5 Anomalies, instantons and the strong CP problem

While perturbation theory is a powerful and useful tool in understanding field theories, for our exploration of physics beyond the Standard Model an understanding of non-perturbative physics will be crucial. There are many reasons for this.

1. One of the great mysteries of the Standard Model is non-perturbative in nature: the smallness of the θ parameter.
2. Strongly interacting field theories will figure in many proposals to understand other mysteries of the Standard Model.
3. The interesting dynamical properties of supersymmetric theories, both those directly related to possible models of nature and those which provide insights into broad physics issues, are non-perturbative in nature.
4. If string theory describes nature, non-perturbative effects are necessarily of critical importance.

We have introduced lattice gauge theory, which is perhaps our only tool for doing systematic calculations in strongly coupled theories. But, as a tool, its value is quite limited. Only a small number of calculations are tractable in practice, and the difficult numerical challenges sometimes obscure the underlying physics. Fortunately, there is a surprising amount that one can learn from symmetry considerations, from semiclassical arguments and from our experimental knowledge of one strongly coupled theory, QCD. In each of these, an important role is played by the phenomena known as anomalies and, related to these, a set of semiclassical field configurations known as *instantons*.

Usually, the term "anomaly" is used to refer to the quantum mechanical violation of a symmetry which is valid classically. Instantons are finite-action solutions of the Euclidean equations of motion, typically associated with tunneling phenomena. Anomalies are crucial to understanding the decay of the π^0 in QCD. Anomalies and instantons account for the absence of a ninth light pseudoscalar meson in the hadron spectrum. Within the weak-interaction theory, anomalies and instantons lead to violations of baryon and lepton number; these effects are unimaginably tiny at the current time but were important in the early universe. The absence of anomalies in gauge currents is important to the consistency of theoretical structures, including both field theories and string theories. The cancelation of anomalies within the Standard Model itself is quite non-trivial. Similar constraints on possible extensions of the Standard Model will be very important. The θ parameter of QCD was mentioned in the previous chapter. The θ term seems innocuous, but, owing to anomalies and instantons, its potential effects are real. Because the θ term violates CP, they are also dramatic. The problem of the smallness of the θ parameter – the *strong CP problem* – forcibly suggests new phenomena beyond the Standard Model, and this will be

https://doi.org/10.1017/9781009290883.010 Published online by Cambridge University Press

a recurring theme in this book. In the present chapter we explain how anomalies arise and some of the roles which they play. The discussion is meant to provide the reader with a good working knowledge of these subjects, but it is not encyclopedic. A guide to texts and reviews on the subject appears at the end of the chapter.

5.1 The chiral anomaly

Before discussing real QCD, let us consider a non-Abelian gauge theory theory, with only a single flavor of quark. Before making any field redefinitions, the Lagrangian takes the form:

$$\mathcal{L} = -\frac{1}{4g^2}F_{\mu\nu}^2 + i\bar{q}D^\mu\sigma_\mu\bar{q}^* + iqD^\mu\sigma_\mu q^* + m\bar{q}q + m^*\bar{q}^*q^*. \tag{5.1}$$

The Lagrangian is written here in terms of two-component fermions (see Appendix A). The fermion mass need not be real:

$$m = |m|e^{i\theta}. \tag{5.2}$$

In this chapter it will sometimes be convenient to work with four-component fermions, and it is valuable to make contact with this language in any case. In terms of these, the mass contribution is

$$\mathcal{L}_m = (\mathrm{Re}\, m)\; \bar{q}q + (\mathrm{Im}\, m)\; \bar{q}\gamma_5 q. \tag{5.3}$$

In order to bring this mass contribution to the conventional form, with no γ_5s, one could try to redefine the fermions; switching back to the two-component notation we have

$$q \to e^{-i\theta/2}q, \quad \bar{q} \to e^{-i\theta/2}\bar{q}. \tag{5.4}$$

However, in field theory transformations of this kind are potentially fraught with difficulties because of the infinite number of degrees of freedom.

