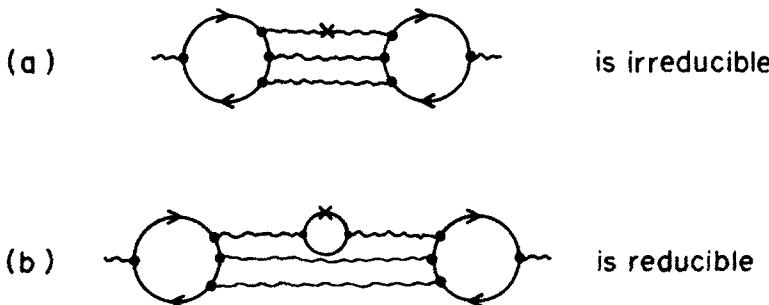


Fig. 9.A5 Illustration of removal of insertions in $T(k^2)$.

^c The external momentum k can be routed through a graph in more than one way, and this gives rise to ambiguities in the definition of a skeleton expansion. The difficulty occurs in graphs of W_μ containing at least 3 closed electron loops, and are therefore at least of order e_0^{14} . For a discussion of the difficulty and a convention of momentum-routing that overcomes the difficulty, see T. T. Wu, *Phys. Rev.* **125**, 1436 (1962).

Therefore,

$$\begin{aligned} a\Lambda_\mu^*[S, D_{\alpha\beta}, \gamma_\nu; e_0^2, p, p'] \\ = \Lambda_\mu^*[a^{-1}S, bD_{\alpha\beta}, a\gamma_\nu; b^{-1}e_0^2, p, p']. \end{aligned} \quad (9.A22)$$

For the functional T^* , a graph of order $2n$ has the structure

$$T_{2n}^* \sim e_0^{2n} S^{2n+\sigma} (D_{\alpha\beta})^{n-\sigma} (\gamma_\nu)^{2n+\sigma} (2ik_\mu)^{1-\sigma}, \quad (9.A23)$$

where σ is an integer that receives an additive contribution 1 from each differentiation of an electron line, and 0 from that of a photon line. Under the transformation (9.A20) supplemented by $2ik_\mu \rightarrow 2ik_\mu/b$,

$$T_{2n}^* \rightarrow b^{n-1} T_{2n}^*. \quad (9.A24)$$

Therefore,

$$\begin{aligned} b^{-1} T^*[S, D_{\alpha\beta}, \gamma_\nu, 2ik_\mu; e_0^2, k^2] \\ = T^*[a^{-1}S, bD_{\alpha\beta}, a\Gamma_\nu, b^{-1}2ik_\mu; b^{-1}e_0^2, k^2]. \end{aligned} \quad (9.A25)$$

5 Renormalization

We are interested in renormalizing the functions S' , $D'_{\alpha\beta}$, and Γ_ν , which satisfy the following coupled functional equations:

$$\begin{aligned} \Gamma_\mu(p, p') &= \gamma_\mu + \Lambda_\mu^*[S', D'_{\alpha\beta}, \Gamma_\nu; e_0^2, p, p'], \\ W_\mu(k) &= 2ik_\mu + ik_\mu \Gamma^*[S', D'_{\alpha\beta}, \Gamma_\nu, W_\lambda; e_0^2, k^2], \\ [S'(p)]^{-1} &= [S'(p_0)]^{-1} + (p - p_0)^\mu \Gamma_\mu(p, p_0), \\ [D'(k^2)]^{-1} &= \int_0^1 dx \, k^\mu W_\mu(xk). \end{aligned} \quad (9.A26)$$

These yield divergent functions as solutions, because the functionals Λ_μ^* and T^* are divergent. However, since they are only logarithmically divergent, they can be rendered convergent through one subtraction.

Using the abbreviations $\Lambda_\mu^*(p, p')$ and $T^*(k^2)$ for the functionals, we define two finite functionals by^d

$$\begin{aligned} \widetilde{\Lambda}_\mu(p, p') &\equiv \Lambda_\mu^*(p, p') - [\Lambda_\mu^*(p_0, p_0)]_{p_0=m}, \\ \widetilde{T}(k^2) &\equiv T^*(k^2) - T^*(\mu^2), \end{aligned} \quad (9.A27)$$

where μ is an arbitrary invariant mass of the photon, and p_0 is the momentum of an electron on mass shell, with $p_0^2 = m^2$. The subscript $p_0 = m$ instructs us to commute p_0 all the way to the right, and then replace it by m . Thus,

$$[\Lambda_\mu^*(p_0, p_0)]_{p_0=m} = L \gamma_\mu, \quad (9.A28)$$

^d For simplicity we have chosen to subtract Γ_μ at a mass-shell momentum p_0 . Actually the subtraction can be made at any momentum, whose invariant mass would then serve as an extra floating renormalization point in addition to μ .

where L is a power series in e_0^2 with logarithmically divergent coefficients. The same is true of $T^*(\mu^2)$.

We now define finite renormalized functions \widetilde{S} , $\widetilde{D}_{\alpha\beta}$, $\widetilde{\Gamma}_\nu$, \widetilde{W}_μ as solutions of the functional equations obtained from (9.A26) by replacing Λ_μ^* and T^* by Λ_μ and \widetilde{T} respectively, and replacing e_0^2 by an appropriate number e^2 :

$$\begin{aligned}\widetilde{\Gamma}_\mu(p, p') &= \gamma_\mu + \widetilde{\Lambda}_\mu[\widetilde{S}, \widetilde{D}_{\alpha\beta}, \widetilde{\Gamma}_\nu; e^2, p, p'], \\ \widetilde{W}_\mu(k) &= 2ik_\mu + ik_\mu \widetilde{T}[\widetilde{S}, \widetilde{D}_{\alpha\beta}, \widetilde{\Gamma}_\nu, \widetilde{W}_\lambda; e^2, k^2], \\ [\widetilde{S}(p)]^{-1} &= [\widetilde{S}(p_0)]^{-1} + (p - p_0)^\mu \widetilde{\Gamma}_\mu(p, p_0), \\ [\widetilde{D}(k^2)]^{-1} &= \int_0^1 dx \, k^\mu W_\mu(xk).\end{aligned}\quad (9.A29)$$

We fix the normalization of \widetilde{S} by the condition

$$[\widetilde{S}(p_0)]^{-1} = \not{p}_0 - m. \quad (9.A30)$$

Then

$$[\widetilde{S}(p)]^{-1} = \not{p} - m + (p - p_0)^\mu \widetilde{\Lambda}_\mu(p, p_0), \quad (9.A31)$$

with the property

$$[(\not{p} - m)\widetilde{S}(p)]_{p=p_0} = 1. \quad (9.A32)$$

The normalization of \widetilde{D} is such that

$$[ik^2 \widetilde{D}(k^2)]_{k^2=\mu^2} = 1. \quad (9.A33)$$

We now show that the renormalized quantities are proportional to the unrenormalized ones. Note that $\widetilde{\Gamma}_\mu$ can be rewritten as follows:

$$\begin{aligned}\widetilde{\Gamma}_\mu &= \gamma_\mu + \Lambda_\mu^* - L\gamma_\mu = (1 - L) \left(\gamma_\mu + \frac{1}{1 - L} \Lambda_\mu^* \right) \\ &= Z' \left\{ \gamma_\mu + \frac{1}{Z'} \Lambda_\mu^* [\widetilde{S}, \widetilde{D}_{\alpha\beta}, \widetilde{\Gamma}_\nu; e^2, p, p'] \right\},\end{aligned}\quad (9.A34)$$

where

$$Z' = 1 - L. \quad (9.A35)$$

This shows that a subtraction is equivalent to rescaling. Similarly, we can write

$$\begin{aligned}\widetilde{W}_\mu &= 2ik_\mu + ik_\mu [T^*(k^2) - T^*(\mu^2)] \\ &= Z \left\{ 2ik_\mu + \frac{1}{Z} ik_\mu T^*[\widetilde{S}, \widetilde{D}_{\alpha\beta}, \widetilde{\Gamma}_\nu, \widetilde{W}_\lambda; e^2, k^2] \right\},\end{aligned}\quad (9.A36)$$

where

$$Z = 1 - \frac{1}{2} T^*(\mu^2). \quad (9.A37)$$

Using the scaling properties (9.A22) and (9.A25), we obtain

$$\begin{aligned}\frac{1}{Z'} \widetilde{\Gamma}_\mu &= \gamma_\mu + \Lambda_\mu^* \left[Z' \widetilde{S}, Z \widetilde{D}_{\alpha\beta}, \frac{1}{Z} \widetilde{\Gamma}_\nu; \frac{e^2}{Z}, p, p' \right], \\ \frac{1}{Z} \widetilde{W}_\mu &= 2ik_\mu + ik_\mu T^* \left[Z' \widetilde{S}, Z \widetilde{D}_{\alpha\beta}, \frac{1}{Z'} \widetilde{\Gamma}_\nu, \frac{1}{Z} \widetilde{W}_\lambda, \frac{e^2}{Z}, k^2 \right].\end{aligned}\quad (9.A38)$$

Thus, the system of equations (9.A29) can be reduced to (9.A26) by putting

$$\begin{aligned}\Gamma_\mu &= \frac{1}{Z'} \widetilde{\Gamma}_\mu, \\ W_\mu &= \frac{1}{Z} \widetilde{W}_\mu, \\ S' &= Z' \widetilde{S}, \\ D'_{\mu\nu} &= Z D_{\mu\nu}, \\ e_0^2 &= \frac{e^2}{Z}.\end{aligned}\quad (9.A39)$$

The last statement is known as “charge renormalization”.^c

The renormalized quantities are presumably finite, because they satisfy the finite equations (9.A29). To show that they are actually finite, one has to show that (9.A29) has finite solutions. That is, the restoration of renormalized insertions into a renormalized skeleton should produce a convergent integral. That this is true has been shown by Weinberg.¹³

6 Gauge Invariance and the Photon Mass

As we have stated earlier, gauge invariance (or current conservation) requires that the vacuum polarization tensor $\Pi_{\mu\nu}$ be of the form

$$\Pi_{\mu\nu}(k) = (g_{\mu\nu} k^2 - k_\mu k_\nu) \Pi(k^2). \quad (9.A40)$$

This leads to a photon propagator whose gauge-invariant part reads

$$iD'_{\mu\nu}(k) = \left(g_{\mu\nu} - \frac{k_\mu k_\nu}{k^2} \right) \frac{1}{ik^2 [1 + \Pi(k^2)]}. \quad (9.A41)$$

The pole at $k^2 = 0$ corresponds to a massless photon. This pole can be shifted only if $\Pi(k^2)$ develops a pole at $k^2 = 0$, which cannot happen in perturbation theory. Thus, in perturbation theory, the masslessness of the photon is an automatic consequence of gauge invariance, and there is no need for mass renormalization.

^c It should be emphasized that the renormalization scheme is based on perturbation theory, and both Z and Z^{-1} must be regarded as power series in α_0 with divergent coefficients. If we terminate the power series to any finite order, as we must do in practice, then both Z and Z^{-1} are divergent, though $Z Z^{-1} = 1$ to the order considered. Formal considerations lead one to believe that $0 \leq Z \leq 1$, and probably $Z \rightarrow 0$ as $\Lambda \rightarrow \infty$ [G. Källen, *Helv. Phys. Acta*, **25**, 417 (1952)].

¹³ S. Weinberg, *Phys. Rev.* **118**, 838 (1960).

However, a naive calculation of $\Pi_{\mu\nu}$ to the lowest order fails to verify (9.A40). One finds instead that the leading divergence of $\Pi_{\mu\nu}$ is quadratic and proportional to $g^{\mu\nu}$. This means that naively one would obtain

$$iD'_\mu(k) = \frac{g_{\mu\nu}}{i[k^2 + R(k^2)]}, \quad (9.A42)$$

where $R(k^2)$ appears to be a quadratically divergent mass term.

The source of the difficulty is the failure of current conservation, due to the singular nature of the current

$$j^\mu(x) = ie\bar{\psi}(x)\gamma^\mu\psi(x), \quad (9.A43)$$

which involves the product of $\psi(x)$ and its canonical conjugate $\psi^\dagger(x)$ at the same space-time point. There are several ways to rectify the calculation.

A method to overcome the problem, known as the “point-splitting method”,¹⁴ is to re-define the current as

$$j^\mu(x) = \lim_{\epsilon \rightarrow 0} ie\bar{\psi}(x + \frac{1}{2}\epsilon)\gamma^\mu\psi(x - \frac{1}{2}\epsilon) \exp\left[ie \int_{x-\epsilon/2}^{x+\epsilon/2} dy^\mu A_\mu(y)\right]. \quad (9.A44)$$

The exponential factor above is necessary to maintain gauge invariance. This factor gives a non-vanishing contribution even in the limit $\epsilon \rightarrow 0$, and cures the problem.

A simpler remedy is to expand the naively calculated $\Pi_{\mu\nu}(k)$ in a Taylor series about $k = 0$, and subtract a sufficient number of leading terms until one gets the desired form. In practice one subtraction suffices. The justification for this procedure is as follows. One argues that the violation of gauge invariance occurs in the real part of $\Pi_{\mu\nu}$ but not in the imaginary part, because the latter describes physical instead of virtual processes. Thus, gauge invariance can be restored by subtracting an appropriate polynomial in k , which has no imaginary part. The practical recipe then is to replace the naive $\Pi_{\mu\nu}(k)$ by $\Pi_{\mu\nu}(k) - \Pi_{\mu\nu}(0)$. One may note that the quadratically divergent constant $\Pi_{\mu\nu}(0)$ can be cancelled by a counter term in the Lagrangian density, of the form $C F^{\mu\nu} F_{\mu\nu}$. Thus it can be absorbed by re-scaling $F^{\mu\nu}$. However, such a procedure is purely formal, and does not justify the recipe any better.

The failure of gauge invariance can occur even in finite Feynman graphs, for example, in the scattering of light by light to lowest order. Here, the sum of the $4!$ graphs, corresponding to all possible permutations of the external photon momenta, gives a convergent integral (although each individual graph is logarithmically divergent). However, the integral is not gauge invariant, and must be modified by the methods mentioned above.

¹⁴ J. Schwinger, *Phys. Rev.* **82**, 664 (1951); K. Johnson, in *Particles and Field Theory*, Eds. S. Deser and K. W. Ford (Prentice-Hall, Englewood Cliffs, N. J., 1965).

CHAPTER 10

METHOD OF EFFECTIVE POTENTIAL

10.1 Spontaneous Symmetry Breaking

In this chapter we shall discuss renormalization in the presence of spontaneous symmetry breaking. To put things in perspective, it should be pointed out that spontaneous symmetry breaking is a purely theoretical concept that has nothing to do with experiments. The vacuum expectation value of the Higgs field, for example, is not a directly observable quantity. It can be deduced indirectly from observational data only within the framework of a theoretical model. There is no experimental way in which we can tell whether masses arise from “mechanical mass terms” or from spontaneous symmetry breaking.

One might then wonder why we should use this concept at all. The reason, as we have learned in the Weinberg-Salam model, is that it enables us to construct a renormalizable theory of massive vector bosons with which we can do practical calculations. Thus, we should know something about renormalization with spontaneous symmetry breaking, which is most conveniently done with the help of the “effective potential”.

To introduce the technique, we shall discuss the mathematical example of a self-coupled scalar field. The relevance to physics comes when the scalar field is coupled to other fields, such as a gauge field. A simple example of the latter will be described, but we forego a full treatment of renormalization in a realistic theory¹.

10.2 The Effective Action²

Consider a real scalar field with Lagrangian density

$$\begin{aligned}\mathcal{L}(x) &= \frac{1}{2} \partial_\mu \phi(x) \partial^\mu \phi(x) - V(\phi(x)), \\ V(\phi) &= \frac{\alpha_0}{4!} \phi^4 - \frac{m_0^2}{2} \phi^2.\end{aligned}\tag{10.1}$$

The classical value of the vacuum field is at the minimum ρ_0 of $V(\phi)$, but the vacuum expectation value $\langle\phi\rangle$ of the quantum field is not necessarily the same,

¹ See B. W. Lee and J. Zinn-Justin, *Phys. Rev.* D5, 3137 (1974); D7, 1049 (1973).

² J. Goldstone, A. Salam, and J. Weinberg, *Phys. Rev.* 127, 965 (1962); G. Jona-Lasinio, *N. Cimento* 34, 1790 (1964).

being defined by

$$\langle \phi \rangle \equiv \lim_{J \rightarrow 0} \frac{\langle 0^+ | \phi_{\text{op}} | 0^- \rangle_J}{\langle 0^+ | 0^- \rangle_J}, \quad (10.2)$$

where the notation is that of Sec. 7.3.

Let us recall the definition of the generating functional $W[J]$:

$$\begin{aligned} \exp \frac{i}{\hbar} W[J] &= \mathcal{N} \int (\mathcal{D}\phi) \exp \frac{i}{\hbar} \{S[\phi] + (J, \phi)\}, \\ S[\phi] &= \int d^4x \mathcal{L}(x), \quad (J, \phi) = \int d^4x J(x) \phi(x). \end{aligned} \quad (10.3)$$

As a book-keeping device, we keep \hbar as displayed in (10.3), but set $\hbar = c = 1$ everywhere else. By expanding in powers of \hbar , we obtain an expansion in terms of the number of closed loops in the Feynman graphs. If we restore all \hbar and c , but keep $\mathcal{L}(x)$ exactly as written in (10.1), we would find that $\hbar c \alpha$ is the only dimensionless parameter in the theory. Thus the loop expansion is actually an expansion in powers of $\hbar c \alpha$, but it performs an infinite re-summation of Feynman graphs.

The vacuum expectation value of the field in the presence of an external source $J(x)$ is given by

$$\phi_c(x) \equiv \frac{\langle 0^+ | \phi_{\text{op}}(x) | 0^- \rangle_J}{\langle 0^+ | 0^- \rangle_J} = \frac{\delta W[J]}{\delta J(x)}. \quad (10.4)$$

The vacuum expectation value $\langle \phi \rangle$ is the limit of $\phi_c(x)$ as $J \rightarrow 0$. Here, ϕ_c is determined by the external source function J . We now ask the question: What source function J will produce a given function ϕ_c ? To answer the question, we shall reformulate the problem so that it becomes convenient to use ϕ_c as the independent variable, instead of J .

We define an “effective action” $\Gamma[\phi_c]$ by making a Legendre transformation:

$$\Gamma[\phi_c] = W[J] - (J, \phi_c), \quad (10.5)$$

where J is to be eliminated in terms of ϕ_c , through solving (10.4). We call $\Gamma[\phi_c]$ an effective action, because it is a functional of a classical field ϕ_c , and hence akin to $S[\phi]$. We note that

$$\begin{aligned} \frac{\delta \Gamma[\phi_c]}{\delta \phi_c(x)} &= \frac{\delta W[J]}{\delta \phi_c(x)} - J(x) - \int d^4y \frac{\delta J(y)}{\delta \phi_c(x)} \phi_c(y), \\ \frac{\delta W[J]}{\delta \phi_c(x)} &= \int d^4y \frac{\delta W[J]}{\delta J(y)} \frac{\delta J(y)}{\delta \phi_c(x)} = \int d^4y \frac{\delta J(y)}{\delta \phi_c(x)} \phi_c(y). \end{aligned}$$

Substituting the latter equation into the former, we obtain

$$\frac{\delta \Gamma[\phi_c]}{\delta \phi_c(x)} = -J(x). \quad (10.6)$$

If we put $J = 0$, ϕ_c should become a constant, by translational invariance.

Hence $\langle \phi \rangle$ is the root of the equation

$$\left. \frac{d\Gamma[\phi_c]}{d\phi_c} \right|_{\phi_c=\langle \phi \rangle} = 0. \quad (10.7)$$

We recall that $W[J]$ is the generating functional for connected Green's functions. We now show that $\Gamma[\phi_c]$ is the generating functional for proper, or one-particle irreducible (1-PI), Green's functions. We can always expand $\Gamma[\phi_c]$ in powers of ϕ_c , as follows:

$$\Gamma[\phi_c] = \sum_{n=0}^{\infty} \frac{1}{n!} \int d^4x_1 \dots d^4x_n \Gamma_n(x_1, \dots, x_n) \phi_c(x_1) \dots \phi_c(x_n). \quad (10.8)$$

We now show that Γ_n is the n -point 1-PI Green's function, *i.e.*, the sum of all connected 1-PI Feynman graphs with n external legs (with external propagators omitted).

Consider the quantity

$$\mathcal{N} \int (D\phi) \exp \frac{i}{a} \{\Gamma[\phi] + (J, \phi)\}. \quad (10.9)$$

As $a \rightarrow 0$, the integrand is dominated by its value at a saddle point, which occurs at $\phi = \phi_c$, for (10.6) is precisely the saddle-point condition. Thus

$$\mathcal{N} \int (D\phi) \exp \frac{i}{a} \{\Gamma[\phi] + (J, \phi)\} \xrightarrow{a \rightarrow 0} \mathcal{N} \exp \frac{i}{a} \{\Gamma[\phi_c] + (J, \phi_c)\}. \quad (10.10)$$

By (10.5), this statement is the same as

$$\exp \frac{i}{a} W[J] \xrightarrow{a \rightarrow 0} \mathcal{N} \int (D\phi) \exp \frac{i}{a} \{\Gamma[\phi] + (J, \phi)\}. \quad (10.11)$$

In the limit $a \rightarrow 0$, the right side is the sum of all tree graphs of a new theory whose action is $\Gamma[\phi]$. A vertex in one of these tree graphs is some Γ_n , by the definition (10.8). This fact is illustrated in Fig. 10.1 (a). On the other hand, $W[J]$ is by definition the sum of all connected Green's functions of the original theory, and a connected Green's function can be dissected into 1-PI components, as shown in Fig. 10.1 (b). Therefore, Γ_n is a 1-PI Green's function of the original theory. ■

10.3 The Effective Potential

We may alternatively expand $\Gamma[\phi_c]$ in terms of ϕ_c and its derivatives, as follows:

$$\Gamma[\phi_c] = \int d^4x [-U(\phi_c(x)) + \tfrac{1}{2}\partial_\mu \phi_c(x) \partial^\mu \phi_c(x) Z(\phi_c(x)) + \dots]. \quad (10.12)$$

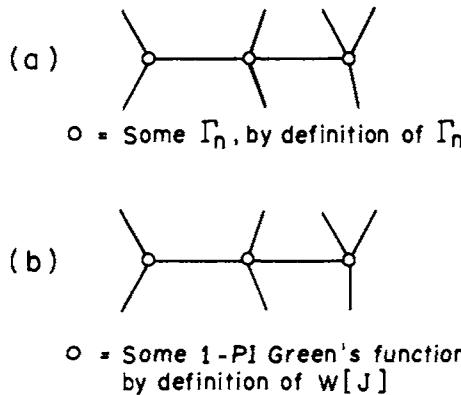


Fig. 10.1 Aid for proving that the effective action generates 1P-I Green's functions.

The function U is called the “effective potential”. It may be extracted from (10.12) by putting $\phi_c(x) = \rho$ (constant). Under this condition, all terms in (10.12) vanish except the first, and we have

$$\Gamma[\rho] = -\Omega U(\rho), \quad (10.13)$$

where Ω is the total volume of space-time.

Let the Fourier transforms of ϕ_c and Γ_n be denoted by

$$\tilde{\phi}_c(k) = \int d^4x e^{ik \cdot x} \phi_c(x) \quad (10.14)$$

$$\begin{aligned} \widetilde{\Gamma}_n(k_1, \dots, k_n) &= (2\pi)^4 \delta^4(k_1 + \dots + k_n) \\ &= \int d^4x_1 \dots d^4x_n \Gamma_n(x_1, \dots, x_n) \exp[-(ik_1 \cdot x_1 + \dots + ik_n \cdot x_n)]. \end{aligned}$$

In terms of these, we can rewrite (10.8) as

$$\begin{aligned} \Gamma[\phi_c] &= \sum_{n=0}^{\infty} \frac{1}{n!} \int \frac{d^4k_1}{(2\pi)^4} \dots \frac{d^4k_n}{(2\pi)^4} (2\pi)^4 \delta^4(k_1 + \dots + k_n) \widetilde{\Gamma}_n(k_1, \dots, k_n) \\ &\quad \cdot \tilde{\phi}_c(k_1) \dots \tilde{\phi}_c(k_n). \end{aligned} \quad (10.15)$$

Now put

$$\phi_c(x) = \rho \quad (\text{constant}), \quad (10.16)$$

so that

$$\tilde{\phi}_c(k) = (2\pi)^4 \delta^4(k) \rho. \quad (10.17)$$

Then

$$\Gamma[\rho] = \Omega \sum_{n=0}^{\infty} \frac{\rho^n}{n!} \tilde{\Gamma}_n(0), \quad (10.18)$$

where $\Omega = (2\pi)^4 \delta^4(0)$, and

$$\tilde{\Gamma}_n(0) \equiv \tilde{\Gamma}_n(0, 0, \dots, 0). \quad (10.19)$$

Comparing (10.18) with (10.13), we obtain

$$U(\rho) = - \sum_{n=0}^{\infty} \frac{\rho^n}{n!} \tilde{\Gamma}_n(0). \quad (10.20)$$

Thus we can make the identification

$$-\left. \frac{d^n U(\rho)}{d\rho^n} \right|_{\rho=0} = \tilde{\Gamma}_n(0). \quad (10.21)$$

We should remember that Γ_n is defined in terms of the field $\phi(x)$. When there is spontaneous symmetry breaking, it is more appropriate to use the shifted field

$$\eta(x) = \phi(x) - \langle \phi \rangle. \quad (10.22)$$

The 1-PI Green's functions $\Gamma_n^{(s)}$ defined with respect to $\eta(x)$ are linear combinations of Γ_n , and may be obtained from the shifted version of (10.20):

$$U(\rho - \langle \phi \rangle) = - \sum_{n=0}^{\infty} \frac{[\rho - \langle \phi \rangle]^n}{n!} \tilde{\Gamma}_n^{(s)}(0). \quad (10.23)$$

Physically useful quantities can be obtained directly from the effective potential:

$$\begin{aligned} [dU(\rho)/d\rho]_{\rho=\langle \phi \rangle} &= 0 && \text{(condition for } \langle \phi \rangle\text{)}, \\ [d^2U(\rho)/d\rho^2]_{\rho=\langle \phi \rangle} &= m^2 && \text{(renormalized mass)}, \\ -[d^4U(\rho)/d\rho^4]_{\rho=\langle \phi \rangle} &= \alpha && \text{(renormalized coupling constant)}. \end{aligned} \quad (10.24)$$

The quantities m^2 and α are free parameters of the renormalized theory. Since we must require $m^2 \geq 0$, the first two conditions say that $\langle \phi \rangle$ is at a minimum of $U(\rho)$, just as ρ_0 is at a minimum of $V(\rho)$. We shall proceed to calculate $U(\rho)$ explicitly in the one-loop approximation.

The vacuum expectation value $\rho(J)$ of the field in the presence of a constant external source J may be likened to the magnetization of a ferromagnet in a constant external magnetic field. Its qualitative behavior is illustrated in Fig. 10.2. The behavior $\lim_{J \rightarrow 0} \rho(J) \neq 0$ corresponds to ferromagnetism. To obtain the effective potential, we note that $dU(\rho)/d\rho = J$. Thus, $dU(\rho)/d\rho$ can be obtained by inverting Fig. 10.2, as shown in Fig. 10.3a. The effective potential is then obtained through integration, as illustrated in Fig. 10.3b.

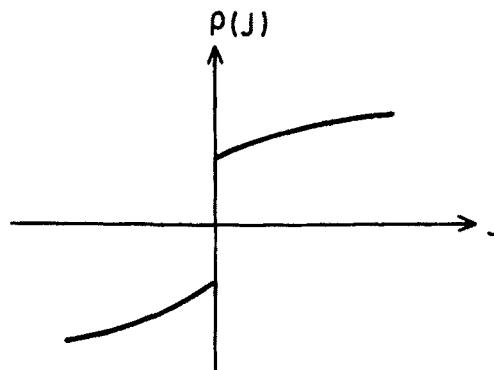


Fig. 10.2 Vacuum expectation value as function of external source.

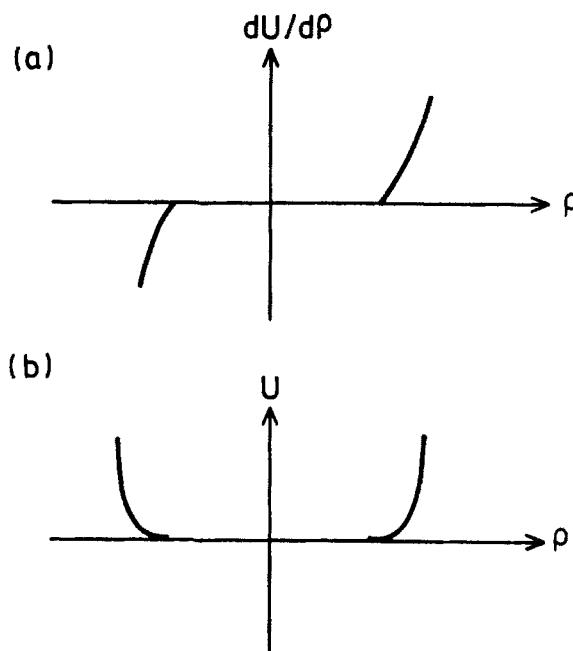


Fig. 10.3 Graphical construction to obtain the effective potential.

10.4 The Loop Expansion³

The loop expansion of $W[J]$ to one-loop order may be obtained from (10.3) by the method of saddle-point integration. The saddle-point is at $\phi = \phi_0$, with ϕ_0 satisfying

³ R. Jackiw, *Phys. Rev. D9*, 1686 (1974).

$$\left. \frac{\delta S[\phi]}{\delta \phi(x)} \right|_{\phi=\phi_0} = -J(x). \quad (10.25)$$

Note that $\phi_0(x)$ is a function of x , and is also a functional of J . When $J \rightarrow 0$, $\phi_0(x)$ becomes a solution to the classical equation of motion.

Expanding $S[\phi]$ about ϕ_0 , we have

$$\begin{aligned} S[\phi + \phi_0] &= S[\phi_0] + \int d^4x \phi(x) \{ \delta S[\phi]/\delta \phi(x) \}_{\phi=\phi_0} \\ &\quad + \frac{1}{2} \int d^4x d^4y \phi(x) \phi(y) \{ \delta^2 S[\phi]/\delta \phi(x) \delta \phi(y) \}_{\phi=\phi_0} \\ &\quad + S_2[\phi, \phi_0], \end{aligned} \quad (10.26)$$

where S_2 includes all higher terms. We introduce a propagator function $\Delta[\phi_0]$, which is a functional of ϕ_0 , by the definition

$$\langle x | i\Delta^{-1}[\phi_0] | y \rangle = \left. \frac{\delta^2 S[\phi]}{\delta \phi(x) \delta \phi(y)} \right|_{\phi=\phi_0}. \quad (10.27)$$

Then

$$S[\phi + \phi_0] = S[\phi_0] - (J, \phi) + \frac{1}{2}(\phi, i\Delta^{-1}[\phi_0] \phi) + S_2[\phi, \phi_0]. \quad (10.28)$$

Now we write

$$\begin{aligned} \exp \frac{i}{\hbar} W[J] &= \mathcal{N} \int (D\phi) \exp \frac{i}{\hbar} \{ S[\phi + \phi_0] + (\phi + \phi_0, J) \} \\ &= \mathcal{N} \int (D\phi) \exp \frac{i}{\hbar} \{ S[\phi_0] + \frac{1}{2}(\phi, i\Delta^{-1}[\phi_0] \phi) \\ &\quad + S_2[\phi, \phi_0] + (\phi_0, J) \} \\ &= \exp \frac{i}{\hbar} \{ S[\phi_0] + (\phi_0, J) \} \cdot \left\{ \mathcal{N} \int (D\phi) \exp \frac{i}{2\hbar} (\phi, i\Delta^{-1}[\phi_0] \phi) \right\} \\ &\quad \cdot \left\langle \exp \frac{i}{\hbar} S_2 \right\rangle, \end{aligned} \quad (10.29)$$

where $\langle \rangle$ denotes functional average with weighting function

$$\exp \frac{i}{2\hbar} (\phi, i\Delta^{-1}[\phi_0] \phi).$$

The middle factor within curly brackets in (10.29) is equal to $\{\det i\Delta^{-1}[\phi_0]\}^{-1/2}$. Thus

$$W[J] = S[\phi_0] + (\phi_0, J) + \frac{i\hbar}{2} \ln \det\{i\Delta^{-1}[\phi_0]\} + W_2[J], \quad (10.30)$$

where

$$W_2[J] = -i\hbar \ln \left\langle \exp \frac{i}{\hbar} S_2 \right\rangle. \quad (10.31)$$

We can show that W_2 is of higher order in \hbar than first. To do this, rescale ϕ by putting $\phi = \hbar^{1/2}\tilde{\phi}$. Then

$$\begin{aligned} \hbar^{-1}S_2[\phi, \phi_0] &= \hbar^{-1}\phi\phi\phi[\delta^2S/\delta\phi \delta\phi \delta\phi]_{\phi=\phi_0} + \dots \\ &= \hbar^{1/2}\tilde{\phi}\tilde{\phi}\tilde{\phi}[\delta^2S/\delta\phi \delta\phi \delta\phi]_{\phi=\phi_0} + \dots. \end{aligned} \quad (10.32)$$

Thus,

$$W_2 = -i\hbar \ln[1 + O(\hbar^{1/2})] \sim O(\hbar^{3/2}). \quad (10.33)$$

In fact $W_2 \sim O(\hbar^2)$, because we know from Sec. 7.7 that $W[J]$ admits a loop expansion in powers of \hbar , and there are no fractional powers. Thus the first two terms in (10.30) represent the tree approximation, and the second term is the complete one-loop correction.

Substitution of (10.30) into (10.5) will give the loop expansion of the effective action $\Gamma[\phi_c]$. However, we must first express $S[\phi_0]$ as a functional of ϕ_c . It is easily seen that $\phi_c = \phi_0$ in the tree approximation. Hence $(\phi_c - \phi_0) \sim O(\hbar)$:

$$\phi_c(x) = \frac{\delta W[J]}{\delta J(x)} = \phi_0(x) + \phi_1(x), \quad \phi_1(x) \sim O(\hbar). \quad (10.34)$$

Now we expand $S[\phi_0]$ in the following manner:

$$\begin{aligned} S[\phi_0] &= S[\phi_c - \phi_1] \\ &= S[\phi_c] - \int d^4x \phi_1(x) \{\delta S[\phi]/\delta\phi(x)\}_{\phi=\phi_0} + O(\hbar^2) \\ &= S[\phi_c] + (\phi_1, J) + O(\hbar^2), \end{aligned} \quad (10.35)$$

where in the last step we have used (10.25). Substituting the last relation into (10.30), and the latter into (10.5), we obtain

$$\Gamma[\phi_c] = S[\phi_c] + \frac{i\hbar}{2} \ln \det\{i\Delta^{-1}[\phi_c]\} + O(\hbar^2).$$

Now we put $J = \text{constant}$. Then $\phi_c(x) = \rho$, and

$$\Gamma[\rho] = S[\rho] + \frac{i\hbar}{2} \ln \det\{i\Delta^{-1}[\rho]\} + O(\hbar^2), \quad (10.37)$$

where ρ is a constant. We know, on the other hand, that

$$\begin{aligned}\Gamma[\rho] &= -\Omega U(\rho), \\ S[\rho] &= \left[\int d^4x \mathcal{L}(x) \right]_{\phi=\rho} = -\Omega V(\rho).\end{aligned}\quad (10.38)$$

Therefore

$$U(\rho) = V(\rho) - \frac{i\hbar}{2} \Omega^{-1} \ln \det\{i\Delta^{-1}[\rho]\} + O(\hbar^2). \quad (10.39)$$

To calculate the second term, we note that

$$\begin{aligned}\ln \det\{i\Delta^{-1}[\rho]\} &= \text{Tr} \ln\{i\Delta^{-1}[\rho]\} \\ &= \int \frac{d^4k}{(2\pi)^4} \ln \langle k | i\Delta^{-1}[\rho] | k \rangle.\end{aligned}\quad (10.40)$$

Hence

$$U(\rho) = V(\rho) - \frac{i\hbar}{2} \int \frac{d^4k}{(2\pi)^4} \Omega^{-1} \ln \langle k | i\Delta^{-1}[\rho] | k \rangle + O(\hbar^2). \quad (10.41)$$

10.5 One-Loop Effective Potential

In the Lagrangian density, put $\phi(x) = \rho + \phi_1(x)$, where ρ is a constant. Then

$$\begin{aligned}S[\phi] &= \int d^4x [\frac{1}{2} \partial_\mu \phi_1 \partial^\mu \phi_1 - V(\phi_1 + \rho)] \\ &= -\frac{1}{2} \int d^4x \phi_1 [\square^2 + V''(\rho)] \phi_1 + \dots,\end{aligned}\quad (10.42)$$

where the omitted terms involve higher powers of ϕ_1 . By (10.27), we have

$$\begin{aligned}\langle x | i\Delta^{-1}[\rho] | y \rangle &= \left. \frac{\delta^2 S[\phi]}{\delta \phi_1(x) \delta \phi_2(y)} \right|_{\phi_1=0} \\ &= -[\square^2 + V''(\rho)] \delta^4(x - y).\end{aligned}\quad (10.43)$$

Hence

$$\Omega^{-1} \langle k | i\Delta^{-1}(\rho) | k \rangle = k^2 - V''(\rho). \quad (10.44)$$

Substituting this into (10.41), we obtain

$$U(\rho) = V(\rho) + \frac{\hbar}{2} \int \frac{d^4k_E}{(2\pi)^4} \ln [k_E^2 + V''(\rho)] + O(\hbar^2), \quad (10.45)$$

where we have rotated into Euclidean momentum space. The integral above is divergent. We cut it off at $k_E^2 = \Lambda^2$, getting

$$\int d^4 k_E \ln(k_E^2 + V'') = \pi^2 \left[\Lambda^4 \left(\ln \Lambda - \frac{1}{4} \right) + \Lambda^2 V'' \right. \\ \left. + \frac{1}{2} (V'')^2 \left(\ln \frac{V''}{\Lambda^2} - \frac{1}{2} \right) \right] + O\left(\frac{1}{\Lambda^2}\right). \quad (10.46)$$

Using this result, we have

$$U(\rho) = V(\rho) + \hbar V_1(\rho) + O(\hbar^2), \quad (10.47)$$

where

$$V_1(\rho) = \frac{\Lambda^2}{32\pi^2} V''(\rho) + \frac{[V''(\rho)]^2}{64\pi^2} \left[-\frac{1}{2} + \ln \frac{V''(\rho)}{\Lambda^2} \right]. \quad (10.48)$$

We have dropped a constant of the order of $\Lambda^4 \ln \Lambda$, but this has no effect on the location of the minimum of $U(\rho)$.

To separate $V_1(\rho)$ into its divergent and convergent parts, we have to introduce an arbitrary scale parameter μ , so that we can write

$$\ln \frac{V''}{\Lambda^2} = \ln \frac{V''}{\mu^2} + \ln \frac{\mu^2}{\Lambda^2}.$$

With this, we have

$$V_1(\rho) = \frac{1}{32\pi^2} \left[\Lambda^2 V'' - \frac{1}{2} (V'')^2 \ln \frac{\Lambda^2}{\mu^2} \right] + \frac{(V'')^2}{64\pi^2} \left(-\frac{1}{2} + \ln \frac{V''}{\mu^2} \right). \quad (10.49)$$

The first term is divergent and the second term is convergent.

10.6 Renormalization

1 General scheme

The divergences in $V_1(\rho)$ have to be eliminated through renormalization. For this to be possible, the divergent terms must have the same forms as those present in the Lagrangian, so that they can be absorbed. Put another way, the theory is renormalizable if the counter terms needed to cancel the divergences are of the same forms as terms present in the Lagrangian.

For a general $V(\rho)$ that is a polynomial of degree n , $V''(\rho)$ is a polynomial of degree $n - 2$, and the divergent part of $V_1(\rho)$ is a polynomial of degree $2n - 4$. Thus, the theory is renormalizable if $2n - 4 \leq n$, or $n \leq 4$. This condition is satisfied by our choice of $V(\rho)$.

To display the counter terms, we rewrite the parameters in $\mathcal{L}(x)$ in the following forms:

$$\alpha_0 = \alpha_1 + \delta\alpha, \\ m_0^2 = m_1^2 + \delta m^2, \quad (10.50)$$

where $\delta\alpha$ and δm^2 may be divergent in perturbation theory, but α_1 and m_1^2 are required to be finite parameters. Since $\delta\alpha$ and δm^2 arise only in quantum theory, we assume for the purpose of the loop expansion that

$$\delta\alpha \sim \delta m^2 \sim O(\hbar). \quad (10.51)$$

We now rewrite

$$\begin{aligned} \mathcal{L}(x) &= \frac{1}{2} \partial_\mu \phi \partial^\mu \phi + V(\phi) + \frac{\delta m^2}{4} \phi^2 - \frac{\delta\alpha}{4!} \phi^4, \\ V(\phi) &= \frac{\alpha_1}{4!} \phi^4 - \frac{m_1^2}{4} \phi^2. \end{aligned} \quad (10.52)$$

In the corresponding classical theory, the vacuum value of the field is given by

$$\rho_0 = m_1(3/\alpha_1)^{1/2}. \quad (10.53)$$

The last two terms in $\mathcal{L}(x)$ are the counter terms^a. Since they are of order \hbar , they may be simply added to the one-loop effective potential. Thus we have

$$U(\rho) = V(\rho) + \hbar V_1(\rho) + \frac{\delta\alpha}{4!} \rho^4 - \frac{\delta m^2}{4} \rho^2. \quad (10.54)$$

We choose $\delta\alpha$ and δm^2 so as to cancel the divergent part of $V_1(\rho)$:

$$\begin{aligned} \delta\alpha &= \frac{3\alpha_1^2 \hbar}{32\pi^2} \ln \frac{\Lambda^2}{\mu^2}, \\ \delta m^2 &= \frac{\alpha_1 \hbar}{32\pi^2} \left(\Lambda^2 + \frac{m_1^2}{2} \ln \frac{\Lambda^2}{\mu^2} \right). \end{aligned} \quad (10.55)$$

The right side of these equations are of course ambiguous up to additive finite terms, but they can always be absorbed into the scale parameter μ . Thus we have

$$U(\rho) = V(\rho) + \frac{\hbar}{64\pi^2} [V''(\rho)]^2 \left[-\frac{1}{2} + \ln \frac{V''(\rho)}{\mu^2} \right]. \quad (10.56)$$

The minimum occurs at $\rho = \langle\phi\rangle$, where

$$U'(\langle\phi\rangle) = 0. \quad (10.57)$$

We write

$$\langle\phi\rangle = \rho_0 + \rho_1, \quad (10.58)$$

where ρ_0 is defined by (10.53). The renormalized mass m and the renormalized coupling constant α are defined respectively by

^a We have left out wave function renormalization, whereby the kinetic term is replaced by $\frac{1}{2}Z\partial_\mu\phi\partial^\mu\phi$. This provides a counter term needed to render finite the overall normalization of the full propagator. The graphs to be cancelled by the counter term come from the momentum-dependent part of the proper self-energy, and contain at least two closed loops. Hence $Z = 1 + O(\hbar^2)$.

$$\begin{aligned} m^2 &= U''(\rho_0 + \rho_1), \\ \alpha &= U'''(\rho_0 + \rho_1). \end{aligned} \quad (10.59)$$

For reference we record formulas for the derivatives of U :

$$\begin{aligned} U' &= V' + kV''V''' \ln(V''/\mu^2), \\ U'' &= V'' + k[(V'')^2 + \alpha_1 V''] \ln(V''/\mu^2) + (V'')^2, \\ U''' &= V''' + k[3\alpha_1 V''' \ln(V''/\mu^2) + [(V'')^3/V''] + 3\alpha_1 V'''], \\ U'''' &= \alpha_1 + \alpha_1^2 \{3 \ln(V''/\mu^2) + (6\alpha_1 \rho^2/V'') - [\alpha_1^2 \rho^4/(V'')^2] + 3\}, \end{aligned} \quad (10.60)$$

where

$$k = \hbar/32\pi^2, \quad (10.61)$$

and

$$\begin{aligned} V' &= \frac{\alpha_1}{6} \rho^3 - \frac{m_1^2}{2} \rho, \\ V'' &= \frac{\alpha_1}{2} \rho^2 - \frac{m_1^2}{2}, \\ V''' &= \alpha_1 \rho, \\ V'''' &= \alpha_1. \end{aligned} \quad (10.62)$$

2 Massive case

Suppose $m_1 > 0$. Then we may assume $\rho_1 \ll \rho_0$. Keeping ρ_1/ρ_0 only to first order, we obtain (setting $\hbar = 1$)

$$\frac{\rho_1}{\rho_0} = -\frac{\alpha_1}{32\pi^2} \ln \frac{m_1^2}{\mu^2} + O(\alpha_1^2). \quad (10.63)$$

Using this, we obtain

$$\begin{aligned} \langle \phi \rangle &= m_1 \left(\frac{3}{\alpha_1} \right)^{1/2} \left[1 - \frac{\alpha_1}{32\pi^2} \ln \frac{m_1^2}{\mu^2} + O(\alpha_1^2) \right], \\ m^2 &= m_1^2 \left[1 + \frac{3\alpha_1}{32\pi^2} \left(1 + \frac{1}{3} \ln \frac{m_1^2}{\mu^2} \right) + O(\alpha_1^2) \right], \\ \alpha &= \alpha_1 \left[1 + \frac{3\alpha_1}{32\pi^2} \left(1 + \ln \frac{m_1^2}{\mu^2} \right) + O(\alpha_1^2) \right]. \end{aligned} \quad (10.64)$$

3 Massless case

For the case $m_1 = 0$, there is no intrinsic mass scale, and the coupling constant must be defined at a floating renormalization point μ . From (10.60) we find

$$U'''(\rho) = \alpha + \frac{\alpha^2 \hbar}{64\pi^2} \left(11 + 3 \ln \frac{\alpha \rho^2}{2\mu^2} \right). \quad (10.65)$$

Let us define a new scale parameter μ_0 by

$$U'''(\mu_0) = \alpha. \quad (10.66)$$

Its value is given by

$$\ln \frac{\alpha \mu_0^2}{2\mu^2} = -\frac{11}{3}. \quad (10.67)$$

We now define the running coupling constant as

$$\alpha(\rho) = U'''(\rho) = \alpha + \frac{3\hbar\alpha^2}{32\pi^2} \ln \frac{\rho^2}{\mu_0^2}. \quad (10.68)$$

Thus, the parameter α is the value of the running coupling constant at $\rho = \mu_0$. Rewriting (10.68) in the form

$$\frac{1}{\alpha(\rho)} = \frac{1}{\alpha(\mu_0)} - \frac{3}{32\pi^2} \ln \frac{\rho^2}{\mu_0^2}, \quad (10.69)$$

we recognize that this is just (9.69), re-derived by a different method.

The really interesting question in the massless case is whether spontaneous symmetry breaking can occur. In other words, can $\langle\phi\rangle$ be non-zero, purely due to radiative corrections? If the answer is yes, we would have a model of dynamical mass generation, and that would be physically interesting.

From (10.60) we have

$$U'(\rho) = \frac{\rho^4}{4!} \left[\alpha + \frac{3\hbar\alpha^2}{32\pi^2} \left(\ln \frac{\rho^2}{\mu^2} + \text{const.} \right) \right]. \quad (10.70)$$

It appears that there is a root $\langle\phi\rangle$, satisfying

$$\hbar\alpha \ln \frac{\langle\phi\rangle^2}{\mu^2} = -\frac{32\pi^2}{3}. \quad (10.71)$$

However, this requires that a quantity of order \hbar in the loop expansion be equated with a quantity of order 1. Hence this root must be rejected, since it lies beyond the validity of the approximation used.

The technical reason for the failure of (10.71) is that the two terms of (10.70) are of order α and α^2 respectively, and thus cannot cancel each other within the region of validity of the formula. However, if we could modify the model in such a manner that α and α^2 are replaced by two independent coupling constants, then we might have a cancellation, which would then lead to spontaneous symmetry breaking by radiative corrections. Such a model can indeed be found, the simplest example being “massless” scalar electrodynamics.

10.7 Dimensional Transmutation

Coleman and E. Weinberg⁴ showed that, in the renormalized version of “massless” scalar electrodynamics, the scalar field has an *arbitrary* vacuum expectation value $\langle\phi\rangle$, even though classically $\langle\phi\rangle = 0$. The consequences are dramatic: $\langle\phi\rangle$ emerges as a spontaneous generated physical mass scale, and both the scalar and the vector particle develop dynamic masses proportional to $\langle\phi\rangle$. Thus, there is actually no such thing as massless scalar electrodynamics. They named this phenomenon “dimensional transmutation”.

We shall merely give a sketch of the derivation of the Coleman-Weinberg result. The Lagrangian density is

$$\mathcal{L} = -\frac{1}{4} F^{\mu\nu} F_{\mu\nu} + [(\partial_\mu - ieA_\mu)\phi^*][(\partial^\mu + ieA^\mu)\phi] - \frac{\lambda_0}{6} (\phi^*\phi)^2. \quad (10.72)$$

The self-coupling λ_0 is necessary to renormalize the scalar-scalar scattering amplitude. We shall immediately transform away the phase of ϕ by going to unitary gauge, and discard the phase as being irrelevant to our purpose. Thus we work with the Lagrangian density

$$\mathcal{L} = -\frac{1}{4} F^{\mu\nu} F_{\mu\nu} + \frac{1}{2} e^2 A_\mu A^\mu \phi^2 + \frac{1}{2} \partial_\mu \phi \partial^\mu \phi - \frac{\lambda_0}{4!} \phi^4, \quad (10.73)$$

where we have rescaled ϕ by a factor of $2^{-1/2}$, in order to facilitate comparison with (10.1). The vector field is subject to the subsidiary condition

$$\partial_\mu A^\mu = 0. \quad (10.74)$$

We take e to be a renormalized coupling constant, ignoring counter terms for its renormalization. The coupling constant λ_0 is a bare coupling constant and is to be renormalized.

The action of the theory is

$$S[A, \phi] = \int d^4x \left[\frac{1}{2} A^\mu (\square^2 + e^2 \phi^2) A_\mu - \frac{1}{2} \phi \square^2 \phi - \frac{\lambda_0}{4!} \phi^4 \right]. \quad (10.75)$$

We shall integrate out the vector fields in the path integral of the theory, to obtain an action for ϕ alone:

$$\exp iS[\phi] = \mathcal{N} \int (DA) \delta[\partial_\mu A^\mu] \exp i S[A, \phi]. \quad (10.76)$$

Each component of the vector field integrated out contributes a factor

$$\begin{aligned} [\det(\square^2 + e^2 \phi^2)]^{-1/2} &= \exp[-\frac{1}{2} \ln \det(\square^2 + e^2 \phi^2)] \\ &= \exp[-\frac{1}{2} \text{Tr} \ln(\square^2 + e^2 \phi^2)]. \end{aligned} \quad (10.77)$$

⁴ S. Coleman and E. Weinberg, *Phys. Rev.* D7, 1888 (1973).

Since there are three independent components, we have

$$S[\phi] = - \int d^4x \left(\frac{1}{2} \phi \square^2 \phi + \frac{\lambda_0}{4!} \phi^4 \right) - \frac{3}{2} \text{Tr} \ln(\square^2 + e^2 \phi^2). \quad (10.78)$$

Thus, the effective potential for ϕ , to lowest order in λ_0 , is given by

$$U(\rho) = \frac{\lambda_0}{4!} \rho^4 + \frac{3}{2} \int \frac{d^4 k_E}{(2\pi)^4} \ln(k_E^2 + e^2 \rho^2). \quad (10.79)$$

The second term corresponds to a sum of all one-loop graphs with external ϕ lines, with only closed vector boson loops. The scalar loops are regarded as higher order effects.

We can now take over the development following (10.45). The renormalized effective potential (up to an additive constant) can be read off (10.56):

$$U(\rho) = \rho^4 \left[\frac{\lambda}{4!} + \frac{3e^4}{64\pi^2} \left(-\frac{1}{2} + \ln \frac{e^2 \rho^2}{\mu^2} \right) \right], \quad (10.80)$$

where λ is the running coupling constant at the renormalization point μ . Even though the e^4 term is a one-loop result, while the λ term is a zero-loop result, the two terms can be comparable, because λ and e are independent coupling constants.

Differentiating (10.80), we have

$$U'(\rho) = 4\rho^3 \left(\frac{\lambda}{4!} + \frac{3e^4}{64\pi^2} \ln \frac{e^2 \rho^2}{\mu^2} \right), \quad (10.81)$$

which has a non-zero root $\langle\phi\rangle$, satisfying

$$\frac{3e^4}{64\pi^2} \ln \frac{e^2 \langle\phi\rangle^2}{\mu^2} = -\frac{\lambda}{4!}. \quad (10.82)$$

This relation holds within the validity of our approximation. We can use it to eliminate μ from the theory. The mass scale is then set by $\langle\phi\rangle$. Like any renormalized parameter, $\langle\phi\rangle$ is a free parameter of the theory. The effective potential can now be written as

$$U(\rho) = \frac{3e^4 \rho^4}{64\pi^2} \left(-\frac{1}{2} + \ln \frac{\rho^2}{\langle\phi\rangle^2} \right). \quad (10.83)$$

The mass of the scalar particle is then given by

$$m_S^2 \equiv U''(\langle\phi\rangle) = \frac{3^4}{8\pi^2} \langle\phi\rangle^2. \quad (10.84)$$

The vector fields develops mass through the Higgs mechanism in the usual way:

$$m_V^2 = e^2 \langle\phi\rangle^2. \quad (10.85)$$

There is nothing unusual about the emergence of a mass scale in renormalization, even when the theory has no intrinsic mass parameter. As we know, in order for the unrenormalized theory to make sense, we have to introduce a cutoff. The cutoff momentum represents a hidden scale of the theory. Upon renormalization, this hidden scale is replaced by an arbitrary finite scale, in the form of a floating renormalization point. In a “massless” theory, one might expect that all renormalization points are equivalent. What is remarkable here is that there is a preferred renormalization point, by virtue of the fact that the particles develop mass dynamically. The term “dimensional transmutation” refers to the emergence of a *preferred* renormalization point in a theory without intrinsic mass scale: the infinite cutoff momentum has been “transmuted” into physical mass.

It is widely believed that dimensional transmutation takes place in quantum chromodynamics with massless quarks. The reasoning is as follows. The “massless” quantum chromodynamics should give a fair account of the observed mass spectrum of the ordinary hadrons, which are bound states of the nearly massless quarks u and d . Since there is no intrinsic mass scale in that theory, dimensional transmutation must take place.

10.8 A Non-Relativistic Example

An instructive example of dimensional transmutation in non-relativistic quantum mechanics is given by Thorn⁵. Consider the two-dimensional Schrödinger equation with an attractive δ -function potential:

$$[-\nabla^2 - \lambda_0 \delta^2(\mathbf{x})]\psi(\mathbf{x}) = E\psi(\mathbf{x}). \quad (10.86)$$

By dimensional analysis, λ_0 is dimensionless. Thus the Hamiltonian does not contain an intrinsic energy scale. Nevertheless, it is possible for the system to have a bound state. To see this, let $\phi(\mathbf{k})$ be the Fourier transform of $\psi(\mathbf{x})$:

$$\psi(\mathbf{x}) = \int \frac{d^2k}{(2\pi)^2} e^{i\mathbf{k}\cdot\mathbf{x}} \phi(\mathbf{k}). \quad (10.87)$$

By Fourier analyzing both sides of (10.86), we have

$$(k^2 + B)\phi(\mathbf{k}) = \lambda_0\psi(0), \quad B = -E. \quad (10.88)$$

The solution is

$$\phi(\mathbf{k}) = \frac{\lambda_0\psi(0)}{k^2 + B}. \quad (10.89)$$

Integrating both sides over \mathbf{k} , we obtain the eigenvalue condition for the binding energy B :

$$1 = \frac{\lambda_0}{4\pi^2} \int \frac{d^2k}{k^2 + B}. \quad (10.90)$$

⁵ C. Thorn, *Phys. Rev. D* **19**, 639 (1979), Eq. (4.11).

The integral on the right-hand side is divergent. Introducing a cutoff at $|\mathbf{k}| = \Lambda$, we obtain

$$1 = \frac{\lambda_0}{4\pi} \ln \left(\frac{\Lambda^2}{B} + 1 \right). \quad (10.91)$$

Thus, for large Λ , we have

$$B = \Lambda^2 e^{-4\pi/\lambda_0}. \quad (10.92)$$

If λ_0 is considered to be an unrenormalized coupling constant, then we can demand that it depends on Λ in such a way that B remains finite as $\Lambda \rightarrow \infty$. The infinite cutoff is hereby transmuted into an *arbitrary* binding energy B .

The binding energy now fixes the energy scale of the renormalized system. Consider, for example, the scattering solution

$$\phi(\mathbf{k}) = 4\pi^2 \delta^2(\mathbf{k} - \mathbf{k}_0) + \frac{\lambda_0 \psi(0)}{k^2 - k_0^2 - i\varepsilon} \quad (10.93)$$

Integrating both sides over \mathbf{k} , we have the condition

$$\psi(0) = 1 + \frac{\lambda_0 \psi(0)}{4\pi^2} \int \frac{d^2 k}{k^2 - k_0^2 - i\varepsilon}. \quad (10.94)$$

Again introducing the cutoff Λ , we find

$$\psi(0) = \left[1 - \frac{\lambda_0}{4\pi} \ln \left(\frac{\Lambda^2}{-k^2 - i\varepsilon} + 1 \right) \right]^{-1}. \quad (10.95)$$

Eliminating Λ with the help of (10.92), we obtain

$$\lambda_0 \psi(0) = -4\pi \left[\ln \frac{-k_0^2 - i\varepsilon}{B} \right]^{-1}. \quad (10.96)$$

Thus, the renormalized wave function is

$$\phi(\mathbf{k}) = 4\pi^2 \delta^2(\mathbf{k} - \mathbf{k}_0) - \frac{1}{\pi} \left[(k^2 - k_0^2 + i\varepsilon) \ln \left(\frac{-k_0^2 - i\varepsilon}{B} \right) \right]^{-1}, \quad (10.97)$$

from which we can obtain the total scattering cross section at energy $E = k_0^2$:

$$\sigma_{\text{tot}}(E) = \frac{64}{\pi\sqrt{E}} \left[\left(\ln \frac{E}{B} \right)^2 + \pi^2 \right]^{-1}. \quad (10.98)$$

Naive dimensional analysis would lead us to expect $\sigma_{\text{tot}} \propto E^{-1/2}$; but actually $\sigma_{\text{tot}} \propto E^{-1/2} (\ln E)^{-2}$ at high energies.

The nice things about this model is that we can understand the meaning of the cutoff Λ in a physical way. Let us consider the δ -function potential to be the limit of a suitably chosen square well:

$$[-\nabla^2 + V(r)]\psi(\mathbf{x}) = E\psi(\mathbf{x}) \quad (10.99)$$

where $r = |\mathbf{x}|$, and

$$V(r) = \begin{cases} -\lambda_0/\pi a^2 & (r < a), \\ 0 & (r > a). \end{cases} \quad (10.100)$$

There is a bound state in the square well, with binding energy

$$B = \frac{c_0}{a^2} e^{-4\pi/\lambda_0} \quad (c_0 = 4e^{-2\gamma+1/2}), \quad (10.101)$$

where $\gamma = 0.5772 \dots$ is Euler's constant. To maintain a bound state at fixed binding energy B as $a \rightarrow 0$, the well depth must increase according to

$$\frac{\lambda_0}{\pi a^2} = \frac{4}{a^2} \left(\ln \frac{c_0}{a^2 B} \right)^{-1}. \quad (10.102)$$

Thus, the parameter λ_0 must approach zero, so that the well depth increases more slowly than a^{-2} . Otherwise, the binding energy would diverge, and the Hamiltonian would not be bounded from below. The cutoff Λ in the previous treatment can be identified as

$$\Lambda = \frac{\sqrt{c_0}}{a}. \quad (10.103)$$

10.9 Application to Weinberg-Salam Model

If the Higgs field in the Weinberg-Salam model is taken to be a dynamical field, then the Higgs potential will have radiative corrections. Write the Higgs potential in the Weinberg-Salam Lagrangian density in the form

$$V(\rho) = \frac{\lambda_0}{4!} \rho^4 + \frac{m_0^2}{2} \rho^2 + \text{const.}, \quad (10.104)$$

where the parameter m_0^2 is allowed to have either sign. [We are in unitary gauge, and the field ρ is $\sqrt{2}$ times that appearing in (6.32)]. We shall include radiative corrections coming from all one-loop graphs containing only gauge vector boson loops. The effective potential can be obtained immediately from (10.80):

$$U(\rho) = \frac{\lambda}{4!} \rho^4 + \frac{m^2}{2} \rho^2 + \frac{3\rho^4}{64\pi^2} \sum_V e_V^4 \left(-\frac{1}{2} + \ln \frac{e_V^2 \rho^2}{\mu^2} \right), \quad (10.105)$$

where the sum over V extends over all vector bosons coupled to the Higgs field, namely, W^+ , W^- , and Z . Each vector boson develops mass through the Higgs mechanism according to

$$m_V^2 = e_V^2 \langle \phi \rangle^2. \quad (10.106)$$

This relation can be used to eliminate the coupling constant e_V . We can also re-define μ to absorb all ρ^4 terms in the effective potential. Thus we write

$$U(\rho) = \frac{m^2}{2} \rho^2 + K\rho^4 \ln \frac{\rho^2}{\mu^2},$$

$$K = \frac{3}{64\pi^2 \langle \phi \rangle^4} (2m_W^4 + m_Z^4). \quad (10.107)$$

The derivatives of $U(\rho)$ are given by

$$U'(\rho) = m^2\rho + 4K\rho^3 \left(\frac{1}{2} + \ln \frac{\rho^2}{\mu^2} \right),$$

$$U''(\rho) = m^2\rho + 12K\rho^2 \left(\frac{7}{6} + \ln \frac{\rho^2}{\mu^2} \right). \quad (10.18)$$

If radiative corrections are ignored, we must take $m^2 < 0$ in order to have spontaneous symmetry breaking. When radiative corrections are taken into account, spontaneous symmetry breaking occurs even if $m^2 = 0$, and the Higgs field mass is given by the generalization of (10.84). Now, imagine that we increase m^2 from zero. For sufficiently small positive m^2 , we would still have spontaneous symmetry breaking, although the Higgs mass should become smaller. If we keep increasing m^2 , there will come a point when spontaneous symmetry breaking disappears, and at this point the Higgs mass will have its smallest possible value. The qualitative features of $U(\rho)$ are shown in Fig. 10.4, for different values of m^2 .

From (10.108) we find that $U'(\rho)$ has a non-zero root at $\langle \phi \rangle$, satisfying

$$\langle \phi \rangle^2 \left(\frac{1}{2} + \ln \frac{\langle \phi \rangle^2}{\mu^2} \right) = -\frac{m^2}{4K}. \quad (10.109)$$

At this root we have

$$U(\langle \phi \rangle) = -K\langle \phi \rangle^4 \left(1 + \ln \frac{\langle \phi \rangle^2}{\mu^2} \right),$$

$$m_H^2 = U''(\langle \phi \rangle) = 8K\langle \phi \rangle^2 \left(\frac{3}{2} + \ln \frac{\langle \phi \rangle^2}{\mu^2} \right), \quad (10.110)$$

where m_H is the mass of the Higgs boson. It is clear that $\langle \phi \rangle$ will cease to be the lowest minimum if $U(\langle \phi \rangle)$ becomes positive. Hence we must have $\ln \langle \phi \rangle^2 / \mu^2 \geq -1$. This gives a theoretical lower bound to the Higgs mass:⁶

$$m_H^2 \geq \frac{3}{16\pi^2 \langle \phi \rangle^2} (2m_W^4 + m_Z^4). \quad (10.111)$$

⁶ S. Weinberg, *Phys. Rev. Lett.* **36**, 294 (1976).

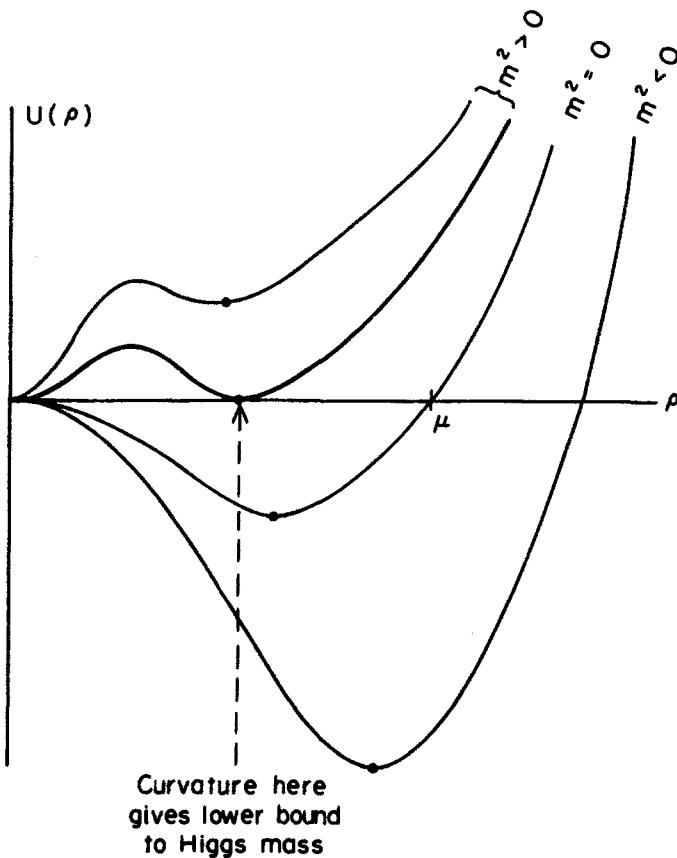


Fig. 10.2 One-loop effective potential in Weinberg-Salam model including only effects of closed gauge boson loops. The potentials are shown for various choices of the mass parameter m^2 in the potential term in the Lagrangian.

From (6.28) and (6.36) we have

$$m_W^2 = \frac{\pi\alpha\langle\phi\rangle^2}{\sin^2\theta_W} \quad (\alpha \approx 1/137), \quad (10.112)$$

$m_Z/m_W = \sin\theta_W.$

Hence

$$m_H \geq \frac{\sqrt{3}\alpha\langle\phi\rangle}{4\sin^2\theta_W} (2 + \sec^4\theta_W)^{1/2}. \quad (10.113)$$

Using $\langle\phi\rangle = 247 \text{ GeV}/c^2$, $\sin^2\theta_W = 0.22$, we obtain

$$m_H \geq 6.8 \text{ GeV}/c^2. \quad (10.114)$$

CHAPTER 11

THE AXIAL ANOMALY

11.1 Origin of the Axial Anomaly

We are familiar with the fact that, for a free Dirac particle, the left and right chiral states L and R decouple from each other in the massless limit [see (6.5)]. Since the electromagnetic interaction does not couple L to R [see (6.3)], one might expect that such a decoupling also takes place for a charged Dirac particle. This is of course true if one starts with a massless theory; but if one starts with a massive theory, then L and R remain coupled to each other in the massless limit¹. The reason is that, even in the massless limit, a charged Dirac particle of given *helicity* can make a transition into a virtual state of the opposite helicity, by emitting a real photon. This effect, which we shall discuss in detail later, is the physical origin of the axial anomaly.

The masslessness of a Dirac field theory is formally expressed by the invariance of the Lagrangian density under a chiral transformation

$$\psi(x) \rightarrow e^{-i\omega\gamma_5}\psi(x). \quad (11.1)$$

According to Noether's theorem, chiral invariance implies the existence of a conserved axial-vector current, usually taken to be the gauge-invariant chiral current^a

$$j_5^\mu = \bar{\psi}\gamma^\mu\gamma_5\psi. \quad (11.2)$$

Using the equations of motion for Heisenberg fields, one can derive the formal operator identity

$$\partial_\mu j_5^\mu = 2mj_5, \quad (11.3)$$

where m is the mass, and j_5 is the chiral density:

$$j_5 = i\bar{\psi}\gamma_5\psi. \quad (11.4)$$

^a Definitions of currents do not include the factors representing the unrenormalized coupling constants. We remind the reader that $\gamma_5 = -i\gamma^0\gamma^1\gamma^2\gamma^3$, and $\epsilon^{0123} = -\epsilon_{0123} = 1$. These differ by a sign from the convention used by Adler (ref. 2).

¹ T. D. Lee and M. Nauenberg, *Phys. Rev.* **133B**, 1549 (1964). These authors also point out that massless spinor electrodynamics is a pathological theory: the S -matrix elements of the theory are infinite, due to "mass singularities". To cancel these singularities, one has to define the S -matrix elements with respect to appropriately chosen statistical ensembles of degenerate states (rather than pure states). For this reason, a massless Dirac theory should always be considered to be the limit of a massive theory.

Thus, one expects $\partial_\mu j_5^\mu \rightarrow 0$ when $m \rightarrow 0$. However, this is false. As we shall show later, a more careful analysis gives instead the “anomalous” result

$$\partial_\mu j_5^\mu = 2mj_5 + \frac{\alpha_0}{2\pi} \tilde{F}^{\mu\nu} F_{\mu\nu}, \quad (11.5)$$

where $\tilde{F}^{\mu\nu} = \frac{1}{2}\epsilon^{\mu\nu\rho\beta}F_{\rho\beta}$, and α_0 is the unrenormalized fine-structure constant. [The operators in (11.5) are unrenormalized field operators]. Thus, j_5^μ is not conserved in the massless limit. This fact does not contradict Noether’s theorem, for $\tilde{F}^{\mu\nu} F_{\mu\nu}$ is the 4-divergence of a vector, and hence it is possible to define a new axial vector current that is conserved. Unfortunately, the new current is not gauge invariant, and hence cannot be coupled to physical fields. (We expand on this point later). The last term in (11.5) is the axial anomaly. Its presence reflects a conflict between gauge invariance and chiral invariance.

Why should there be a difference between the results of formal and “careful” reasoning? In our discussion of vacuum polarization in Chapter 9, we learned that currents like $j^\mu = \bar{\psi}\gamma^\mu\psi$ and $j_5^\mu = \bar{\psi}\gamma^\mu\gamma_5\psi$ are singular operators, due to the fact that ψ is coupled to its canonical conjugate ψ^\dagger at the same space-time point. As a consequence, formal statements about currents may be ambiguous. For example, through formal use of the equations of motion, one could seemingly show that all matrix elements of $\partial_\mu j^\mu$ vanish; but what one really gets is an indeterminate quantity like $\infty - \infty$. To obtain unambiguous results, the theory must be made well-defined, through extra requirements (such as the condition $k_\mu \Pi^{\mu\nu} = 0$ for the vacuum polarization tensor). The conflict between gauge invariance and chiral invariance arises from the fact that no physically acceptable requirement can be found that would maintain $\partial_\mu j^\mu = 0$ and $\partial_\mu j_5^\mu = 0$ simultaneously. The axial anomaly arises when one insists upon gauge invariance, through the definition $\partial_\mu j^\mu \equiv 0$. The justification for the latter ultimately rests with experiments.

The manipulation of singular operators is a delicate matter. Special tools have been devised for that purpose, including such items as “point splitting method”, “Schwinger term”, “seagull term”, and “T* product”. We avoid their use to keep things simple, and refer the interested reader to the reviews by Adler² and Jackiw.³

11.2 The Triangle Graph

The most elementary manifestation of the axial anomaly is the triangle graph: a closed fermion loop with one axial-vector vertex and two vector vertices, as shown in Fig. 11.1. The two graphs there differ only in the labelling of the external photons, and are collectively referred to as “the triangle graph”. Adler⁴ and Bell and Jackiw⁵ were the first to discuss the anomaly arising therefrom, hence also the name “Adler-Bell-Jackiw anomaly”.

² S. Adler, in *Lectures on Elementary Particles and Quantum Field Theory*, eds. S. Deser, M. Grisaru, and H. Pendleton (MIT Press, Cambridge, 1970).

³ R. Jackiw, in S. B. Treiman, R. Jackiw, and D. J. Gross, *Lectures on Current Algebra and Its Applications* (Princeton University Press, Princeton, 1972).

⁴ S. Adler, *Phys. Rev.* **177**, 2426 (1969).

⁵ J. S. Bell and R. Jackiw, *N. Cimento* **60A**, 47 (1969).

The triangle graph can occur in different physical theories. In quantum electrodynamics, it is the lowest-order description of the creation of two photons by an external axial-vector source. In the Weinberg-Salam model, it gives the lowest-order amplitude for the virtual process $Z \rightarrow 2\gamma$, when one sums over all possible internal fermion loops. (The amplitude of the corresponding real process vanishes, because a spin 1 state cannot decay into two real photons⁶). In both cases, the graph is given by the same mathematical expression. The difference lies in the values of the coupling constants, which give an overall multiplicative factor, and the external propagators or wave functions that one chooses to attach to the graph.

Omitting all coupling constants and external propagators, we denote the triangle graph by

$$t_{\alpha\beta\mu}(k_1, k_2) = s_{\alpha\beta\mu}(k_1, k_2) + s_{\beta\alpha\mu}(k_2, k_1), \quad (11.6)$$

where

$$s_{\alpha\beta\mu}(k_1, k_2)$$

$$= -i \int \frac{d^4 p}{(2\pi)^4} \text{Tr} \left(\frac{1}{p + k_1 - m} \gamma_\alpha \frac{1}{p - m} \gamma_\beta \frac{1}{p - k_2 - m} \gamma_\mu \gamma_5 \right). \quad (11.7)$$

This is a linearly divergent integral; but, as we shall see later, the sum of the two terms in (11.6) is convergent.

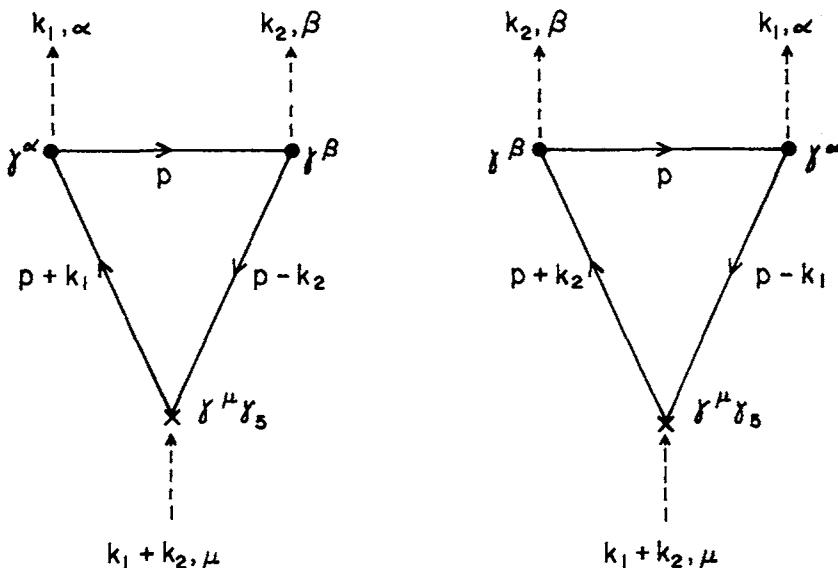


Fig. 11.1 The triangle graph

⁶ C. N. Yang, *Phys. Rev.* **77**, 242 (1950); L. D. Landau, *Dokl. Akad. Nauk. (USSR)* **60**, 207 (1948).

Gauge invariance requires that

$$k_1^\alpha t_{\alpha\beta\mu}(k_1, k_2) = k_2^\beta t_{\alpha\beta\mu}(k_1, k_2) = 0. \quad (11.8)$$

The naive statement (11.3) leads us to expect

$$(k_1 + k_2)^\mu t_{\alpha\beta\mu}(k_1, k_2) = 2m v_{\alpha\beta}(k_1, k_2), \quad (11.9)$$

where

$$\begin{aligned} v_{\alpha\beta}(k_1, k_2) &= \int \frac{d^4 p}{(2\pi)^4} \text{Tr} \left(\frac{1}{p + k_1 - m} \gamma_\alpha \frac{1}{p - m} \gamma_\beta \frac{1}{p - k_2 - m} \gamma_5 \right) \\ &\quad + (k_1 \leftrightarrow k_2, \alpha \leftrightarrow \beta). \end{aligned} \quad (11.10)$$

We shall now check these statements.

Contracting (11.7) with $(k_1 + k_2)^\mu$, we have

$$\begin{aligned} (k_1 + k_2)^\mu s_{\alpha\beta\mu}(k_1, k_2) &= \int \frac{d^4 p}{(2\pi)^4} \text{Tr} \left[\frac{1}{p + k_1 - m} \gamma_\alpha \frac{1}{p - m} \gamma_\beta \frac{1}{p - k_2 - m} (k_1 + k_2) \gamma_5 \right]. \end{aligned} \quad (11.11)$$

The trace above can be rewritten as

$$\begin{aligned} \text{Tr} \left(\frac{1}{p + k_1 - m} \gamma_\alpha \frac{1}{p - m} \gamma_\beta \gamma_5 \right) &+ \text{Tr} \left(\gamma_\alpha \frac{1}{p - m} \gamma_\beta \frac{1}{p - k_2 - m} \gamma_5 \right) \\ &+ 2m \text{Tr} \left(\frac{1}{p + k_1 - m} \gamma_\alpha \frac{1}{p - m} \gamma_\beta \frac{1}{p - k_2 - m} \gamma_5 \right). \end{aligned} \quad (11.12)$$

This should give a pseudotensor when the p^μ integration is performed. Thus, the first two terms vanish upon integration, because each depends on only one external momentum, out of which we cannot construct a pseudotensor. The last term survives, and verifies (11.9). Hence naive chiral invariance is respected in the limit $m \rightarrow 0$.

In a similar fashion we find

$$\begin{aligned} k_1^\alpha t_{\alpha\beta\mu}(k_1, k_2) &= i \int \frac{d^4 p}{(2\pi)^4} \text{Tr} \left(\frac{1}{p + k_1 - m} \gamma_\beta \frac{1}{p - k_2 - m} \gamma_\mu \gamma_5 \right. \\ &\quad \left. - \frac{1}{p + k_2 - m} \gamma_\beta \frac{1}{p - k_1 - m} \gamma_\mu \gamma_5 \right). \end{aligned} \quad (11.13)$$

Let

$$F_{\beta\mu}(p) = \text{Tr} \left(\frac{1}{p - m} \gamma_\beta \frac{1}{p - k_1 - k_2 - m} \gamma_\mu \gamma_5 \right). \quad (11.14)$$

Then

$$k_1^\alpha t_{\alpha\beta\mu}(k_1, k_2) = i \int \frac{d^4 p}{(2\pi)^4} \left[F_{\beta\mu}(p + k_1) - F_{\beta\mu}(p + k_2) \right]. \quad (11.15)$$

This would vanish, if the two terms on the right-hand side were at worst logarithmically divergent, for then we could make independent shifts of the integration variables. But the two terms are in fact linearly divergent. We note that

$$\int d^4 p F_{\beta\mu}(p + a) = \int d^4 p \left[F_{\beta\mu}(p) + a^\lambda \frac{\partial F_{\beta\mu}(p)}{\partial p^\lambda} + \dots \right], \quad (11.16)$$

where the omitted terms give vanishing surface integrals; but the second term does not vanish, because asymptotically $F_{\beta\mu} \sim p^{-3}$. Thus,

$$k_1^\alpha t_{\alpha\beta\mu}(k_1, k_2) = \frac{i}{(2\pi)^4} (k_2 - k_1)^\lambda \int d^4 p \frac{\partial F_{\beta\mu}(p)}{\partial p^\lambda}. \quad (11.17)$$

The integral above can be calculated by transforming to Euclidean momentum space:

$$\begin{aligned} I &= a^\lambda b^\beta c^\mu \int d^4 p \frac{\partial F_{\beta\mu}(p)}{\partial p^\lambda} = i a_E^\lambda b_E^\beta c_E^\mu \int d^4 p_E \frac{\partial F_{\beta\mu}(p_E)}{\partial p_E^\lambda} \\ &= i a_E^\lambda b_E^\beta c_E^\mu \int dS_E^\lambda F_{\beta\mu}(p_E), \end{aligned} \quad (11.18)$$

where

$$\begin{aligned} dS_E^\lambda &= \frac{p_E^\lambda}{p_E} dS_E, \\ dS_E &= p_E^3 d\Omega, \\ \int d\Omega &= 2\pi^2. \end{aligned} \quad (11.19)$$

We shall need only the asymptotic form of $F_{\beta\mu}(p_E)$:

$$b_E^\beta c_E^\mu F_{\beta\mu}(p_E) \xrightarrow[p_E \rightarrow \infty]{} \frac{4i}{p_E^4} b_E^\beta c_E^\mu p_E^\alpha (k_1 + k_2)_E^\nu \epsilon_{\alpha\beta\mu\nu}. \quad (11.20)$$

Thus,

$$I = -4a_E^\lambda b_E^\beta c_E^\mu (k_1 + k_2)_E^\nu \epsilon_{\alpha\beta\mu\nu} \int d\Omega \frac{p_E^\lambda p_E^\alpha}{p_E^2}. \quad (11.21)$$

Noting that

$$\int d\Omega \frac{p_E^\lambda p_E^\alpha}{p_E^2} = \frac{1}{4} \delta_{\lambda\alpha} \int d\Omega = \frac{\pi^2}{2} \delta_{\lambda\alpha},$$

we have

$$I = -2\pi^2 a_E^\alpha b_E^\beta c_E^\mu (k_1 + k_2)_E^\nu \epsilon_{\alpha\beta\mu\nu}. \quad (11.22)$$

Using this in (11.17), and transforming back to Minkowski space, we finally obtain

$$k_1^\alpha t_{\alpha\beta\mu}(k_1, k_2) = -\frac{i}{4\pi^2} k_1^\alpha k_2^\nu \epsilon_{\alpha\beta\mu\nu}. \quad (11.23)$$

Similarly,

$$k_2^\beta t_{\alpha\beta\mu}(k_1, k_2) = \frac{i}{4\pi^2} k_1^\alpha k_2^\nu \epsilon_{\alpha\beta\mu\nu}. \quad (11.24)$$

We see that gauge invariance is violated.

As in the case of vacuum polarization, we can enforce gauge invariance by subtracting a suitable polynomial in the external momenta (see Appendix, Chapter 9). It is clear that the following redefined amplitude satisfies the requirements of gauge invariance:

$$T_{\alpha\beta\mu}(k_1, k_2) = t_{\alpha\beta\mu}(k_1, k_2) - \frac{i}{4\pi^2} (k_1 - k_2)^\nu \epsilon_{\alpha\beta\mu\nu}. \quad (11.25)$$

We take this to be the correct expression of the triangle graph.

With (11.25), and noting (11.9), we find

$$(k_1 + k_2)^\mu T_{\alpha\beta\mu}(k_1, k_2) = 2m\nu_{\alpha\beta}(k_1, k_2) + \frac{i}{2\pi^2} k_1^\mu k_2^\nu \epsilon_{\alpha\beta\mu\nu}, \quad (11.26)$$

which does not vanish in the limit $m \rightarrow 0$. This shows that chiral invariance is violated by the triangle graph, when gauge invariance is enforced. The last term in (11.26) is the axial anomaly.

Note that $T_{\alpha\beta\mu}$ is symmetric under the interchange of the two external photons (Bose symmetry), and that the anomaly is independent of m .

The triangle graph depends only on two independent momenta k_1, k_2 . Let

$$q^\mu = (k_1 + k_2)^\mu. \quad (11.27)$$

By Lorentz invariance and Bose symmetry, $T_{\alpha\beta\mu}$ must be an invariant function times a pseudotensor, which can only be one of the following:

$$\begin{aligned} & \epsilon_{\alpha\beta\lambda\nu} k_1^\lambda k_2^\nu q_\mu, \\ & \epsilon_{\alpha\beta\lambda\nu} k_1^\lambda k_2^\nu (k_1 - k_2)^\mu (k_1^2 - k_2^2), \\ & \epsilon_{\alpha\beta\lambda\nu} (k_1 - k_2)^\nu, \\ & (\epsilon_{\alpha\mu\lambda\sigma} k_{1\beta} - \epsilon_{\beta\mu\lambda\sigma} k_{2\alpha}) k_1^\lambda k_2^\sigma, \\ & (\epsilon_{\alpha\mu\lambda\sigma} k_{2\beta} - \epsilon_{\beta\mu\lambda\sigma} k_{1\alpha}) k_1^\lambda k_2^\sigma. \end{aligned} \quad (11.28)$$

All but the first two are ruled out by gauge invariance.

We shall consider only the case $k_1^2 = k_2^2$, so that only the first case in (11.28) survives:

$$T_{\alpha\beta\mu}(k_1, k_2) = i\epsilon_{\alpha\beta\lambda\nu} k_1^\lambda k_2^\nu q_\mu R(q^2). \quad (11.29)$$

The fact that (11.23) is finite shows, through (11.25), that R is finite. By similar arguments we can write

$$\nu_{\alpha\beta}(k_1, k_2) = i\epsilon_{\alpha\beta\lambda\nu} k_1^\lambda k_2^\nu S(q^2). \quad (11.30)$$

Thus, (11.26) gives the relation

$$q^2 R(q^2) = 2mS(q^2) + \frac{1}{2\pi^2}. \quad (11.31)$$

In the limit $m \rightarrow 0$, we must have

$$q^2 R(q^2) \xrightarrow[m \rightarrow 0]{} \frac{1}{2\pi^2}. \quad (11.32)$$

Therefore

$$T_{\alpha\beta\mu}(k_1, k_2) \xrightarrow[m \rightarrow 0]{} \frac{i}{2\pi^2} \epsilon_{\alpha\beta\lambda\nu} k_1^\lambda k_2^\nu \frac{q_\mu}{q^2 + i\epsilon}. \quad (11.33)$$

In the massless limit, the entire contribution to the triangle graph comes from the anomaly, which is manifested as a pole at $q^2 = 0$ (the “anomaly pole”). We shall give a physical explanation of this result later. Note that, according to (11.31), the anomaly pole at $q^2 = 0$ is always present, regardless of m . But it is below the physical threshold $q^2 = 4m^2$, and touches the physical region only when $m \rightarrow 0$.

The second-order radiative corrections to the triangle graph are represented by the Feynman graphs shown in Fig. 11.2. Adler and Bardeen⁷ have shown that their net contribution to the axial anomaly vanishes. They argue that higher radiative corrections are also absent. This leads to the remarkable result that the last term in (11.26) represents the exact anomaly, to all orders of perturbation theory.