A simple calculation uncovers one of the simplest manifestations of an anomaly. Suppose, first, that m is very large, $m \to M$. In that case we need to integrate out the quarks and obtain a low-energy effective theory. To do this, we study the path integral (see Appendix C)

$$Z = \int [dA_\mu] \int [dq][d\bar{q}]e^{iS}. \tag{5.5}$$

Suppose that $M = e^{i\theta}|M|$. In order to make M real, we can again make the transformations $q \to qe^{-i\theta/2}, \bar{q} \to \bar{q}e^{-i\theta/2}$ (in four-component language, this is $q \to e^{-i\theta/2\gamma_5}q$)). The result of integrating out the quark, i.e. of performing the path integral over q and $\bar{q}$, can be written in the form

$$Z = \int [dA_\mu] \int e^{iS_{\rm eff}}. \tag{5.6}$$

Here $S_{\rm eff}$ is the effective action which describes the interactions of gluons at scales well below M.

https://doi.org/10.1017/9781009290883.010 Published online by Cambridge University Press

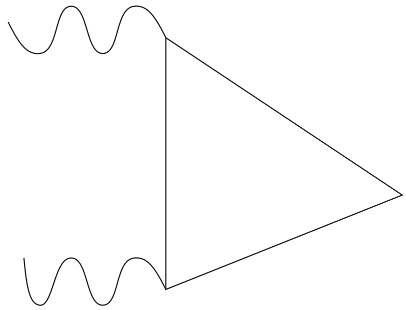

Fig. 5.1 The triangle diagram associated with the four-dimensional anomaly. At the right-hand vertex, one has insertions of the axial current and the chiral density.

Because the field redefinition which eliminates θ amounts to just a change of variables in the path integral, one might expect that there can be no θ-dependence in the effective action. But this is not the case. To see this, suppose that θ is small and, instead of redefining the field treat the θ term as a small perturbation by expanding the exponential. Now consider a term in the effective action with two external gauge bosons. This is obtained from the Feynman diagram in Fig. 5.1. The corresponding term in the action is given by (see Eq. (2.17))

$$\delta\mathcal{L}_{\rm eff} = -i\frac{\theta}{2}M\,\mathrm{Tr}(T^aT^b)\int\frac{d^4p}{(2\pi)^4}\,\mathrm{Tr}\left(\gamma_5\frac{1}{\not p+\not k_1-M}\,\not\epsilon_1\frac{1}{\not p-M}\,\not\epsilon_2\frac{1}{\not p-\not k_2-M}\right). \quad (5.7)$$

Here, the k_is are the momenta of the two gluons, while the ϵs are their polarizations and a and b are their color indices. Introducing Feynman parameters and shifting the p integral gives

$$\begin{aligned}\delta\mathcal{L}_{\rm eff} = &-i\theta g^2M\,\mathrm{Tr}(T^aT^b)\int d\alpha_1 d\alpha_2\int\frac{d^4p}{(2\pi)^4}\mathrm{Tr}\Big(\gamma_5(\not p-\alpha_1\not k_1+\alpha_2\not k_2+\not k_1+M)\\ &\times\frac{\not\epsilon_1(\not p-\alpha_1\not k_1+\alpha_2\not k_2+M)\,\not\epsilon_2(\not p-\alpha_1\not k_1+\alpha_2\not k_2-\not k_2+M)}{\left[p^2-M^2+O(k_i^2)\right]^3}\Big).\end{aligned} \quad (5.8)$$

For small k_i we can neglect the k-dependence of the denominator. The trace in the numerator is easy to evaluate, since we can drop terms linear in p. This gives, after performing the integrals over the αs,

$$\delta\mathcal{L}_{\rm eff} = g^2M^2\theta\,\mathrm{Tr}(T^aT^b)\,\epsilon_{\mu\nu\rho\sigma}k_1^\mu k_2^\nu\epsilon_1^\rho\epsilon_2^\sigma\int\frac{d^4p}{(2\pi)^4}\frac{1}{(p^2-M^2)^3}. \quad (5.9)$$

This corresponds to a term in the effective action, which, after performing the integral over p and including a combinatoric factor two from the different ways to contract the gauge bosons, is given by

$$\delta\mathcal{L}_{\rm eff} = \frac{1}{32\pi^2}\theta\,\mathrm{Tr}(F\tilde F). \quad (5.10)$$

Now why does this happen? On the one hand, at the level of the path integral the transformation would seem to amount to a simple change of variables, and it is hard to see why this should have any effect. On the other hand, if one examines the diagram of Fig. 5.1 then one sees that it contains terms which are linearly divergent and thus it should be regulated. A simple way to regulate this diagram is to introduce a Pauli–Villars regulator, which means that one subtracts off a corresponding amplitude with some very large mass Λ. However, our expression above is independent of Λ. So the θ-dependence from the regulator fields cancels that of Eq. (5.10). This sort of behavior is characteristic of an anomaly.

Consider now the case where $m \ll \Lambda_{\rm QCD}$. In this case we should not integrate out the quarks, but we still need to take into account the regulator diagrams. So, if we redefine the fields so, that the quark mass is real (γ_5-free, in the four-component description), the low-energy theory contains light quarks and the θ term of Eq. (5.10).

We can describe this in a fashion which indicates why this is referred to as an *anomaly.* For small m the classical theory has an approximate symmetry under which

$$q \to e^{i\alpha} q, \quad \bar{q} \to e^{i\alpha}\bar{q} \tag{5.11}$$

(in four-component language, $q \to e^{i\alpha\gamma_5} q$). In particular we can define a current

$$j_5^\mu = \bar{q}\gamma_5\gamma_\mu q \tag{5.12}$$

and, classically,

$$\partial_\mu j_5^\mu = m\bar{q}\gamma_5 q. \tag{5.13}$$

Under a transformation by an infinitesimal angle α one would expect that

$$\delta L = \alpha\partial_\mu j_5^\mu = m\alpha\bar{q}\gamma_5 q. \tag{5.14}$$

But the divergence of the current contains another, m-independent, term:

$$\partial_\mu j_5^\mu = m\bar{q}\gamma_5 q + \frac{1}{32\pi^2} F\tilde{F}. \tag{5.15}$$

The first term follows from the equations of motion. To see why the second term is present, we will study a three-point function involving the current and two gauge bosons A_μ and will ignore the quark mass:

$$\Gamma^{AAj} = T\langle \partial_\mu j^{5\mu} A_\rho A_\sigma \rangle. \tag{5.16}$$

This is essentially the calculation we encountered above. Again the diagram is linearly divergent and requires regularization. Let us first consider the graph without the regulator mass. The graph of Fig. 5.1 actually implies two graphs, because we must include the interchange of the two external gluons. The combination is easily seen to vanish, by the sorts of manipulations one usually uses to prove Ward identities:

$$\frac{g^2}{(2\pi)^4} \int d^4p \, {\rm Tr} \left(\not{q}\gamma_5 \frac{1}{\not{p} + \not{k}_1} \not{\epsilon}_1 \frac{1}{\not{p}} \not{\epsilon}_2 \frac{1}{\not{p} - \not{k}_2} + (1 \leftrightarrow 2) \right). \tag{5.17}$$

https://doi.org/10.1017/9781009290883.010 Published online by Cambridge University Press

Writing

$$\not{q}\gamma_5 = -\gamma_5(\not{k}_1 + \not{p}) - (\not{p} - \not{k}_2)\gamma_5 \tag{5.18}$$

and using the cyclic property of the trace, one can cancel a propagator in each term. This leaves

$$\int d^4p \,\mathrm{Tr}\left(-\gamma_5\, \not{\epsilon}_1 \frac{1}{\not{p}}\, \not{\epsilon}_2 \frac{1}{\not{p} - \not{k}_2} - \gamma_5 \frac{1}{\not{p} + \not{k}_1}\, \not{\epsilon}_1 \frac{1}{\not{p}}\, \not{\epsilon}_2 + (1 \leftrightarrow 2)\right). \tag{5.19}$$

Now making the shift $p \to p + k_2$ in the first term and $p \to p + k_1$ in the second, one finds a pairwise cancelation.

These manipulations, however, are not reliable. In particular, in a highly divergent expression the shifts do not necessarily leave the result unchanged. With a Pauli–Villars regulator the integrals are convergent and the shifts are reliable, but the regulator diagram is non-vanishing and gives the anomaly equation above. One can see this by a direct computation or relate it to our previous calculation, including the masses for the quark and noting that $\not{q}\gamma_5$, in the diagrams with massive quarks, can be replaced by $M\gamma_5$.

This anomaly can be derived in a number of other ways. One can define, for example, the current by *point splitting*, i.e. separating the two fields in the current by an amount ϵ and inserting a Wilson line to ensure gauge invariance.

$$j_5^\mu = \bar{q}(x + \epsilon) \exp\left(i \int_x^{x+\epsilon} dx^\mu A_\mu\right) q(x). \tag{5.20}$$

Because the operators in quantum field theory are singular at short distances, the Wilson line makes a finite contribution. Expanding the exponential carefully, one recovers the same expression for the current. We will do this shortly in two dimensions, leaving the four-dimensional case for the end-of-chapter exercises. A beautiful derivation, closely related to that performed above, is due to Fujikawa. Here one considers the anomaly as arising from a lack of invariance of the path integral measure. One carefully evaluates the Jacobian associated with the change of variables $q \to q(1 + i\gamma_5\alpha)$ and shows that it yields the same result. We will do a calculation along these lines in a two-dimensional model shortly, leaving the four-dimensional case for the exercises.