The argument for the absence of radiative corrections is as follows. Any radiative correction to the basic triangle graph must contain at least one internal photon line, and two extra vertices on the basic triangular loop. If one regulates

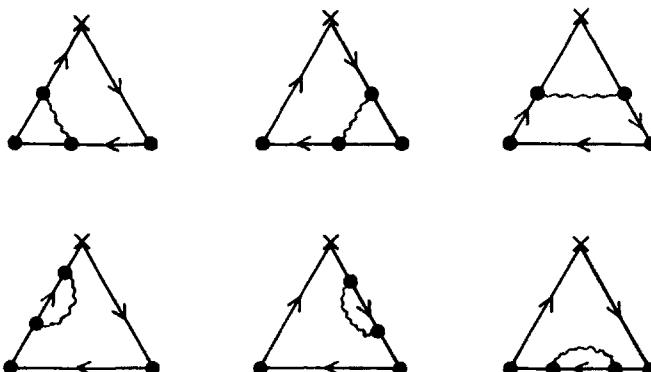


Fig. 11.2 Second-order radiative corrections to the triangle graph. Their anomalies have been shown to cancel one another.

⁷ S. L. Adler and W. A. Bardeen, *Phys. Rev.* **182**, 1517 (1969). See also A. Zee, *Phys. Rev. Lett.* **29**, 1198 (1972).

the **photon** propagator (thus preserving chiral invariance at the expense of gauge invariance), then one would have a finite integral, there being at least two extra electron propagators along the basic triangular loop to make the integral converge. Therefore, no anomaly occurs, because one could freely shift integration variables. The only graph without internal photon lines is the basic triangle graph, which is solely responsible for the axial anomaly.

11.3 Anomalous Divergence of the Chiral Current

We can deduce (11.5) from (11.26) as follows. The gauge-invariant amplitude $T_{\alpha\beta\mu}$ may be expressed as the matrix element of the unrenormalized Heisenberg operator j_5^μ between the physical vacuum state and a physical two-photon state:

$$e_0^2 T_{\alpha\beta\mu}(k_1, k_2) = \langle k_1, \alpha; k_2; \beta | j_{5\mu}(0) | 0 \rangle, \quad (11.34)$$

where it is understood that the right-hand side is calculated only to the lowest order in e_0^2 . Correspondingly, we have

$$-ie_0^2(k_1 + k_2)^\mu T_{\alpha\beta\mu}(k_1, k_2) = \langle k_1, \alpha; k_2, \beta | \partial^\mu j_{5\mu}(0) | 0 \rangle. \quad (11.35)$$

The result (11.26) can be reproduced by taking

$$\partial^\mu j_{5\mu}(x) = 2mj_5(x) + \frac{\alpha_0}{2\pi} \tilde{F}^{\mu\nu}(x) F_{\mu\nu}(x), \quad (11.36)$$

where

$$\begin{aligned} F^{\mu\nu} &= \partial^\mu A^\nu - \partial^\nu A^\mu, \\ \tilde{F}^{\mu\nu} &= \frac{1}{2} \epsilon^{\mu\nu\alpha\beta} F_{\alpha\beta}, \\ \alpha_0 &= e_0^2/4\pi. \end{aligned} \quad (11.37)$$

The unrenormalized fine-structure constant α_0 in (11.36) is replaced by the renormalized one, if all operators there are replaced by renormalized operators. Absence of radiative corrections to (11.26) would mean that (11.36) is an operator identity, valid to all orders of perturbation theory.

Noting that

$$\begin{aligned} \tilde{F}^{\mu\nu} F_{\mu\nu} &= \partial^\mu X_\mu, \\ X_\mu &= 2\epsilon_{\mu\alpha\beta\gamma} A^\alpha \partial^\beta A^\gamma, \end{aligned} \quad (11.38)$$

we can define a new axial vector current

$$J_5^\mu \equiv \bar{\psi} \gamma^\mu \gamma_5 \psi - \frac{\alpha_0}{2\pi} X^\mu, \quad (11.39)$$

which satisfies

$$\partial_\mu J_5^\mu = 2mj_5, \quad (11.40)$$

and is therefore conserved in the limit $m \rightarrow 0$. The new current is not gauge invariant, and hence cannot be coupled to physical fields. However, the corresponding charge Q_5 , which by virtue of (11.38) is a constant of motion in the limit $m \rightarrow 0$, is gauge invariant:

$$Q_5 = \int d^3x J_5^0 = \int d^3x \left(\psi^\dagger \gamma_5 \psi - \frac{\alpha_0}{\pi} \mathbf{A} \cdot \nabla \times \mathbf{A} \right). \quad (11.41)$$

A gauge transformation $\mathbf{A} \rightarrow \mathbf{A} + \nabla \chi$ gives rise to a surface integral that vanishes. Thus Q_5 can be a physical quantity. In fact,

$$[Q_5, \psi(x)] = -\gamma_5 \psi(x), \quad (11.42)$$

which shows that Q_5 is the generator of infinitesimal chiral transformations:

$$\psi \rightarrow (1 - i\omega \gamma_5) \psi = \psi + i\omega [Q_5, \psi]. \quad (11.43)$$

We mention in passing that the above situation changes if A^μ is replaced by a non-Abelian gauge field $A_a{}^\mu$. In that case (11.36) is generalized, with $\tilde{F}^{\mu\nu} F_{\mu\nu}$ replaced by $\tilde{F}_a{}^{\mu\nu} F_{a\mu\nu}$.⁸ The quantity X_μ in (11.38) is replaced by (5.6). If we continue to use (11.39), then Q_5 is not gauge invariant under “large” gauge transformations (see Sec. 8.6), owing to the existence of the topological charge. Physical consequences of this fact will be discussed in Sec. 12.5. A deviation of the anomaly via functional integrals is given in Sec. 12.6.

11.4 Physical Explanation of the Axial Anomaly

To understand the origin of the axial anomaly physically, we consider the process:

Axial-vector source \rightarrow Two real photons,

of which the triangle graph gives the lowest order description. We shall examine its “absorptive part”, which is obtained by “cutting” the graph as shown in Fig. 11.3. This means that we replace the cut electron propagators with their imaginary parts, which are δ -functions that force the electrons to go on their mass shell. An electron propagating in the opposite sense to the arrow on the propagator is defined to be a positron. Thus, the absorptive part is a product of two amplitudes describing a succession of two physical processes:

1. Source $\rightarrow e^+ + e^-$,
2. $e^+ + e^- \rightarrow \gamma + \gamma$.

⁸ W. A. Bardeen, *Phys. Rev.* **184**, 1848 (1969).

⁹ A. D. Dolgov and V. I. Zakharov, *Nucl. Phys.* **B27**, 525 (1971); J. Hořejši, *J. Phys. G: Nucl. Phys.* **L7** (1986).

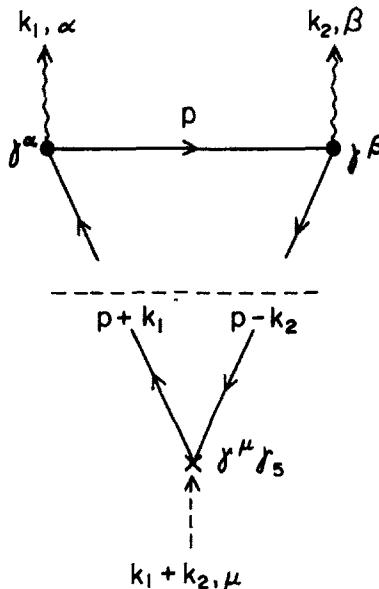


Fig. 11.3 The absorptive part of the triangle graph is obtained by “cutting” the graph.

The triangle graph can be obtained from its absorptive part through a dispersion relation¹⁰. The absorptive part determines the imaginary part of the invariant function $R(q^2)$ in (11.29). For complex values of the argument, $R(s)$ is given by

$$R(s) = \frac{1}{\pi} \int_{4m^2}^{\infty} ds' \frac{\text{Im } R(s')}{s' - s}, \quad (11.44)$$

plus possible “subtraction terms” which are polynomials in s . From (11.25) we see that $T_{\alpha\beta\mu}$ and $t_{\alpha\beta\mu}$ share the same absorptive part, which is gauge invariant and unambiguous. The axial anomaly comes from the fact that the absorptive part does not vanish in the limit $m \rightarrow 0$. In fact, we see from (11.32) that

$$\text{Im } R(q^2) \xrightarrow{m \rightarrow 0} -\frac{1}{2\pi} \delta(q^2). \quad (11.45)$$

Our purpose is to explain this result physically.

It seems at first sight that the absorptive part vanishes in the massless limit, by chiral invariance. The reasoning runs as follows.

¹⁰ For reference on dispersion relations and the analytic properties of Feynman graphs see R. J. Eden, P. V. Landshoff, D. I. Olive, and J. C. Polkinghorne, *The Analytic S-Matrix* (Cambridge University Press, Cambridge, England, 1966).

Go to the center-of-mass frame of the two final photons. In the initial process, the axial-vector source must make an e^+e^- pair of total spin 0, because a spin 1 state cannot subsequently go into two real photons⁶. The source can manage this via the interaction $\gamma_0\gamma_5$. Thus e^+ and e^- must have the same helicity, and hence opposite chirality in the massless limit. Since $\bar{R}\gamma_0\gamma_5L = 0$, the initial process is forbidden in the massless limit.

In the final process, the e^+e^- pair annihilates into two photons by going through a virtual intermediate state, one of which is depicted in Fig. 11.4. (Another possible intermediate state is obtained by reversing the helicity of the virtual electron; still others are obtained by replacing the outgoing virtual electron by an incoming virtual positron.) The transition at vertex B is allowed; but that at A is forbidden in the limit $m \rightarrow 0$. The reason is that vertex A calls for helicity flip, which is the same as chirality flip in the massless limit, and the electromagnetic interaction conserves chirality. For any other intermediate state, there is always one allowed and one forbidden vertex. Thus the final process is also forbidden.

However, this is not the whole story. As $m \rightarrow 0$, the virtual state in the final process approaches a real state, because in that limit the transition $e \rightarrow e + \gamma$ is no longer forbidden by energy-momentum conservation. If the electron has non-zero momentum p , the photon can be emitted only in the forward direction; but if $p \rightarrow 0$, the photon can be emitted in any direction. Thus, although the matrix elements vanish, so does the energy denominator, and they may compensate each other. The situation is not as simple as it appears at first glance: we must not take the limit $m \rightarrow 0$ too soon.

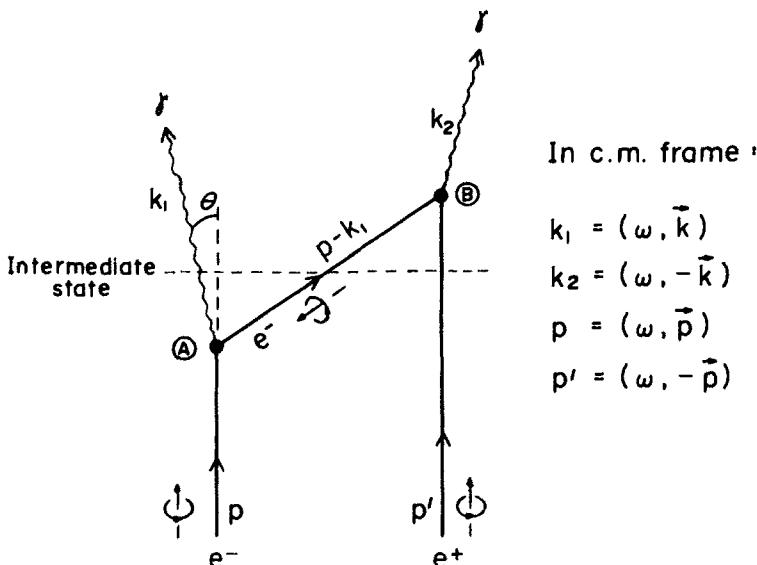


Fig. 11.4 An intermediate virtual state that contributes to the absorptive part of the triangle graph. The anomaly arises from the fact that in the massless limit the virtual state can become real, and gives a non-vanishing contribution at threshold.

We shall estimate the absorptive part through the following crude argument. The matrix element of the initial process (source $\rightarrow e^+e^-$) carries a factor m . The matrix element of the final process ($e^+e^- \rightarrow \gamma\gamma$), when summed over all possible intermediate states, is m times the Feynman propagator of the virtual electron. We multiply together the matrix elements and integrate over all directions of emission of one of the final photons. The phase-space factor from this integration is proportional to $|\mathbf{p}| = (\omega^2 - m^2)^{1/2}$. Thus we have

$$\begin{aligned} \text{Im } R(\omega, m) &\propto \frac{m^2}{\omega^2} \left(1 - \frac{m^2}{\omega^2}\right)^{1/2} \int d(\cos \theta) \frac{1}{(p - k_1)^2 - m^2}, \quad (\omega > m), \\ \text{Im } R(\omega, m) &= 0, \quad (\omega < m), \end{aligned} \quad (11.46)$$

where ω is the c. m. energy of either photon, θ is the c. m. emission angle of one of the photons, and p and k_1 are 4-vectors defined in Fig. 11.4. A factor ω^{-2} has been supplied to make the dimension come out right. This factor is unique up to a pure-number constant, because there is no other energy scale in the problem, barring m . To compare with (11.45), note that the invariant mass of the two-photon system is

$$q^2 = 4\omega^2. \quad (11.47)$$

A little algebra gives

$$(p - k_1)^2 - m^2 = -2p \cdot k_1 = -2\omega^2 \left[1 - \left(1 - \frac{m^2}{\omega^2}\right)^{1/2} \cos \theta \right]. \quad (11.48)$$

Before doing the integral, let us see whether the vanishing of the energy denominator can compensate for the vanishing of the matrix elements. Consider first the kinematic domain $m/\omega \ll 1$ and $\theta \ll 1$, where (11.48) becomes

$$-\omega^2 \left(\theta^2 + \frac{m^2}{\omega^2} \right).$$

The integrand in (11.46) becomes proportional to

$$\frac{(m/\omega)^2}{\theta^2 + (m/\omega)^2} \xrightarrow[m \rightarrow 0]{} \begin{cases} 1 (\theta = 0) \\ 0 (\theta \neq 0). \end{cases} \quad (11.49)$$

This gives zero when integrated over angles. Hence no compensation occurs here; the absorptive part vanishes as $m \rightarrow 0$, for any value of ω above threshold.

Next, consider the threshold value $\omega = m$. Here (11.48) becomes $-2m^2$, independent of θ . Thus, as $m \rightarrow 0$, the energy denominator diverges like m^{-2} , and exactly compensates for the vanishing matrix elements. We expect the absorptive part to be peaked at threshold energy, and diverge there in the massless limit (because of the extra factor ω^{-2} from dimensional analysis).

Substituting (11.48) into (11.46), we obtain

$$\text{Im } R(\omega, m) = C \frac{m^2}{\omega^4} \ln \frac{1 - (1 - m^2/\omega^2)^{1/2}}{1 + (1 - m^2/\omega^2)^{1/2}}, \quad (11.50)$$

where C is a constant pure number. A detailed calculation shows that this is in fact the right answer.⁹ In Fig. 11.5 we plot $\text{Im } R$ as a function of ω^2 , for different values of m . We can see that

$$\text{Im } R(\omega, m) \xrightarrow[m \rightarrow 0]{} \text{Const. } \delta(\omega^2). \quad (11.51)$$

This explains (11.43).

We can also understand the behavior of $\text{Im } R$ from the standpoint of analytic properties. We see from (11.31) that the anomaly is manifested through the fact that R has a fixed pole at $q^2 = 0$, with fixed residue (the anomaly pole). It touches the physical region $q^2 \geq 4m^2$ only in the limit $m \rightarrow 0$. If we increase m from 0, the pole remains intact, but the physical region moves away from it. The presence of the pole is then manifested only indirectly, by the fact that $\text{Im } R$ increases towards threshold.

11.5 Cancellation of Anomalies

In analogy with the vertex $\Gamma''(p_1, p_2)$ discussed in section 9.2, we define an axial vertex function $\Gamma_5''(p_1, p_2)$ to be the sum of all connected proper Feynman graphs (with external propagators omitted) that describe the creation of a fermion-anti-fermion pair by an external axial-vector source. Some of these graphs are shown in Fig. 11.6. We are especially interested in the last graph, which is “anomalous” in the sense that it harbors a triangle subgraph. The graph

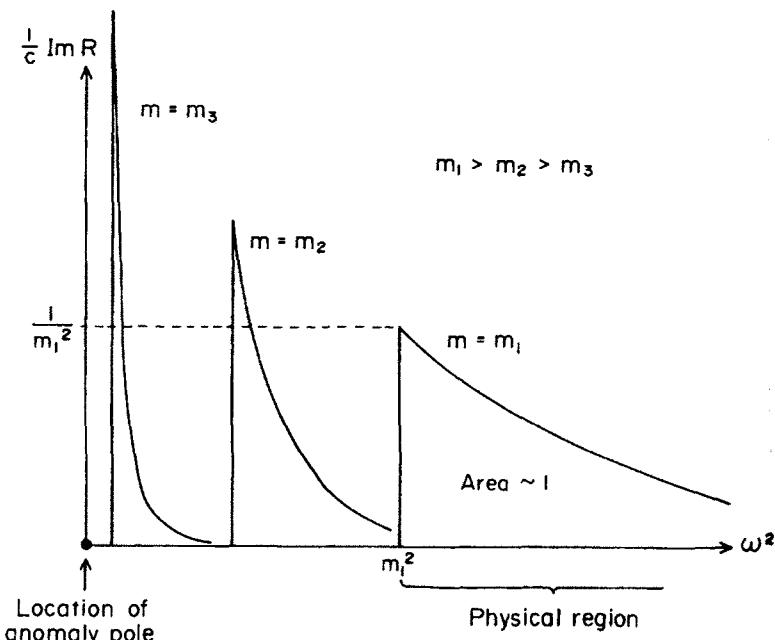


Fig. 11.5. Absorptive part of the triangle graph. The areas under the curves remain finite as $m \rightarrow 0$. Hence they approach $\delta(\omega^2)$ as $m \rightarrow 0$.

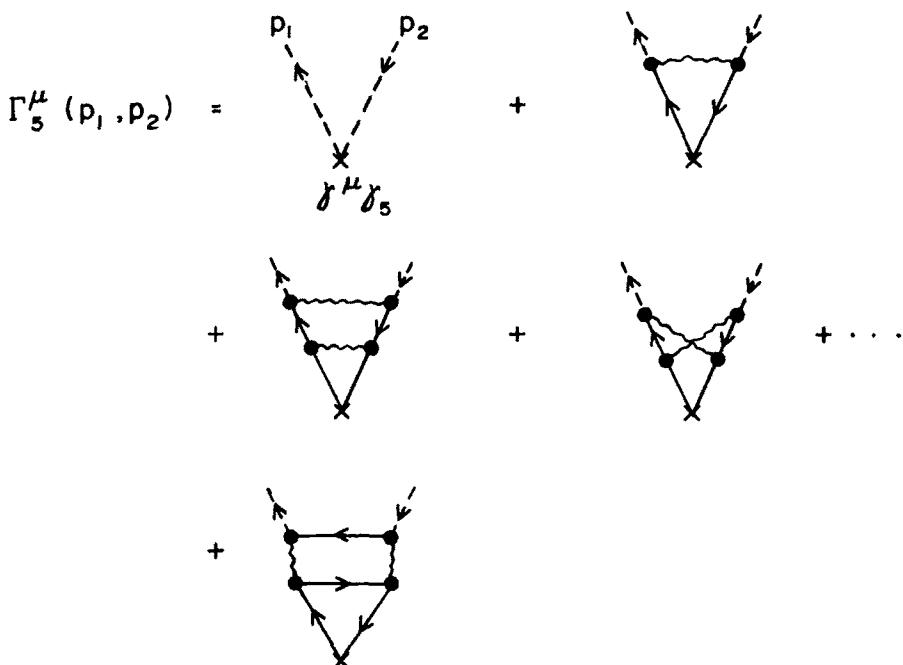


Fig. 11.6 Feynman graphs for the axial vertex function. The last graph is “anomalous”, in that it has different scaling properties from the others, leading to non-renormalizability.

is logarithmically divergent (even though the triangle subgraph is finite) because the triangle graph grows linearly with its external momenta. The divergence of the anomalous graph makes Γ_5^μ non-renormalizable.

We recall that renormalizability depends on a scaling property, such as (9.A21) in the case of Γ^μ , which enables us to subtract off the divergent part of a Feynman graph and re-express the operation as infinite rescaling. In quantum electrodynamics, the scaling property of a “normal” Feynman graph is determined by the fact that each vertex touches two electron lines and a photon line, and carries a factor e_0 . The triangle graph, however, has a *different* scaling property, as we can see indirectly by referring to (11.36). The first term there is normal, but the anomalous second term corresponds to a Feynman graph in which two photon lines are attached to a single vertex carrying a factor e_0^2 . This means that anomalous graphs in Γ_5^μ scale differently from the normal ones, and there is no renormalization procedure that could simultaneously get rid of both the divergence from anomalous graphs and that from normal graphs. Therefore Γ_5^μ is truly divergent.

It should be noted that the mass-dependent part of the triangle graph is normal, and would have been renormalizable in the usual manner all by itself. The anomalous part, which destroys renormalizability, is independent of the mass.

The non-renormalizability of Γ_5^μ is of no consequence in quantum electrodynamics, because Γ_5^μ is not a physical quantity in that theory. However, it does occur in the Weinberg-Salam model, for example in the graph shown in Fig. 11.7, which contributes to $e-\nu$ scattering. Though of high order, this graph by itself would give the disastrous prediction that the scattering cross section is infinite. But we have to sum over all possible fermion triangular loops. As we shall see below, when the contributions of all members in the first family e, ν, u, d are added together, the anomalous parts miraculously cancel one another, leaving a renormalizable mass-dependent residue. The latter is of such a high order in the coupling constants that we can ignore it in practice.

Let the possible fermions in the triangular loop be numbered by n , and let $T_{\alpha\beta\mu}^{(n)}$ denote the contribution of the n th fermion. Then the amplitude for the virtual process $Z \rightarrow \gamma\gamma$ is given by

$$M_{\alpha\beta\mu} = \sum_n Q_n^2 (Q'_R - Q'_L)_n T_{\alpha\beta\mu}^{(n)}, \quad (11.52)$$

where Q_n is the electric charge, and $Q'_{R,n}$ and $Q'_{L,n}$ the right and left-handed neutral charges of the n th fermion. These quantum numbers have been given in Table 6.2. Using (11.26), we obtain

$$\begin{aligned} q^\mu M_{\alpha\beta\mu} &= 2 \sum_n Q_n^2 (Q'_R - Q'_L)_n m_n v_{\alpha\beta}^{(n)} \\ &\quad + \frac{i}{2\pi^2} k_1^\mu k_2^\nu \epsilon_{\alpha\beta\mu\nu} \sum_n Q_n^2 (Q'_R - Q'_L)_n. \end{aligned} \quad (11.53)$$

The last term above is the anomaly, which is independent of masses. In Fig. 11.8, we show the relevant vertices for the fermions in the first family, from which follows

$$\sum_n Q_n^2 (Q'_L - Q'_R)_n = \frac{1}{\sin 2\theta_w} [0 + 1 + 3(-\frac{4}{9} + \frac{1}{3})] = 0. \quad (11.54)$$

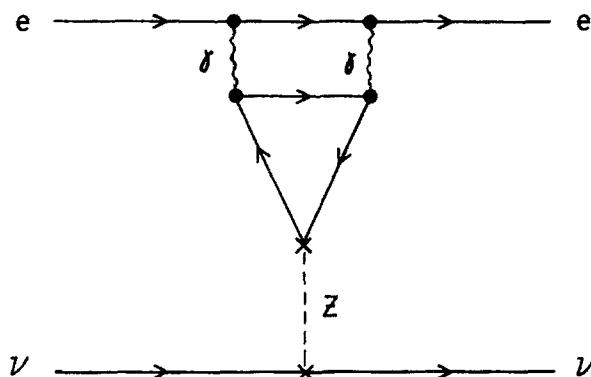


Fig. 11.7 An infinite non-renormalizable contribution to electron-neutrino scattering in the Weinberg-Salam model. Renormalizability requires cancellation of axial anomalies among all possible triangular fermion loops.

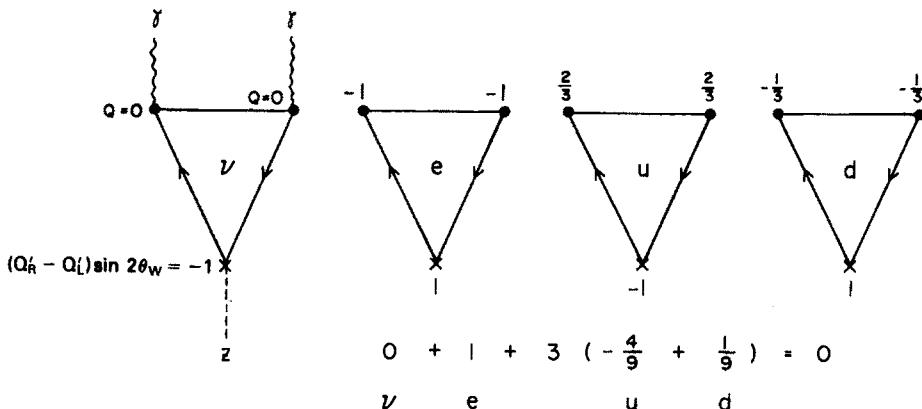


Fig. 11.8 Cancellation of axial anomalies in the virtual process $Z \rightarrow 2\gamma$ in the Weinberg-Salam model.

Note that each quark flavor must be counted thrice to take into account the color quantum number. Thus, the anomaly cancels for any value of the Weinberg angle, and $M_{\alpha\beta\mu}$ is renormalizable in the usual manner.

There are possible triangle graphs in the Weinberg-Salam model, corresponding to $Z \rightarrow W^+W^-$, $\gamma \rightarrow W^+W^-$, etc. The anomalies in all such graphs must cancel, in order that the theory be renormalizable. In these more general graphs, the fermion can change identity as it goes around the triangular loop, so that the fermion propagators along the loop in general have different masses. Again, because the anomalies are independent of the masses, their cancellation depends only on the quantum numbers of the fermions. For a particular process, the condition that the anomalies cancel is

$$\text{Tr}[(V_1 V_2 + V_2 V_1) A] = 0, \quad (11.55)$$

where V_1 , V_2 are the charge matrices that occur at the vector vertices, and A is the axial-charge matrix occurring at the axial-vector vertex. Instead of trying to verify the above case by case, we shall formulate a general rule for anomaly cancellation.

Consider any gauge theory. Let the fermion multiplet be denoted by the column vector ψ , in which the entries are Dirac spinors. The conserved currents coupled to the gauge fields are

$$j_a^\mu = \bar{\psi} \gamma^\mu t_a \psi, \quad (11.56)$$

where t_a ($a = 1, \dots, N$) are the generators of the gauge group, represented by matrices that act on ψ . If the right and left-handed components of ψ transform differently under the gauge group, then t_a will be represented by different matrices t_a^R , t_a^L , when it acts on the right- or left-handed component. We write accordingly

$$j_a^\mu = \frac{1}{2} \bar{\psi} \gamma^\mu t_a^R (1 + \gamma_5) \psi + \frac{1}{2} \bar{\psi} \gamma^\mu t_a^L (1 - \gamma_5) \psi. \quad (11.57)$$

In any triangle graph in which all three vertices involve these currents, the fermion loop must be either completely left-handed or completely right-handed, because there is no coupling between right and left. If a particular left-handed loop is allowed, so is the corresponding right-handed loop, whose contribution differs only by a sign at the $\gamma^\mu \gamma_5$ vertex. Thus, the condition that no anomalies arise from the gauge interactions is that, for all a, b, c ,

$$\text{Tr}[\{t_a^L, t_b^L\}t_c^L] - \text{Tr}[\{t_a^R, t_b^R\}t_c^R] = 0. \quad (11.58)$$

Of course, the theory may have other conserved currents (e.g., those related to ungauged global symmetries) that give rise to anomalies. But these do not make the theory unrenormalizable, just as the anomalies in Γ_5^μ have no relevance to the renormalizability of quantum electrodynamics.

We now check (11.58) for the Weinberg-Salam model. There are 4 generators t_0 and t_k ($k = 1, 2, 3$), with

$$\begin{aligned} [t_0, t_k] &= 0, \\ \text{Tr } t_0 &= \text{Tr } t_k = 0. \end{aligned} \quad (11.59)$$

Recall that

$$\begin{aligned} t_k^R &\equiv 0, \\ t_k^L &= \frac{1}{2}\tau_k, \end{aligned} \quad (11.60)$$

and that both members of a left-handed doublet have the same t_0 . It is easy to verify that for both right and left-handed representations,

$$\begin{aligned} \text{Tr}[\{t_i, t_j\}t_k] &= 0, \\ \text{Tr}[\{t_i, t_j\}t_0] &= 0, \\ \text{Tr}(t_i t_0^2) &= 0. \end{aligned} \quad (11.61)$$

The only thing left to be checked is the trace of t_0^3 . By direct calculation using Table 6.2, we find (remembering that quarks have color),

$$\text{Tr}(t_0^L)^3 = \text{Tr}(t_0^R)^3 = -\frac{2}{3}. \quad (11.62)$$

Thus, all anomalies cancel in the Weinberg-Salam model.

The sufficient conditions that enable the cancellation to occur are

- that the anomaly is independent of masses,
- that the anomaly has no radiative corrections.

The cancellations then follows purely from the quantum number assignment within the fermion family.

Since the three fermion families in the Weinberg-Salam model are identical in multiplet structure, all anomalies cancel within each family. The necessity for these cancellations to occur is the best theoretical argument we have for the standard family structure. It strongly suggests that the yet unobserved members of the third family, *i.e.* ν'' and t , should exist. More generally, the requirement that all anomalies cancel imposes a severe constraint on possible physical models.¹¹

¹¹ D. J. Gross and R. Jackiw, *Phys. Rev. D6*, 477 (1972).

11.6 't Hooft's Principle

With the proliferation of quarks and leptons, it is natural to ask whether they can be reduced to something simpler. One ordinarily thinks in a language one knows. So it is not surprising that the first thought is to make them out of other fermions—"preons", perhaps¹². The first puzzle we face is how the electron could be so light when it is so small. Suppose the electron is a bound state of massless preons confined within the electron's intrinsic radius, which has an experimental upper bound of $a = 10^{-6}$ cm. Naive intuition would lead us to expect a mass greater than $a^{-1} \sim 10^6 m_e$. If the bound-state picture is qualitatively correct, then some principle must be at work to suppress the electron mass. 't Hooft¹³ suggests that the principle is chiral invariance, as expressed in the following form:

A composite particle must reproduce the axial anomaly due to its fermionic constituents.

The principle seems obvious. We know that the electron must exhibit the axial anomaly, which in a composite picture has to arise from its more point-like constituents. However, the consequences are non-trivial, as we shall see later in detail. Briefly it works as follows. In the idealized limiting case of massless preons, the anomaly pole (which, regardless of mass, is always located at $q^2 = 0$) occurs in the physical region. If preons are confined and unobservable, then the anomaly pole must appear to be due to a physical bound state, which must therefore have zero mass. If chiral symmetry is not spontaneously broken, we identify the bound state with the electron. The actual observed mass of the electron is then attributed to the fact that preons have mass (for whatever reason). Although the principle does not enable us to calculate the electron mass, it gives a qualitative explanation for its smallness.

To discuss 't Hooft's principle in more detail, we adopt a model for the preons, hoping that the results will be more general than the model. The preons are associated with a Dirac field ψ coupled to gauge fields G_a^μ , which correspond to the generators L_a of a simple gauge group, say, $SU(N)$. The Lagrangian density is

$$\mathcal{L} = i\bar{\psi}(\gamma_\mu \partial^\mu + igL_a G_a^\mu)\psi - \kappa\bar{\psi}\psi, \quad (11.63)$$

where the limit $\kappa \rightarrow 0$ is to be taken. This model has the same structure as quantum chromodynamics (QCD), with possible differences coming only from the choice of the gauge group, and the flavor multiplet structure of the fermions. We exploit this similarity by drawing upon the folklores of QCD.

Like the quarks in QCD, the preons are assumed to be permanently confined. Thus, dimensional transmutation should take place, giving rise to a confinement scale, namely, the size of an electron. The theory should be asymptotically free, which means that preon flavors are limited. For simplicity take one flavor of preons in the fundamental representation of $SU(N)$.

¹² The name was suggested by J. C. Pati, A. Salam, and J. Strathdee, *Phys. Lett.* **59B**, 265 (1975).

¹³ G. 't Hooft, in *Recent Development in Gauge Theories*, eds. G. 't Hooft et al. (Plenum Press, New York, 1980).

In this model there are two ungauged symmetries, the “electromagnetic” $U(1)$ and the chiral $U(1)$, which we assume to be unbroken^b. They give rise to two conserved currents:

$$\begin{aligned} j^\mu &= \bar{\psi} \gamma^\mu \psi, \\ j_5^\mu &= \bar{\psi} \gamma^\mu \gamma_5 \psi. \end{aligned} \quad (11.64)$$

Although these currents are not coupled to any dynamical fields in the model, we can imagine turning on external sources that couple to them, and measuring the three-current correlation function

$$C_{\alpha\beta\mu}(x, y) = i\langle 0 | T j_\alpha(x) j_\beta(y) j_{5\mu}(0) | 0 \rangle. \quad (11.65)$$

This describes vacuum fluctuations of the system in the absence of external sources. Existence of the axial anomaly implies that there are definite correlations in the vacuum fluctuations of j_α , j_β , and $j_{5\mu}$, at different space-time points. The absorptive part of $C_{\alpha\beta\mu}$ contains information about the physical excitations of the system that can transmit signals from one correlated point to another. Thus, $C_{\alpha\beta\mu}$ is an intrinsic property of the preon system.

We assume that the short-distance behavior of $C_{\alpha\beta\mu}$ can be studied by treating the gauge interactions of the preons in perturbation theory, taking the preons to be free particles in the zeroth-order approximation. This assumption is the same as that underlying the parton picture in QCD. Lacking an understanding of confinement from first principles, we cannot really prove its validity. However, we can offer the following plausibility arguments:

1. Owing to asymptotic freedom, preons should be seen as free particles by an agent that delivers a momentum transfer much higher than that set by the confinement scale.

2. Asymptotic perturbation theory in QCD has yielded results consistent with experiments.

According to this assumption, the Fourier transform of $C_{\alpha\beta\mu}$ for asymptotically high momenta is given by the usual triangle graph, with preons in the triangular loop:

$$\begin{aligned} T_{\alpha\beta\mu}(k_1, k_2) &= \frac{iN}{2\pi^2} \epsilon_{\alpha\beta\lambda\nu} k_1^\lambda k_2^\nu \frac{q_\mu}{q^2 + i\epsilon}, \\ q &\equiv k_1 + k_2, \end{aligned} \quad (11.66)$$

where $k_1 \rightarrow \infty$, $k_2 \rightarrow \infty$. Note that the asymptotic domain includes the point $(k_1 + k_2)^2 = 0$, and hence covers the anomaly pole. A factor N is supplied, corresponding to the number of preon “colors”. Assuming that there are no radiative corrections, we take (11.66) to be valid to all orders in the gauge coupling g .

^b In this respect the model differs from QCD, where chiral symmetry is believed to be spontaneously broken, giving rise to a pseudoscalar Goldstone boson identifiable with the pion (see Sec. 12.4). We have no understanding from first principles why this should happen. Thus we cannot say what constraints must be placed on the preon model in order that the spontaneous breakdown does not happen.

We recall that (11.66) is obtained through enforcing the “electromagnetic” $U(1)$ symmetry, at the expense of chiral $U(1)$. There is no necessity for doing it here. We can add to (11.66) a polynomial in k_1, k_2 without changing its physical content, which resides in the absorptive part:

$$\text{Ab } T_{\alpha\beta\mu} \propto \delta(q^2). \quad (11.67)$$

This is non-zero, because the physical region in the asymptotic domain is $q^2 \geq 0$, the preons being massless. The point $q^2 = 0$ lies both in the asymptotic and the finite-momentum region. Thus, having obtained (11.67) in the asymptotic domain, we can continue to use it in the finite-momentum region. In fact, the amplitude $T_{\alpha\beta\mu}$ is given by (11.66) for all k_1, k_2 , because it must have the general form (11.29), by symmetry arguments.

The reasoning from here on¹⁴ relies on the assumptions that

- (a) preons are confined,
- (b) chiral symmetry is not spontaneously broken.

The confinement assumption means that, in the finite-momentum region, preons cannot exist as physical states, and hence cannot contribute to the absorptive part of $T_{\alpha\beta\mu}$. Therefore, there must be physical bound states that do. Since the absorptive is non-vanishing only at $q^2 = 0$, there must be massless physical bound states.

These massless bound states cannot be spin 0 particles, for if there were such particles in this model, they would have to be Goldstone bosons associated with the spontaneous breakdown of chiral symmetry. We have decreed that the latter does not occur.^c

The anomaly pole cannot be due to particles of spin 1 or higher, because such particles can be coupled to external sources only via effective derivative couplings. Consequently, their contributions to the absorptive part of the triangle graph vanish at threshold. But the absorptive part is non-vanishing only at threshold.

The only remaining possibility is a massless spin 1/2 bound state, which, as we know, can produce the anomaly pole. This is identified as the electron.

We can represent the matrix elements of j^μ and j_5^μ between electron states in the forms

$$\begin{aligned} \langle e_2 | j^\mu | e_1 \rangle &= g_V \bar{u}(\mathbf{p}_2, s_2) \gamma^\mu u(\mathbf{p}_1, s_1), \\ \langle e_2 | j_5^\mu | e_1 \rangle &= g_A \bar{u}(\mathbf{p}_2, s_2) \gamma^\mu \gamma_5 u(\mathbf{p}_1, s_1), \end{aligned} \quad (11.68)$$

where $u(\mathbf{p}, s)$ are Dirac spinors, and g_V and g_A are constants. In order that the electron reproduce the anomaly (11.66), we must have

$$g_V^2 g_A = N, \quad (11.69)$$

^c If chiral symmetry breaks down spontaneously, then 't Hooft's principle says something about the coupling of the associated Goldstone boson to the currents j^μ and j_5^μ . Applied to QCD, it reproduces the well-known calculation of the lifetime of the decay $\pi^0 \rightarrow 2\gamma$, which we discuss in Sec. 12.4.

¹⁴ For more detailed discussions see Y. Frishman, A. Schwimmer, T. Banks, and S. Yankielowicz, *Nucl. Phys.* B177, 157 (1981); S. Coleman and B. Grossman, *Nucl. Phys.* B203, 205 (1982).

which is a concrete result of 't Hooft's principle. In a more realistic model with multi-flavored preons, the above condition generalized to a relation between the multiplet structure constants of the preons and those of the bound states. One may then use it to see what type of preon flavor structures can give rise to bound states having the observed multiplet structure of quarks and leptons.¹⁵

For a space-time view of 't Hooft's principle, calculate the Fourier transform of (11.66) to obtain the three-current correlation function:

$$\begin{aligned} C_{\alpha\beta\mu}(x, y) &= \frac{1}{2\pi^2} \epsilon_{\alpha\beta\mu\lambda} \left[\frac{\partial}{\partial r_\lambda} \delta^4(r) \right] \frac{1}{\rho} \frac{d}{d\rho} D_F(\rho), \\ r &\equiv \frac{1}{2}(x + y), \\ \rho &\equiv x - y. \end{aligned} \quad (11.70)$$

Here $D_F(\rho)$ is the Feynman propagator for a massless particle:

$$D_F(\rho) \equiv - \int \frac{d^4 q}{(2\pi)^2} \frac{e^{-iq\cdot\rho}}{q^2 + i\epsilon}. \quad (11.71)$$

The imaginary part of $iD_F(\rho)$ vanishes outside the light cone $\rho^2 = 0$. Hence the imaginary part of the correlation function is non-zero only if $x = y$ and $x^2 = 0$. This means that there must be physical excitations traveling on the light cone to produce the correlation. They must therefore be massless excitations. Consider a ray on the light cone, as indicated by the line OP in Fig. 11.9. The short-distance domain is a neighborhood about O in which asymptotic freedom tells us that the massless excitations are free preons. If we go sufficiently far away from O along the ray OP, we eventually go beyond the confinement scale. Massless excitations in that region must be bound states—electrons.

If we extend the model by giving mass to the preons, the situation becomes complicated, and reliance on QCD folklore is not enough. The complicating circumstance is that the anomaly pole, which still lies at $q^2 = 0$, is beyond the reach of the physical region, and does not directly contribute to the absorptive part of $T_{\alpha\beta\mu}$. The latter will generally have non-vanishing values for all q^2 above threshold, as depicted qualitatively in Fig. 11.5. To determine it requires a dynamical calculation. In particular, the location of the threshold becomes a dynamical question: It could be higher than the kinematic threshold for free preon creation, because the absorptive part can vanish in that neighborhood, for dynamical reasons. Thus, we cannot derive the electron mass in any simple way.

¹⁵ S. Dimopoulos, S. Raby, and L. Susskind, *Nucl. Phys.* B173, 208 (1980).

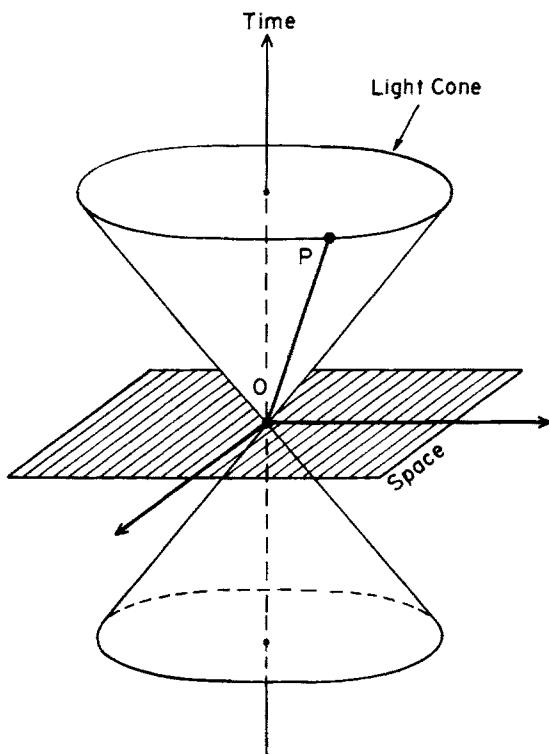


Fig. 11.9 Space-time view of 't Hooft's principle. There must exist massless excitations that move on the light cone, in order to account for the three-current correlations implied by the existence of the axial anomaly. Near O these excitations are preons. At P, beyond the confinement scale, they are massless bound states—electrons.

CHAPTER 12

QUANTUM CHROMODYNAMICS

12.1 General Properties

1 Lagrangian Density

We have discussed in Chapter 2 the physical motivation for introducing color as a quantum number for quarks. By gauging the symmetry group $[SU(3)]_{\text{color}}$, we obtain quantum chromodynamics (QCD), the currently accepted model of the strong interactions.

The 8 generators of $[SU(3)]_{\text{color}}$ are represented in the fundamental representation by the Gell-Mann matrices $\lambda_a/2$ ($a = 1, \dots, 8$) listed in Table 2.2, with the normalization

$$\text{Tr}(\lambda_a \lambda_b) = 2 \delta_{ab}. \quad (12.1)$$

The gauge fields, called “gluon fields,” are denoted by G_a^μ ($a = 1, \dots, 8$). We use the notation (see Chapter 4)

$$\begin{aligned} G^\mu(x) &= \tfrac{1}{2} \lambda_a G_a^\mu(x), \\ \mathcal{F}^{\mu\nu}(x) &= \partial^\mu G^\nu(x) - \partial^\nu G^\mu(x) + ig_0[G^\mu(x), G^\nu(x)], \end{aligned} \quad (12.2)$$

where g_0 is a dimensionless number—the unrenormalized gauge coupling constant.

The matter fields consists of spinor quark fields denoted collectively by $q(x)$, with components $q_\alpha^{fi}(x)$, where

$$\begin{aligned} i &= 1, 2, 3 \quad (\text{color index: red, yellow, green}), \\ f &= 1, \dots, 6 \quad (\text{flavor index: } u, d, c, s, t, b), \\ \alpha &= 1, \dots, 4 \quad (\text{spinor index}). \end{aligned}$$

We usually suppress the spinor index, and sometimes all indices. Instead of using the flavor index, we sometimes denote the quark fields of various flavors by their conventional names:

$$\begin{array}{ccccccc} u & & d & & c & & s \\ (\text{up}) & & (\text{down}) & & (\text{charm}) & & (\text{strange}) \end{array} \quad \begin{array}{c} t \\ (\text{top}) \end{array} \quad \begin{array}{c} b \\ (\text{bottom}) \end{array}$$

The existence of the top quark has not been confirmed experimentally, but is suggested by the internal consistency of the Weinberg-Salam model of electroweak interactions. There may be other flavors yet undiscovered.

The complete Lagrangian density of QCD is

$$\begin{aligned}\mathcal{L} &= -\frac{1}{4} \mathfrak{F}_a^{\mu\nu} \mathfrak{F}_{a\mu\nu} + \bar{q}(iD - M)q, \\ D^\mu &= \partial^\mu + ig_0 G^\mu,\end{aligned}\quad (12.3)$$

where M is a color-independent mass matrix in the flavor indices, which will be discussed in detail in Sec. 12.6. The theory is renormalizable, but we shall not present the proof.¹

2 Feynman Rules

It is generally believed that quarks and gluons are confined, i.e., colored states do not exist in the physical sector of the Hilbert space. Thus, the basis states for the S -matrix of QCD consists not of single quark or gluon states, but color-singlet states containing physical hadrons. This is a non-perturbative effect that we shall discuss in more detail in Chapter 13. For many applications, it is useful to define Feynman graphs in terms of free quark and gluon propagators.

The generating functional for connected Green's functions $W[J, K]$ is given by (see Chapter 8)

$$\begin{aligned}\exp iW[J, K] &= \mathcal{N} \int (DG)(D\bar{q})(Dq)(D\eta^*)(D\eta) \\ &\cdot \exp i \int d^4x [\mathcal{L}_{\text{eff}} + J_a^\mu G_{a\mu} + (Kq) + (\bar{q}K)],\end{aligned}\quad (12.4)$$

where $J_a^\mu(x)$ is a c-number vector source, and $K(x)$ is an anticommuting c-number spinorial source. The effective Lagrangian density \mathcal{L}_{eff} is given in a general Lorentz gauge by

$$\begin{aligned}\mathcal{L}_{\text{eff}} &= \mathcal{L} + \frac{1}{2\alpha} (\partial^\mu G_{a\mu})(\partial^\nu G_{a\nu}) \\ &+ \frac{1}{2} \eta_a^*(\delta_{ab} \square^2 + g_0 f_{abc} G_c^\mu \partial_\mu) \eta_b,\end{aligned}\quad (12.5)$$

where λ is the gauge parameter [see (8.62)], η_a^* and η_a are anticommuting c-number ghost fields, and f_{abc} are the $SU(3)$ structure constants given in Table 2.3.

We divide the effective Lagrangian density into a “free” part and an “interaction” part:

$$\mathcal{L}_{\text{eff}} = \mathcal{L}_0 + \mathcal{L}'.\quad (12.6)$$

The free Lagrangian density is taken to be

$$\begin{aligned}\mathcal{L}_0 &= \frac{1}{2} G_{a\mu} \left[g^{\mu\nu} \square^2 - \left(1 - \frac{1}{\alpha}\right) \partial^\mu \partial^\nu \right] G_{a\nu} \\ &+ i\bar{q} \gamma^\mu \partial_\mu q + \frac{1}{2} \eta_a^* \square^2 \eta_a.\end{aligned}\quad (12.7)$$

¹ G. 't Hooft, *Nucl. Phys.* **B33**, 173 (1971); **B35**, 167 (1971).

The interaction Lagrangian density is given by

$$\mathcal{L}' = \mathcal{L}'_{\text{gluon}} + \mathcal{L}'_{\text{quark}} + \mathcal{L}'_{\text{ghost}}, \quad (12.8)$$

$$\begin{aligned} \mathcal{L}'_{\text{gluon}} &= \frac{1}{2} g_0 f_{abc} (\partial^\mu G_a^\nu - \partial^\nu G_a^\mu) G_{b\mu} G_{c\nu} \\ &\quad - \frac{1}{4} g_0^2 f_{abc} f_{ab'c'} G_b^\mu G_c^\nu G_{b'\mu} G_{c'\nu}, \end{aligned} \quad (12.9)$$

$$\mathcal{L}'_{\text{quark}} = -\frac{1}{2} g_0 (\bar{q} \gamma_\mu \lambda_a q) G_a^\mu, \quad (12.10)$$

$$\mathcal{L}'_{\text{ghost}} = \frac{1}{2} g_0 f_{abc} (\eta_a^* \partial_\mu \eta_b) G_c^\mu. \quad (12.11)$$

The Feynman rules for Green's functions can be read off the above formulas, and are given in Fig. 12.1.

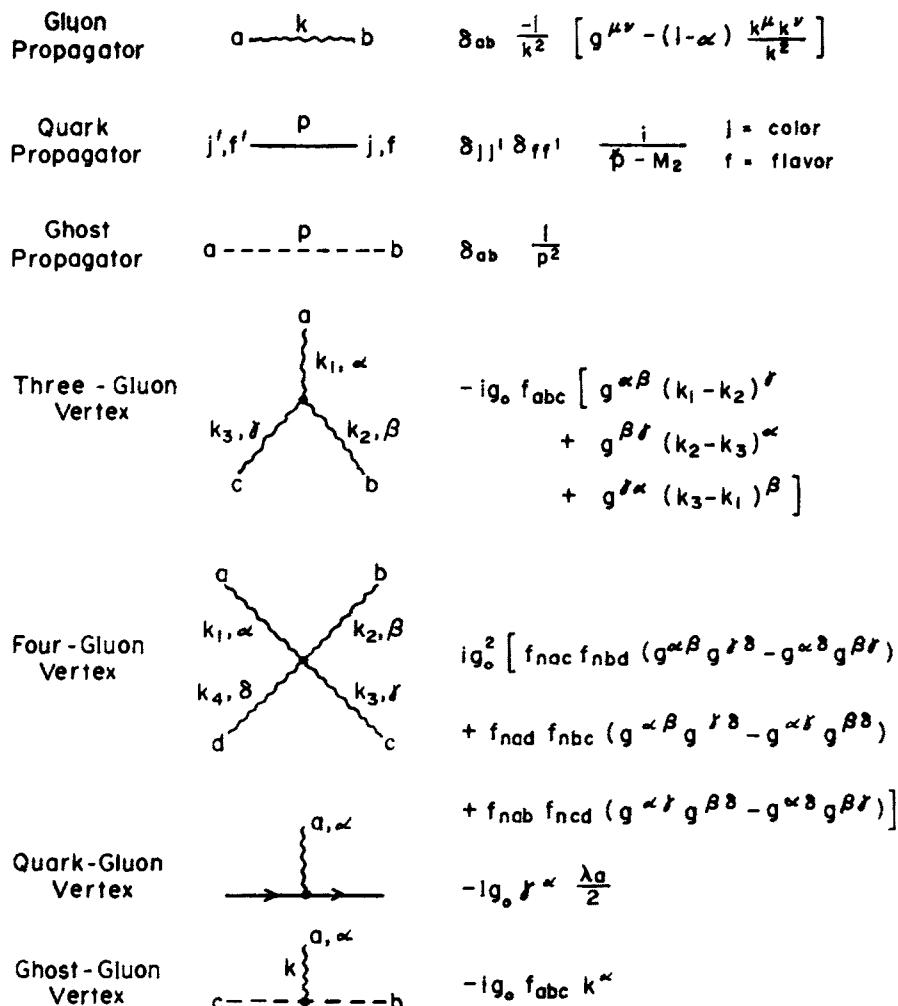


Fig. 12.1 Feynman rules for QCD. All gluon momenta flow into the vertex.

To illustrate combinatorial considerations in the derivation of the Feynman rules, we discuss the three-gluon vertex, which arises from the first term in (12.9). Suppose the three external gluon lines have the following labels:

$$\begin{array}{ll} \text{Color index:} & a, b, c, \\ \text{Momentum:} & k_1, k_2, k_3, \\ \text{Polarization index:} & \alpha, \beta, \gamma, \end{array}$$

with all momenta flowing into the graph. The first term in (12.9) can be rewritten as

$$g_0 f_{abc} (\partial^\mu G_{a'}{}^\nu) G_{b'\mu} G_{c'\nu},$$

due to the fact that f_{abc} is antisymmetric in its indices. We must pick up all terms in the above sum that contribute to the vertex graph. First consider $a' = a$. Then either $b' = b$, $c' = c$, or $b' = c$, $c' = b$. These two possibilities give the contribution

$$-ig_0 f_{abc} (k_1{}^\beta g^{\alpha\gamma} - k_1{}^\gamma g^{\alpha\beta}).$$

The complete vertex is obtained by adding two other contributions corresponding to the alternatives $a' = b$ and $a' = c$.

3 Quark-Gluon Interactions

Let us recombine the Gell-Mann matrices into the following set of matrices:

$$\begin{aligned} \lambda_3 &= \begin{pmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 0 \end{pmatrix}, & \lambda_8 &= \frac{1}{\sqrt{3}} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -2 \end{pmatrix}, \\ \tau_{12} &= \frac{1}{2}(\lambda_1 + i\lambda_2) = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \\ \tau_{13} &= \frac{1}{2}(\lambda_4 + i\lambda_5) = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \\ \tau_{23} &= \frac{1}{2}(\lambda_6 + i\lambda_7) = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}. \end{aligned} \tag{12.12}$$

The matrix τ_{ij} is a color “raising” matrix, which changes a quark of color j into one of color i . The hermitian conjugate $\tau_{ij}^\dagger = \tau_{ji}$ does the reverse. Correspond-

ingly, define

$$\begin{aligned} A^\mu &\equiv G_3^\mu, \\ B^\mu &\equiv G_8^\mu, \\ X^\mu &\equiv 2^{-1/2}(G_1^\mu + iG_2^\mu), \\ Y^\mu &\equiv 2^{-1/2}(G_4^\mu + iG_5^\mu), \\ Z^\mu &\equiv 2^{-1/2}(G_6^\mu + iG_7^\mu). \end{aligned} \quad (12.13)$$

Then

$$\begin{aligned} \mathcal{L}'_{\text{quark}} &= -\frac{g_0}{2} \bar{q}(\lambda_3 A + \lambda_8 B)q \\ &\quad - \frac{g}{\sqrt{2}} [\bar{q}(\tau_{21}X + \tau_{31}Y + \tau_{32}Z)q + \text{c.c.}], \end{aligned} \quad (12.14)$$

which shows that the quarks have two kinds of “charges”: a color ‘isotopic charge’ corresponding to an eigenvalue of $\lambda_3/2$, and a color ‘hypercharge’ corresponding to an eigenvalue of $\lambda_8/2$. The gluons X , Y , Z also carry these charges, for a quark can change color by absorbing or emitting one of these gluons, as illustrated in Fig. 12.2. The charge assignments for quarks and gluons are listed in Table 12.1.

4 Gluon Self-Interactions

Before we examine the gluon self-interactions in QCD, it is instructive to study those in a simpler case: $SU(2)$ pure-gauge theory, with gauge fields G_1^μ ,

Table 12.1 COLOR CHARGES

All charges are in units of g_0 .

Q_A = Color isotopic charge (source of A field)

Q_B = Color hypercharge (source of B field)

		Q_A	Q_B
Quarks	1 red	$\frac{1}{2}$	$\frac{1}{2\sqrt{3}}$
	2 yellow	$-\frac{1}{2}$	$-\frac{1}{2\sqrt{3}}$
	3 green	0	$-\frac{1}{\sqrt{3}}$
Charged Gluons	X	-1	0
	Y	$-\frac{1}{2}$	$-\frac{\sqrt{3}}{2}$
	Z	$\frac{1}{2}$	$-\frac{\sqrt{3}}{2}$

G_2^μ , G_3^μ , and field tensor

$$\mathcal{F}_a^{\mu\nu} = \partial^\mu G_a^\nu - \partial^\nu G_a^\mu - e\epsilon_{abc}G_b^\mu G_c^\nu \quad (a = 1, 2, 3). \quad (12.15)$$

The analog of (12.13) is

$$A^\mu \equiv G_3^\mu, \\ X^\mu \equiv 2^{-1/2}(G_1^\mu + iG_2^\mu), \quad (12.16)$$

which defines a real “photon” field A^μ , and a charged vector boson field X^μ . Writing the field tensor in terms of A^μ and X^μ , using the shorthand notation (6.51) for 4-tensors, we have

$$\begin{aligned} \mathcal{F}_1 &= \frac{1}{\sqrt{2}}(D \times X + \text{c.c.}), \\ \mathcal{F}_2 &= \frac{1}{i\sqrt{2}}(D \times X - \text{c.c.}), \\ \mathcal{F}_3 &= F + ieX^* \times X, \end{aligned} \quad (12.17)$$

where

$$\begin{aligned} F^{\mu\nu} &\equiv \partial^\mu A^\nu - \partial^\nu A^\mu, \\ D^\mu &\equiv \partial^\mu - ieA^\mu. \end{aligned} \quad (12.18)$$

Thus, the field X^μ has charge $-e$. The Lagrangian density is

$$\begin{aligned} \mathcal{L} &= -\frac{1}{4}\mathcal{F}_a^{\mu\nu}\mathcal{F}_{a\mu\nu} \\ &= -\frac{1}{4}F^{\mu\nu}F_{\mu\nu} - ieF^{\mu\nu}X_\mu^*X_\nu \\ &\quad - \frac{1}{2}(D^\mu X^\nu - D^\nu X^\mu)^*(D_\mu X_\nu - D_\nu X_\mu) \\ &\quad + \frac{1}{2}e^2[(X^* \cdot X)^2 - (X \cdot X)^*(X \cdot X)]. \end{aligned} \quad (12.19)$$

Apart from the last term, this is the Lagrangian density of massless vector electrodynamics, with an “anomalous” magnetic moment.

The last term, a quartic self-interaction of the charged field, marks the difference between massless vector electrodynamics and full $SU(2)$ gauge

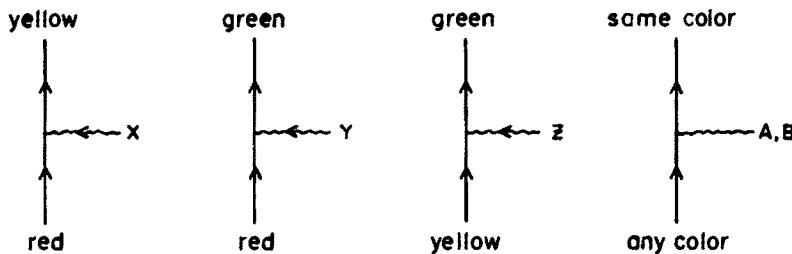


Fig. 12.2 Quark-gluon interactions

theory. Without it, the theory is invariant only under a smaller gauge group $U(1)$. It is also not renormalizable, due to the logarithmic divergences of the skeleton graphs for charge-charge scattering [Fig. 12.3(a)]. The quartic self-interaction [Fig. 12.3(b)] provides the necessary counter term to cancel these divergences, making the theory renormalizable. This interaction is the analog of $\lambda(\phi^*\phi)^2$ in scalar electrodynamics; but its coefficient is not an independent coupling constant, due to spin constraints.

In the Lagrangian density (12.19), the term

$$-ieF^{\mu\nu}X_\mu^*X_\nu \quad (12.20)$$

is not necessary for $U(1)$ gauge invariance, and appears to be a “non-minimal” interaction from the point of view of electrodynamics, giving rise to an “anomalous” magnetic moment over and above the orbital moment. If the charged vector field were massive, the resulting total magnetic moment would correspond to a gyromagnetic ratio $g = 2$ [see (6.55)]. As we shall show in the next section, that $g - 2 = 0$ is a necessary condition for a massive charged vector theory to have a massless limit. From the point of view of $SU(2)$ gauge invariance, the term (12.20) is part of the minimal interaction, for it arises from the cross product between the two terms of $\mathcal{F}_3^{\mu\nu}$ in (12.17). Thus, the value $g - 2 = 0$ is natural, and of kinematical origin (just as in the Dirac equation). Indeed, it is needed for internal consistency: if $g - 2 \neq 0$, then the vacuum polarization tensor of the A field would diverge quadratically rather than logarithmically, spoiling renormalizability.

We now return to QCD. According to (12.13), there are two “photon” fields A and B , coupled respectively to color isotopic charge and color hypercharge. There are three charged vector fields X, Y, Z . To express the gluon field tensor in terms of these new fields, it is helpful to have all the structure constants f_{abc} displayed for easy reference. This is provided by Fig. 12.4. We obtain in a straightforward manner (again using a shorthand tensor notation) the following

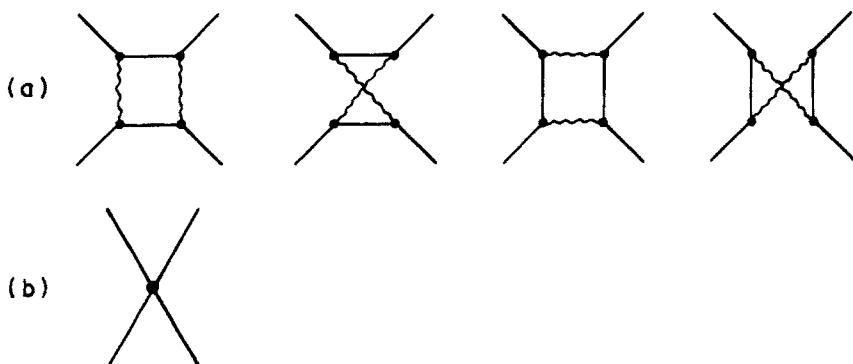


Fig. 12.3.

- (a) Logarithmically divergent skeleton graphs for the scattering of charged vector bosons through exchange of photons.
- (b) Quartic self-interaction that cancels the divergences in (a).

expressions:

$$\mathfrak{F}_3 = \partial \times A + ig_0(X^* \times X + \frac{1}{2}Y^* \times Y + \frac{1}{2}Z^* \times Z),$$

$$\mathfrak{F}_8 = \partial \times B + ig_0 \frac{\sqrt{3}}{2} \cdot (Y^* \times Y + Z^* \times Z),$$

$$\frac{1}{\sqrt{2}}(\mathfrak{F}_1 + i\mathfrak{F}_2) = D \times X + \frac{g_0}{\sqrt{2}} Y \times Z^*, \quad (12.21)$$

$$\frac{1}{\sqrt{2}}(\mathfrak{F}_4 + i\mathfrak{F}_5) = D \times Y + \frac{ig_0}{2} Z \times X,$$

$$\frac{1}{\sqrt{2}}(\mathfrak{F}_6 + i\mathfrak{F}_7) = D \times Z + \frac{g_0}{\sqrt{2}} X^* \times Y,$$

where the covariant derivatives are defined by

$$\begin{aligned} DX &= (\partial - ig_0 A)X, \\ DY &= (\partial - \frac{ig_0}{2} A - ig \frac{\sqrt{3}}{2} B)Y, \\ DZ &= (\partial + \frac{ig_0}{2} A - ig \frac{\sqrt{3}}{2} B)Z, \end{aligned} \quad (12.22)$$

from which we can read off the various charges, confirming the assignments given in Table 12.1.

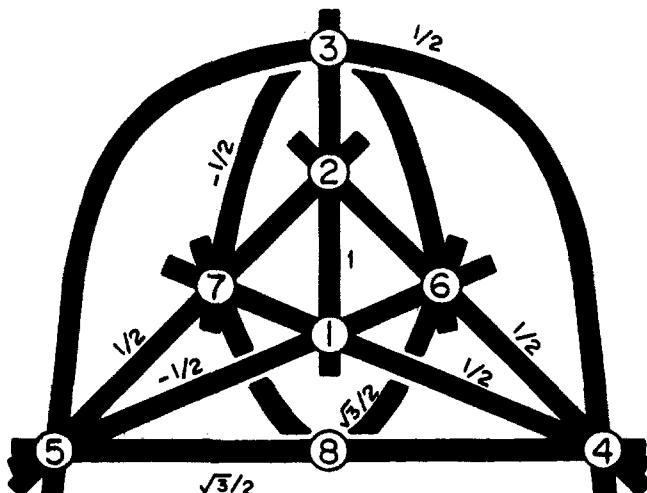


Fig. 12.4 $SU(3)$ structure constants f_{abc} . Any triplet of numbers abc joined by a black belt in the figure corresponds to a non-vanishing element f_{abc} . The number labelling the black belt is the value of f_{abc} for the ordering $a > b > c$. The value of f_{acb} etc. can be obtained by noting that f_{abc} is completely antisymmetric in its indices.

We indicate the structure of the Lagrangian density for pure QCD as follows:

$$\begin{aligned}\mathcal{L}_{\text{pure QCD}} &= -\frac{1}{4} \mathfrak{F}_a^{\mu\nu} \mathfrak{F}_{a\mu\nu} \\ &= \mathcal{L}_{\text{kin}} + \mathcal{L}_{\text{mag. mom.}} + \mathcal{L}_{\text{int}}, \\ \mathcal{L}_{\text{kin}} &= -\frac{1}{4}(\partial \times A)^2 - \frac{1}{4}(\partial \times B)^2 - \frac{1}{2}[|D \times X|^2 + |D \times Y|^2 + |D \times Z|^2], \\ \mathcal{L}_{\text{mag. mom.}} &= -ig_0(\partial^\mu A^\nu - \partial^\nu A^\mu)(X_\mu^* X_\nu + \frac{1}{2}Y_\mu^* Y_\nu + \frac{1}{2}Z_\mu^* Z_\nu) \\ &\quad -ig_0 \frac{\sqrt{3}}{2}(\partial^\mu B^\nu - \partial^\nu B^\mu)(Y_\mu^* Y_\nu + Z_\mu^* Z_\nu).\end{aligned}\tag{12.23}$$

The term \mathcal{L}_{kin} taken alone would describe a theory of three charged vector bosons X , Y , Z interacting in minimal electromagnetic fashion with two “photon” fields A and B . The term $\mathcal{L}_{\text{mag. mom.}}$ endows each of the charged vector bosons with color gyromagnetic ratio $g = 2$, with respect to both the A field and the B field. The rest of the Lagrangian density is lumped into \mathcal{L}_{int} , which we shall not write out. It contains interactions involving at least three charged vector bosons at the same point, and makes the theory renormalizable.

12.2 The Color Gyromagnetic Ratio

We shall now discuss the color gyromagnetic ratio, which was mentioned earlier in connection with the nature of gluon interactions, and which will be relevant to our discussion of asymptotic freedom in the next section.

As a thought experiment, consider the behavior of the charged gluons X , Y , Z in an external color magnetic field associated with the A or B gluon field. Ignoring all gluon interactions except those with the imposed external field, we reduce the problem to that of massless charged vector bosons interacting independently with an external Abelian magnetic field. The non-Abelian nature of QCD enters only in so far as it endows the vector bosons with charges.

We first consider a *massive* charged spinning particle, of charge e and mass m , interacting with a weak homogeneous external magnetic field \mathbf{B} . Let ζ be the average spin, defined as the expectation value of the spin operator in the rest frame of a one-particle state. The gyromagnetic ratio g is conventionally defined through the equation

$$\frac{d\zeta}{dt} = g \frac{e}{2m} \zeta \times \mathbf{B}.\tag{12.24}$$

We shall try to transform this equation to an arbitrary Lorentz frame, and then take the limit $m \rightarrow 0$. If the limit exists, then the resulting equation defines g in the massless case.

To do this, we define the polarization 4-vector S^μ by

$$S^\mu = (0, \zeta) \quad (\text{in rest frame}).\tag{12.25}$$

Let u^μ be the 4-velocity (of the center of a wave packet, whose motion can be

treated classically):

$$u^\mu \equiv \frac{1}{\sqrt{1 - v^2}} (1, \mathbf{v}), \quad (12.26)$$

where \mathbf{v} is the 3-velocity. In every Lorentz frame we have $S \cdot u = 0$, or

$$S^0 = \mathbf{S} \cdot \mathbf{v}. \quad (12.27)$$

The time derivative of this, evaluated in the rest frame, together with (12.24), give the equation of motion for S^μ in the rest frame:

$$\begin{aligned} \frac{dS^0}{dt} &= \mathbf{S} \cdot \frac{d\mathbf{v}}{dt} \quad (\text{in rest frame}), \\ \frac{d\mathbf{S}}{dt} &= g \frac{e}{2m} \mathbf{S} \times \mathbf{B} \quad (\text{in rest frame}). \end{aligned} \quad (12.28)$$

The covariant generalization of the above is²

$$\frac{dS^\mu}{d\tau} = g \frac{e}{2m} [F^{\mu\nu} S_\nu + u^\mu (S \cdot F \cdot u)] - u^\mu \left(\frac{du^\nu}{d\tau} S_\nu \right), \quad (12.29)$$

where τ is the proper time:

$$dt = \frac{d\tau}{\sqrt{1 - v^2}}. \quad (12.30)$$

The electromagnetic tensor $F^{\mu\nu}$ is related to the electric and magnetic fields by (3.28). We use the notation $(S \cdot F \cdot u) = S_\alpha F^{\alpha\beta} u_\beta$. We can verify (12.29) by noting that the right-hand side contains all possible 4-vectors that can contribute, and reduces to (12.28) in the rest frame with no electric field. In a homogeneous electromagnetic field, u^μ obey the equation of motion

$$\frac{du^\mu}{d\tau} = \frac{e}{m} F^{\mu\nu} u_\nu. \quad (12.31)$$

Substituting this into (12.29), we obtain

$$\frac{dS^\mu}{d\tau} = \frac{e}{2m} [g F^{\mu\nu} S_\nu + (g - 2) u^\mu (S \cdot F \cdot u)]. \quad (12.32)$$

In terms of the time t of a fixed observer, this reads

$$\frac{dS^\mu}{dt} = \frac{g}{2} \frac{e}{E} F^{\mu\nu} S_\nu + \frac{e}{2m} (g - 2) u^\mu (S \cdot F \cdot u), \quad (12.33)$$

where $v^\mu \equiv (1, \mathbf{v})$, and E is the energy of the particle:

$$E \equiv \frac{m}{\sqrt{1 - v^2}}. \quad (12.34)$$

² V. Bargman, L. Michel, and V. L. Telegdi, *Phys. Rev. Lett.* **2**, 435 (1959).

As $m \rightarrow 0$, (12.33) approaches a well-defined limit only if

$$g - 2 = 0, \quad (12.35)$$

which is therefore a necessary constraint for a spinning charged massless particle.

In the absence of an electric field, the spatial component of (12.32) reads

$$\frac{d\mathbf{S}}{dt} = \frac{g}{2} \frac{e}{E} \mathbf{S} \times \mathbf{B} - \frac{e}{2m} (g - 2) \mathbf{v} (\mathbf{S} \cdot \mathbf{v} \times \mathbf{B}). \quad (12.36)$$

The instantaneous spin ζ is related to \mathbf{S} through a Lorentz transformation, and can be shown to obey the equation of motion³

$$\frac{d\zeta}{dt} = \left[\frac{e}{E} + \frac{e}{2m} (g - 2) \right] \zeta \times \mathbf{B} + \frac{e}{2m} (g - 2) \frac{E}{E + m} (\mathbf{v} \cdot \mathbf{B}) \mathbf{v} \times \zeta. \quad (12.37)$$

For $g - 2 = 0$, ζ and \mathbf{S} become identical, with

$$\frac{d\mathbf{S}}{dt} = \frac{e}{E} \mathbf{S} \times \mathbf{B}, \quad (12.38)$$

which is the equation describing the spin precession of a massless charged particle.

From the above discussion, we see that the gluons X, Y, Z all have color gyromagnetic ratio $g - 2 = 0$. The quarks also have $g - 2 = 0$, by virtue of the Dirac equation. (The value of $g - 2$ here refers only to the “kinematic” value, and does not include radiative corrections.)

12.3 Asymptotic Freedom

1 The Running Coupling Constant

The running coupling constant of QCD is defined through the renormalized gluon propagator^a. The unrenormalized full propagator may be represented in the form

$$D'_{ab}^{\mu\nu}(k) = \delta_{ab} \left(g^{\mu\nu} - \frac{k^\mu k^\nu}{k^2} \right) \frac{d'(k^2/\Lambda^2, \alpha_0)}{ik^2} + (\text{gauge-dependent terms}), \quad (12.39)$$

where Λ is a cutoff momentum, and

$$\alpha_0 \equiv g_0^2 / 4\pi. \quad (12.40)$$

^a We refer to Chapter 9 for concepts not fully explained here.

³ V. B. Berestetskii, E. M. Lifshitz, and L. P. Pitaevskii, *Relativistic Quantum Theory*, Part 1 (Pergamon, Oxford, England, 1971), p. 127.

The renormalized propagator is obtained by putting

$$d'\left(\frac{k^2}{\Lambda^2}, \alpha_0\right) = Z\left(\frac{\lambda^2}{\Lambda^2}, \alpha_0\right) d\left(\frac{k^2}{\lambda^2}, \alpha_\lambda\right), \quad (12.41)$$

where d is a finite function, and

$$\alpha_\lambda = \alpha_0 Z\left(\frac{\lambda^2}{\Lambda^2}, \alpha_0\right). \quad (12.42)$$

We call this the running coupling constant corresponding to the renormalization point λ .^b The definition of d is made unique by the normalization condition

$$d(1, \alpha_\lambda) = 1. \quad (12.43)$$

From (12.41) and (12.42) we see that

$$\alpha_\lambda d\left(\frac{k^2}{\lambda^2}, \alpha_\lambda\right) = \alpha_0 d'\left(\frac{k^2}{\Lambda^2}, \alpha_0\right). \quad (12.44)$$

Thus, the left-hand side is a renormalization-group invariant:

$$\alpha_\lambda d\left(\frac{k^2}{\lambda^2}, \alpha_\lambda\right) = \alpha_\nu d\left(\frac{k^2}{\nu^2}, \alpha_\nu\right), \quad (12.45)$$

where λ and ν are two arbitrary renormalization points. Putting $\nu^2 = k^2$, we obtain

$$\alpha_k = \alpha_\lambda d\left(\frac{k^2}{\lambda^2}, \alpha_\lambda\right). \quad (12.46)$$

The β -function is defined by

$$\beta(\alpha_\lambda) = \lambda^2 \frac{d}{d\lambda^2} \left[\alpha_0 Z\left(\frac{\lambda^2}{\Lambda^2}, \alpha_0\right) \right], \quad (12.47)$$

where the right-hand side is to be re-expressed as a function of α_λ with the help of (12.42). Under a change in the renormalization point λ , α_λ changes according to

$$\frac{d\alpha_\lambda}{d \ln(\lambda^2)} = \beta(\alpha_\lambda). \quad (12.48)$$

The right-hand side of (12.44) is a power series in α_0 beginning with the power α_0^2 . Hence the β -function has a power series expansion of the form

$$\beta(\alpha) = \beta_0 \alpha^2 + \beta_1 \alpha^3 + \dots \quad (12.49)$$

Gross and Wilczek⁴, Politzer,⁵ were the first to point out that $\beta_0 < 0$, and that

^b The numbers Λ^2 and λ^2 are negative, being the squared invariant masses of Euclidean momenta. The limit $\lambda \rightarrow \infty$ is understood to mean $-\lambda^2 \rightarrow \infty$.

⁴ D. J. Gross and F. Wilczek, *Phys. Rev. Lett.* **30**, 1343 (1973); *Phys. Rev. D8*, 3633 (1973).

⁵ H. D. Politzer, *Phys. Rev. Lett.* **30**, 1346 (1973).

this implies

$$\alpha_\lambda \xrightarrow[\lambda \rightarrow \infty]{} 0. \quad (12.50)$$

This means, for example, that the interaction between two quarks due to one-gluon exchange vanishes in the limit of infinite momentum. This phenomenon is called “asymptotic freedom”, and furnishes a basis for the parton model of electron-proton deep inelastic scattering.

Using the lowest-order term in (12.49), we obtain from (12.48)

$$\frac{1}{\alpha_k} = \frac{1}{\alpha_\lambda} - \beta_0 \ln \frac{k^2}{\lambda^2} + O(\alpha_\lambda), \quad (12.51)$$

which is the analog of (9.18) in quantum electrodynamics, except that this is exact in the limit $k \rightarrow \infty$. Using (12.46) we can also write

$$d\left(\frac{k^2}{\lambda^2}, \alpha_\lambda\right) = 1 - \beta_0 \alpha_\lambda \ln \frac{k^2}{\lambda^2} + O(\alpha_\lambda^2). \quad (12.52)$$

The β -function can be obtained directly from (12.47), by calculating the Feynman graphs for the gluon propagator⁴. The lowest-order coefficient β_0 can be obtained from the graphs shown in Fig. 12.5, with the result

$$\beta_0 = -\frac{1}{6\pi} \left(\frac{33}{2} - N_f \right), \quad (12.53)$$

where N_f is the number of quark flavors. The quark contribution is $N_f/2$ times the value $(3\pi)^{-1}$ in quantum electrodynamics [see (9.40)]^c. For β_0 to be

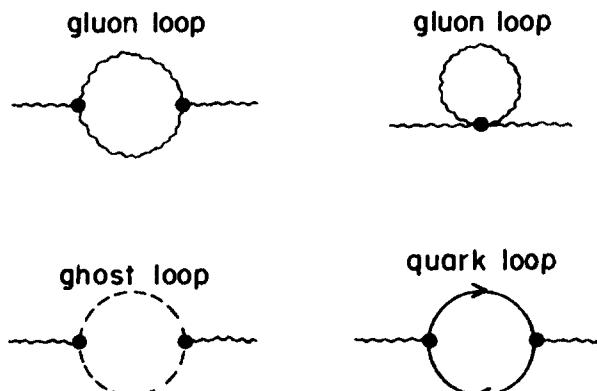


Fig. 12.5 Feynman graphs contributing to the GL function to lowest order.

^c The factor 1/2 arises as follows. To find the quark contribution to β_0 , it suffices to consider the self-energy graph of the A field with one quark loop. For any given flavor, we see from Table 12.1 that two colors contribute, and the charges are both 1/2. Thus, $2(1/2)^2 = 1/2$. The same result is obtained by considering the B field, because $\sum Q_A^2 = \sum Q_B^2$, as we can see from Table 12.1.

negative, it is necessary for the gluon self-interactions to dominate, which requires

$$N_f < 17. \quad (12.54)$$

The next coefficient is given by⁶

$$\beta_1 = -\frac{1}{8\pi^2} \left(51 - \frac{19}{3} N_f \right). \quad (12.55)$$

We note that (12.51) admits the solution

$$\alpha_k = -\frac{1}{\beta_0 \ln(k^2/\bar{\Lambda}^2)}, \quad (12.56)$$

where $\bar{\Lambda}^2$ is an arbitrary scale parameter, which can be determined by fitting QCD predictions to experimental data from electron-proton deep inelastic scattering. Existing data corresponding to $-k^2$ between 10 and 100 (GeV)² can be fitted with $\bar{\Lambda} \sim 0.5$ GeV.⁷ This gives $\alpha_k \approx 0.3$ at $(-k^2) = (10 \text{ GeV})^2$, for $N_f = 6$.

2 The Vacuum as Magnetic Medium

To establish asymptotic freedom, we only need to know β_0 , which we shall calculate by an elementary method due to Nielsen⁸. In this method, one views the vacuum state as a magnetic medium, and finds β_c through the response of the medium to an external magnetic field. The method is applicable to any field theory, and brings out the interesting view that asymptotic freedom is the outcome of a competition between Landau diamagnetism and Pauli paramagnetism, and depends crucially on the spins of the fields in the theory. We develop this point of view in a version due to Johnson.⁹

The running coupling constant α_k determines the k th Fourier transform of the color electrostatic potential between two static charges (isotopic charge or hypercharge) placed in the vacuum [see (9.20)]:

$$V(k) = \frac{\alpha_k}{k^2}, \quad (12.57)$$

where k is a space-like momentum. Let us write, by (12.46),

$$\alpha_k = \alpha_\Lambda d\left(\frac{k^2}{\Lambda^2}, \alpha_\Lambda\right), \quad (12.58)$$

and consider $\Lambda \rightarrow \infty$. We look upon Λ as a cutoff, and α_Λ as a bare coupling

⁶ W. E. Caswell, *Phys. Rev. Lett.* **33**, 244 (1974); D. R. T. Jones, *Nucl. Phys.* **B75**, 531 (1974).

⁷ A. J. Buras, *Rev. Mod. Phys.* **52**, 199 (1980).

⁸ N. K. Nielsen, *Am. J. Phys.* **49**, 1171 (1981).

⁹ K. Johnson, in *Asymptotic Realms of Physics*, eds. A. Guth, K. Huang, and R. L. Jaffe (MIT Press, Cambridge, 1983).

constant. The dielectric constant of the vacuum may be defined as

$$\epsilon_\Lambda(k) \equiv 1/d\left(\frac{k^2}{\Lambda^2}, \alpha_\Lambda\right). \quad (12.59)$$

Introducing the electric polarizability $P_\Lambda(k)$ by

$$\epsilon_\Lambda(k) = 1 + P_\Lambda(k), \quad (12.60)$$

we find from (12.52) that

$$P_\Lambda(k) = \beta_0 \alpha_\Lambda \ln \frac{\Lambda^2}{k^2} + O(\alpha_\Lambda^2). \quad (12.61)$$

Thus, asymptotic freedom means that the vacuum is a medium with negative electric polarizability.

In a Lorentz-invariant theory, the dielectric constant and the magnetic permeability are inverses of each other^d. Hence the latter is given by

$$\mu_\Lambda(k) = d\left(\frac{k^2}{\Lambda^2}, \alpha_\Lambda\right). \quad (12.62)$$

Introducing the magnetic susceptibility of the vacuum $\chi_\Lambda(k)$ by

$$\mu_\Lambda(k) = 1 + \chi_\Lambda(k), \quad (12.63)$$

we find

$$\chi_\Lambda(k) = -\beta_0 \alpha_\Lambda \ln \frac{\Lambda^2}{k^2} + O(\alpha_\Lambda^2). \quad (12.64)$$

Paramagnetism corresponds to $\beta_0 < 0$ (asymptotic freedom), and diamagnetism corresponds to $\beta_0 > 0$.

To calculate χ_Λ , we first calculate the vacuum energy in the presence of an external color magnetic field associated with the neutral gluon field A or B . To lowest order in α_Λ , all quarks and charged gluons can be treated as free particles, except for their interactions with the external field. The non-Abelian nature of the theory comes in only through the fact that gluons carry charge. The magnetic susceptibility will receive independent additive contributions from the quarks of different colors and flavors, and from the charged gluons.

We only need to consider the following prototype system: a charged massless free field (boson or fermion), of charge e and spin S , interacting with a weak homogeneous external magnetic field. The single-particle energies are the eigenvalues of the single-particle Hamiltonian H , whose square is given by

$$H^2 = |\mathbf{p} - e\mathbf{A}|^2 - 2e\mathbf{S} \cdot \mathbf{B}, \quad (12.65)$$

where $\mathbf{p} = -i\nabla$, $\mathbf{B} = \nabla \times \mathbf{A}$ is the external magnetic field, and the components of \mathbf{S} are spin matrices, each having only two possible eigenvalues $\pm S$. (For example, $\mathbf{S} = \sigma/2$ for the spin 1/2 case). The last term in (12.65) is designed to

^d This is because the velocity of light in the QCD vacuum is $[\epsilon_\Lambda(k)\mu_\Lambda(k)]^{-1/2}$.

give the equation of motion

$$[\mathbf{S}, H^2] = 2ie\mathbf{S} \times \mathbf{B}, \quad (12.66)$$

whose one-particle expectation leads to (12.38).

Let us choose \mathbf{B} to be a homogeneous field pointing along the z -axis:

$$\mathbf{B} = \hat{\mathbf{z}}B. \quad (12.67)$$

The two terms in (12.65) commute with each other, and the first term is of the form of a non-relativistic Hamiltonian (not squared Hamiltonian) of a particle in a magnetic field. Its eigenvalues have the well-known Landau spectrum:¹⁰

$$\text{Landau eigenvalue} = p_z^2 + 2eB(n + \frac{1}{2}) \quad (n = 0, 1, 2, \dots), \quad (12.68)$$

$$\text{Degeneracy} = \Omega^{2/3}eB/2\pi,$$

where p_z is the momentum along the direction of the magnetic field, and Ω is the total spatial volume. Each value of the quantum number n corresponds to a circular classical orbit of the particle. The degeneracy arises from the fact that the center of the orbit could be anywhere in a plane normal to \mathbf{B} . The single-particle energies are labelled by the three quantum numbers p_z , n , S_z :

$$\begin{aligned} E(p_z, n, S_z) &= [p_z^2 + 2eB(n + \frac{1}{2} - S_z)]^{1/2}, \\ n &= 0, 1, 2, \dots, \\ S_z &= \pm S. \end{aligned} \quad (12.69)$$

The vacuum energy per unit volume is

$$\mathcal{E}_{\text{vac}} = (-1)^{2S} \frac{eB}{2\pi} \int \frac{dp_z}{2\pi} \sum_n \sum_{S_z} \left[p_z^2 + 2eB(n + \frac{1}{2} - S_z) \right]^{1/2}, \quad (12.70)$$

where the factor $(-1)^{2S}$ takes into account the connection between spin and statistics^e. The magnetic susceptibility χ is defined by

$$\mathcal{E}_{\text{vac}} = -\frac{1}{2}\chi B^2. \quad (12.71)$$

For $S > 1/2$, (12.69) has complex eigenvalues, corresponding to the unstable modes studied by Nielsen and Olesen¹¹. However, they do not contribute to χ .

We cut off the divergent expression (12.70) by restricting the values of p_z and n such that

$$E(p_z, n, S_z) < \Lambda. \quad (12.72)$$

^e We take the vacuum energy to be the zero-point energy of the system. For a single boson field, this is $1/2$ the sum of all single-particle energies. For a charged boson field, the factor $1/2$ is cancelled by the fact that the charged field has two components. For a fermion field, the vacuum energy is the sum of all negative single-particle energies.

¹⁰ K. Huang, *Statistical Mechanics*, 2nd ed. (Wiley, New York, 1987), Chap. 11.

¹¹ N. K. Nielsen and P. Olesen, *Nucl. Phys.* B144, 376 (1978).

In the limit $\Lambda \rightarrow \infty$ and $B \rightarrow 0$, (12.70) can be evaluated by approximating the higher terms in the n -sum by the Euler formula¹²

$$\sum_{n=N}^K f(n + \frac{1}{2}) \cong \int_N^K dx f(x) - \frac{1}{24} [f'(K) - f'(N)], \quad (12.73)$$

where N is chosen to be large enough for the approximation to be valid. Its precise value is unimportant; the terms for $n < N$ will affect only the scale of Λ . The calculation is done in great detail in Ref. 7, and will not be repeated here. We only note that the continuum approximation to the n -sum does not contribute to χ ; only the second term in (12.73) is relevant. This is related to the fact that a classical system does not exhibit diamagnetism¹⁰. The result for the vacuum energy density is

$$\mathcal{E}_{\text{vac}} = -(-1)^{2S} \frac{e^2 B^2}{16\pi^2} [(2S)^2 - \frac{1}{3}] \ln \frac{\Lambda^2}{ceB} + \text{const.}, \quad (12.74)$$

where c is some number unimportant for our purpose. From this we obtain

$$\chi = (-1)^{2S} \frac{e^2}{8\pi^2} [(2S)^2 - \frac{1}{3}] \ln \frac{\Lambda^2}{ceB}, \quad (12.75)$$

which is the vacuum magnetic susceptibility in a weak static external magnetic field ($k^2 = ceB \rightarrow 0$).