5.1.1 Applications of the anomaly in four dimensions

The anomaly has a number of important consequences for real physics.

- π^0 *decay* The divergence of the axial isospin current

$$(j_5^3)^\mu = \bar{u}\gamma_5\gamma^\mu\bar{u} - \bar{d}\gamma_5\gamma^\mu d \tag{5.21}$$

  has an anomaly due to electromagnetism. This gives rise to a coupling of the π^0 to two photons, and the correct prediction of the lifetime was one of the early triumphs of the color theory of quarks. The computation of the π^0 decay rate appears in the exercises.
- *Anomalies in gauge currents signal an inconsistency in a theory* They mean that gauge invariance, which is crucial to the whole structure of gauge theories (e.g. to the fact that they are simultaneously unitary and Lorentz invariant) is lost. The absence of

https://doi.org/10.1017/9781009290883.010 Published online by Cambridge University Press

gauge anomalies is one of the striking ingredients of the Standard Model, and it is also crucial in extensions such as string theory.

- *The anomaly considered here*, as we have indicated above, accounts for the absence of a ninth axial Goldstone boson in the QCD spectrum.

5.1.2 Return to QCD

What we have just learned is that if in our simple model above we require that the quark masses are real then we must allow for the possible appearance, in the Lagrangian of the Standard Model, of the θ term in Eq. (5.10). In weak interactions this term does not have physical consequences. At the level of the renormalizable terms, we have seen that the theory respects separate B and L symmetries; B, for example, is anomalous. So, if we simply redefine the quark fields by a B transformation, we can remove θ from the Lagrangian.

For the θ angles of QCD and QED we have no such symmetry. In the case of QED we do not really have a non-perturbative definition of the theory, and the effects of θ are hard to assess, but one might expect that, when embedded in any consistent structure (such as a grand unified theory (GUT) or string theory) they will be very small, possibly zero. As we saw, $F\tilde{F}$ gives a total divergence. The right-hand side of Eq. (4.24) is not gauge invariant, however, so one might imagine that it could be important. But, as long as A falls off at least as fast as $1/r$ (i.e. F falls off faster than $1/r^2$), the surface term behaves as $1/r^4$ and so vanishes.

In the case of non-Abelian gauge theories, the situation is more subtle. It is again true that $F\tilde{F}$ can be written as a total divergence:

$$F\tilde{F} = \partial^\mu K_\mu, \quad K_\mu = \epsilon_{\mu\nu\rho\sigma}\left(A_\nu^a F_{\rho\sigma}^a - \frac{2}{3} f^{abc} A_\nu^a A_\rho^b A_\sigma^c\right). \tag{5.22}$$

However, the statement that F falls off faster than $1/r^2$ does not permit an equally strong statement about A. We will see shortly that there are finite-action classical solutions for which $F \sim 1/r^4$ but $A \to 1/r$, so that the surface term cannot be neglected. These solutions are instantons. This is the reason that θ can have real physical effects.

5.2 A two-dimensional detour

There are many questions in four dimensions which we cannot answer except by using numerical lattice calculation. These include the problem of dimensional transmutation and the effects of the anomaly on the hadron spectrum. There is a class of models in two dimensions which are asymptotically free and in which one can study these questions in a controlled approximation. Two dimensions often form a poor analog for four but, for some of the issues we are facing here, the parallels are extremely close. In these two-dimensional examples the physics is more manageable, but still rich. In four dimensions,

https://doi.org/10.1017/9781009290883.010 Published online by Cambridge University Press

the calculations are qualitatively similar; they are only more difficult because the Dirac algebra and the various integrals are more involved.

5.2.1 The anomaly in two dimensions

First we investigate the anomaly in the quantum electrodynamics of a massless fermion in two dimensions; this will be an important ingredient in the full analysis. The point-splitting method is particularly convenient here. Just as in four dimensions, we write

$$j_5^\mu = \bar{\psi}(x+\epsilon)\exp\left(i\int_x^{x+\epsilon} A_\rho dx^\rho\right)\gamma^\mu\gamma^5\psi(x). \tag{5.23}$$

Naively, one can set $\epsilon = 0$ and then the divergence vanishes by the equations of motion. In quantum field theory, however, products of operators become singular as the operators come close together. For very small ϵ we can pick up the leading singularity in the product of $\psi(x+\epsilon)\psi(x)$ by using the operator product expansion (OPE). The OPE states that the product of two operators at short distances can be written as a series of local operators of progressively higher dimension, with coefficients that are less and less singular. For our case this means that

$$\bar{\psi}(x+\epsilon)\gamma^\mu\gamma^5\psi(x) = \sum \frac{c_n}{\epsilon^{1-n}}\mathcal{O}_n(x), \tag{5.24}$$

where $\mathcal{O}_n$ is an operator of dimension n. The leading term comes from the unit operator. To evaluate its coefficient we can take the vacuum expectation value of both sides of this equation. On the left-hand side, this is just the propagator.