3 The Nielsen-Hughes Formula

By comparing (12.75) with (12.64), we find

$$\beta_0 = -\frac{(-1)^{2S}}{2\pi} [(2S)^2 - \frac{1}{3}]. \quad (12.76)$$

This is valid for a massless charged field of any spin S (for which $g - 2 = 0$). It has also been derived by Hughes¹³ from a rather different point of view. We call it the “Nielsen-Hughes formula”.

When (12.76) is used in (12.49) and (12.48), we have the rate of change of the running coupling constant $\alpha = e^2/4\pi$, where e is the charge with respect to the “electromagnetic” field used in deriving the magnetic susceptibility. If the running coupling constant is defined as $\alpha = g^2/4\pi$, and the charge referred to above is $e = Qg$, then (12.76) should be multiplied by Q^2 .

The term $(2S)^2$ in (12.76) represents the effect due to an enhancement of the external magnetic field coming from spin alignment, while the term $-1/3$ represents the effect of orbital motion, which produces a magnetic field that tends to cancel the imposed field. These terms are respectively the analogs of Pauli paramagnetism and Landau diamagnetism in solid-state physics, where the respective susceptibilities also bear the ratio (for $S = 1/2$)¹⁰

$$\chi_{\text{Landau}}/\chi_{\text{Pauli}} = -\frac{1}{3}. \quad (12.77)$$

¹² M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions* (National Bureau of Standards, Washington, 1964), p. 806.

¹³ R. J. Hughes, *Phys. Lett.* **97b**, 246 (1980); *Nucl. Phys.* **B186**, 376 (1981).

In the present case, an additional effect arises due to fact that fermions have negative vacuum energy, as expressed by the factor $(-1)^{2S}$. Thus, quantum electrodynamics would have been asymptotically free, were it not for the necessity for “hole theory”.

To apply the Nielsen-Hughes formula to QCD, we add the contributions from the quarks of various colors and flavors, and from the gluons X, Y, Z . Taking the external magnetic field to be that associated with a neutral gluon field (A or B), we find with the help of Table 12.1 that

$$\begin{aligned} (\beta_0)_{\text{quarks}} &= N_f/6\pi, \\ (\beta_0)_{\text{gluons}} &= -11/4\pi. \end{aligned} \quad (12.78)$$

Therefore

$$\beta_0 = -\frac{1}{6\pi} \left(\frac{33}{2} - N_f \right), \quad (12.79)$$

as stated in (12.53).

12.4 The Pion as Goldstone Boson

1 The Low-Energy Domain

The scale $\bar{\Lambda} \sim 0.5$ GeV associated with asymptotic freedom defines a high-momentum (or short-distance) regime

$$\ln(k^2/\bar{\Lambda}^2) \gg 1,$$

in which quarks and gluons can be treated as weakly interacting particles in perturbative QCD. At the other end of the scale ($k^2/\bar{\Lambda}^2 \lesssim 1$) is low-energy hadronic physics, in which the interacting units are not individual quarks and gluons, but hadrons. So far we cannot solve QCD in this domain. However, from the vast amount of experimental information available, some highly fruitful ideas had been developed long before QCD was invented. Translated into expected properties of QCD, some of these ideas become concrete statements that are much easier to comprehend. They also serve as clues to the mathematical structure of QCD. We are referring to what is known as “current algebra”.¹⁴ In this section we translate a piece of that wisdom called “PCAC” (partially conserved axial-vector current hypothesis).

2 Chiral Symmetry: an Idealized Limit

The mass matrix M in the QCD Lagrangian density is a phenomenological quantity whose origin is unknown. It is the same as the quark mass matrix in the Weinberg-Salam model discussed in Chapter 6. If we adopt the latter to describe the electroweak interactions, then renormalizability requires that M arise from the

¹⁴ See S. B. Treiman, in S. B. Treiman, R. Jackiw, and D. J. Gross, *Lectures in Current Algebra and Its Applications* (Princeton University Press, Princeton, 1972); S. L. Adler and R. F. Dashen, *Current Algebras* (Benjamin, New York, 1968).

vacuum expectation value of the Higgs field; but it also contains arbitrary constants, and remains a phenomenological quantity. Whatever the origin, M can always be brought to diagonal form in the manner indicated in (6.72), through flavor-mixing transformations.^f Thus, the mass term in the Lagrangian density can be written as

$$\mathcal{L}_{\text{mass}} = - \sum_{i=1}^3 \sum_{f=1}^6 m_f \bar{q}^{fi} q^{fi}. \quad (12.80)$$

If QCD leads to quark confinement, as we shall assume, then the mass parameters m_f are not observable quantities. As we shall see later, however, they can be determined in terms of observable hadronic masses through "current algebra" methods. They are called "current" quark masses, to distinguish them from "constituent" quark masses, which are parameters used in phenomenological quark models of hadronic structure.

The approximate unitary symmetry of the strong interactions implies that $m_u \cong m_d \cong m_s$. Since isospin conservation is a much better symmetry than the whole of flavor $SU(3)$, the relation $m_u \cong m_d$ should hold to a higher degree of accuracy than $m_d \cong m_s$. The mass parameters for c, b, t should all be much larger than those for u, d, s , because we see no trace of flavor $SU(4)$ or higher symmetries in the hadronic spectrum.

If we put $m_u = m_d$, then isospin will become exactly conserved. In this limit, n and p will have the same mass, and so will π^+ , π^- , and π^0 . But this does not explain the smallness of the pion mass ($m_\pi/m_n = 0.14$), which makes the pion very special among hadrons. To understand this, Nambu¹⁵ and Chou¹⁶ suggested that there is a limit in which the pion is a massless Goldstone boson associated with spontaneous symmetry breaking. This limit is of course an idealized theoretical construct. As an example, Nambu and Jona-Lasinio¹⁷ gave a non-renormalizable model in which the fundamental fields are massless nucleon fields, and thus possesses chiral symmetry, whose spontaneous breakdown gives rise to nucleon mass, and a massless Goldstone boson identified with the pion.

To translate these ideas into QCD, we consider the idealized limit corresponding to

$$m_u = m_d = 0. \quad (12.81)$$

The Lagrangian density exhibiting the chiral symmetry in this limit is

$$\mathcal{L}_{\text{chiral}} = -\frac{1}{4} G_a^{\mu\nu} G_{a\mu\nu} + \bar{\psi} i \not{D} \psi, \quad (12.82)$$

where

$$\psi = \begin{pmatrix} u \\ d \end{pmatrix}. \quad (12.83)$$

The quark color index has been suppressed.

^f We temporarily ignore the complications relating to CP violation discussed in Sec. 12.6.

¹⁵ Y. Nambu, *Phys. Rev. Lett.* **4**, 380 (1960).

¹⁶ Chou Kuang-chao, *Soviet Phys. JETP* **12**, 492 (1961).

¹⁷ Y. Nambu and G. Jona-Lasinio, *Phys. Rev.* **122**, 345 (1961); **124**, 246 (1961).

The full QCD Lagrangian density can be written in the form

$$\mathcal{L} = \mathcal{L}_{\text{chiral}} - (m_u \bar{u}u + m_d \bar{d}d) + \mathcal{L}_{\text{scbt}}, \quad (12.84)$$

where $\mathcal{L}_{\text{scbt}}$ contains terms pertaining only to the quarks s, c, b, t . We shall regard $\mathcal{L}_{\text{chiral}}$ as an “unperturbed” Lagrangian density. The idealized limit of chiral symmetry corresponds to the unperturbed problem.

As we shall show later, m_u and m_d are in fact small compared to the nucleon mass. Hence $\mathcal{L}_{\text{chiral}}$ should, by itself, give a reasonably good description of “ordinary” hadronic physics, in which strangeness and higher flavors do not play a direct role. This assumption implies that hadronic masses arise from dimensional transmutation, since there are no intrinsic mass parameters in $\mathcal{L}_{\text{chiral}}$.

Our “unperturbed” system, as described by $\mathcal{L}_{\text{chiral}}$, is invariant under the following global symmetry groups:

$$\begin{aligned} [SU(2)]_V: \psi &\rightarrow e^{-i\mathbf{r} \cdot \boldsymbol{\omega}/2} \psi, \\ [U(1)]_V: \psi &\rightarrow e^{-i\alpha} \psi, \\ [SU(2)]_A: \psi &\rightarrow e^{-i\gamma_5 \mathbf{r} \cdot \boldsymbol{\theta}/2} \psi, \\ [U(1)]_A: \psi &\rightarrow e^{-i\gamma_5 \beta} \psi, \end{aligned} \quad (12.85)$$

where $\boldsymbol{\omega}, \boldsymbol{\theta}, \alpha, \beta$ are arbitrary real constants. The subscripts V and A stand respectively for “vector” and “axial-vector”. The associated Noether currents are respectively

$$\begin{aligned} J^k_\mu &= \bar{\psi} \gamma_\mu \tau_k \psi \quad (k = 1, 2, 3) \quad (\text{isospin current}), \\ j_\mu &= \bar{\psi} \gamma_\mu \psi \quad (\text{baryonic current}), \\ J_{5\mu}^k &= \bar{\psi} \gamma_\mu \gamma_5 \tau_k \psi \quad (k = 1, 2, 3), \\ j_{5\mu} &= \bar{\psi} \gamma_\mu \gamma_5 \psi. \end{aligned} \quad (12.86)$$

The baryonic current is conserved even in the perturbed system defined by \mathcal{L} . The isospin current is conserved as long as $m_u = m_d$. The axial-vector currents are conserved only in the chiral-symmetric limit apart from possible anomalies discussed later.

How are these symmetries manifested in nature? First of all, $[SU(2)]_V$ and $[U(1)]_V$ are manifested directly as isospin conservation and baryon number conservation, respectively. In particular, hadrons fall into easily recognizable isospin multiplets.

On the other hand, a direct manifestation of $[SU(2)]_A$ would require that each isospin multiplet be accompanied by a mirror multiplet of the same mass, but with opposite parity. No hint of this can be found in the hadronic spectrum. For example, there is not even an approximate mirror image of the nucleon iso-doublet. Assuming that the real world is well-approximated by the chiral-symmetric limit, we must conclude that $[SU(2)]_A$ is spontaneously broken. This calls for an $I = 1$ pseudoscalar Goldstone boson, which we identify with the “unperturbed” pion. The physical pion then corresponds to the perturbed state of the Goldstone boson, whose mass comes from m_u and m_d .

The question of how $[U(1)]_A$ is manifested presents a puzzle, and will be discussed separately later.

What is the dynamical reason that $[SU(2)]_A$ manifests itself in the Goldstone mode? We do not have an answer. In the non-renormalizable pre-QCD model of Nambu and Jona-Lasinio¹⁷, the cause of spontaneous symmetry breakdown is a direct nucleon-nucleon attraction built into the model, in analogy with the effective electron-electron attraction responsible for the formation of Cooper pairs in the theory of superconductivity⁸. A direct quark-quark interaction is ruled out by renormalizability in QCD; but an effective interaction could arise, just as the effective electron-electron attraction in superconductivity arises from the more fundamental electron-phonon interaction: It has been suggested that instantons play a role in such an effective quark-quark interaction.¹⁸

Whatever the dynamical mechanism, the spontaneous breakdown of chiral symmetry will lead to non-vanishing vacuum expectation values $\langle \bar{u}u \rangle$ and $\langle \bar{d}d \rangle$, which furnish the scale for hadronic masses in the chiral limit (dimensional transmutation). One could also attribute to u and d some sort of effective masses—what one would term “constituent” quark masses. Thus we see that the dynamical problem of chiral symmetry breaking cannot be dissociated from that of quark confinement.

3 PCAC

Let the pion state be denoted by $|\pi^j\rangle$, where $j = 1, 2, 3$ is the isospin index. The operator $J_{S\mu}^k$ can annihilate states of the same quantum numbers as the pion, and hence connects the pion state to the vacuum. By Lorentz invariance and isospin conservation, we can write

$$\langle 0 | J_{S\mu}^k(x) | \pi^j \rangle = i \delta_{jk} f_\pi p_\mu e^{-ip \cdot x}, \quad (12.87)$$

where p_μ is the pion 4-momentum, and f_π is a constant. We immediately obtain

$$\langle 0 | \partial^\mu J_{S\mu}^k(x) | \pi^j \rangle = \delta_{jk} f_\pi m_\pi^2 e^{-ip \cdot x}. \quad (12.88)$$

This is consistent with the view that the pion is a Goldstone boson in the chiral limit, for $\partial^\mu J_{S\mu}^k = 0$ implies $m_\pi = 0$.

Define

$$\phi_\pi^k(x) \equiv \frac{1}{m_\pi^2 f_\pi} \partial^\mu J_{S\mu}^k(x). \quad (12.89)$$

Then

$$\langle 0 | \phi_\pi^k(x) | \pi^j \rangle = \delta_{jk} e^{-ip \cdot x}. \quad (12.90)$$

Thus, $\phi_\pi^k(x)$ can be used as a pion field operator. It is composed of quark operators, reflecting the bound-state nature of the pion. The content of “PCAC” is a rule for using $\phi_\pi^k(x)$ in the chiral limit ($m_\pi \rightarrow 0$).

⁸ In the theory of superconductivity, the Goldstone boson is “eaten” by the electromagnetic field through the Higgs mechanism, and the photon becomes massive in a superconductor (Meissner effect).

¹⁸ R. D. Carlitz, *Phys. Rev.* **D17**, 3225 (1978).

Consider the reaction $a \rightarrow b + \pi^k$. Using the reduction formula¹⁹

$$\langle \pi^k, b^{\text{out}} | a^{\text{in}} \rangle = i \int d^4x e^{ip \cdot x} (\square^2 + m_\pi^2) \langle b | \phi_\pi^k(x) | a \rangle, \quad (12.91)$$

where p is the pion 4-momentum, we can obtain the transition amplitude by extracting the coefficient of $i(2\pi)^4 \delta^4(p + p_b - p_a)$:

$$\text{Amp } (a \rightarrow b + \pi^k) = (m_\pi^2 - p^2) \langle a | \phi_\pi^k(0) | b \rangle. \quad (12.92)$$

The implementation of “PCAC” consists of defining the off-mass-shell amplitude to be

$$T_{ab}^k(p^2) \equiv \frac{m_\pi^2 - p^2}{f_\pi p^2} \langle a | \partial^\mu J_{S\mu}^k(0) | b \rangle. \quad (12.93)$$

The chiral limit is approached by first letting $m_\pi \rightarrow 0$ to obtain the off-mass-shell amplitude in the chiral limit, and then taking the mass-shell limit $p^2 \rightarrow 0$ eventually. Thus,

$$T_{ab}^k(p^2) \xrightarrow[\text{chiral limit}]{} -\frac{1}{f_\pi} \langle a | \partial^\mu J_{S\mu}^k(0) | b \rangle. \quad (12.94)$$

We now determine the constant f_π . In QCD, none of the currents in (12.86) are coupled to dynamical fields. When we enlarge the system to include the electroweak interactions, J_μ^k and $J_{S\mu}^k$ ($k = 1, 2$) become part of the charge-changing weak currents coupled to W_μ^k ($k = 1, 2$). Thus, the components $k = 1, 2$ in (12.87) are involved in the matrix element for charged pion decay ($\pi \rightarrow \mu + \nu'$). Consequently, f_π can be determined from the charged pion lifetime. Without going through the derivation, we merely quote the result:¹⁴

$$\text{Rate } (\pi \rightarrow \mu + \nu') = \frac{1}{4\pi} f_\pi^2 (G \cos \theta)^2 m_\pi m_\mu^2 \left(1 - \frac{m_\mu^2}{m_\pi^2}\right)^2, \quad (12.95)$$

where G is the Fermi constant, θ the Cabibbo angle, and m_π and m_μ are respectively the physical mass of the pion and the muon. Using the value 2.6×10^{-8} s for the lifetime of the charged pion, one obtains

$$f_\pi = 93 \text{ MeV}. \quad (12.96)$$

This is called the “pion decay constant”.

4 The Decay $\pi^0 \rightarrow 2\gamma$

We apply (12.94) to the electromagnetic decay of π^0 into two photons. The π^0 operator is proportional to

$$J_{S\mu}^3 = \bar{u} \gamma_\mu \gamma_5 u - \bar{d} \gamma_\mu \gamma_5 d, \quad (12.97)$$

¹⁹ K. Huang and H. A. Weldon, *Phys. Rev.* D11, 257 (1975) show that the reduction formula is valid for bound-state operators.

which has an axial anomaly. From (11.36) we obtain

$$\partial^\mu J_{5\mu}^3 = 2m_\mu(\bar{u}\gamma_5 u) - 2m_d(\bar{d}\gamma_5 d) + \frac{\xi\alpha}{2\pi} \tilde{F} \cdot F, \quad (12.98)$$

where

$$\xi = 3 \left[\left(\frac{2}{3} \right)^2 - \left(\frac{1}{3} \right)^2 \right] = 1. \quad (12.99)$$

The factor 3 accounts for the quark colors. In the chiral limit we put $m_u = m_d = 0$, so that only the anomaly survives in (12.98).

Let the 4-momenta and polarization vectors of the two final photons be denoted respectively by k_1, ϵ_1 and k_2, ϵ_2 . The off-mass-shell decay amplitude in the chiral limit is, by (12.94) and (12.98),

$$\begin{aligned} T(\omega) &= -\frac{1}{f_\pi} \frac{\alpha}{2\pi} \langle k_1 \epsilon_1, k_2 \epsilon_2 | \tilde{F} \cdot F | 0 \rangle \\ &= -\frac{2\alpha}{\pi f_\pi} \epsilon_{\mu\nu\alpha\beta} \epsilon_1^\mu \epsilon_2^\nu k_1^\alpha k_2^\beta \\ &= -\frac{4\alpha}{\pi f_\pi} \omega^2 |\epsilon_1 \times \epsilon_2|, \end{aligned} \quad (12.100)$$

where the 3-vectors are those in the c.m. frame, with

$$\begin{aligned} k_1 &= (\omega, \mathbf{k}), & k_2 &= (\omega, -\mathbf{k}), & |\mathbf{k}| &= \omega, \\ \epsilon_1 &= (0, \boldsymbol{\epsilon}_1), & \epsilon_2 &= (0, \boldsymbol{\epsilon}_2). \end{aligned} \quad (12.101)$$

Using (12.100) we obtain, after a standard calculation,

$$\text{Rate } (\pi^0 \rightarrow 2\gamma) = \frac{\alpha^2}{64\pi^3} \left(\frac{m_\pi}{f_\pi} \right)^2 m_\pi, \quad (12.102)$$

which gives for the π^0 lifetime

$$\tau = (8.5 \text{ eV})^{-1}, \quad (12.103)$$

in agreement with the experimental value²⁰

$$\begin{aligned} \tau_{\text{expt}} &= [(7.95 \pm 0.55) \text{ eV}]^{-1} \\ &= 0.828 \times 10^{-16} \pm 0.057 \text{ s}. \end{aligned} \quad (12.104)$$

If we had not known about the anomaly, we would have obtained zero for the decay rate in the chiral limit, and would have been puzzled by the failure of “PCAC” here, when it had been successful in other applications. The puzzle did exist historically, and was the main motivation behind the discovery of the axial anomaly.

²⁰ Particle Data Group, *Rev. Mod. Phys.* **52**, S1 (1980).

If we had forgotten about color, the decay rate we calculated would have been too small by a factor of 3. However, this is not evidence for color by itself, for we could have gotten the right answer by making the current out of nucleons instead of quarks.

We can view the amplitude (12.100) from another point of view, namely it is the 4-divergence of the triangle graph with respect to the axial-vector vertex. In the chiral limit $m_u = m_d = 0$, the anomaly due to the u and d quark loops must be reproduced by massless bound states if quarks are confined, according to 't Hooft's principle (Sec. 11.7). We have assumed that the limiting chiral symmetry is spontaneously broken. Hence the massless bound state must be the Goldstone boson π^0 , whose coupling to two photons is entirely determined by the anomaly. The triangle graph also receives contributions from quarks of higher flavors, but they do not contribute to the anomaly that the π^0 has to reproduce, because they remain massive in the chiral limit we are considering. This is just a formal way of saying that the pion is composed mostly of $\bar{u}u$ and $\bar{d}d$, but very little $\bar{s}s$, $\bar{c}c$, etc.

5 Extension to Pion Octet

Since there is strong evidence for approximate unitary symmetry in hadronic physics, we might entertain the idea of an extended chiral-symmetric limit corresponding to

$$m_u = m_d = m_s = 0. \quad (12.105)$$

By the same reasoning as before, we would conclude that this extended chiral symmetry is spontaneously broken, and is manifested through the existence of a Goldstone boson identifiable with the pion octet. To consider this limit, all we have to do is to enlarge (12.82) and (12.83) by the re-definition

$$\Psi = \begin{pmatrix} u \\ d \\ s \end{pmatrix}, \quad (12.106)$$

and replace the Pauli matrices τ_k in (12.85) and (12.86) by the Gell-Mann matrices λ_a ($a = 1, \dots, 8$). We then have an "unperturbed" system with global symmetry group

$$[SU(3)]_V \times [SU(3)]_A \times [U(1)]_V \times [U(1)]_A, \quad (12.107)$$

with $[SU(3)]_V$ realized directly in the "eight-fold way", and $[SU(3)]_A$ realized in the Goldstone mode.

The mass splittings in the pion octet, as well as those in all other flavor $SU(3)$ multiplets, arise from the perturbation Lagrangian density

$$-(m_u \bar{u}u + m_d \bar{d}d + m_s \bar{s}s), \quad (12.108)$$

whose effect is usually treated by "chiral perturbation theory"²¹, which is a combination of "current algebra" methods and "extended PCAC". Using such

²¹ H. Pagels, *Phys. Reports* 16C, 219 (1975).

techniques, one can relate quark mass ratios to those involving pions and kaons. We quote the results of Weinberg:²²

$$\frac{m_d}{m_u} \cong \frac{m^2(K^0) - m^2(K^+) + m^2(\pi^+)}{2m^2(\pi^0) + m^2(K^+) - m^2(K^0) - m^2(\pi^+)} = 1.80, \quad (12.109)$$

$$\frac{m_s}{m_d} \cong \frac{m^2(K^0) + m^2(K^+) - m^2(\pi^+)}{m^2(K^0) - m^2(K^+) + m^2(\pi^+)} = 20.1.$$

Using the observed mass splittings in flavor $SU(3)$ multiplets as a guide for an estimate of m_s , Weinberg suggests²³

$$\begin{aligned} m_s &= 150 \text{ MeV}, \\ m_d &= 7.5 \text{ MeV}, \\ m_u &= 4.2 \text{ MeV}. \end{aligned} \quad (12.110)$$

12.5 The $U(1)$ Puzzle

If the symmetry $[U(1)]_A$ were manifested directly, then in the chiral limit all massless hadrons would have a massless partner of opposite parity. In the real world we would expect a scalar counterpart of the pion, of roughly the same mass. Since there does not appear to be such a particle, we assume that the symmetry is spontaneously broken. But then there should be an $I = 0$ pseudoscalar Goldstone boson, whose perturbed state should have about the same mass as the pion. Using chiral perturbation theory, Weinberg²⁴ has estimated the mass to be less than $\sqrt{3}m_\pi$. Among the known hadrons, the only candidates with the right quantum numbers are $\eta(549)$ and $\eta'(985)$. Both violate the Weinberg bound. Besides, $\eta(549)$ has already been claimed by the pion octet. The $U(1)$ puzzle is: *where is the extra Goldstone boson?*

't Hooft removed the puzzle by showing that the expected Goldstone boson is not a physical particle, due to instanton effects. We shall not go into the detailed analysis, but merely mention some relevant points.

The current j_5^μ is not conserved, due to a QCD axial anomaly:^b

$$\partial_\mu j_5^\mu = \frac{N_f g_0^2}{16\pi^2} \tilde{\mathcal{F}} \cdot \mathcal{F}, \quad (12.111)$$

where N_f is the number of quark flavors taken into account in the chiral limit ($N_f = 2$ or 3), and we use the abbreviation $\tilde{\mathcal{F}} \cdot \mathcal{F} = \tilde{\mathcal{F}}_a^{\mu\nu} \mathcal{F}_{a\mu\nu}$. Now, we know

^b See the remarks at the end of Sec. 11.3. The form of the anomaly is an obvious generalization from the Abelian case. The coefficient is $N_f/2$ times that in the Abelian case—the same factor that appears in the β function, as discussed in footnote c.

²² S. Weinberg, in *A Festschrift for I. I. Rabi*, Ed. L. Motz (New York Academy of Sciences, New York, 1977).

²³ For another suggestion see T. D. Lee, *Particle Physics and Introduction to Field Theory* (Harwood Publishers, Chur, Switzerland, 1981), p. 584.

²⁴ S. Weinberg, *Phys. Rev.* D11, 3583 (1975).

that $\tilde{\mathcal{J}} \cdot \mathcal{J}$ is a total 4-divergence [see (5.6)]:

$$\tilde{\mathcal{J}} \cdot \mathcal{J} = \partial_\mu \bar{X}^\mu, \quad (12.112)$$

whose integral over all space-time is proportional to the topological charge. We can define a conserved but non-gauge-invariant current

$$J_5^\mu \equiv j_5^\mu - \frac{N_f g_0^2}{16\pi^2} X^\mu, \quad (12.113)$$

which is the non-Abelian generalization of (11.39). The generator of the $[U(1)]_A$ symmetry may be taken to be

$$Q_5 = \int d^3x J_5^0 = \int d^3x \left[\psi^\dagger \gamma_5 \psi - \frac{N_f g_0^2}{16\pi^2} X^0 \right]. \quad (12.114)$$

In the Abelian case, Q_5 is gauge invariant, because of the absence of topological charge. This is no longer true here, and Q_5 is not a physical quantity. In fact, Q_5 is not even conserved, because of the existence of instantons. To see this, integrate the equation $\partial_\mu J_5^\mu$ over Euclidean 4-space. The result can be presented in the form

$$\int_{-\infty}^{\infty} dt \frac{dQ}{dt} = 2N_f q[G], \quad (12.115)$$

where

$$q[G] = \frac{g_0^2}{32\pi^2} \int d^4x \tilde{\mathcal{J}} \cdot \mathcal{J} \quad (12.116)$$

is the topological charge, a functional of the gauge field G_a^μ . For G_a^μ corresponding to one instanton, $q[G] = 1$. Therefore, in this case, the boundary values of Q_5 in Euclidean time differ by

$$\Delta Q_5 = 2N_f q[G]. \quad (12.117)$$

This can be attributed to the fact that an instanton interpolates (in Euclidean time) between two gauge-field configurations differing by one unit of topological charge. (See Sec. 8.6.) Thus, there is no reason to expect $[U(1)]_A$ to have physical manifestations. 't Hooft's detailed analysis explains what becomes of the would-be Goldstone boson, which we shall not go into.

The $U(1)$ puzzle is not a mathematical paradox, but rather a frustration of cherished beliefs. The work of 't Hooft did remove the puzzle in its original form, but generates new questions. For example, does "instanton physics" alter conventional thinking in low-energy hadronic physics, in particular the spontaneous breaking of $[SU(2)]_A$? To go into these questions would embroil us in unsettled and controversial arguments. In the next section, we discuss only one

relatively well-understood aspect of instanton physics—what become of the θ -worlds in QCD.

12.6 θ -Worlds in QCD

1 Euclidean Action

We recall that there are “large” and “small” gauge transformations. (See Sec. 8.6). We can make the vacuum state invariant under all gauge transformations, including the “large” ones, by adding to the action a term proportional to the topological charge. [See (8.140)]. Ignoring the heavy quarks c, b, t , we consider the Minkowski action

$$S_\theta = \int d^4x [-\frac{1}{4}\mathcal{F}^2 + \bar{\psi}(iD - M)\psi] - \theta q[G], \quad (12.118)$$

where ψ is given by (12.106), and M is a mass matrix to be discussed in detail later. The parameter θ multiplying the topological charge $q[G]$ is an unknown constant. In quarkless QCD, it labels different θ -worlds that are physically distinct. We shall see that introducing quarks into the theory changes the meaning of θ , both mathematically and physically.

We shall work in Euclidean 4-space. To continue from Minkowski space to the latter, we replace the quantities in the left column of Table 12.2 by the

Table 12.2
CONTINUATION FROM
MINKOWSKI TO EUCLIDEAN 4-SPACE

	Minkowski	Euclidean
Coordinate	x^0 x^k	$-ix_E^4$ $x_E^k \quad (k = 1, 2, 3)$
Momentum	p^0 p^k	ip_E^4 p_E^k
Gauge fields	G_a^0 G_a^k	$i(G_E)_a^4$ $(G_E)_a^k$
Invariants	\mathcal{F}^2 $\mathcal{F} \cdot \mathcal{F}$	\mathcal{F}_E^2 $i\mathcal{F}_E \cdot \mathcal{F}_E$
Topological charge	$q[G]$	$q[G_E]$
Dirac matrices	γ^0 γ^k $\gamma^\mu \gamma^\nu + \gamma^\nu \gamma^\mu = 2g^{\mu\nu}$ $\gamma_5 = -i\gamma^0 \gamma^1 \gamma^2 \gamma^3$ $\not{p} = \gamma^\mu p_\mu$	γ_E^4 $-i\gamma_E^k$ $\gamma_E^\mu \gamma_E^\nu + \gamma_E^\nu \gamma_E^\mu = 2\delta_{\mu\nu}$ $(\gamma_E)_5 = \gamma_E^1 \gamma_E^2 \gamma_E^3 \gamma_E^4$ $i\not{p}_E = i\gamma_E^\mu p_E^\mu$
θ -Action	S_θ	$i(S_E)_\theta$

corresponding ones in the right column. Note that there is no distinction between Euclidean upper and lower indices. The Euclidean Dirac matrices are all hermitian, and γ_5 is represented by the same matrix in both spaces. From now on we drop the subscript E denoting Euclidean quantities. The Euclidean action is written as

$$S_\theta = \int d^4x [\frac{1}{4}\bar{\psi}^2 + \bar{\psi}(D + M)\psi] + i\theta q[G]. \quad (12.119)$$

The “partition function”, from which Euclidean quark Green’s functions can be obtained, is given by

$$Z_\theta[\eta, \eta'] = N \int (DG)(D\bar{\psi})(D\psi) \exp[-S_\theta - (\bar{\psi}, \eta) - (\bar{\eta}, \psi)], \quad (12.120)$$

where η and η' are anti-commuting c-number sources, and we have left understood gauge-fixing and ghost terms in the exponent. We use the abbreviation $(f, g) = \int d^4x f \cdot g$, where $f \cdot g$ is a product contracted in all indices, if any.

Note that the θ -term in (12.119) is pure-imaginary. The topological charge always contribute a phase factor to the path integral, whether in Minkowskian or Euclidean space-time.

2 The Axial Anomaly and the Index Theorem

In the massless limit, S_θ is invariant under chiral transformations. Since all quantities in (12.120) are classical, one might expect to find the classical result $\partial_\mu j_5^\mu = 0$, and wonder where the axial anomaly comes from. Fujikawa²⁵ showed that it comes from the non-chiral-invariance of the fermionic measure

$$d\mu = (D\bar{\psi})(D\psi). \quad (12.121)$$

To see this, define a complete set of spinor eigenfunctions of D :ⁱ

$$\begin{aligned} D\phi_n(x) &= E_n\phi_n(x), \\ \int d^4x \phi_n^\dagger(x)\phi_m(x) &= \delta_{nm}, \\ \sum_n \phi_n(x)\phi_n^\dagger(y) &= \delta^4(x - y), \end{aligned} \quad (12.122)$$

where $\phi_n(x)$ is a functional of G_a^μ , by virtue of the dependence of D on the latter. We expand the quark fields ψ and $\bar{\psi}$ of each flavor as follows (with the flavor index omitted for brevity):

$$\psi(x) = \sum_n a_n\phi_n(x), \quad (12.123)$$

$$\bar{\psi}(x) = \sum_n b_n\phi_n^\dagger(x),$$

ⁱ We normalize ϕ_n to unity in a 4-dimensional Euclidean volume $\Omega = 1$. The infinite-volume limit is taken by letting the unit of volume approach zero.

²⁵ K. Fujikawa, *Phys. Rev. Lett.* **42**, 1195 (1979).

where $\{a_n\}$ and $\{b_n\}$ are sets of independent anti-commuting c-numbers. We take

$$d\mu = \prod_f \prod_n db_n da_n, \quad (12.124)$$

where f denotes flavor. Under a local chiral transformation,

$$\begin{aligned} \psi(x) &\rightarrow \psi'(x) = e^{-i\gamma_5\alpha(x)} \psi(x), \\ \bar{\psi}(x) &\rightarrow \bar{\psi}'(x) = \bar{\psi}(x) e^{-i\gamma_5\alpha(x)}, \end{aligned} \quad (12.125)$$

and the coefficients a_n and b_n transform linearly:

$$\begin{aligned} a_n &\rightarrow a'_n = \sum_m C_{nm} a_m, \\ b_n &\rightarrow b'_n = \sum_m C_{nm} b_m, \end{aligned} \quad (12.126)$$

where C_{nm} is an ordinary number:

$$C_{nm} = \int d^4x \phi_n^\dagger(x) e^{-i\gamma_5\alpha(x)} \phi_m(x). \quad (12.127)$$

Since γ_5 anticommutes with \not{D} , we can choose ϕ_n to be states of definite chirality. For each $E_n \neq 0$, there are two degenerate states of opposite chirality. In such a “chiral representation” C_{nm} is diagonal, and we can easily see that

$$\prod_n da'_n = (\det C)^{-1} \prod_n da_n. \quad (12.128)$$

For infinitesimal $\alpha(x)$ we have

$$\begin{aligned} (\det C)^{-1} &= \sum_n \left[1 - i \int d^4x \delta\alpha(x) \phi_n^\dagger(x) \gamma_5 \phi_n(x) \right] \\ &= \exp -i \int d^4x \alpha(x) \sum_n \phi_n^\dagger(x) \gamma_5 \phi_n(x). \end{aligned} \quad (12.129)$$

This result is independent of the representation. Thus, under an *infinitesimal local chiral transformation*,

$$d\mu \rightarrow d\mu' = e^{i\Delta} d\mu, \quad (12.130)$$

where

$$\Delta = -2N_f \int d^4x \alpha(x) \sum_n \phi_n^\dagger(x) \gamma_5 \phi_n(x), \quad (12.131)$$

where N_f is the number of quark flavors under consideration. This shows the non-invariance of the fermionic measure.

The sum in the integral of (12.131) is ambiguous, and we define it as

$$\sum_n \phi_n^\dagger(x) \gamma_5 \phi_n(x) = \lim_{\Lambda \rightarrow \infty} \sum_n e^{-E_n^2/\Lambda^2} \phi_n^\dagger(x) \gamma_5 \phi_n(x), \quad (12.132)$$

where Λ is an energy cutoff, which respects chiral symmetry. The calculation of the right-hand side is straightforward. We refer the reader to Ref. 26, and just quote the result:

$$\sum_n \phi_n^\dagger(x) \gamma_5 \phi_n(x) = \frac{g_0^2}{32\pi^2} \mathfrak{F} \cdot \mathfrak{F}. \quad (12.133)$$

Substituting this into (12.131), we obtain

$$\Delta = \frac{N_f g_0^2}{16\pi^2} \int d^4x \alpha(x) \mathfrak{F} \cdot \mathfrak{F}. \quad (12.134)$$

In the absence of sources, Z_θ should be invariant under a local chiral transformation, since all the spinor components of ψ and $\bar{\psi}$ are independently integrated over. Using (12.134), we find that, under an infinitesimal local chiral transformation, Z_θ changes by

$$\begin{aligned} \delta Z_\theta = i & \int (DG) d\mu e^{-S_0} \\ & \times \int d^4x \alpha(x) \left[-\partial_\mu j_5^\mu + 2\bar{\psi} M \gamma_5 \psi + \frac{N_f g_0^2}{16\pi^2} \mathfrak{F} \cdot \mathfrak{F} \right]. \end{aligned} \quad (12.135)$$

Hence the quantity in the brackets must vanish, and we have the axial anomaly as expressed in (12.111).

The above derivation of the axial anomaly is no more rigorous than that of Chapter 11, for we have not shown that the result is independent of the cutoff procedure. Thus, we have not improved on the argument (given in Sec. 11.3) that there are no radiative corrections to the anomaly.

Let us go back to (12.133) and integrate both sides over Euclidean 4-space. The left-hand side, being made well-defined by the cutoff procedure (12.132), receives contributions only from states with $E_n = 0$, which we call “zero-modes”. For $E_n \neq 0$, the contributions of the two degenerate chiral states cancel each other. The right-hand side gives the topological charge. Hence

$$n_+ - n_- = q, \quad (12.136)$$

where n_\pm are respectively the number of zero-modes of chirality ± 1 (of a given flavor), in the presence of a background gauge field G_a^μ , and q is the topological charge of the background field. This is the *Atiyah-Singer index theorem*²⁶, for which our derivation is of the nature of a “poor man’s proof”. The theorem is rigorous.

²⁶ M. Atiyah and I. Singer, *Ann. Math.* **87**, 484 (1968).

For a background field consisting of one instanton, we have $q = 1$, and the index theorem tells us $n_+ - n_- = 1$. This shows that there is at least one zero-mode for each quark flavor in a one-instanton field. 't Hooft²⁷ has shown that there is exactly one zero-mode with positive chirality in the instanton field, with normalizable wave function

$$\phi_{\text{0-mode}}(x) = (1 + x^2)^{-3/2} u, \quad (12.137)$$

where $x^2 = x^\mu x^\mu$, and u is a constant Dirac spinor.

3 Chiral Limit: Collapse of the θ -Worlds

The chiral limit is defined by setting $M = 0$ in the action:

$$S_\theta^{(0)} = \frac{1}{4}(\mathcal{F}, \mathcal{F}) + (\bar{\psi}, \not{D}\psi) + i\theta q[G]. \quad (12.138)$$

The partition function is then

$$\begin{aligned} Z_\theta^{(0)}[\eta, \bar{\eta}] &= \mathcal{N} \int (DG) \exp\{-\frac{1}{4}(\mathcal{F}, \mathcal{F}) - i\theta q[G]\} f_0[G, \eta, \bar{\eta}], \\ f_0[G, \eta, \bar{\eta}] &= \int d\mu \exp[-(\bar{\psi}, \not{D}\psi) - (\eta, \bar{\psi}) - (\bar{\eta}, \psi)]. \end{aligned} \quad (12.139)$$

Changing the variables of integration by performing an *infinitesimal global chiral transformation*, we have

$$\begin{aligned} &\int d\mu \exp[-(\bar{\psi}, \not{D}\psi) - (\eta, \bar{\psi}) - (\bar{\eta}, \psi)] \\ &= \int d\mu' \exp[-(\bar{\psi}', \not{D}\psi') - (\eta, \bar{\psi}') - (\bar{\eta}, \psi')] \\ &= e^{2i\Delta} \int d\mu \exp[-(\bar{\psi}, \not{D}\psi) - (\eta', \bar{\psi}) - (\bar{\eta}', \psi)], \end{aligned} \quad (12.140)$$

where

$$\begin{aligned} \eta' &= e^{-i\alpha\gamma_5} \eta, \\ \bar{\eta}' &= \bar{\eta} e^{-i\alpha\gamma_5}. \end{aligned} \quad (12.141)$$

Therefore

$$f_0[G, \eta, \bar{\eta}] = e^{2i\Delta} f_0[G, \eta', \bar{\eta}'], \quad (12.142)$$

where, by (12.134),

$$\Delta = 2\alpha N_f q[G]. \quad (12.143)$$

²⁷ G. 't Hooft, *Phys. Rev. Lett.* **37**, 8 (1976); *Phys. Rev. D* **14**, 3432 (1976). See also Coleman in *The Ways of Subnuclear Physics*, Ed. A. Zichichi (Plenum, New York, 1980).

With this, we have

$$\begin{aligned} Z_{\theta}^{(0)}[\eta, \bar{\eta}] &= \mathcal{N} \int (\mathrm{D}G) \exp\{-\frac{1}{4}(\mathfrak{F}, \mathfrak{F}) - i(\theta - 2\alpha N_f)q[G]\} \\ &\quad \times f_0[G, \eta', \bar{\eta}']. \end{aligned} \quad (12.144)$$

Hence

$$Z_{\theta}^{(0)}[\eta, \bar{\eta}] = Z_{\theta-2\alpha N_f}^{(0)}[\eta', \bar{\eta}']. \quad (12.145)$$

This is also valid for a finite global chiral transformation, because the α 's corresponding to successive infinitesimal transformations are additive, owing to the group property of chiral transformations. Therefore, we can always reduce θ to zero by making a global chiral transformation on the fermion sources, with $\alpha = \theta/2N_f$. Such a chiral transformation does not change the physics, because all Green's functions are evaluated in the sourceless limit. Thus, all θ -worlds are physically equivalent to one another, when the theory contains massless quarks. In particular, the theory is invariant under CP, in contradistinction to quarkless QCD with $\theta \neq 0$.

Any quark field left out in the above discussion has no effect on the conclusion, for they are spectators that do not participate in the chiral transformation. Thus, for the θ -worlds to "collapse", thereby restoring CP, it is sufficient to have one massless quark.

4 Quark Mass Matrix

We now take into account the mass term

$$\mathcal{L}_{\text{mass}} = -\bar{\psi} M \psi. \quad (12.146)$$

The partition function becomes

$$\begin{aligned} Z_{\theta}[\eta, \bar{\eta}; M] &= \mathcal{N} \int (\mathrm{D}G) \exp\{-\frac{1}{4}(\mathfrak{F}, \mathfrak{F}) - i\theta q[G]\} f_M[G, \eta, \bar{\eta}], \\ f_M[G, \eta, \bar{\eta}] &= \int d\mu \exp[-(\bar{\psi}, (\mathcal{D} + M)\psi) - (\eta, \bar{\psi}) - (\bar{\eta}, \psi)]. \end{aligned} \quad (12.147)$$

The angle θ can again be absorbed into the fermion sector, through the chiral transformation (12.141); but now it has a physical presence in the mass matrix. The relation (12.142) is now replaced by

$$f_M[G, \eta, \bar{\eta}] = e^{2i\Delta} f_{M'}[G, \eta', \bar{\eta}'], \quad (12.148)$$

where M' is the transformed mass matrix defined by

$$M' = M e^{-2i\alpha\gamma_5}. \quad (12.149)$$

Thus,

$$Z_{\theta}[\eta, \bar{\eta}; M] = Z_{\theta-2\alpha N_f}[\eta', \bar{\eta}'; M']. \quad (12.150)$$

By choosing $\alpha = \theta/2N_f$, we have

$$Z_\theta[\eta, \bar{\eta}; M] = Z_0[\eta', \bar{\eta}'; M']. \quad (12.151)$$

Since η' , $\bar{\eta}'$ are arbitrary sources, the physical meaning of θ resides solely in the θ -dependence of M' .

To parametrize the mass matrix, it is convenient to decompose ψ into its right and left-handed components R and L . The most general mass term can be written in the form

$$\mathcal{L}_{\text{mass}} = -\bar{L}\mathcal{M}R' - \bar{R}\mathcal{M}^\dagger L, \quad (12.152)$$

where \mathcal{M} is an arbitrary complex 3×3 matrix. The mass matrix then takes the form

$$M = A + i\gamma_5 B, \quad (12.153)$$

where A and B are hermitian matrices given by

$$\begin{aligned} A &= \frac{1}{2}(\mathcal{M} + \mathcal{M}^\dagger), \\ B &= \frac{i}{2}(\mathcal{M} - \mathcal{M}^\dagger). \end{aligned} \quad (12.154)$$

It is clear that the transformed mass matrix M' is also of the form (12.153):

$$M' = A' + i\gamma_5 B', \quad (12.155)$$

where A' and B' are 3×3 hermitian matrices.

The symmetry group $[SU(3) \times U(1)]_V \times [SU(3) \times U(1)]_A$ of the unperturbed problem is equivalent to $[SU(3) \times U(1)]_L \times [SU(3) \times U(1)]_R$, which acts in the following manner:

$$\begin{aligned} [SU(3) \times U(1)]_L: L &\rightarrow e^{-i\omega_a \lambda_a/2 - i\delta} L, \\ [SU(3) \times U(1)]_R: R &\rightarrow e^{-i\rho_a \lambda_a/2 - i\epsilon} R. \end{aligned} \quad (12.156)$$

If there were no restrictions on these transformations, then through them we can make \mathcal{M} diagonal, with non-negative diagonal elements. In particular, the term $i\gamma_5 B$ in the mass matrix, which violates CP invariance, can be transformed away. This was what we did in Chapter 6 in the context of the Weinberg-Salam model. We now recognize that there are constraints previously ignored, arising from

- (a) the existence of the topological charge,
- (b) the assumption that $[SU(3)]_A$ is spontaneously broken.

The existence of the topological charge leads to the term $i\theta q[G]$ in the gauge-field sector, and the transformation (12.148) in the fermion sector. Thus, the theory can be CP-conserving only if the value of θ is chosen in a special way in conjunction with parameters in the original mass matrix M . In general, therefore, the theory violates CP.

The constraint imposed by the spontaneous breaking of $[SU(3)]_A$ was pointed out by Dashen²⁸ and Nuyts²⁹. In the chiral limit, the vacuum is infinitely degenerate, labelled by a phase angle ξ . An $[SU(3)]_A$ transformation changes ξ , and takes one vacuum state into another. The perturbation $\mathcal{L}_{\text{mass}}$ explicitly violates $[SU(3)]_A$, and singles out a particular value of ξ . When we turn off the perturbation, the vacuum is left in the state labelled by that ξ . If we choose a vacuum with the wrong value of ξ , then $\mathcal{L}_{\text{mass}}$ cannot be considered a small perturbation, for its iterative effect will tend to rotate the vacuum to the correct one. Thus, to be able to use chiral perturbation theory, we must start with the correct vacuum, out of the infinitely many degenerate ones.

An analogy may be made with a ferromagnet. In the absence of an external field, the magnetization can point along any direction in space. The effect of an external magnetic field, however small, can be treated as a small perturbation only if its direction coincides with that of the unperturbed magnetization.

In the present context, the condition that $\mathcal{L}_{\text{mass}}$ can be treated as a small perturbation is that it must not create Goldstone bosons from the unperturbed vacuum:

$$\langle 0 | \mathcal{L}_{\text{mass}} | \pi^k \rangle = 0, \quad (12.157)$$

where π^k ($k = 1, \dots, 8$) refers to any member of the unperturbed pion octet π, η, κ . Since π^k is a pseudoscalar particle, this imposes a constraint only on the term $i\gamma_5 B'$ in (12.155):

$$\langle 0 | i\bar{\psi} \gamma_5 B' \psi | \pi^k \rangle = 0 \quad (k = 1, \dots, 8). \quad (12.158)$$

This requires $\bar{\psi} i\gamma_5 B' \psi$ to be a singlet with respect to $[SU(3)]_A$. Hence B' must be proportional to the unit 3×3 matrix:

$$B' = \omega \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}. \quad (12.159)$$

Thus

$$\begin{aligned} \bar{\psi} M' \psi &= \bar{\psi} (A' + i\gamma_5 \omega) \psi \\ &= \bar{L}(A' + i\omega) R + \bar{R}(A' - i\omega)L. \end{aligned} \quad (12.160)$$

We are no longer free to make $[SU(3)]_A$ transformations, but we can still make $[SU(3)]_V$ transformations, through which A' can be brought to diagonal form, with eigenvalues a_f ($f = 1, 2, 3$). The current quark masses are then $(a_f^2 + \omega^2)^{1/2}$ ($f = 1, 2, 3$).

To determine ω , we note that the transformation (12.149), which takes θ out of the gauge-field sector into the fermion sector, gives rise to a phase factor in the determinant of the left-handed part of M' , given by

$$e^{2i\alpha N_f} = e^{i\theta}. \quad (12.161)$$

²⁸ R. Dashen, *Phys. Rev.* D3, 1879 (1971).

²⁹ J. Nuyts, *Phys. Rev. Lett.* 26, 1604 (1971).

There may be other phase factors coming from the parameters in the original mass matrix M ; but we shall ignore them, because they merely cause a shift of θ , which is arbitrary anyway. Diagonalizing A' , and treating ω as small, we have

$$\begin{aligned}\det(A' - i\omega) &= (m_u - i\omega)(m_d - i\omega)(m_s - i\omega) + O(\omega^2) \\ &= m_u m_d m_s - i\omega(m_u m_d + m_u m_s + m_d m_s) + O(\omega^2).\end{aligned}\quad (12.162)$$

Equating the phase factor of the above to $e^{i\theta}$, we obtain

$$\omega = -\frac{\theta}{\frac{1}{m_u} + \frac{1}{m_d} + \frac{1}{m_s}} + O(\theta^2). \quad (12.163)$$

Hence

$$\mathcal{L}_{\text{mass}} = -(m_u \bar{u} u + m_d \bar{d} d + m_s \bar{s} s) + \mathcal{L}_{\text{CP}}, \quad (12.164)$$

with the CP violating term \mathcal{L}_{CP} given to lowest order in θ by

$$\begin{aligned}\mathcal{L}_{\text{CP}} &= i\zeta\theta(\bar{u}\gamma_5 u + \bar{d}\gamma_5 d + \bar{s}\gamma_5 s), \\ \zeta &= \frac{m_u m_d m_s}{m_u m_d + m_u m_s + m_d m_s}.\end{aligned}\quad (12.165)$$

If one of the quark masses vanishes, then $\mathcal{L}_{\text{CP}} = 0$, as expected. The coefficient ζ was first derived in pre-QCD language by Bég³⁰, and in the present form by Baluni³¹.

5 Strong CP Violation

The term \mathcal{L}_{CP} in (12.164) gives rise to CP violation by the strong interactions, which may be avoided in two ways:

(a) We may set one of the quark masses equal to zero, and it would be most natural to take $m_u = 0$. While not ruled out by experimental facts, such a choice is deemed extremely unpalatable from a phenomenological point of view.³²

(b) We have remarked that θ may be shifted by parameters in the original mass matrix M . By constructing M from a Higgs field, θ can be banished from QCD into the Higgs sector, with attendant delights for theorists³³. However, there is as yet no indication that such a mechanism operates in nature.

Taking \mathcal{L}_{CP} as given by (12.165), one can calculate its observable effects through the use of standard chiral perturbation theory. We merely quote certain results exact in the chiral limit.³⁴

³⁰ M. A. B. Bég, *Phys. Rev.* **D4**, 3810 (1971).

³¹ V. Baluni, *Phys. Rev.* **D19**, 2227 (1979).

³² P. Langacker and H. Pagels, *Phys. Rev.* **D19**, 2070 (1979).

³³ R. D. Peccei and H. R. Quinn, *Phys. Rev. Lett.* **38**, 1440 (1977); S. Weinberg, *Phys. Rev. Lett.* **40**, 223 (1978); F. Wilczek, *Phys. Rev. Lett.* **40**, 279 (1978).

³⁴ R. J. Crewther, P. DiVecchia, G. Veneziano, and E. Witten, *Phys. Lett.* **88B**, 123 (1979).

One of the effects of \mathcal{L}_{CP} is to induce a CP violating term in the effective pion-nucleon coupling:

$$\begin{aligned}\mathcal{L}_{\pi NN} &= g_{\pi NN}(\bar{N}\gamma_5\tau N) \cdot \boldsymbol{\pi} + g'_{\pi NN}(\bar{N}\tau N) \cdot \boldsymbol{\pi}, \\ g'_{\pi NN} &= -\theta\zeta f_\pi^{-1}(m_\Xi - m_N),\end{aligned}\quad (12.166)$$

where m_Ξ and m_N are respectively the mass of Ξ and the nucleon. Numerically,

$$\begin{aligned}|g_{\pi NN}| &\cong 13.4, \\ |g'_{\pi NN}| &\cong 0.038|\theta|,\end{aligned}\quad (12.167)$$

where (12.109) has been used to calculate ζ .

Using the effective pion-nucleon coupling, one can calculate the neutron electric dipole moment D_n exactly in the chiral limit. The physical reason is as follows: the neutron can dissociate virtually into a proton and a pion (among other possibilities). The charge separation in such a virtual state contributes to the neutron electric dipole moment, because there is a CP-violating pion-nucleon vertex. In the chiral limit $m_\pi \rightarrow 0$, this virtual state dominates over all others, because the virtual pion travels far from the virtual proton before recombining, thus maximizing the contribution to the electric dipole moment. The result of a calculation based on such a picture gives

$$\frac{D_n}{m_N} = \frac{1}{4\pi^2} g_{\pi NN} g'_{\pi NN} \ln \frac{m_N}{m_\pi}, \quad (12.168)$$

with numerical value

$$D_n = 5.2 \times 10^{-16} \theta \text{ cm}. \quad (12.169)$$

Comparison with the experimental bound

$$|D_n| < 10^{-24} \text{ cm} \quad (12.170)$$

yields

$$|\theta| < 10^{-9}. \quad (12.171)$$

CHAPTER 13

LATTICE GAUGE THEORY

13.1. Wilson's Lattice Action

In this chapter we use discrete space-time as a way to regularize a quantum field theory. Consider a large but finite 4-dimensional rectangular lattice with lattice spacing a . Lattice sites are denoted by x , y , and lattice directions are denoted by μ , ν . Without causing confusion, we can use the same symbols μ , ν to denote the corresponding unit vectors. Thus, for example, $x + \mu$ is a site displaced from x by one unit along direction μ . Since a unit vector is also a link between two sites, μ also denotes a directed link. We can uniquely associate 4 directed links with each site.

To go from Minkowski to Euclidean space-time, we put $x^0 = -ix^4$, where x^0 and x^4 are respectively the Minkowski and the Euclidean time. A similar transformation is made in the time components of all 4-vectors. The action of the continuum Euclidean field theory is given by

$$S_E(A) = \frac{1}{4} \int d^4x F_b^{\mu\nu} F_b^{\mu\nu} = \frac{1}{4\kappa} \int d^4x \text{Tr}(F^{\mu\nu} F^{\mu\nu}), \quad (13.1)$$

$$F^{\mu\nu} = F_b^{\mu\nu} L_b, \quad \text{Tr}(L_b L_c) = \kappa \delta_{bc}.$$

Here $x = (x^1, x^2, x^3, x^4)$, and there is no distinction between upper and lower indices.

An obvious way to lattice the theory is to replace derivatives by finite differences. This method, however, does not preserve gauge invariance, which is an essential attribute of the theory. We shall use a gauge-invariant lattice action introduced by Wilson.^{1,2} Recall that the gauge field enables parallel transport, which is implemented through the path representation of the gauge group G . [See (4.64).] The analog on the lattice can be constructed as follows. Let the lattice gauge field be denoted by $A_x^\mu = A_x^{a\mu} L_a$. The most elementary paths on the lattice are links, out of which any path can be constructed. With each link on the lattice we associate a group element, called a “link variable”, of the form

$$U_x^\mu = \exp(-igaA_x^\mu), \quad (13.2)$$

where the link is identified by the site x and the direction μ , and g is the gauge coupling constant. The eigenvalues of gaA_x^μ are angle variables; they lie between $\pm \pi$. The range of the eigenvalues of A_x^μ becomes infinite only in the limit $a \rightarrow 0$.

¹ K. G. Wilson, *Phys. Rev.* D10, 2445 (1974).

² See C. Rebbi, *Lattice Gauge Theories and Monte Carlo Simulations*, (World Scientific, Singapore, 1983) for a collection of reprints on lattice gauge theory.

The link variable associated with a link opposite to μ is denoted by

$$U_x^{-\mu} = (U_x^\mu)^{-1}. \quad (13.3)$$

Suppose ψ_x is a field attached to sites. Under a local gauge transformation $g_x \in G$,

$$\psi_x \rightarrow g_x \psi_x, \quad U_x^\mu \rightarrow g_x U_x^\mu g_{x+\mu}^{-1}. \quad (13.4)$$

The combination $\psi_x^\dagger U_x^\mu \psi_{x+\mu}$ is therefore gauge invariant. In this sense, the link variable transports ψ_x from one end of the link to the other end, in a "parallel" manner.

In the continuum the field strength tensor was defined in terms of an infinitesimal closed path. [See (4.68).] By analogy we define it on the lattice in terms of a "plaquette", a square bounded by four links. The field tensor $F_x^{\mu\nu}$, which is associated with the oriented plaquette specified by the links $\{\mu, \nu\}$ attached to site x , is defined through the relation

$$U_4 U_3 U_2 U_1 = \exp(-iga^2 F_x^{\mu\nu}), \quad (13.5)$$

where the U 's correspond to the links shown in Fig. 13.1, and are explicitly given by

$$\begin{aligned} U_1 &= \exp(-igaA_x^\mu), \\ U_2 &= \exp(-igaA_{x+\mu}^\nu), \\ U_3 &= \exp(igaA_{x+\nu}^{\mu}), \\ U_4 &= \exp(igaA_x^\nu). \end{aligned} \quad (13.6)$$

The plaquette with the opposite orientation is labeled by $\{\nu, \mu\}$, and corresponds to $(U_4 U_3 U_2 U_1)^{-1}$. Thus $F_x^{\mu\nu} = -F_x^{\nu\mu}$. When $a \rightarrow 0$ we recover the continuum field tensor:

$$F_x^{\mu\nu} = \frac{1}{a} [(A_{x+\mu}^\nu - A_x^\nu) - (A_{x+\nu}^\mu - A_x^\mu) + ig[A_x^\mu, A_x^\nu] + O(a). \quad (13.7)$$

The proof is entirely analogous to that for (4.68).

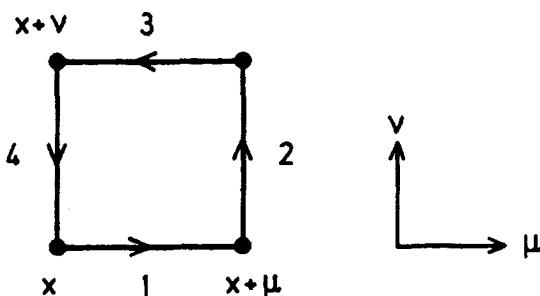


Fig. 13.1 A plaquette on a lattice.

To construct the lattice action, note that, from (13.5),

$$\sum_x \sum_{\mu, \nu} \text{Tr}(U_4 U_3 U_2 U_1) = \sum_x \sum_{\mu, \nu} \text{Tr} \left[1 - ig a^2 F_x^{\mu\nu} - \frac{1}{2} g^2 a^4 (F_x^{\mu\nu})^2 + \dots \right], \quad (13.8)$$

where μ, ν are summed from 1 to 4. The first term on the right side is a constant, and the second term vanishes because $F_x^{\mu\nu}$ is antisymmetric in μ, ν . Note that the sum over x, μ, ν on the left side is twice the sum over all plaquettes, because interchanging μ and ν changes the orientation of the designated plaquette, but the summand is independent of the orientation. Thus we have

$$\sum_P \text{Tr } U_P \xrightarrow[a \rightarrow 0]{} -\frac{1}{4} g^2 \int d^4x \sum_{\mu, \nu} \text{Tr}(F_x^{\mu\nu})^2 + \text{const.}, \quad (13.9)$$

where $U_P = U_4 U_3 U_2 U_1$ is the product of link variables around the plaquette P , and the sum extends over all plaquettes in the lattice. In this sum, plaquettes of opposite orientations are not counted as distinct. Comparing this with (13.1), we see that the lattice action can be chosen to be

$$S(U) = -\frac{1}{g^2 \kappa} \sum_P \text{Re Tr } U_P, \quad (13.10)$$

which is clearly gauge invariant. Lorentz invariance, however, is necessarily lost, and one hopes to recover it in the continuum limit.

The quantized theory is defined through a Feynman path integral, using the classical action obtained above. We impose definite boundary conditions on the bounding surfaces. That is, the link variables on all the surfaces are fixed. The partition function, or vacuum to vacuum amplitude, is given by

$$Z = \int dU e^{-S(U)}, \quad dU = \prod_{x, \mu} d\mu(U_x^\mu), \quad (13.11)$$

where $d\mu(U)$ is the invariant group measure at the group element U . Thus, a configuration for the field is specified by assigning a group element to each link in the lattice, and the integration over configurations consists of independent group integrations for each link. The invariant group measure insures the gauge invariance of Z .

The integration in (13.11) goes over redundant regions, because of the gauge invariance of the action; but, for a finite lattice, the partition function is finite for compact groups such as $SU(N)$. There is thus no necessity for gauge fixing. Of course, the partition diverges in the infinite-volume limit, as in statistical mechanics; but that does not affect averages of physical quantities.

In some applications, gauge fixing is called for. In that event we may proceed as follows. To completely fix the gauge, consider an open path made up of the ordered links l_1, \dots, l_n , where $l_n \neq l_1$. Let g_i denote a gauge transformation made at the end site of l_i . We can set the link variables on the path to arbitrary values by successively choosing g_1, g_2, \dots, g_n appropriately. This would have been impossible if the path were closed, for then g_n would affect both l_n and l_1 . By an obvious extension, we can set the link variables on any tree-like set of

links, i.e., a set of links containing no closed loops. A “maximal tree” is a tree with the property that the addition of one more link will cause the appearance of a closed loop. Gauge fixing to a maximal degree can be done by fixing the link variables belonging to a maximal tree. The links not in the maximal tree are called free links, and the corresponding link variables are the only ones integrated over in (13.11).

As an example of gauge fixing, choose one of the axes of the Euclidean lattice as the “time” axis. A maximal tree consists of all “worldlines” that start from the initial time, and terminate one time step short of the final time. This is illustrated in Fig. 13.2. The free links are those in the last time step, plus all space-like links not on the boundary. Clearly, adding any free link to the tree produces a closed loop. This example corresponds to temporal gauge, or axial gauge if we rename our “time” direction as a spatial direction.

To close this section, we emphasize that the basic variables on the lattice are the link variables, which are elements of the gauge group. In contrast, the basic variables of the continuum theory, $A^\mu = A_a^\nu L_a$, are elements of the Lie algebra. Thus, we can consider discrete gauge groups on the lattice, but not in the continuum formulation. More significantly, this points up the fact that the lattice theory has a richer content. For example, the groups $SU(2)$ and $SO(3)$ both share the same Lie algebra, but they are quite different groups: the former is simply-connected while the latter is multiply connected. These groups have the same local properties, but differ in their global properties. This difference is reflected in the fact that there are topological solitons in the $SO(3)$ theory but not in the $SU(2)$ theory. Numerical studies indicate that, for finite lattice spacing, the $SO(3)$ theory has a phase transition (as the gauge coupling constant is varied) caused by a condensation of topological solitons, whereas the theory based on $SU(2)$ does not have such a transition.³ It is true that the $SU(2)$ and $SO(3)$ theories both approach the same theory in the naive continuum limit $a \rightarrow 0$. However, the physically relevant continuum limit involves renormalization, and it is not at all obvious that the difference in phase structure should disappear in that limit.

13.2 Transfer Matrix

We shall derive the Hamiltonian for the lattice gauge field from the action given in the last section. Before we do this, however, let us illustrate the procedure with a simple example—the one-dimensional harmonic oscillator. Consider discretized pure-imaginary time, with lattice spacing a . The coordinate at time t is denoted by q_t , which, for simplicity, is assumed to be periodic: $q_{N+1} \equiv q_1$. The partition function is

$$Z = \int dq_1 \cdots dq_N e^{-S(q)}, \quad S(q) = \frac{a}{2} \sum_{t=1}^N \left[\left(\frac{q_{t+1} - q_t}{a} \right)^2 + q_t^2 \right]. \quad (13.12)$$

³ I. G. Halliday and A. Schwimmer, *Phys. Lett.* **101B**, 327 (1981).

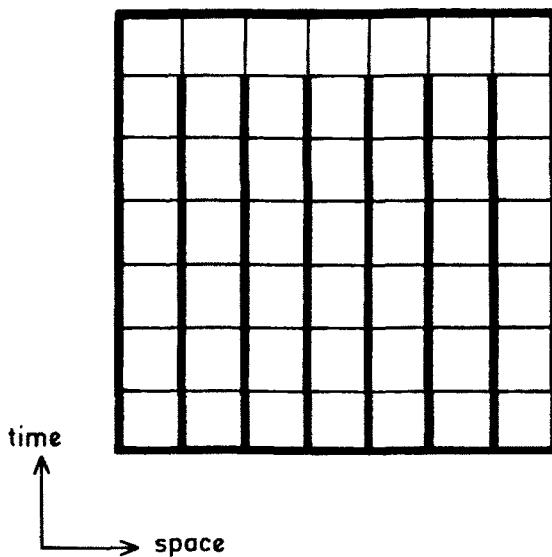


Fig. 13.2 A “maximal tree” on which the link variables are fixed, giving a maximal degree of gauge-fixing.

The fact that only successive times are coupled together makes it useful to define a “transfer matrix”:

$$\langle q' | T | q \rangle \equiv \exp \left[-\frac{1}{2a} (q' - q)^2 - \frac{a}{2} q^2 \right]. \quad (13.13)$$

We can then write

$$\begin{aligned} Z &= \int dq_1 \cdots dq_N \langle q_{N+1} | T | q_N \rangle \cdots \langle q_2 | T | q_1 \rangle \\ &= \int dq_1 \langle q_1 | T^N | q_1 \rangle = \text{Tr } T^N. \end{aligned} \quad (13.14)$$

As the notation implies, we have introduced a Hilbert space spanned by $|q\rangle$, which are eigenstates of a coordinate operator \hat{q} , with the usual properties:

$$\hat{q}|q\rangle = q|q\rangle, \quad \langle q'|q\rangle = \delta(q' - q), \quad \int dq |q\rangle\langle q| = 1. \quad (13.15)$$

We also define a momentum operator \hat{p} , such that

$$[\hat{p}, \hat{q}] = -i, \quad |q + r\rangle = e^{-i\hat{p}r}|q\rangle. \quad (13.16)$$

The transfer operator \hat{T} , which takes the system from one time to the next, can be explicitly calculated. Using (13.15) and (13.16), we find

$$\begin{aligned}
 \hat{T} &= \int dq dq' |q\rangle\langle q| T |q'\rangle\langle q'| \\
 &= \int dq dq' |q\rangle\langle q'| \exp \left[-\frac{1}{2a} (q' - q)^2 - \frac{a}{2} q^2 \right] \\
 &= \int dq dq' |q\rangle\langle q| \exp \left[-i\hat{p}(q' - q) - \frac{1}{2a} (q' - q)^2 - \frac{a}{2} q^2 \right] \\
 &= (2\pi a)^{1/2} \exp \left[-\frac{a}{2} (\hat{p}^2 + \hat{q}^2) \right].
 \end{aligned} \tag{13.17}$$

Therefore the Hamiltonian is

$$\hat{H} = \frac{1}{2} (\hat{p}^2 + \hat{q}^2). \tag{13.18}$$

With this, we can describe the system in real time instead of pure-imaginary time.

13.3 Lattice Hamiltonian⁴

To find the Hamiltonian for the lattice gauge theory, it is necessary to forbid time-dependent gauge transformations, for otherwise the time evolution of the system will not have a unique description. To this end, we shall impose temporal gauge, $A_x^4 = 0$. This is always possible, as one can show using the same arguments as in the continuum case. [See (4.69).] In this gauge all time-like link variables are trivial: $U_x^4 = 1$.

There are two types of plaquettes: space-like ones, in which all 4 links are space-like; and time-like ones, in which two links are time-like and two are space-like. These are illustrated in Fig. 13.3. We are interested in the limit in

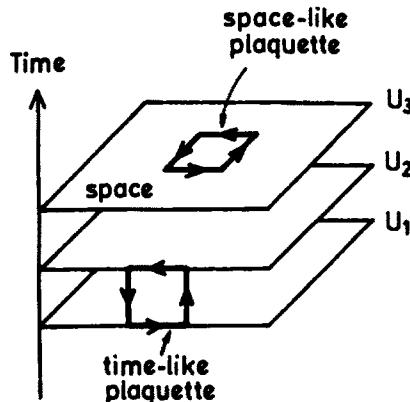


Fig. 13.3 Space-like and time-like plaquettes. U_n denotes the configuration of the system at time t_n .

⁴ J. Kogut and L. Susskind, *Phys. Rev.* D11, 395 (1975); M. Creutz, *Phys. Rev.* D15, 1128 (1977).

which the Euclidean time becomes a continuous variable. Thus we should start with a lattice whose spacing a_4 in the time direction is different from a . This means that (13.8) has to be amended, by replacing $a^2 F_x^{4\mu}$ by $a_4 a F_x^{4\mu}$. Also, $\int d^4x$ corresponds to a sum over x multiplied by $a_4 a^3$ instead of a^4 . With these modifications, the action (13.10) is replaced by

$$S(U) = -\frac{1}{g^2 \kappa} \operatorname{Re} \operatorname{Tr} \left[\frac{1}{a_4} \sum_{P \text{ time}} U_P + \frac{a_4}{a} \sum_{P \text{ s-like}} U_P \right], \quad (13.19)$$

where s denotes space, and t denotes time.

We shall introduce a new notation to facilitate the discussion of time evolution. First, divide the lattice into time slices. In one time slice, let $U(l)$ be the link variable associated with the spatial link l . The link opposite to l is denoted by $-l$, with $U(-l) = U(l)^{-1}$. The set of all spatial link variables is denoted by U :

$$U = \{\dots, U(l), \dots\}, \quad (13.20)$$

which specifies the configuration of the system on one time slice. The time evolution is described by a sequence of configurations $\{U_1, U_2, \dots\}$ for successive time slices, as depicted in Fig. 13.3. Let $\Omega[U]$ be the sum of all space-like plaquettes on a time slice, and $K(U', U)$ be the sum of all time-like plaquettes between two successive time slices (all up to a factor):

$$\begin{aligned} \Omega(U) &= \frac{1}{g^2 \kappa} \sum_{\text{plaq.}} \operatorname{Re} \operatorname{Tr}[U(l_1)U(l_2)U(l_3)U(l_4)], \\ K(U', U) &= \frac{1}{g^2 \kappa} \sum_l \operatorname{Re} \operatorname{Tr}[U'(-l)U(l)]. \end{aligned} \quad (13.21)$$

The action can now be written in the form

$$S(U) = -\sum_{t=1}^N \left[\frac{a}{a_4} K(U_{t+1}, U_t) + \frac{a_4}{a} \Omega(U_t) \right], \quad (13.22)$$

where, for simplicity, we have assumed a periodic time evolution of period N , with $U_{N+1} \equiv U_1$. The transfer matrix is

$$\langle U' | T | U \rangle = \exp \left[\frac{a}{a_4} K(U', U) + \frac{a_4}{a} \Omega(U) \right]. \quad (13.23)$$

Now we can cast the partition function in the form

$$\begin{aligned} Z &= \int dU_1 \cdots dU_N \langle U_{N+1} | T | U_N \rangle \cdots \langle U_2 | T | U_1 \rangle = \operatorname{Tr} T^N, \\ dU_t &= \prod_l d\mu(U_t(l)). \end{aligned} \quad (13.24)$$

Our goal is to define a Hilbert space, and an operator \hat{T} on this space, such that its matrix elements form the transfer matrix. (We put a caret over

all Hilbert-space operators.) It might be tempting to imitate continuum gauge theory, and take as basis states $|A\rangle$, which are eigenstates of a field operator $\hat{A}_b(l)$. The momentum operator, which generates a translation in the field variable, would then be the electric field:

$$\hat{A}_b(l)|A\rangle = A_b(l)|A\rangle, \quad \hat{E}_b(l) = i \frac{\partial}{\partial A_b(l)}. \quad (13.25)$$

This is not satisfactory, however, because the basic variables on the lattice are the link variables rather than the gauge fields. Accordingly we define a link operator

$$\hat{U}(l) = \exp[-ig a L_b \hat{A}_b(l)], \quad (13.26)$$

and use its eigenstates as a basis for our Hilbert space:

$$\begin{aligned} |U\rangle &\equiv |\dots, U(l), \dots\rangle, \\ \hat{U}(l)|U\rangle &= U(l)|U\rangle, \\ \langle U' | U \rangle &= \prod_l \delta(U'(l) - U(l)), \\ \int d\mu(U) |U\rangle \langle U| &= 1, \quad d\mu(U) = \prod_l d\mu(U(l)). \end{aligned} \quad (13.27)$$

We now turn to the momentum operator. By definition, it should generate a displacement of $U(l)$, i.e., multiplication by a group element g . The electric field operator defined earlier fits the bill, for it generates a displacement of $A(l)$, and thus has the effect of multiplying $U(l)$ by some group element. We shall rederive this result in a more general fashion. Let $\hat{R}_l(g)$ be a Hilbert-space operator that multiplies the l -th link variable by g , while leaving all others unaltered:

$$\begin{aligned} \hat{R}_l(g)|U\rangle &= |U'\rangle, \\ U'(l) &= gU(l), \quad U'(l') = U(l'), \quad (l' \neq l). \end{aligned} \quad (13.28)$$

It represents the gauge group, for

$$\hat{R}_l(g)\hat{R}_l(g') = \hat{R}_l(gg'). \quad (13.29)$$

Therefore it can be parametrized in the form

$$\begin{aligned} \hat{R}_l(g) &= \exp[i\omega_b \hat{E}_b(l)], \\ [\hat{E}_a(l), \hat{E}_b(l)] &= -iC_{abc}\hat{E}_c(l). \end{aligned} \quad (13.30)$$

The generators $\hat{E}_b(l)$ may be called the lattice electric field. To relate the parameters ω_b to g , we note that, by definition,

$$\exp[i\omega_b \hat{E}_b(l)]|\dots, U(l), \dots\rangle = |\dots, gU(l), \dots\rangle, \quad (13.31)$$

which can be true only if

$$g = \exp(i\omega_b L_b). \quad (13.32)$$

We can easily calculate the following commutators, which take the place of canonical commutators:

$$\begin{aligned} [\hat{E}_b(l), \hat{U}(l)] &= -L_b \hat{U}(l), \\ [\hat{E}_b(l), \hat{U}_b(-l)] &= \hat{U}(-l)L_b. \end{aligned} \quad (13.33)$$

For $G = SU(N)$, a local Casimir operator is

$$\hat{E}^2(l) \equiv \hat{E}_b(l)\hat{E}_b(l), \quad (13.34)$$

with the properties

$$[\hat{E}^2(l), \hat{E}_b(l)] = 0, \quad \hat{E}^2(l) = \hat{E}^2(-l). \quad (13.35)$$

We can now calculate the transfer operator:

$$\begin{aligned} \hat{T} &= \int dU dU' |U'\rangle\langle U'| T |U\rangle\langle U| \\ &= \int dU dU' \exp \left[\frac{a}{a_4} K(U', U) + \frac{a_4}{a} \Omega(U) \right] |U'\rangle\langle U|. \end{aligned} \quad (13.36)$$

Now parametrize U' by $\{\omega_b(l)\}$:

$$\begin{aligned} |U'\rangle &= \prod_l \exp[i\omega_b(l)\hat{E}_b(l)]|U\rangle, \\ U'(l) &= \exp[i\omega_b(l)L_b]U(l), \\ \int dU' &= \prod_l \int d\mu(\omega(l)). \end{aligned} \quad (13.37)$$

We then have

$$\begin{aligned} \hat{T} &= \prod_l \int d\mu(\omega(l)) \exp \left\{ i\omega_b(l)\hat{E}_b(l) + \frac{a}{a_4 g^2 \kappa} \text{Tr} \cos[\omega_b(l)L_b] \right\} \\ &\quad \times \exp \left\{ \frac{a_4}{a} \Omega[\hat{U}] \right\}. \end{aligned} \quad (13.38)$$

Now take the limit $a_4 \rightarrow 0$. The integral is dominated by the neighborhood of $\omega_b(l) = 0$. Thus we expand about this point:

$$\begin{aligned} \text{Tr} \cos(\omega_b L_b) &\rightarrow n - \frac{\kappa}{2} \omega_b \omega_b, \\ d\mu(\omega(l)) &\rightarrow \prod_b d\omega_b(l), \end{aligned} \quad (13.39)$$

where n is the dimension of the representational matrices. The last relation comes from the fact that the group measure is flat near the identity element. Putting together everything, it is a matter of doing a Gaussian integral to get the answer

$$\hat{T} = C \exp[-a_4 \hat{H} + O(a_4^2)], \quad (13.40)$$

where C is a constant, and \hat{H} is the Hamiltonian:

$$\hat{H} = \frac{g^2}{2a} \sum_l \hat{E}^2(l) - \frac{1}{g^2 \kappa a} \sum_{\text{plaq.}} \text{Re Tr}[\hat{U}(l_1)\hat{U}(l_2)\hat{U}(l_3)\hat{U}(l_4)]. \quad (13.41)$$

There is still a residual gauge freedom, which we can remove by imposing the lattice analog of Gauss' law; but we shall not dwell on it.

We remarked in Sec. 4.4 that the information carried by $F^{\mu\nu}$ is incomplete, while that carried by A^μ is redundant. The path representative of the gauge group, on the other hand, contains just the right amount of information. In the continuum, however, we have to use A^μ as coordinates, and this leads to the difficulties discussed in Secs. 8.7 and 8.8. In fact, it is not clear that there exists a consistent quantization scheme in the continuum. Here, on the lattice, we do use the path representatives—the link variables—as coordinates. The Hamiltonian formulation given above demonstrates that canonical quantization can be carried out in terms of these coordinates.

13.4 Lattice Fermions

In continuous Euclidean space-time, a fermion field $\psi(x)$ has the action

$$\int d^4x \bar{\psi}(x) (i\gamma^\mu D^\mu - m) \psi(x), \quad (13.42)$$

where D^μ is the covariant derivative, and γ^μ ($\mu = 1, 2, 3, 4$) are anti-hermitian Dirac matrices, with $\gamma^4 = i\gamma^0$. On the lattice this translates into

$$S_f = a^4 \sum_d \left[\frac{1}{2ia} \sum_\mu \bar{\psi}_x \gamma^\mu (U_x^\mu \psi_{x+\mu} - U_{x-\mu}^{-\mu} \psi_{x-\mu}) + m \bar{\psi}_x \psi_x \right]. \quad (13.43)$$

The fermion partition function is given by $\int D\psi D\bar{\psi} \exp(-S_f)$. For a periodic lattice of N lattice sites, we Fourier analyze ψ_x as follows:

$$\begin{aligned} \psi_x &= N^{-1/2} \sum_p e^{ip \cdot x} \eta_p, \\ p^\mu &= \frac{2\pi n^\mu}{L}, \quad \left(n^\mu = 0, \pm 1, \pm 2, \dots, \pm \frac{L}{2a} \right). \end{aligned} \quad (13.44)$$

Note that each component of the lattice momentum is bounded by $\pm \pi/a$. The free-particle part of the action now reads

$$S_f^{(0)} = a^{-1} \sum_p \bar{\eta}_p [\gamma^\mu \sin(p^\mu a) + ma] \eta_p, \quad (13.45)$$

from which the free fermion propagator $K(p)$ can be read off:

$$K_f^{-1}(p) = \sum_\mu \gamma^\mu \frac{\sin p^\mu a}{a} + m. \quad (13.46)$$

Within the allowed range $-\pi < p^\mu a \leq \pi$, the zeros of $\sin p^\mu a$ are 0 and π . If $m = 0$, therefore, $K_f(p)$ has $2^4 = 16$ poles, corresponding to the 4 components of p being chosen independently as 0 or π/a . If $m \neq 0$, the positions of the poles will be shifted, but there remains 16 fermions of the same mass. In general, the action S_f in D dimensions describes 2^D fermions with degenerate mass.

This “species doubling” does not happen in a scalar theory. To see this, consider the action

$$\begin{aligned} S_{sc} &= \frac{a^4}{2} \sum_x \sum_\mu (\phi_{x+\mu} - \phi_x)^2 + \frac{1}{2} m^2 \phi_x^2 \\ &= \sum_k \tilde{\phi}_k \left[\sum_\mu \frac{2(1 - \cos k^\mu a)}{a^2} + m^2 \right] \tilde{\phi}_k, \end{aligned} \quad (13.47)$$

which leads to the propagator

$$K_{sc}^{-1}(k) = \sum_\mu \frac{2(1 - \cos k^\mu a)}{a^2} + m^2. \quad (13.48)$$

Since $(1 - \cos x)$ has only one root $x = 0$ in the interval $-\pi < x \leq \pi$, there is only one particle.

To remove the fermion degeneracy, Wilson modifies the action to generate a scalar-like term $(1 - \cos p^\mu a)$ in the propagator:

$$S_W = S_f + \frac{r}{2a} a^4 \sum_{x,\mu} [\bar{\psi}_x (U_x^\mu \psi_{x+\mu} + U_{x-\mu}^{-\mu} \delta_{x-\mu}) - 2\bar{\psi}_x \psi_x], \quad (13.49)$$

where r is an arbitrary real parameter. With this addition, the fermion propagator becomes

$$K_W^{-1}(p) = \sum_\mu \gamma^\mu \frac{\sin p^\mu a}{a} + m + \frac{r}{a} (1 - \cos p^\mu a). \quad (13.50)$$

For $r \neq 0$, only one root ($p^\mu = 0$) corresponds to a particle of mass m . The remaining 15 roots correspond to masses of order $1/a$. Thus, there are still 16 fermions, but the mass degeneracy is lifted. In the limit $a \rightarrow 0$, only one mass remains finite, while all others become infinite. The price for avoiding mass degeneracy is to give up chiral symmetry. While S_f is invariant under chiral transformations when $m = 0$, the term $-2\bar{\psi}_x \psi_x$ in the Wilson addition destroys the symmetry, even when $m \rightarrow 0$. Presumably chiral invariance will re-emerge in the continuum limit.

There is an alternative scheme for lattice fermions due to Kogut and Susskind,⁴ in which the “large” and “small” components of the Dirac spinor are placed on different neighboring lattice sites. This results in the so-called “staggered fermions”.

It has been suggested⁵ that species doubling arises from the need for anomaly cancellation: Since the lattice provides a chiral-invariant ultraviolet cutoff, there

⁵ L. H. Karsten and J. Smit, *Nucl. Phys.* **B183**, 103 (1981).

can be no axial anomaly. Thus, copies of fermions are required, so that their anomalies cancel. For Wilson fermions, the cancellation occurs among fermions of different masses. In the continuum limit, when some of the masses become infinite, the anomaly will re-emerge, and this is physically acceptable in a theory like QED or QCD, where there is no physical coupling to the chiral current.

The lattice fermions described here cannot be used in the Weinberg-Salam model, in which left- and right-handed particles can have different quantum numbers. For that, we need to construct lattice fermions of definite chirality, with no partner of the opposite chirality. Unfortunately, this seems to be impossible in the conventional framework. Nielsen and Ninomiya⁶ have shown, through topological arguments, that any locally gauge-invariant lattice theory with short-ranged interactions must have an equal number of left- and right-handed chiral fermions. How to put the Weinberg-Salam model on a lattice remains an open question.

13.5 Wilson Loop and Confinement

We return to pure gauge theory on a Euclidean lattice. It is natural to consider correlation functions of the type

$$G(P) = \text{Tr}\langle U(l_1)U(l_2)\dots U(l_n)\rangle, \quad (13.51)$$

where $U(l)$ is the link variable for the directed link l , (either space-like or time-like,) and $P = \{l_1, l_2, \dots, l_n\}$ is a connected directed path made up of the links indicated in the ordered list. The average is defined by:

$$\langle O \rangle = \frac{\int dU O e^{-S(U)}}{\int dU e^{-S(U)}}. \quad (13.52)$$

Without gauge fixing, $G(P)$ vanishes unless P is a closed path. This may be seen as follows. Under a gauge transformation,

$$\text{Tr}[U(l_1)\dots U(l_n)] \rightarrow \text{Tr}[g_{x_1} U(l_1) \cdots U(l_n) g_{x_n}^{-1}], \quad (13.53)$$

where x_1 and x_n are respectively the starting and ending site of the path. This quantity is not gauge invariant for an open path. Its average is therefore zero, since the integration measure and the weight factor are both gauge invariant. It is non-vanishing in a fixed gauge, and describes quark propagation. We shall discuss this in detail in the Appendix, Chapter 14.

We consider here the case of closed-loop paths. The corresponding gauge-invariant Green's functions are called "Wilson loops":

$$W(C) = \text{Tr}\langle U(l_1)U(l_2)\dots U(l_n)\rangle, \quad (13.54)$$

$$C = \{l_1, l_2, \dots, l_n\} \quad (\text{closed loop}).$$

⁶ H. B. Nielsen and M. Ninomiya, *Nucl. Phys.* **B185**, 20 (1981).

In the naive continuum limit, the above approaches the form

$$W(C) \rightarrow \text{Tr} \left\langle \mathcal{P} \exp \left[-ig \oint_C dx^\mu A^\mu(x) \right] \right\rangle, \quad (13.55)$$

where \mathcal{P} is the path-ordering operator that orders the matrices L_a in $A^\mu(x) = L_a A_a^\mu(x)$ along the path C . This makes the Wilson loop gauge invariant. Time ordering of $A^\mu(x)$ is implied in the functional integral. For a physical interpretation of the Wilson loop, we rewrite

$$\oint_C dx^\mu A^\mu(x) = \int d^4x j^\mu(x) A^\mu(x), \quad (13.56)$$

where $j^\mu(x)$ is a c -number function that vanishes everywhere except along the curve C , on which it has a δ -function singularity. It may be regarded as the current density of non-dynamical particles, whose only role is to act as sources for the gauge field. We call these particles “external quarks” in general. In a $U(1)$ gauge theory we would call them external electrons. In a time-sliced view of the lattice, the external quark current describes the creation of a quark-antiquark pair at an initial Euclidean time, their subsequent propagation, and eventually annihilation at a later time. This process is illustrated in Fig. 13.4. The exponent in (13.55) is the action coming from the interaction of the gauge field with the external quarks. Thus, $W(C)$ gives the relative probability amplitude for the process described above, as a function of the shape of C .

For very large loops, $\ln W(C)$ generally exhibits two types of behavior: It decreases either as the perimeter or the area of C . In the former case, expansive loops such as C_1 in Fig. 13.3 are allowed, and quark and antiquark can be far apart from each other. In the latter case, thin loops such as C_2 in Fig. 13.4 are favored, and quark and antiquark propagate as a bound state. This is the Wilson criterion for confinement⁷:

$$W(C) \sim \begin{cases} e^{-KL(C)}, & \text{no confinement;} \\ e^{-K\Sigma(C)}, & \text{confinement.} \end{cases} \quad (13.57)$$

Here, $L(C)$ is the perimeter of C , $\Sigma(C)$ is the minimal area enclosed by C , and K is a constant. The above are statements about the response of the pure-gauge vacuum to external perturbations.

Consider the rectangular loop C_3 in Fig. 13.4, with sides L and T in the space and time directions, respectively. In the confining regime, we have

$$W(C_3) \sim e^{-KLT}. \quad (13.58)$$

This describes the propagation of an external static quark-antiquark pair in Euclidean time. We can interpret the exponent as $-ET$, where E is the potential energy of the pair:

⁷ K. G. Wilson, *op. cit.*

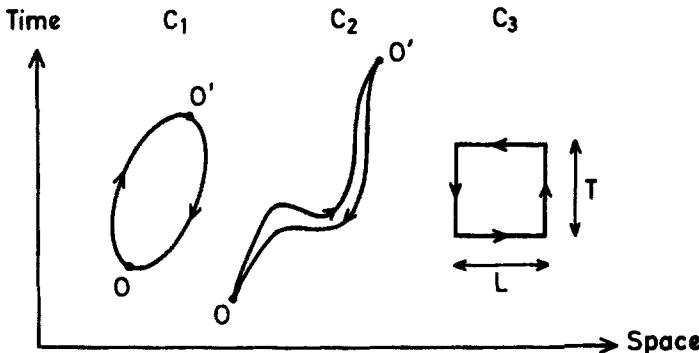


Fig. 13.4 Wilson loops of various shapes.

$$E(L) = KL. \quad (13.59)$$

The coefficient K is the force between the pair, and is independent of separation. In the confinement regime, then, a static external quark-antiquark pair behaves as if it were held together by a string of tension K . For dynamical quarks, the string can break, owing to the possibility of pair creation. When that happens, the original quark-antiquark pair becomes two pairs. The idea of a static potential breaks down when dynamical quarks are included, and the Wilson criterion becomes inapplicable. One hopes, however, that if a pure-gauge theory is confining according to the Wilson criterion, then it is an indication that dynamical quarks will be confined when they are introduced. The Wilson criterion also inspires physical pictures of the QCD vacuum, which we shall discuss in the next chapter.

The motivation for the Wilson criterion came from the strong-coupling expansion on the lattice. Expanding the action in inverse powers of the coupling constant g , we have

$$W(C) = \frac{1}{Z} \int dU \text{Tr}[U(l_1) \dots U(l_n)] \sum_{k=0}^{\infty} \frac{1}{k!(g^2 \kappa)^k} \left(\sum_P \text{Re Tr } U_P \right)^k. \quad (13.60)$$

A typical term in the above expansion involves the group integral

$$\int dU [U(l_1) \dots U(l_n)] [U_{P_1} U_{P_2} \dots U_{P_k}], \quad (13.61)$$

where $\{P_1, \dots, P_k\}$ labels a set of k plaquettes. The link variable on each link is being integrated over the group independently. For each link, the invariant group integration $\int dU f(U)$ gives zero unless $f(U)$ belongs to the singlet representation, i.e., unless it is a multiple of the identity matrix. Thus, given the curve C , the integral (13.61) is zero unless the set of plaquettes $\{P_1, \dots, P_k\}$

forms a surface whose boundary coincides with C , but runs in the opposite sense. That is, the links on the boundary of this surface must be the ordered set $\{-l_n, -l_{n-1}, \dots, -l_1\}$. The contributions from links not on the boundary cancel among themselves. This is illustrated in Fig. 13.5. There are many such surfaces, and at least one will have the smallest surface area $\Sigma(C)$. If the minimal surface is sufficiently large, the number of plaquettes contained in it will be

$$k_0 = \frac{\Sigma(C)}{a^2}. \quad (13.62)$$

This number is also the minimum order to which we must expand, before we get a non-vanishing contribution. Therefore

$$W(C) \sim (g\kappa)^{-2k_0} = e^{-K\Sigma(C)}, \quad K = \frac{2}{a^2} \ln(\kappa g), \quad (g \gg 1). \quad (13.63)$$

If the strong-coupling expansion converges, the higher terms in the expansion can be neglected as $\Sigma(C) \rightarrow \infty$. The expansion in powers of g^{-1} presumably has a radius of convergence. Assuming this radius is not zero, we will always have confinement on the lattice for a sufficiently large g . When g decreases to the radius of convergence, higher-order terms will become important and invalidate the area law, and a perimeter law is expected to take over. The argument for this behavior will be given in the next chapter. Using the language of statistical mechanics, we call the change at the radius of convergence a ‘phase transition’ between a strong-coupling (high-temperature) and a weak-coupling (low-temperature) regime.

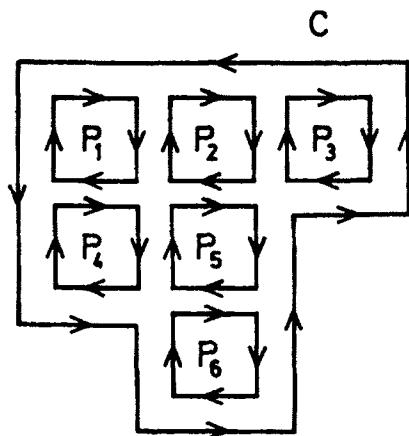


Fig. 13.5 Plaquettes forming a surface spanning the closed contour C .

For $U(1)$ gauge theory, it can be shown that a phase transition exists.⁸ Since the electron is not confined, one would presume that the fine-structure constant lies in the weak-coupling region. For the non-Abelian case, there are no proven results. Early numerical studies are consistent with the assumption that there is no phase transition for $SU(2)$ and $SU(3)$.⁹ If the assumption is correct, it would mean that one is always in the confining phase.

13.6 Continuum Limit

The physical continuum limit is different from the naive limit $a \rightarrow 0$, because one must hold the values of physical quantities fixed while letting $a \rightarrow 0$. To do this, we measure a physical quantity of dimension length. For definiteness, let us consider the correlation length ξ between two plaquettes. The lattice spacing enters only as a scale factor:

$$\xi = a\xi_{\text{latt}}, \quad (13.64)$$

where ξ_{latt} is the correlation length measured in lattice units, and is dimensionless. Since ξ is a physical quantity, it must remain finite when $a \rightarrow 0$, and this is possible only if

$$\xi_{\text{latt}} \xrightarrow[a \rightarrow 0]{\longrightarrow} \infty. \quad (13.65)$$

Thus, the continuum limit is signaled by the divergence of a physical length, when it is measured in lattice units. Equivalently, we can define a mass in lattice units:

$$M_{\text{latt}} = \xi_{\text{latt}}^{-1}, \quad (13.66)$$

which vanishes in the continuum limit.

We can calculate ξ_{latt} as a function of the bare coupling g as follows. The plaquette-plaquette correlation function $G(r)$ should have an asymptotic form

$$G(r) \xrightarrow[r \rightarrow \infty]{} e^{-r/\xi}. \quad (13.67)$$

In lattice units, we have $r = na$, and $r/\xi = n/\xi_{\text{latt}}$. Therefore

$$\xi_{\text{latt}} = - \lim_{n \rightarrow \infty} \frac{n}{\ln G(na)}. \quad (13.68)$$

By definition, $G(na)$ is the correlation at n lattice spacings. It does not depend on a , whose presence is just a convention of writing. Thus, ξ_{latt} depends only on the bare coupling constant g . The continuum limit corresponds to the critical value $g = g_c$ at which $\xi_{\text{latt}} \rightarrow \infty$. This is analogous to the divergence of the correlation length at a critical temperature in a statistical system.

The assumption that all physical quantities are finite in the continuum limit means that they depend on g in a definite way. Suppose

⁸ A. Guth, *Phys. Rev.* D21, 2291 (1980).

⁹ See Ref. 2.

$$\xi_{\text{latt}} = f(g). \quad (13.69)$$

We shall show that the g -dependence of any other physical quantity is determined by $f(g)$. Let Q be a physical quantity of dimension (length) d . Putting $Q = a^d Q_{\text{latt}}$, we have

$$Q \xi^{-d} = Q_{\text{latt}} \xi_{\text{latt}}^{-d}. \quad (13.70)$$

Assuming that $Q \xi^{-d}$ is finite in the continuum limit, we must have

$$Q_{\text{latt}} = C[f(g)]^d, \quad (13.71)$$

where C is a constant.

From the strong-coupling expansion we obtained the string tension K in (13.63). Using the fact that K has dimension (length) $^{-2}$, we obtain

$$f(g) = \frac{1}{(a^2 K)^2} = \frac{C}{\ln^2(\kappa g)}, \quad (g \gg 1). \quad (13.72)$$

In the weak coupling region, we have used perturbation theory to calculate the β -function, and obtained the coupling constant [Eq. (12.56)]:

$$\frac{g^2}{4\pi} = \frac{1}{\gamma_0 \ln(M^2/\Lambda^2)}, \quad \gamma_0 = \frac{33}{12\pi}. \quad (13.73)$$

where M is a mass scale. From this formula we obtain

$$f(g) = C' \exp[1/(2\gamma_0 g^2)] = C' \exp[8\pi^2/(11g^2)], \quad (g \ll 1). \quad (13.74)$$

Since $f(g)$ diverges at $g = 0$, the continuum limit occurs at $g_c = 0$. A qualitative sketch of $f(g)$ is shown in Fig. 13.6. We see that the continuum limit is very far from the strong-coupling regime, where confinement can be established. As we have mentioned, numerical studies indicate that no phase transition separates the two regimes, but the evidence is not conclusive.

13.7 Monte Carlo Methods

One of the advantages of lattice gauge theory is that it can be simulated through Monte Carlo methods. To discuss the procedure, it is convenient to use the language of statistical mechanics. Let us put $S(U) = \beta H(U)$, where $\beta = 1/g^2$, and regard $H(U)$ as a Hamiltonian. Consider a statistical ensemble, in which the configurations U are distributed with probability $P(U)$. The ensemble average of $O(U)$ is then given by

$$\langle O \rangle = \int dU O(U) P(U). \quad (13.75)$$

The equilibrium ensemble is the canonical ensemble, characterized by the Boltzmann distribution:

$$P_0(U) = \frac{1}{Z} e^{-\beta H(U)}, \quad Z = \int dU e^{-\beta H(U)}. \quad (13.76)$$

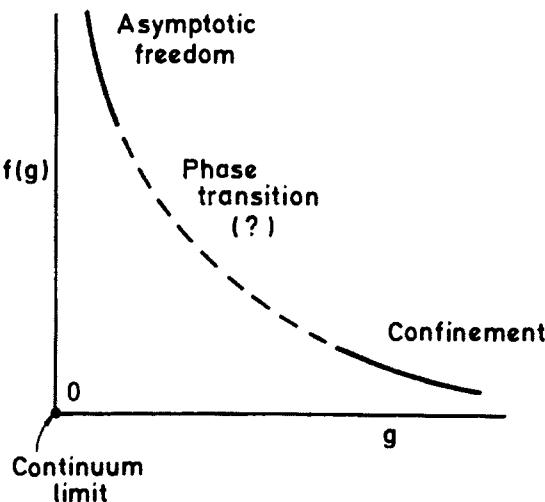


Fig. 13.6. Behavior of the scaling function (13.69) in QCD, as a function of the bare coupling constant g .

In the Monte Carlo method, our goal is to generate a “time” sequence of configurations, such that, after a sufficiently long time (the warmup time), the configuration U occurs with probability $P(U)$. Thus, the time average of any quantity, taken after the warmup time, will be the same as its average over the canonical ensemble. This is illustrated in Fig. 13.7.

The time sequence is generated by a stochastic process: A configuration U is updated to U' with a transition probability $T(U', U)$, which has the following general properties:

1. $T(U', U) \geq 0$,
2. $\int dU' T(U', U) = \int dU T(U, U') = 1$, (13.77)
3. $T(U', U) e^{-\beta H(U)} dU = T(U, U') e^{-\beta H(U')} dU'$.

The first two are just properties that any probability should have. The last is known as the condition of “detailed balance”. Integrating over U' on both sides of this condition, we obtain

$$e^{-\beta H(U)} = \int dU' T(U, U') e^{-\beta H(U')}, \quad (13.78)$$

which states that the Boltzmann distribution is an eigenfunction of $T(U', U)$. This means that the transition probability preserves the equilibrium ensemble. We now show that a non-equilibrium ensemble does not evolve away from the equilibrium ensemble.

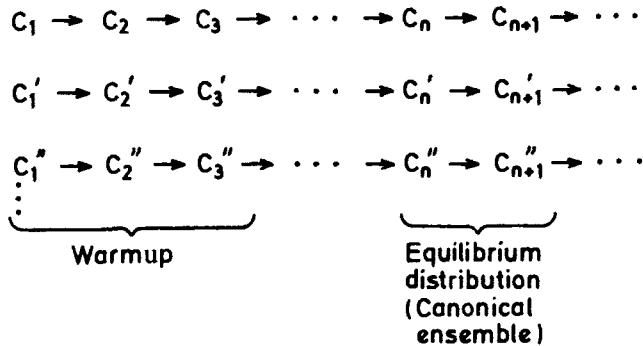


Fig. 13.7 Time sequence of configurations generated through a stochastic process. The configurations in any one sequence after a given time form an ensemble. So do the configurations belonging to different sequences at a given instant of time.

Imagine that we simultaneously generate an infinite number of sequences. The instantaneous configurations form an ensemble. (See Fig. 13.7.) We update this ensemble according to the transition probability described above. This means that a distribution $P(U)$ is replaced in the next time step by

$$P'(U) = \int dU' T(U', U) P(U). \quad (13.79)$$

Let the “distance” between two distributions P_1, P_2 be defined as

$$d(P_1, P_2) = \int dU |P_1(U) - P_2(U)|. \quad (13.80)$$

The distance between $P'(U)$ and the equilibrium distribution is then

$$\begin{aligned} \int dU |P'(U) - P_0(U)| &= \int dU \left| \int dU' T(U, C') P(U') - P_0(U) \right| \\ &= \int dU \left| \int dU' T(U, C') [P(U') - P_0(U')] \right| \\ &\leq \int dU \int dU' T(U, C') |P(U') - P_0(U')| \\ &\leq \int dU' |P(U') - P_0(U')| \end{aligned} \quad (13.81)$$

where we have used (13.78) in the second step. Thus

$$d(P', P_0) \leq d(P, P_0), \quad (13.82)$$

which means that P' can only get closer to the equilibrium distribution. Thus we expect the ensemble to approach the canonical ensemble eventually. In any single sequence, the collection of all configurations after the N -th one forms an ensemble. We expect such an ensemble to approach the canonical ensemble as $N \rightarrow \infty$.

The above proof depends only on the general properties (13.77) of the transition amplitude. In practice there is great freedom in the choice of the transition probability. We describe two popular choices that have proven to be simple and efficient for implementation on a computer.

In the so-called "Metropolis method", the transition probability is given by

$$T(U', U) = \begin{cases} 1, & \text{if } H(U') < H(U); \\ e^{-\beta[H(U') - H(U)]}, & \text{if } H(U') > H(U). \end{cases} \quad (13.83)$$

It is easily shown that this satisfies (13.77). The updating procedure is as follows. Choose a configuration at random, as candidate for the new configuration. If it lowers the energy, accept it; otherwise accept it with probability $\exp(-\beta\epsilon)$, where ϵ is the increase in energy. This simulates thermal fluctuations, which can kick a system into a state of higher energy. In practice, one updates the lattice configuration one link at a time. Thus, the computation of energy differences involves only neighboring links to the one being updated, namely those involved in plaquettes containing the link being updated.

In the "heat-bath method", a link variable is replaced by a new value whose probability is given by a local Boltzmann distribution, with the rest of the link variables held fixed. The method is so named, because in effect one thermalizes the link being updated, by touching it with a heat bath determined by the surrounding links.

In both the Metropolis and the heat-bath method, one has to generate link variables at random. In so doing, care must be taken to ensure that the link variables, which are elements of a continuous group, are chosen with uniform probability over the group manifold.

The heat bath method is equivalent to performing an infinite number of Metropolis updates on a link before passing to the next, and is in principle more efficient. In actual implementation, however, generating the local Boltzmann distribution is demanding on computer time and memory, except under special circumstances.¹⁰

¹⁰ M. Creutz, *Phys. Rev. D* **21**, 2308 (1980) gives an efficient algorithm for implementing the heat bath method for $SU(2)$ gauge theory, which relies on special properties of the group).

CHAPTER 14

QUARK CONFINEMENT

The idea of quark confinement is based on the experimental fact that quarks have not been detected in isolation, but only as constituents of hadrons. One abstracts from this the rule that all physical states are color singlets. It is widely believed that this is a consequence of QCD, and numerical studies have lent support to the belief; but as yet no proof exists. In this chapter, we present some intuitive pictures of confinement from a particular viewpoint.¹ There are many other views that we cannot discuss here. Among them are the bag model of hadrons,² the color-dielectric vacuum,³ and instantons and merons.⁴

14.1 Wilson Criterion and Electric Confinement

We start by looking at Wilson loops from different points of view. We shall use a continuum notation, assuming that all relevant length scales will be much larger than the cutoff length. Our discourse will be mainly heuristic, and renormalization will not be discussed.

In Euclidean space-time the Wilson loop operator is given by

$$W(C) = \text{Tr} \left[\mathcal{P} \exp ig \oint_C dx^\mu A^\mu(x) \right] \quad (A^\mu = A_a^\mu L_a), \quad (14.1)$$

where C is a directed closed path, and \mathcal{P} is the path-ordering operator, which orders the matrices L_a along the path. If C is not an equal-time loop, then the operators $A_a^\mu(x)$ should be time-ordered (and not path-ordered). The time ordering is automatic if we represent matrix elements of $W(C)$ by path-integrals.

The Wilson criterion is a statement about the vacuum of quarkless QCD. It says that, if the vacuum expectation of $W(C)$ behaves for large C like

$$\langle W(C) \rangle \sim e^{-K\Sigma(C)}, \quad [\Sigma(C) = \text{area enclosed by } C], \quad (14.2)$$

then a static quark and antiquark, inserted into the vacuum far apart from each other, will attract each other with a linear potential. Because of Lorentz invariance, the property (14.2) cannot depend on the choice of coordinate frames in Euclidean space-time. For a large contour C of gentle curvature, it should also be

¹ K. Huang, "Superconductivity and Quark Confinement", in *The New Aspects of Subnuclear Physics*, ed. A. Zichichi (Plenum Publishing, New York, 1980).

² A. Chodos, R. L. Jaffe, K. Johnson, C. B. Thorn, and V. F. Weisskopf, *Phys. Rev.* **D9**, 3471 (1974).

³ T. D. Lee, *Particle Physics and Introduction to Field Theory*, (Harwood Academic Publishers, Switzerland, 1981), Chap. 17.

⁴ C. G. Callan, R. Dashen, and D. Gross, *Phys. Rev.* **D22**, 2478 (1980).

independent of the shape of C . To explore the physical meaning of the Wilson criterion, let us choose C to be a large rectangle, and examine the statement (14.2) in different coordinate frames.

First suppose that the plane of C contains the (imaginary) time axis, as shown in Fig. 14.1(a). We have considered this case earlier in Sec. 13.5. In this view, the vacuum expectation $\langle W(C) \rangle$ is a Green's function describing the propagation of an electric flux tube of length L , for a time interval T . We expect

$$\langle W(C) \rangle \sim e^{-TE(L)}, \quad (14.3)$$

where $E(L)$ is the energy of the pair. The Wilson criterion states that $E(L) \propto L$, which can be interpreted to mean that there is an electric flux tube, with finite energy per unit length between the quark and antiquark. A flux tube that does not end on itself must end on equal and opposite charges (by definition). These charges are not dynamical objects, but merely points where we allow the gauge field to be singular, i.e., external sources. (If we round off the corners of the rectangle C , these point sources will be smeared out.) Thus the Wilson criterion states that a flux tube of finite constant energy per unit length is the sole carrier of color electric flux from an external quark to an external antiquark, giving rise to a linear attractive potential, or constant force, between them.

Now let us flip the rectangle C so that its plane lies in 3-space, say the x - y plane, as shown in Fig. 14.1(b). The effect of $W(C)$ on the vacuum state is to create instantaneously an infinitely thin tube of color electric flux along the closed curve C . This is because contributions to the contour integral comes only from the transverse part of \mathbf{A} , which is conjugate to the electric field. What happens to the electric flux ring? We distinguish only three cases, which are illustrated in Fig. 14.2:

- (a) The flux spreads out and diffuses over all space, as would be the case in the QED vacuum. This would yield $\langle W(C) \rangle \sim 1$.

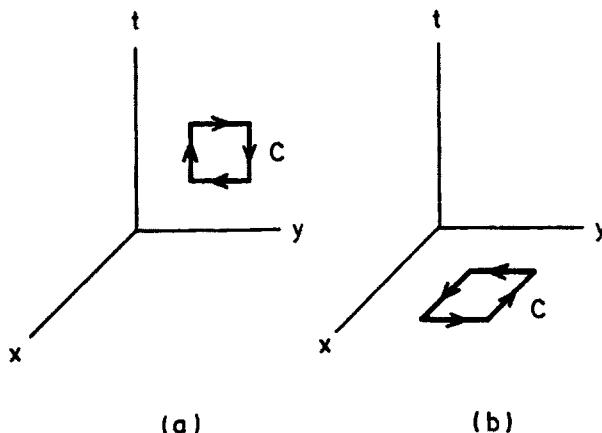


Fig. 14.1 The Wilson contour: (a) Equal-time view. (b) "Time-sliced" view.

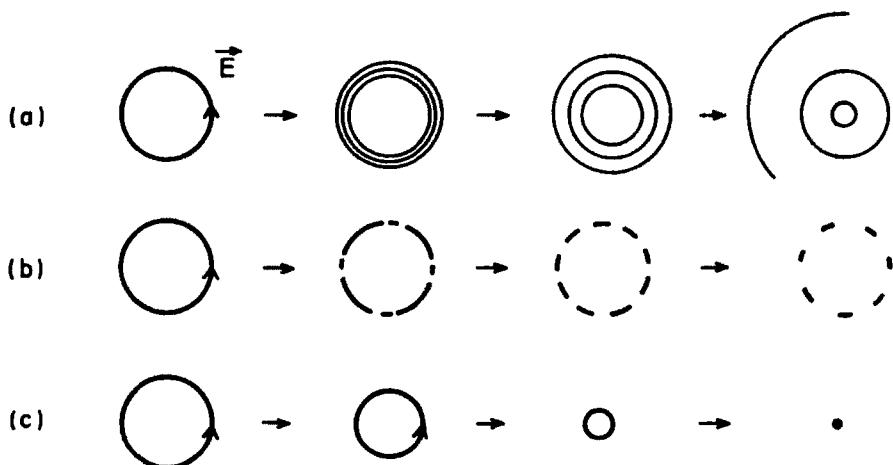


Fig. 14.2 What happens to a thin electric flux ring in a medium?: (a) In the QED vacuum it diffuses; (b) In a charge plasma it is absorbed; (c) In the QCD vacuum it contracts, presumably.

- (b) The flux quickly disappears, being locally absorbed into the vacuum. This would be the case in an electron plasma. It can also happen in the vacuum of a non-Abelian gauge field, because the latter carries charge. Assuming a constant probability for absorption per unit length of the flux ring per unit time, we expect that $\langle W(C) \rangle \sim e^{-KL(C)}$, where $L(C)$ is the perimeter of C .
- (c) The flux tends to spreads out a little, but remains confined to a thin tube. The ring of flux contracts, and eventually shrinks to nothing. This satisfies the Wilson criterion, because the probability that the flux ring will shrink to nothing should be proportional to the area swept out.

Thus, according to the Wilson criterion, quark confinement results when the vacuum confines color electric flux to a thin tube. This is reminiscent of superconductivity, except that a superconducting medium confines magnetic flux and not electric flux. We say that the QCD vacuum exhibits “electric confinement”, in contrast to “magnetic confinement” in superconductors.

The different scenarios above correspond to the possible outcomes of a competition among dynamical processes. The winner is the process that takes the shortest time to complete. There are, of course, possibilities other than those cited. For example, the cross section of the flux tube could grow at a rate comparable to the shrinking of the flux ring, and could make the Wilson loop dependent on some fractional power of $\Sigma(C)$. We do not consider these possibilities, because they involve a tie among competitors, and thus require a fine tuning of parameters of the theory. We shall give a more technical justification of the above scenarios in Sec. 14.4.

14.2 String Model of Hadrons

Accepting the existence of an electric flux tube, we can crudely model a meson as a quark-antiquark pair held together by a flux tube—a string with constant tension. This model applies to baryons as well, because, as far as color properties are concerned, an antiquark contains a two-quark component, owing to the fact that $\mathbf{3} \times \mathbf{3} = \mathbf{\bar{3}} + \mathbf{6}$. Consider two point masses m_1, m_2 , attached to the ends of a string, which is rotating with angular frequency ω . We shall let the masses go to zero, so that the ends of the string move at light speed.⁵

To work out the dynamics, consider first a relativistic point mass m moving in a circle, with constant tangential velocity v and angular frequency ω . The “string tension” is defined as the radial force in the instantaneous rest frame (or co-moving frame) of the particle:

$$T_0 = \frac{mv\omega'}{\sqrt{1 - v^2}} = \frac{mv\omega}{1 - v^2}. \quad (14.4)$$

where $\omega' = \omega/\sqrt{1 - v^2}$ is the angular frequency in the co-moving frame. In the limit $m \rightarrow 0$ at fixed ω , we have

$$\sqrt{1 - v^2} \xrightarrow[m \rightarrow 0]{} \frac{m\omega}{T_0}. \quad (14.5)$$

Consider now an infinitesimal element dx of the rotating string. In the co-moving frame its energy is $dE' = T_0 dx$. Hence in the laboratory frame the energy is

$$dE = \frac{T_0}{\sqrt{1 - v^2}} dx. \quad (14.6)$$

The moment of inertia of the string segment is $(v/\omega)^2 dE$, and its angular momentum is therefore

$$dJ = \frac{v^2}{\omega} dE = \frac{T_0 v^2}{\omega^2 \sqrt{1 - v^2}} dx. \quad (14.7)$$

A rotating string with masses m_1, m_2 and velocities v_1, v_2 attached to its ends has total mass M and total angular momentum J given by

$$\begin{aligned} M &= \frac{m_1}{\sqrt{1 - v_1^2}} + \frac{m_2}{\sqrt{1 - v_2^2}} + \frac{T_0}{\omega} \int_{-v_2}^{v_1} \frac{dv}{\sqrt{1 - v^2}}, \\ J &= \frac{m_1}{\omega \sqrt{1 - v_1^2}} + \frac{m_2}{\omega \sqrt{1 - v_2^2}} + \frac{T_0}{\omega^2} \int_{-v_2}^{v_1} \frac{dv}{\sqrt{1 - v^2}}. \end{aligned} \quad (14.8)$$

In the limit $m_1 \rightarrow m_2 \rightarrow 0$, the contributions from the masses vanish, and we can deduce a relation between J and M :

$$J = \alpha' M^2, \quad \alpha' = \frac{1}{2\pi T_0}. \quad (14.9)$$

⁵ K. Johnson and C. Nohl, *Phys. Rev.* D19, 291 (1979).

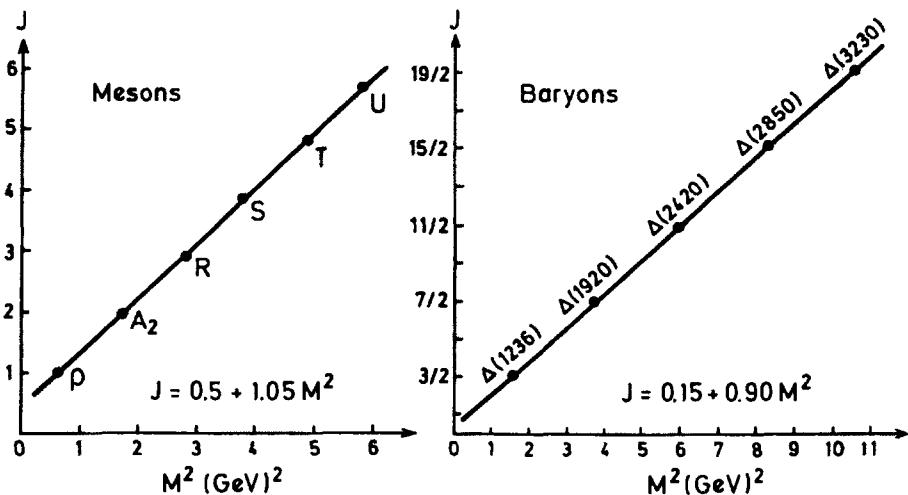


Fig. 14.3 Plots of spin vs. squared mass of hadrons. The particles lie on linear Regge trajectories, with a universal slope of $\sim 1 (\text{GeV})^{-2}$.

We can trust this model only for long strings, i.e., highly excited hadrons. Therefore, the prediction of the model is that $J = \alpha_0 + \alpha' M^2$, where α_0 may depend on the particular hadron we are modeling, but α' is a universal constant.

Experimentally, in a plot of spin vs. squared mass, all known hadrons do fall on straight lines known as "Regge trajectories", as illustrated by the examples in Fig. 14.3. These trajectories differ in intercept, but all share a universal slope

$$\alpha' \approx 0.9 (\text{GeV})^{-2}. \quad (14.10)$$

From this we can calculate the string tension:

$$T_0 \approx 1.4 \times 10^{10} \text{ dynes} = 16 \text{ tons}, \quad (14.11)$$

which is roughly the weight of two elephants.

14.3 Superconductivity: Magnetic Confinement

Superconductivity is similar to quark confinement, but with the roles of electric and magnetic fields reversed. In this sense the phenomena are dual to each other. Since superconductivity is much better understood,⁶ a review of the mechanisms involved may help us to understand quark confinement.

1 Experimental Manifestation

At temperatures below a critical temperature certain metals become superconductors. A salient characteristic is that they abhor magnetic fields, as evidenced

⁶ See P. G. de Gennes, *Superconductivity of Metals and Alloys* (Benjamin, New York, 1966).

by the Meissner effect: An external magnetic field applied to a superconducting sample is either expelled from the body of the sample, or allowed to pass through only in the form of quantized flux tubes, with magnetic flux quantum

$$\Phi_0 = \frac{2\pi\hbar c}{g}, \quad (g = 2e). \quad (14.12)$$

From the surface of the sample, the external field decays exponentially into the body, with a finite penetration depth. A non-dissipative current, the supercurrent, flows in the surface layer. The penetration depth can be viewed as the inverse mass of a photon inside a superconductor, which is therefore a medium in which local gauge invariance is spontaneously broken.

An externally imposed electric field cannot enter a superconducting body, because the field lines terminate on induced surface charges. That is, an electric field is being absorbed rather than expelled.

2 Theory

A very useful phenomenological theory was constructed by Ginzburg and Landau, which is essentially the scalar electrodynamics discussed in Chapter 3. In this model local gauge invariance is spontaneously broken due to a Higgs potential. As shown in Chapter 3, quantized magnetic flux tubes can be formed. If a magnetic monopole and an antimonopole were introduced into an infinite superconducting medium, they would be connected by a flux tube, of finite energy per unit length. Thus, magnetic monopoles are confined due to a linear potential. (If the sample were finite, then depending on energetics, the magnetic flux tube may be thrown out of the sample.)

From a microscopic point of view, there is no doubt that the basic Hamiltonian is just that of non-relativistic electrons and ions, interacting through the Coulomb potential. However, this is too complicated to deal with mathematically. Indeed, one cannot even prove that the ions form a crystal lattice. For a more manageable starting point, one assumes the existence of a lattice. It can then be shown that the electron-lattice interactions give rise to an effective attraction between electrons. However small the attraction, it leads to the Cooper instability: Two electrons near the Fermi surface will form a spin 0 bound state known as a Cooper pair, which is destined to be the carrier of the supercurrent.

The microscopic Bardeen-Cooper-Schrieffer (BCS) theory was the end product of a long chain of developments, spread over decades. It is based on an effective Hamiltonian in which all interactions are ignored, except for a simplified form of the electron-electron attraction. This interaction is in fact orders of magnitude smaller than the neglected screen Coulomb interaction between electrons; but the latter has a very different length scale, and it merely renormalizes the parameters in the BCS Hamiltonian. A variational calculation then shows that the ground state has a Bose condensate of Cooper pairs, which breaks local gauge invariance. The condensate wave function is modeled by the Higgs-like field in the Ginzburg-Landau theory.

3 Mechanism for Monopole Confinement

Suppose a hole is drilled through a superconducting body, and an arbitrary amount of magnetic flux is made to pass through it. After the transients die down, the flux Φ inside the hole will be found to be quantized. How and why will the flux readjust to a quantized value? We shall attempt to answer these questions.

For simplicity, assume that the cross section of the hole is circular. Inside the superconducting body the vector potential is pure-gauge, and non-zero:

$$\mathbf{A} = \nabla \xi, \quad \oint_C d\mathbf{x} \cdot \mathbf{A} = \Phi. \quad (14.13)$$

By cylindrical symmetry, we have

$$\xi = \frac{\Phi \theta}{2\pi}, \quad (14.4)$$

where θ is the azimuthal angle. Thus, the Schrödinger equation for a Cooper pair reads

$$-\frac{\hbar^2}{2m^*} \left(\nabla - \frac{ig}{\hbar c} \nabla \xi \right)^2 \psi = E\psi. \quad (14.15)$$

The wave function has the form $e^{in\theta} R(r)$, where n must be an integer, by continuity. Using this, we can calculate the energy

$$E = \frac{\hbar^2}{2m^*} \int d^3r \left| \left(\nabla - \frac{ig}{\hbar c} \nabla \xi \right) \psi \right|^2 = a + b \left(n - \frac{g}{\hbar c} \frac{\Phi}{2\pi} \right)^2, \quad (14.16)$$

where a, b are positive-definite radial integrals. Thus the energy can assume its minimum possible value only if

$$\Phi = \left(\frac{2\pi\hbar c}{g} \right) n, \quad (n = \text{integer}). \quad (14.17)$$

When this condition is fulfilled, the Cooper pair does not "know" that there is flux in the pipe.⁷

Because there is a macroscopic number of Cooper pairs, the total energy of the superconductor, as a function of the flux Φ , has extremely sharp minima at the quantized values. If Φ was not quantized initially, a transient flow of Cooper pairs will rapidly readjust it to a quantized value, in order to minimize the energy. Similarly, if we thrust a superconducting body into a magnetic field, transient supercurrents will bunch the flux lines into quantized flux tubes of finite energy per unit length, with diameters the order of the penetration depth. Whether these flux tubes will remain inside the body, or be expelled from the body, is again a matter of energetics, and turns out to depend on the ratio of the penetration depth to a characteristic correlation length.

⁷ N. Byers and C. N. Yang, *Phys. Rev. Lett.* 7, (1961) give a more rigorous treatment of flux quantization.

We can now describe the mechanism for monopole confinement. Imagine that a monopole-antimonopole pair was created, and then suddenly pulled apart. A magnetic flux tube will always appear between them momentarily, be the medium a QED vacuum, a normal metal, or a superconductor. The important question is what happens to the flux subsequently. In the QED vacuum or in a normal metal, the flux will spread, and eventually space will be filled by a dipole magnetic field. In a superconductor, on the other hand, the flux will remain confined to a quantized flux tube, which is ringed by a "solenoid" of supercurrent loops, as depicted in Fig. 14.4(a). (It is assumed that the monopoles have suitably quantized charges. Otherwise, the superconductor will go normal.) Since a current loop is a magnetic dipole, the response of the superconductor can also be described as a local neutralization of magnetic charge: A string of magnetic dipoles "transport" the charge of the monopole to the antimonopole, where it is finally neutralized [Fig. 14.4(b)]. Throughout the process of separating the monopole-antimonopole pair, no local magnetic charge density was ever induced in the system.

Note that magnetic confinement is a long-time phenomenon. On a short time scale, before the supercurrents have a chance to do their work, monopoles are not confined. That is, it takes time for local gauge symmetry to break. But time can be translated into inverse temperature, because a Feynman path integral over a Euclidean time interval t is the quantum partition function at temperature $1/t$. Thus, superconductivity can occur only below a certain critical temperature.

A model of linear quark confinement would result, if we could build a theory like superconductivity, but with magnetic and electric fields interchanged. This would be easy if we can merely rename the fields; but there is an absolute distinction between color electric and color magnetic field: color electric charges are defined by the generators of the color gauge group, i.e., they are what quarks carry. Thus, the problem for QCD is to show the existence of electric order in an electric system, in contradistinction to superconductivity, in which magnetic order exists in an electric system.

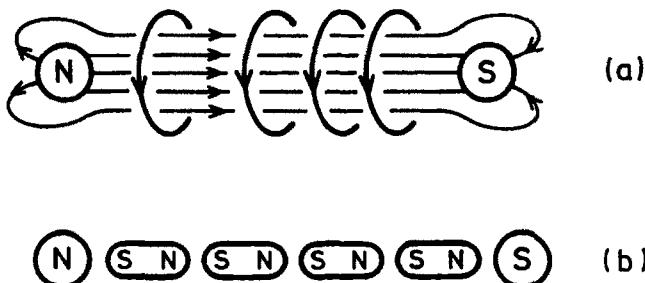


Fig. 14.4 Linear confinement of magnetic monopoles in a superconductor: (a) Formation of flux tube by supercurrent loops. (b) Equivalent string of induced magnetic dipoles.

14.4 Electric and Magnetic Order Parameters⁸

The purpose of this section is twofold: to give a more technical justification of the scenarios depicted in Fig. 14.2, and to discover a relevant symmetry for quark confinement.

To motivate our starting point, let us try to characterize superconductivity by a Wilson-like criterion. To do this we need an operator $M(C)$, which creates an infinitely thin magnetic flux tube coinciding with the closed path C . We may then say that a superconductor is characterized by the behavior

$$\begin{aligned}\langle M(C) \rangle &\sim e^{-K\Sigma(C)}, & (\Sigma(C) = \text{area enclosed by } C), \\ \langle W(C) \rangle &\sim e^{-KL(C)}, & (L(C) = \text{length of } C).\end{aligned}\quad (14.18)$$

The loop-dependent functions $\langle W(C) \rangle$ and $\langle M(C) \rangle$ are order parameters. We say that there is electric order if $-\ln\langle W(C) \rangle$ is ‘‘large’’, i.e., $O(\Sigma(C))$; and magnetic order if $-\ln\langle M(C) \rangle$ is ‘‘large’’. In this section we study some formal properties of these operators for the Abelian case.

The magnetic operator $M(C)$ can be defined more formally as follows. Let Φ denote the magnetic flux in the ring C . The vector potential in the surrounding space is pure-gauge:

$$\mathbf{A}(\mathbf{x}) = \nabla\xi_C(\mathbf{x}), \quad (14.19)$$

where $\xi_C(\mathbf{x})$ is a multi-valued gauge function, with the property that it increases by Φ whenever \mathbf{x} traces a closed path that links C once. (See Fig. 14.5.) An

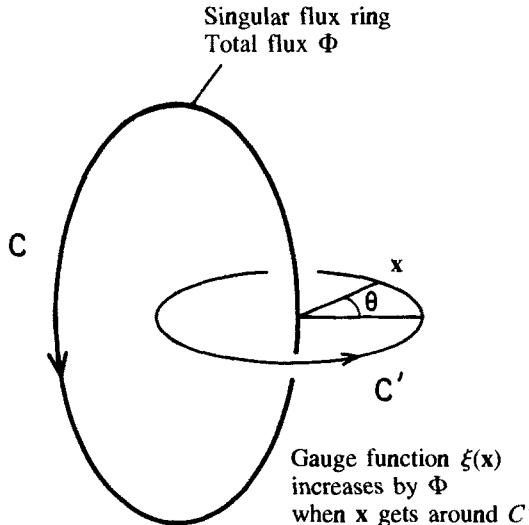


Fig. 14.5 The pure-gauge vector potential due to a magnetic flux ring C is given by $\nabla\xi$.

⁸ G. 't Hooft, *Nucl. Phys.* **B120**, 429 (1977).

explicit form of the gauge function is given in (14.14), from which we can see that, although ξ_C is multi-valued, $\nabla \xi_C$ is single-valued, and thus $\mathbf{A}(\mathbf{x})$ is single-valued. Suppose C' and C are two closed curves that are linked k times. Then

$$\oint_{C'} ds \cdot \nabla \xi_C = k\Phi. \quad (14.20)$$

Let $|\mathcal{A}\rangle$ be an eigenstate of $\mathbf{A}(\mathbf{x})$ with eigenvalue $\mathcal{A}(\mathbf{x})$. We can define $M(C)$ by

$$M(C)|\mathcal{A}\rangle = |\mathcal{A} + \nabla \xi_C\rangle. \quad (14.21)$$

Let us now consider $W(C')$ and $M(C)$, where C' and C are equal-time close curves in 3-space. We have

$$\begin{aligned} M(C)W(C')|\mathcal{A}\rangle &= \text{Tr} \left[\mathcal{P} \exp ig \oint_{C'} ds \cdot \mathbf{A} \right] |\mathcal{A} + \nabla \xi_C\rangle, \\ W(C')M(C)|\mathcal{A}\rangle &= \text{Tr} \left[\mathcal{P} \exp ig \oint_{C'} ds \cdot (\mathbf{A} + \nabla \xi_C) \right] |\mathcal{A} + \nabla \xi_C\rangle. \end{aligned} \quad (14.22)$$

Using (14.20), we obtain

$$W(C')M(C) = M(C)W(C') e^{ikg\Phi}, \quad (14.23)$$

where k is the linkage number between C' and C . Note that the flux quantization condition is just that the exponential factor on the right side is unity. Here, however, Φ must be considered arbitrary, for we are interested in the response of the system to an arbitrary flux.

Let us take the vacuum expectation of both sides of (14.23):

$$\langle W(C')M(C) \rangle = \langle M(C)W(C') \rangle e^{ikg\Phi}. \quad (14.24)$$

A strange thing happens. The linkage number k now loses its geometrical meaning, because the curves C' and C can be deformed into Euclidean space-time, in which linkage has no meaning: *Two curves linked in 3-space can be undone by continuous deformations into the fourth dimension*. To see this, consider two plane curves C and C' linked in 3-space at time $t = 0$. Suppose C' initially lies in the x - y plane. We rotate the plane of C' into the fourth dimension, until it lies in the x - t plane. Then a film sequence of three-space will look like that depicted in Fig. 14.6. The curve C' intersects three-space only

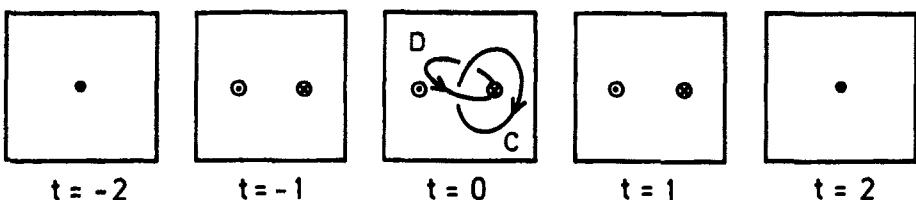


Fig. 14.6 Film sequence showing how to undo a knot by going into the fourth dimension (see text for details).

at two points on the x -axis: where it enters, as marked by \odot , and where it leaves, as marked by \otimes . These two points move apart from a single point at some negative time, and then come together again and disappear at some positive time. The curve C appears only at the instant $t = 0$. We may deform C' continuously by displacing the point \otimes along the space curve D at the instant $t = 0$. It is clear that, if we displace it out of the plane of C , even by an infinitesimal amount, then it is possible to go back continuously to an equal-time situation in which C and C' are unlinked.

Since the linkage number becomes ambiguous only when we take the vacuum expectation, dynamics must restore its meaning. One possibility is that the curves C and C' have long-ranged correlations. This would require the existence of a massless photon. In this case there is neither electric nor magnetic order.

If there are no massless photons, then there must be dynamical surfaces spanning the closed curves, such that k changes by 1 whenever C crosses the surface spanning C' , or vice versa. They would cost energy, and we assume that they cost finite energy per unit area. If C acquires such a surface, then we expect $\langle M(C) \rangle \sim e^{-K\Sigma(C)}$, or magnetic order. If C' does, then we expect $\langle W(C') \rangle \sim e^{-K'\Sigma(C')}$, or electric order. The case of simultaneous electric and magnetic order probably does not occur, because it would cost unnecessary energy. The possible phases described above corresponds to the cases depicted in Fig. 14.2. What our arguments cannot tell us is which one of these phases is actually realized.

Consider now QCD with gauge group $SU(N)$. The term "magnetic flux" has to be made more specific, because $\mathbf{B} \neq \nabla \times \mathbf{A}$. We shall define magnetic flux as the flux of $\nabla \times \mathbf{A}$, which is divergenceless. Let $|\mathcal{A}\rangle$ be the eigenstate of $\mathbf{A}(\mathbf{x}) = L_a \mathbf{A}_a(\mathbf{x})$ with eigenvalue $\mathcal{A}(\mathbf{x})$, in the $A_0 = 0$ gauge. A local gauge transformation is denoted by

$$\begin{aligned} \mathcal{A}(\mathbf{x}) &\rightarrow \mathcal{A}^U(\mathbf{x}), \quad \mathcal{A}^U(\mathbf{x}) = U(\mathbf{x})[\mathcal{A}(\mathbf{x}) + ig\nabla]U^{-1}(\mathbf{x}), \\ U(\mathbf{x}) &= \exp[-igL_a\omega_a(\mathbf{x})]. \end{aligned} \quad (14.25)$$

For a closed curve C in 3-space at a fixed time, $M(C)$ is defined by

$$M(C)|\mathcal{A}\rangle = |\mathcal{A}^{U_C}\rangle, \quad (14.26)$$

where $U_C(\mathbf{x})$ is a multi-valued gauge matrix with the following properties: If \mathbf{x} traces out a closed path C' that links C once, then

$$U_C(\mathbf{x}_f) = zU_C(\mathbf{x}_i), \quad (14.27)$$

where \mathbf{x}_f and \mathbf{x}_i are the final and initial points (which are actually the same point), and z is a matrix characterizing the magnetic flux in C . (See Fig. 14.5.) Since A should be single-valued, we must require

$$U_C \mathcal{A} U_C^{-1} = z U_C \mathcal{A} U_C^{-1} z^{-1}. \quad (14.28)$$

This can be satisfied only if z commutes with every matrix of the gauge group, i.e., it belongs to the center $Z(N)$ of $SU(N)$, the subgroup consisting of the N -th roots of unity:

$$z = e^{i2\pi n/N}, \quad (n = 0, 1, \dots, N-1). \quad (14.29)$$

The point here is that the gauge field is blind to the center, because it transforms according to the adjoint representation, which represents the center by the identity. We say that the gauge field carries no “center charge”. Quarks, on the other hand, do carry center charge, in that they are not invariant under $Z(N)$. The implications of this fact will be explored in the appendix to this chapter.

The magnetic flux Φ in the ring C created by $M(C)$ can be defined by $z = \exp(ig\Phi)$, which leads to the following possible values:

$$\Phi = \frac{2\pi n}{gN}, \quad (n = 0, \dots, N - 1). \quad (14.30)$$

In contrast, the flux in the Abelian case can be any real number. The difference comes from the fact that the center of $U(1)$ is the entire group.

Repeating the development in the Abelian case, we obtain a commutation of the same form:

$$W(C')M(C) = M(C)W(C') e^{ik\Phi} \quad (14.31)$$

where k is the linkage number of the equal-time closed curves C' and C . The value of Φ is restricted to (14.30). As in the Abelian case, this leads to the conclusion that the system may exist in phases with (a) massless photons, (b) electric order, or (c) magnetic order.

In both the Abelian and non-Abelian case, the center of the group is directly relevant to the behavior of the order parameters. In the Abelian case it happens to be the whole group, and its spontaneous breaking at low temperatures leads to superconductivity. By the dual nature of quark confinement, we expect that there is a phase transition in QCD, breaking of $Z(N)$ symmetry at high temperatures, and liberating the quarks. It can be shown that local gauge invariance cannot break down spontaneously in QCD.⁹ Thus, the center is the largest group that can break. In this view, quark confinement is the realization of a “center symmetry”, and deconfinement, its spontaneous breakdown. This will be made plausible in the appendix to this chapter.

14.5 Scenario for Quark Confinement

We now propose a qualitative picture of quark confinement, based on analogies with superconductivity. In the QCD vacuum, electric flux is confined by induced magnetic current loops, which have electric dipole moment. We therefore expect the vacuum fluctuations in QCD to be dominated by that of electric moment density, which has the form

$$\mathbf{d} = \nabla \times \mathbf{j}_m, \quad (14.32)$$

where \mathbf{j}_m is a magnetic current density. Like supercurrent loops in a superconductor, these magnetic current loops should arise from “magnetic Cooper pairs”. What are they, and what could be a mechanism for their formation?

⁹ S. Elitzur, *Phys. Rev. D* **12**, 3978 (1975).

In Abelian electromagnetism there can be no magnetic current density. In QCD, however, it may exist by virtue of the self interactions of the gluons. The non-Abelian nature of the gauge group is therefore essential. Let us first clarify the necessary transformation properties of a magnetic current under color $SU(3)$.

There are three “colors” and eight “charges”. “Color” refers to an axis in the three-dimensional vector space of the fundamental representation, and “charge” refers to a generator of the group. In the fundamental representation $\mathbf{3}$, a group element is represented by a unitary 3×3 matrix

$$U = \exp\left(\frac{i}{2} \lambda_a \omega_a\right), \quad (14.33)$$

where λ_a are the Gell-Mann matrices, and ω_a are real parameters. It effects the transformation $|\mathbf{3}\rangle \rightarrow |\mathbf{3}'\rangle$, with

$$|\mathbf{3}\rangle = \begin{pmatrix} q_1 \\ q_2 \\ q_3 \end{pmatrix}, \quad |\mathbf{3}'\rangle = U|\mathbf{3}\rangle, \quad (14.34)$$

where q_i denotes a quark of color index i ($i = 1, 2, 3$). There is a distinct conjugate representation $\bar{\mathbf{3}}$, with basis vectors

$$\langle \bar{\mathbf{3}}| = (q^1 \ q^2 \ q^3), \quad \langle \bar{\mathbf{3}}'| = \langle \bar{\mathbf{3}}| U^{-1}, \quad (14.35)$$

where q^i denotes an antiquark of color index i ($i = 1, 2, 3$). Clearly

$$\langle \bar{\mathbf{3}}| \mathbf{3}\rangle = q^i q_i \quad (14.36)$$

is invariant under U , i.e., a color singlet. A magnetic current \mathbf{j}_m transforms like $\nabla \times \mathbf{E}$, according to the adjoint representation $\mathbf{8}$:

$$\mathbf{j}'_m = U \mathbf{j}_m U^{-1}. \quad (14.37)$$

Thus it transforms like the traceless part of $|\mathbf{3}\rangle\langle \bar{\mathbf{3}}|$, i.e., the $\mathbf{8}$ in

$$\mathbf{3} \times \bar{\mathbf{3}} = \mathbf{1} + \mathbf{8}. \quad (14.38)$$

It is now clear that, as far as color transformation properties are concerned, an electric flux tube has the structure

$$|\mathbf{3}\rangle\langle \bar{\mathbf{3}}| \mathbf{3}\rangle\langle \bar{\mathbf{3}}| \dots |\mathbf{3}\rangle\langle \bar{\mathbf{3}}|. \quad (14.39)$$

It transports *color* from one end to the other, with no local color density in between. It does for color what the flux tube in superconductivity did for magnetic charge. (See Fig. 14.4.)

The really difficult problem concerns the space-time properties of the flux tube. How do we construct \mathbf{j}_m from the gauge fields? What makes the flux tube stable? We can only make some speculations¹⁰:

- (a) We have shown in (12.56) that the running coupling constant for large momentum is given by

¹⁰ For similar ideas see C. Thorn, *Phys. Rev.* D19, 639 (1979); R. Fukuda, *Phys. Rev.* D21, (1980); C. Peterson, T. H. Hansson, and K. Johnson, *Phys. Rev.* D26, 415 (1982).

$$\frac{g^2(k^2)}{4\pi} = \frac{1}{b \ln(k^2/\bar{\Lambda}^2)}, \quad (14.40)$$

where $b = 33/12\pi$, and $\bar{\Lambda} \approx 0.5$ GeV. The positivity of b is an expression of asymptotic freedom, which arises from non-linearities in the theory. If we could extrapolate this to smaller k^2 , we would find a pole of positive residue at $k = \bar{\Lambda}$, which would correspond to a quark-antiquark bound state, (and not a ghost state). We take this as an indication that there is attraction between $|3\rangle$ and $|3\rangle$, and thus also between $|3\rangle\langle 3|$ and $|3\rangle\langle 3|$. Thus, two gluons attract each other in the color octet channel.

- (b) Since gluons are massless in perturbation theory, any attraction between them will lead to an instability of the bare vacuum. Gluons will be created in order to lower the energy. If nothing counters this tendency, an avalanche of gluons will ensue, making the vacuum a condensate of multi-gluon bound states. The vacuum fluctuations would be closed gluon strings, whose color structure is represented by the trace of (14.39).
- (c) A static quark and antiquark placed in the vacuum, separated by a large distance, will cause one of these closed gluon strings to break, (since an electric field can destroy electric order,) and they will attach themselves to opposite ends of the broken string.
- (d) Asymptotic freedom indicates that quarks are free in a short time scale. Therefore, confinement is a long-time phenomenon. By the correspondence between time and inverse temperature, we expect QCD to exist in a confinement phase at low temperatures, and make a transition to a deconfinement phase at some critical temperature. This has been verified in Monte-Carlo simulations.¹¹
- (e) In superconductivity, local $U(1)$ symmetry is broken at low temperatures. Group structure and duality suggest that, in QCD, symmetry with respect to the center of the gauge group is broken in the high-temperature deconfinement phase.

We close this section by providing Table 14.1, which gives a summary of the correspondences between QCD and a superconductor.

Table 14.1 QCD AND SUPERCONDUCTIVITY

Non-linearities	\leftrightarrow	Electron-lattice interaction
Gluon-gluon attraction	\leftrightarrow	Electron-electron attraction
Instability of bare vacuum	\leftrightarrow	Instability of bare Fermi surface
Multi-gluon bound states	\leftrightarrow	Cooper-pairs
Electric flux tube	\leftrightarrow	Magnetic flux tube
Center symmetry (discrete)	\leftrightarrow	$U(1)$ symmetry (continuous)
Symmetry broken at high temp.	\leftrightarrow	Symmetry broken at low temp.

¹¹ L. D. McLerran and B. Svetitsky, *Phys. Lett.* **98B**, 195 (1981); J. Kuti, J. Polonyi, and K. Szlachanyi, *Phys. Lett.* **98B**, 199 (1981).

Appendix to Chapter 14. Symmetry and Confinement

1 Quark Propagator

The Wilson criterion is not completely satisfactory, because it is not valid in the presence of dynamical quarks. Here we shall use as criterion for confinement the vanishing of the quark propagator in a fixed gauge.

In the continuum theory, the quark propagator

$$G(x, y) = \langle 0 | T\psi(y)\bar{\psi}(x) | 0 \rangle \quad (14.A1)$$

vanishes identically in the absence of gauge fixing, because the vacuum state $|0\rangle$ is locally gauge invariant, whereas the quark operator $\psi(x)$ is not. To obtain a propagator that is not trivially zero, we could replace the above by something gauge-invariant, such as

$$G'(x, y) = \left\langle 0 \left| T\psi(y) \exp \left(-ig \int_x^y dx'_\mu A^\mu(x') \right) \bar{\psi}(x) \right| 0 \right\rangle. \quad (14.A2)$$

This, however, does not describe the propagation of a quark in the vacuum, because the line integral in the exponent introduces an external current concentrated along a path joining x and y . Even if a quark cannot propagate all alone, it can do so by going alongside the current. To study confinement, therefore, we want to study (14.A1) in a fixed gauge. Since gauge-fixing poses serious difficulties in the continuum theory, we shall use lattice regularization. We consider massive quarks, and ignore the problem of species doubling.

Consider a finite Euclidean lattice with N sites, with fixed configurations on the faces of the lattice. A further restriction is that the net baryon number should be zero at the initial time $-T$ and the final time T . Eventually we take the infinite-volume limit $N \rightarrow \infty$ and $T \rightarrow \infty$. To avoid a trivial vanishing of the quark propagator, we must fix the gauge to a maximal degree. As discussed in Sec. 13.5, this can be done by setting all link variables to unity on a maximal tree.

The quark field is denoted by $\psi_{x,s,i,f}$, where x is the site label, s the spinor index, i the color index, and f the flavor index. Labels not explicitly displayed are implied. Thus, ψ , $\underline{\psi}_x$, $\psi_{x,s,i}$ are different ways to denote the quark field. The action is $S(U) + S_f(\bar{\psi}, \psi, U)$, where the first term is the gauge-field action (13.10), and the second term is the fermion action (13.43). We rewrite the latter in the form

$$\begin{aligned} S_f(\bar{\psi}, \psi, U) &= a^4 \sum_{x,y} \bar{\psi}_x K_{xy}(U) \psi_y, \\ K_{xy}(U) &= -\not{D}_{xy} + m \delta_{xy}, \\ \not{D}_{xy} &= \frac{i}{2a} \sum_\mu \gamma^\mu [U_x^\mu \delta_{y-x,\mu} - U_y^{-\mu} \delta_{x-y,\mu}], \end{aligned} \quad (14.A3)$$

where a is the lattice spacing, M the quark mass, and δ_{xy} the Kronecker delta. The problem of species doubling will be ignored. The link variable U_x^μ , also

denoted by $U(l)$, is associated with the link l , which connects x to $x + \mu$, and $U_x^{-\mu} = (U_x^\mu)^{-1}$. Thus $\not{D}_{xy} \neq 0$ only if x and y are nearest neighbors. It describes the hopping of a quark from site to site, with a hopping amplitude proportional to $\gamma^\mu U_x^\mu$.

The quark propagator can be written as

$$G_{xy} = \frac{1}{Z} \int (DU)(D\bar{\psi})(D\psi) e^{-S(U)-S_f(\bar{\psi}, \psi, U)} \psi_y \bar{\psi}_x, \\ Z = \int (DU)(D\bar{\psi})(D\psi) e^{-S(U)-S_f(\bar{\psi}, \psi, U)}, \quad (14.A4)$$

where $\psi, \bar{\psi}$ are Grassmann variables. Only the free links are integrated over, with the invariant group measure. We can immediately integrate over the quark fields, using the method described in Sec. 7.9. [Take the functional derivative $\delta^2/\delta J(x)\delta J(y)$ of (7.152) with respect to the source function $J(x)$, and then set $J(x) = 0$.] The result is

$$G_{xy} = \frac{1}{Z} \int (DU) e^{-S(U)} [\det K(U)] [K^{-1}(U)]_{xy}. \quad (14.A5)$$

In an abbreviated notation, this is written as

$$G = \langle K^{-1} \det K \rangle. \quad (14.A6)$$

The matrix $K^{-1}(U)$ is the free propagator in a background field U . Expanding in powers of m^{-1} , we have

$$K^{-1}(U) = \frac{1}{m - \not{D}} = \frac{1}{m} \sum_{k=0}^{\infty} \left(\frac{\not{D}}{m} \right)^k. \quad (14.A7)$$

Since each factor \not{D} moves the quark exactly one lattice step, the matrix element $[K^{-1}(U)]_{xy}$ is a sum over all possible quark paths between x and y . Of course, only the terms with n not less than the minimum number of steps between x and y can contribute. For the determinantal term, we write

$$\begin{aligned} \det K &= C \det (1 - \not{D}/m) \\ &= C \exp[\ln \det(1 - \not{D}/m)] = C \exp[\text{Tr} \ln(1 - \not{D}/m)] \\ &= C \exp \left[-\text{Tr}(\not{D}/m) - \frac{1}{2} \text{Tr}(\not{D}/m)^2 - \frac{1}{3} \text{Tr}(\not{D}/m)^3 + \dots \right], \end{aligned} \quad (14.A8)$$

where $C = \text{Tr } m$ is a constant. Note that the trace operation includes a sum over sites, and hence $\text{Tr } \not{D}^n$ corresponds to a closed loop made up of n links. The k th term in the exponent is a sum over all quark paths that form a single closed loop made up of k links, and only terms with even $k \geq 4$ give non-vanishing contributions. When the exponent is expanded, we have a sum over all possible combinations of any number of closed loops. The quark propagator is therefore the sum of contributions from all quark paths connecting x to y , in a background of arbitrary distributions of closed-loop paths.

For definiteness choose a maximal tree that corresponds to axial gauge, as illustrated in Fig. 14.7. It consists of all space-like “worldlines” that emanate from one spatial boundary, and terminate one lattice step short of the opposite boundary. All time-like links not on the boundary are free links. There are two kinds of paths connecting two sites x and y : “fixed path”, and “free path”. A fixed path is one that lies entirely on the tree, with fixed hopping amplitudes. A free path is one containing at least one free link. These are illustrated in Fig. 14.7. Corresponding to this division, the propagator decomposes into two terms:

$$\begin{aligned} G &= G^{\text{fixed}} + G^{\text{free}}, \\ G^{\text{fixed}} &= [K^{-1}]^{\text{fixed}} \langle \det K \rangle, \\ G^{\text{free}} &= \langle [K^{-1}]^{\text{free}} \det K \rangle, \end{aligned}$$

where $[K^{-1}]^{\text{fixed}}$ contains contributions only from fixed paths, and $[K^{-1}]^{\text{free}}$ only from free paths. We shall show that, because of a symmetry,

$$[K^{-1}]^{\text{free}} = 0. \quad (14.A9)$$

2 Center Symmetry¹²

Let z_k belong to the center of $SU(3)$:

$$z_k = 1, e^{2\pi i/3}, e^{4\pi i/3}. \quad (14.A10)$$

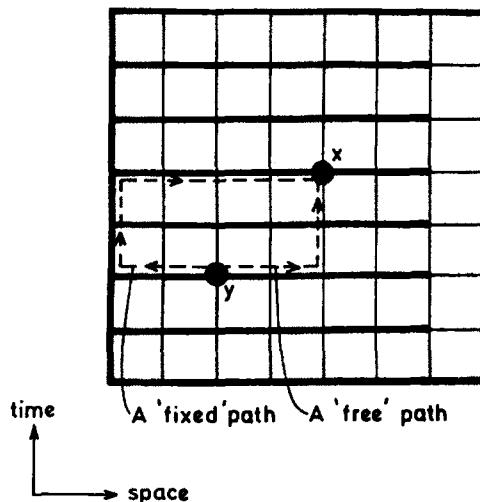


Fig. 14.7 Quark paths on a lattice in axial gauge. The link variables on light-lined links are free to vary, while those on heavy-lined links are set to unity.

¹² J. Polonyi, *Phys. Lett.* **213B**, 340 (1988).

As a convention, we shall associate a time-like link with the smaller of the times at its ends. In our gauge, all time-like links are free, and vice versa. Consider the transformation under which all free links at a particular time t are multiplied by z_k :

$$\begin{aligned} U_x^4 &\rightarrow z_k U_x^4, \quad x = (\mathbf{x}, t) \text{ (}t \text{ fixed),} \\ U_x^{-4} &\rightarrow U_x^{-4} z_k^{-1}. \end{aligned} \quad (14.A11)$$

This is not a gauge transformation, because it assigns group elements to links, not sites. If a plaquette intersects the time slice t at all, it must do so twice, in opposite directions, thus acquiring the factors z_k and z_k^{-1} . These factors cancel each other, because they commute with all group elements, and may always be moved next to each other. Thus a plaquette, and more generally any closed loop, is invariant under a center transformation. The gauge-field action is therefore invariant, except possibly for surface terms involving boundary plaquettes, which can be ignored in the infinite volume limit. Since the group measure is obviously invariant under a center transformation, we have a “center invariance”, which is a symmetry of the pure-gauge system. It arises from the fact that the gauge field has no center charge.

The quark propagator G_{xy} is obviously invariant under a center transformation at time t , as long as t does not lie between x^4 and y^4 . Consider then a center transformation with $x^4 < t < y^4$. Because we can always use zU_x^4 in place of U_x^4 as integration variable, and both the gauge-field action and the integration measure are invariant under this substitution, the quark propagator is still invariant. In fact, G^{fixed} and G^{free} are separately invariant.

We can compute G^{free} in another way. Any path joining x with y must intersect the time slice at t an odd number of times, and hence picks up a factor z_k . All other quark paths contributing to the propagator are closed loops, and therefore invariant. Thus, the transformation multiplies G^{free} by z_k . Since G^{free} is in fact invariant, we have

$$G^{\text{free}} = \frac{1}{3} \sum_k z_k G^{\text{free}} \equiv 0. \quad (14.A12)$$

because $\sum_k z_k \equiv 0$. This result also holds for an Abelian theory, because in that case the center is the whole group.

3 Confinement as Symmetry

We now argue that G_{xy}^{fixed} vanishes for an infinite lattice, in the continuum limit:

- (a) If x and y lie on the same branch of the tree, then there is a direct path connecting them along the branch (a short path). If x and y are lying on different branches of the tree, then all fixed paths have to go via a boundary (a long path). The sum of contributions from short paths are of order $\exp(-m|x - y|)$, while those from long paths are of order $\exp(-mL)$, where L is the size of the lattice, and m is the quark mass.
- (b) In the continuum limit, one must average over neighborhoods of x and y . If each neighborhood contains n lattice sites, then for large n there are the order of n short paths, and n^2 long paths. Thus the average is of order

$$\frac{1}{n^2} (n e^{-m|x-y|} + n^2 e^{-mL}), \quad (14.A13)$$

which vanishes when $L \rightarrow \infty$ (infinite volume limit), and $n \rightarrow \infty$ (continuum limit).

Thus, the gauge-fixed quark propagator is given entirely by G^{free} . Assuming that (14.A12) is valid in the infinite-volume and continuum limit, we conclude that the propagator vanishes.

Some technical points need to be mentioned. First of all, the arguments above depends on a special gauge choice. We have to assume that, if the propagator vanishes in one gauge, then it vanishes in all gauges.

Secondly, the previous arguments hold only for massive quarks. However, it is generally believed that massless quark acquire mass through spontaneous breaking of chiral symmetry. If that is the case, then the above arguments apply to massless quarks as well. This indicates a correlation between quark confinement and chiral symmetry breaking.¹³

Technical points aside, we seem to have arrived at the astounding conclusion that quark confinement is a consequence of symmetry, and therefore a matter of “kinematics”. This is untrue, because in the infinite volume and continuum limit the center symmetric can break down spontaneously, thereby invalidating (14.12). Thus, whether or not quarks are confined is a dynamical question of whether the center symmetry is realized in the unbroken or broken mode.

We can make an analogy with the Ising model, whose Hamiltonian is invariant under a global spin flip. Formally it appears that the average magnetization must be zero, because for every configuration with magnetic moment M , there is an equally probable one of magnetic moment $-M$. Nevertheless, the symmetry can be spontaneously broken, resulting in a non-zero average magnetization. Physically it comes about as follows. The attraction between like spins favors the formation of domains of organized spins, which are however vulnerable to destruction by thermal fluctuations. When the temperature is sufficiently low, large domains become stable, and so they tend to grow, until the system is almost one domain. In any finite time interval, it is very unlikely that we can witness a flipping of the magnetic moment of a large domain due to thermal fluctuations, because that would require a simultaneous spin flip involving many spins. Thus, once a large domain is formed, one will have to wait a long time to see it flip. For a finite system the average waiting time goes to infinity.

By the same token, the vanishing of the propagator, as indicated in (14.A12), is due to the cancellation of contributions from different elements of the center. An incomplete cancellation, for whatever reason, breaks the center symmetry, and liberates the quarks. The analogy with the spin system ends here, however, because the analog of the spin-spin attraction is absent. All we can rely on is the argument that quarks must be free in short times, as indicated by asymptotic freedom. This leads us to conclude that quarks are free in a high-temperature phase, in which the center symmetry must be broken, and that they are confined in a low-temperature phase that restores the center symmetry. This conclusion is consistent with the idea of duality between quark confinement and superconductivity.

¹³ A. Casher, *Phys. Lett.* **83B**, 395 (1979).

INDEX

A

abelian, also $U(1)$ 62
ABJ (Adler–Bell–Jackiw) anomaly, also axial anomaly 231
 adjoint representation,
 general 63, 64
 $SU(2)$ 13–14
 $SU(3)$ 15
 $SU(n)$ 15
 Aharonov–Bohm effect 57
 anomalous dimension 187, 188
 anomaly,
 cancellation 242–246
 pole 236, 239, 242
 anti-quark q^i 22, 24–25, 320
 anti-self dual 92
 asymptotic freedom 9, 180, 262–269
 Atiyah–Singer index theorem 281
 axial anomaly,
 analyticity (anomaly pole) 242
 calculation,
 perturbative, abelian (triangle graphs) 231–237
 non-perturbative, non-abelian, see Fujikawa’s method
 chiral invariance 230–231, 237–238
 index theorem 279–282
 physical basis 238–242
 axial gauge,
 definition 76, 153
 Fadeev–Popov gauge-fixing 163
 temporal gauge relation 161
 axial-vector, 271
 current 230

B

Baker–Hausdorff–Campbell theorem 62
 baryonic current 271
 baryons 3, 15
 baryon number B 8, 12, 14, 22, 27, 49
 BCS (Bardeen–Cooper–Schrieffer) theory, see also superconductivity 313
 β -function, also Gell–Mann–Low function 185, 197
 ϕ^4 -theory 191
 QED 185
 QCD 263–265
 BMT(Bargman–Michel–Telegdi) equation, also spin precession equation 261
 Bose symmetry 235

boson,

 fields 67
 loops 140
 boson–fermion transformation 67
 bottom quark b , see Y 8, 252
 Bott’s theorem 166
BPST (Belavin–Polyakov–Schwartz–Tyupin) instanton 92

C

Cabibbo angle, also θ_1 in K–M matrix 41, 43, 46, 117–118
 Callan–Symanzik equation 187
 canonical quantization 147–150
 Cartan’s theorem 67
 center,
 charge 319
 definition 65, 95
 symmetry 318–319, 324–325
 of $SU(2), Z_2$ 65
 of $SU(n), Z(n)$ 318–319
 charge,
 charges (per generator) 320
 color isotopic 256
 electric 12, 14, 24, 45, 109
 electron 33, 109
 form factor 180
 quantization 49
 charmonium 45
 charm quark c ,
 GIM 43–46
 J/ψ 45
 flavor 252
 chiral invariance,
 axial anomaly 230–231
 lattice fermions 298–299
 t’Hooft’s principle 247
 chirality 6, 41, 106
 chiral,
 limit 272–273, 286–287
 symmetry 269–272
 transformation,
 global 230, 282–283
 local 280
 Coleman–Weinberg mechanism, see dimensional transmutation color, see QCD, $SU(3)$ 7, 9, 30, 32
 charge 256, 320

- experimental evidence 38–39
 gyromagnetic ratio 260–262
 hypercharge 256
 singlet 30–31, 252, 308, 320
 colorless, also color singlet, color neutral 10, 30
 compact group 49, 64
 confinement,
 quark 10, 308–326
 magnetic monopole 10, 314–315, 319–321
 conserved vector current, CVC 42
 constituent quark mass 270, 272
 continuum limit 291, 303–304, 326
 contravariant 10
 Cooper pairs,
 electric (BCS) 313, 314
 Higg's mechanism 55
 magnetic (QCD) 314
 Schrödinger equation 314
 cosets, *see* factor group
 Coulomb gauge,
 definition 58, 79
 Faddeev–Popov 159–160
 geometry (relation to unitary gauge) 101
 instanton singularities 172–173
 relation to temporal gauge 149
 residual gauge freedom 149–150
 counter terms 181
 covariant derivative,
 $U(1)$ 6, 48
 Yang-Mills 69, 74
 covering group 64
 CPT (Charge conjugation-Parity-Time) 118, 119, 167
 CP (Charge conjugation-Parity) 118
 violation,
 strong (θ -angle) 284–287
 weak (δ in K–M matrix) 119
 θ -world 167
 current quark mass
 CVC (conserved vector current) 42
- D**
 decaplet 10 15–16, 27–29
 deep-inelastic scattering,
 electromagnetic 34–35
 hadronic 43
 δ , K–M matrix phase angle 119
 dimensional transmutation 223
 Dirac
 current 33
 field operator 40
 matrices
 Minkowski 11
 Euclidean 278–279
 monopole 102–103
 spinor (helicity-chirality) 41, 105–106
 string 102
 wave function 36
- direct product 62
 doubly connected, *see* $O(3)$ 65
 down quark d 8, 22, 252
 dual tensor,
 gluon $\tilde{F}^{\mu\nu}$ 276
 Yang-Mills \tilde{F}_μ^ν 73
 dynamical symmetry breaking 55
- E**
 effective action 211
 effective potential 213, 227
 electric dipole moment (neutron) 287
 electromagnetic,
 field, *see* Maxwell field
 interactions 33–35, 48–49
 electroweak,
 interactions 9, 40–42
 theory, also Weinberg–Salam model 7, 9, 105–120
 Euclidean space 132–133, 171f, 278
 Euler angles 90
- F**
 factor group,
 definition 81
 homotopy group 95
 $O(3)$ 65
 Faddeev–Popov,
 gauge fixing 152–156
 ghosts 163–164
 Fermi coupling G_F 2, 42, 111
 fermionic measure,
 Berezin integration 143
 chiral invariance, *see* Fujikawa's method
 fermion,
 fields (spinors) 67
 loops 140
 Feynman,
 CVC 42
 gauge 158
 propagator 135–137
 rules (QCD) 253–255
 field strength tensor,
 QCD $\tilde{F}^{\mu\nu}$ 252
 $U(1) F^{\mu\nu}$ 48
 Yang-Mills F_μ^ν 72
 fine structure constant ($\alpha = e^2/4\pi\hbar c$) 2, 33, 109
 fixed points, 185, 197–200
 IR 185
 UV 185
 flavor, 7–8, 22, 31
 $SU(2)_{flavor}$ 44
 $SU(3)_{flavor}$ (“quark model”) 22–28, 31–44
 $SU(4)_{flavor}$ 44, 270
 flux quantization,
 abelian or $U(1)$, Aharonov–Bohm effect 56–58
 non-Abelian 95–98

- flux tube,
 electric (color) 320
 magnetic (monopole) 315
- Fujikawa's method 279–281
- fundamental representation 13–15, 17–18, 63, 69, 252
- Furry's theorem 182
- G**
- gauge choice, also gauge fixing 76, 79–80, 153
 axial 76, 153, 161, 163
 Coulomb 79, 101, 147, 153, 159, 172
 Lorentz 50, 152, 153, 172
 temporal 76, 147, 149, 153, 156, 161, 174
 unitary 54, 82, 101, 107, 110, 153
- gauge fixing,
 Faddeev–Popov 152–156
 Feynman gauge 158
 Landau gauge 158
 lattice gauge theory 290–291
 singularities/discontinuities 150
- gauge
 coupling constant 71
 field 48
 group 6–7
 invariance, global
 abelian 47, 49
 non-abelian 67–68
 invariance, local 47
 abelian 47
 non-abelian 69–71
 orbit 154
 principle 6, 48
 pure 57, 72
 transformations 6
 global 6
 local 76, 78
 large 168, 278
 small 167
 vector bosons 6, 9–10, 83–85, 168
 weak coupling 41–42
- Gauss' Law,
 abelian (quantum) 149
 projection operator 174
 non-abelian,
 classical 7
 quantum 150, 166
- Gaussian integral 135
- Gauss' theorem, failure 173
- Gell-Mann
 CVC 42
 matrices 17, 252, 255
 – Low function, see β -function
 – Okubo mass formula 15, 30
- ghost fields, see Faddeev–Popov ghosts 163–164
 loops 142
- GIM (Glashow–Iliopoulos–Maiani) mechanism 43–45, 103
- global,
 invariance 47, 67–68
 symmetry 49
 transformations 6, 47
- gluons G_a^μ , 8, 10, 252, 321
 bound state 321
 propagator 254, 262, 263
 self-interactions 256–260
- Goldstone, boson or mode 50–53
 longitudinal massive gauge boson 55
 space 82–83
 theorem
- gravitation 2, 6f, 71
- graviton 2
- group manifold,
 $O(3)$ 65
 $SO(3)$ 291
 $SU(2)$ 65, 120, 291
 $U(1)$ 65
- Gribov ambiguity 150, 162, 172–172
- Gribov theorem 172
- H**
- hadrons, see color singlets 3–4, 12, 28, 33, 42
- Hagedorn, ultimate temperature (160 MeV) 3
- heat bath method 307
- helicity 41, 106
- Higgs, 7–8, 10
 boson field 10, 53, 80, 83–85, 110
 doublet 114, 116
 -electron coupling 111
 mass 228
 mechanism 55, 83–84, 100, 224
 mode 53–55
 singlet 54
 space 82, 83
- homotopy,
 1st homotopy group π_1 95
 2nd homotopy group π_2 95, 120
 n-th homotopy group π_n 95
- hypercharge, see also $U(1)$ 12, 23, 45, 108
- I**
- inclusive scattering 33
- index theorem, see Atiyah–Singer theorem 279–282
- independent quark model 28–30, 31
- infinite momentum frame 36
- instanton 88–94
 discontinuity/singularity, see Gribov ambiguity
 gas 169
 tunneling solution 168–70
 $U(1)_A$ 277
- interactions, 2
 electromagnetic 33
 hadronic 42, 44
 strong 9, 254, 257
 weak 40–41
- irreducible representations 12, 13f, 18, 20–23, 25

irrelevant operators 197

isospin

current 271

singlet (strange quark) 7, 45

doublet 7, 13, 42

$SU(2)$,

general 12–14

monopole trap (isospin-spin) 103

$SU(3)$ 23

$SU(6)$ 29–30

isotopic spin, isospin 61

J

Jacobi identity 62, 67

Jacobian 153, 154

J/ψ 45–46

K

kaons 43–44, 69, 119f

kink (one dimensional topological soliton) 104

Kobayashi–Maskawa (K–M) matrix 118–119

L

Landau,

diamagnetism 265, 268

gauge 158

ghost 191, 200

spectrum 267

Landau–Ginzburg 55, 313

lattice gauge theory, 288–307

action 290

chiral invariance 298–299

continuum limit 291, 303–304, 310

field tensor 289

gauge fixing 290–291

gauge invariance 288–290, 299–300

Lorentz invariance 290

phase transitions 291, 302–303

lattice fermions, 287–299

Kogut–Susskind 298

Wilson 298–299

large gauge transformation 167–168, 169, 278

Legendre transformation 211

leptons, 3–4, 8

current 46

flavor 7

number 8, 49, 108

radius 3

Lie algebra 61–62

Lie group 61

linkage number k 317–319

link variable 288, 290, 291

little group,

general 81, 94

spontaneous symmetry breaking 80–81

Weinberg–Salam model, $U(1)_{\text{em}}$ 119

local,

gauge invariance 47–48, 69–71

symmetry 61

transformation 48–49, 60–71, 76, 78

loop expansion 215–218

Lorentz gauge,

definition 50, 153

Faddeev–Popov 156–159

instanton singularity 178

propagator 152

M

magnetic flux, see flux quantization

magnetic monopole 10, 73, 94, 104, 315

mass matrix (Weinberg–Salam model), 114–119

leptons 114–117

quarks, 114

CP violation 283–286

Higgs's couplings (current masses) 116–119

mass scale 221, 224

mass,

Higgs 110–111

leptons 114–117

quark, 269–272

constituent mass 270

current mass 270

W^\pm 110–111

Z^0 110–111

matter fields 67

electromagnetic, Maxwell, photon, $U(1)_{\text{em}}$ field A^μ

classical 23, 83, 47–50

canonical quantized 147–150

gauge boson 2, 48

path integral 150–152

Meissner effect 55, 272f

mesons 3–4, 15

metric space 63

metric tensor 10, 89

Metropolis method 307

monopole–antimonopole pair 315

Monte Carlo methods 304–307

multiply connected 64

$O(3)$ 64

$SO(3)$ 291

N

Nambu–Jona–Lasinio model 270f, 272

neutral currents 9

Nielsen–Hughes formula 268–269

Noether's current 47, 48

non-Abelian, see Yang–Mills fields 62

normal subgroup 81

nucleon doublet 13

O

octets 8 15–16, 27–29

orthogonal group,

$O(3)$,

factor group 65

group manifold 65

- $O(4)$ 133
 $O(n)$ 84
orbit, gauge 154
order parameter,
 electric 316–319
 Landau–Ginzburg (superconductivity) 55
 magnetic 316–319
- P**
- parallel displacement, transport
 continuum 74
 lattice 288–284
parity non-conservation 6, 40–41, 105
parton model 35–38
path integrals,
 bosons, field theory 126–128
 fermions, field theory 142–145
 quantization,
 QED (abelian) 150–162
 Yang-Mills (non-abelian) 162–165
 quantum mechanics 121–127
Pauli matrices,
 2×2 13
 3×3 17
Pauli,
 paramagnetism 265, 268
 principle 30, 32
PCAC (Partially Conserved Axial-vector Current) 269, 272–274
PCAC, extended 275
phase transitions 301–304
photon mass 50, 208
pion,
 decay constant f_π 273
 lifetime 274
 Goldstone boson 269–275
 mass 270–273
 -nucleon coupling 287
 octet 29, 275
 triplet 13, 69
plaquette 289, 293
Poincaré invariance 2–3
point splitting 209
polarization 4-vector 260
preons 247–251
primitive divergence 181
projection operator
 Gauss' Law 174–176
propagator,
 electron 201–202
 gluon (bare) 254, 262 (renormalized) 263
 lattice fermion 287–298
 lattice scalar 298
 photon 202–205
 quark 254, 322–324
 QCD ghost 254
- Q**
- QCD, also quantum chromodynamics 252–287
 Feynman rules 253–255
 superconductor analogy 321
QED,
 canonical quantization 147–150
 path integral quantization 150–162
quark,
 color 30, 252
 confinement 10
 electric flux 308–310, 319–321
 symmetry 325–326
 Wilson loops 299–303
 elementary 12
 flavor, see SU groups 7–9, 256
 hypothesis 15, 22
 independent quark model 28–360
 irreducible representations
 masses 44, 276
 constituent 270
 current 270
 model 22–28
 spinor field 252
 triplet $\bar{3}$, $\bar{\bar{3}}$, 22
quark-gluon,
 coupling 9, 10
 interactions 256–260
quantization, see canonical quantization
path integrals
- R**
- Regge trajectory 312
relevancy 197
relevant operators 198, 199
renormalizable 189
renormalization,
 charge 177–180
 perturbative 201
 real space 196–197
renormalization group RG 183
renormalization group transformation 192–195
running coupling constant,
 RG 184
 QED 179
 QCD 262–265
 ϕ^4 -theory,
 β -function 191
 effective potential 222
- S**
- scaling form 184
scale transformation 183
scaling property 38, 205
Schwinger model 50
self-dual 92
self-energy 182

- semi-simple 62
- simple 62–63
- simply-connected 64, 95
- skeleton graph 182
- small gauge transformation 167
- solitons, 56, 86
 - static 86–88, 100
 - time-dependent 86
 - Weinberg–Salam 119
 - $U(1)$ 58–60
- species doubling 298
- spin precession, see also BMT 260–262
- spontaneous symmetry breaking, 7, 50, 80, 210
 - global, see Goldstone 50–53
 - local, see Higgs 53–55
- standard model 2
- strange quark s 22, 252
- strangeness 12
- string model, hadrons 311–312
- string tension 311–312
- strong interaction, also QCD 9, 254, 257
- structure constants,
 - Lie group 62
 - general 14
 - $SU(3)$ 17–18
- S^2 97
- S^3 89–92, 166
- S^n 95
- $SO(3)$ (special orthogonal group) 291
- SU (special unitary) groups,
 - $SU(2)$, 62
 - adjoint representation 13–14
 - center Z_2 65
 - covering group 64
 - fundamental representation 13–14, 69
 - generators 13
 - group manifold 65, 120, 291
 - multiplets 68–69
 - phase transitions 303
 - spontaneous symmetry breaking 81
 - $SU(2)_{\text{flavor}}$ 44
 - $SU(2)_{\text{weak}}$, see also Weinberg–Salam 1–8, 108
 - $SU(3)$,
 - fundamental representation 17–18
 - generators 17, 19
 - $SU(3)_{\text{color}}$ 30–31
 - $SU(3)_{\text{flavor}}$ 22–28, 31–44
 - $SU(4)_{\text{flavor}}$ 44, 270
 - $SU(6)$ 29–30
 - $SU(n)$,
 - adjoint representation 15
 - center symmetry $Z(n)$ 318–319
 - fundamental representation 15
 - generators 14
 - super-renormalizable 189
 - superstring 10
 - superconductivity, 10, 55, 312–319
 - Landau–Ginzburg 55, 319
- BCS 313
- QCD analogy 321
- symmetry class 18
- symmetry number 139–140
- T
- τ -neutrino 46
- temporal gauge,
 - definition 76, 147
 - residual gauge freedom 149–150, 174
 - relation to axial gauge 161, 174
- θ -worlds, 162, 165–173
- QCD 278–287
- θ -action 171
- top quark t 8, 10, 46, 252
- topological,
 - charge,
 - $SU(2)$ 88–90
 - QCD 277
 - soliton 94–95, 98, 103
 - stability 94
 - tunneling 169–170
- transfer matrix 291–293
- triangle graph, 231–232
- transfer operator, \hat{T} 7, 292
- triplet 3 22
- triviality 191
- t’Hooft, renormalization 105
 - Polyakov monopole 100
 - principle, see chiral invariance 247–252
 - single zero mode 289
- tunneling solution, see instanton 168–169
- U
- unitary gauge,
 - definition 82–83
 - geometry (relation to Coulomb gauge) 101
 - Higg’s singlet 54
 - Higg’s doublet 107, 110
- unitary symmetry 15
- universality 199–200
- universal slope, α' 312
- $U(1)$, also abelian
 - baryon number 12, 14, 22, 49, 271
 - charge 14
 - electromagnetic $U(1)_{\text{em}}$, see electromagnetic field
 - gauging 6, 47
 - group manifold 65
 - hypercharge,
 - strong 12, 23, 45
 - weak $U(1)_{\text{weak}}$, see also Weinberg–Salam 7–8, 108
 - lepton number 49, 108
 - little group 81
 - phase transitions 303
 - strangeness 14, 45
- $U(1)_A$ 271, 274–278
- $U(1)_V$ 271
- up quark u 22, 252

Y, see also *b quark* 45

V

$V-A$ coupling 6

vector bosons, *see* gauge vector bosons

virtual photon 34

vortex lines, *see* vortices

vortices 58, 60

vorticons 120

W

$W^{\pm\mu}$ gauge bosons 8

$W^{\pm\mu}$ gauge bosons, *see* Weinberg–Salam 110–111

Ward-Takahashi identity 183, 201

weak currents 40

weak,

currents 40

hypercharge, *see* $U(1)_{\text{weak}}$

interactions, *see* electroweak

isospin, *see* $SU(2)_{\text{weak}}$

weight diagrams 25–27

Weinberg angle 109

Weinberg–Salam model 7–8, 95, 105

Wilson,

lattice action 288–291

loop 299–300

criterion, for confinement 300–301, 308–311

winding number 90, 91, 166, 170

Y

Yang-Mills fields, also non-Abelian gauge fields 6, 61, 67, 69, 70, 71

Young's tableaux 13

Yukawa couplings 77, 107

Z

$Z^{0\mu}$ gauge boson, *see* Weinberg–Salam 110–111

Z (integers) 95

$Z_2(\pm 1)$ 65

$Z(N)$ 318