It is not hard to work out the fermion propagator in coordinate space in two dimensions. For simplicity we work with space-like separations, so that we can Wick-rotate to Euclidean space. Start with the scalar propagator

$$\begin{aligned}\langle\phi(x)\phi(0)\rangle &= \int \frac{d^2p}{(2\pi)^2}\frac{1}{k^2}e^{-ip\cdot x}\\ &= \frac{1}{2\pi}\ln(|x|\mu),\end{aligned} \tag{5.25}$$

where μ is an infrared cutoff. (When we come to string theory this propagator, with its infrared sensitivity, will play a crucial role.) Correspondingly, the fermion propagator is

$$\langle\bar{\psi}(x+\epsilon)\psi(x)\rangle = \partial\!\!\!/\langle\phi(x)\phi(0)\rangle = \frac{1}{2\pi}\frac{\epsilon\!\!/}{\epsilon^2}. \tag{5.26}$$

Expanding the factor in the exponential to order ϵ gives

$$\partial_\mu j_5^\mu = \text{classical term} + \frac{i}{2\pi}\partial_\mu\epsilon_\rho A^\rho\,\mathrm{Tr}\left(\frac{\epsilon\!\!/}{\epsilon^2}\gamma^\mu\gamma^5\right). \tag{5.27}$$

Evaluating the trace gives $\epsilon_{\mu\nu}\epsilon^\nu$; averaging ϵ over angles ($\langle\epsilon_\mu\epsilon_\nu\rangle = \frac{1}{2}\eta_{\mu\nu}\epsilon^2$) yields

$$\partial_\mu j_5^\mu = \frac{1}{2\pi}\epsilon_{\mu\nu}F^{\mu\nu}. \tag{5.28}$$

https://doi.org/10.1017/9781009290883.010 Published online by Cambridge University Press

This is parallel to the situation in four dimensions. The divergence of the current is itself a total derivative:

$$\partial_\mu j_5^\mu = \frac{1}{2\pi}\epsilon_{\mu\nu}\partial^\mu A^\nu. \tag{5.29}$$

So, it is possible to define a new current which is conserved:

$$J^\mu = j_5^\mu - \frac{1}{2\pi}\epsilon^\mu_\nu A^\nu. \tag{5.30}$$

However, just as in the four-dimensional case, this current is not gauge invariant. There is a familiar field configuration for which A does not fall off at infinity: the field of a point charge. If one has charges $\pm\theta$ at infinity, they give rise to a constant electric field, $F_{0i} = \pm e\theta$. So θ has a very simple interpretation in this theory.

It is easy to see that the physics is periodic in θ. For $\theta > q$ it is energetically favorable to produce a pair of charges from the vacuum which shield the charge at ∞.

5.2.2 Path integral computation of the anomaly

One can also do this calculation using the path integral, following Fujikawa. The redefinition of the fields which eliminates the phase in the fermion mass matrix is, from this point of view, just a change of variables. The question is: what is the Jacobian? The Euclidean path integral is defined by expanding the fields:

$$\psi(x) = \sum a_n \psi_n(x), \tag{5.31}$$

where

$$\not{D}\psi_n(x) = \lambda_n \psi_n(x) \tag{5.32}$$

and the measure is

$$\int \Pi\, da_n da_n^*. \tag{5.33}$$

Here, for normalized functions ψ_n,

$$a_n = \int d^2x\, \psi_n^*(x)\psi(x). \tag{5.34}$$

So, under an infinitesimal γ_5 transformation, we have

$$\delta\psi = i\theta\gamma_5\psi, \tag{5.35}$$

$$\delta a_n = i\theta \int d^2x\, \psi_n(x)\gamma_5\psi_m(x) a_m. \tag{5.36}$$

The required Jacobian is then

$$\det\left(\delta_{nn'} + i\theta \int d^2x\, \bar{\psi}_{n'}\gamma_5\psi_n\right). \tag{5.37}$$

https://doi.org/10.1017/9781009290883.010 Published online by Cambridge University Press

To evaluate this determinant we write $\det(M) = e^{\mathrm{Tr}\log M}$. To linear order in θ, we need to evaluate

$$\mathrm{Tr}\;(i\theta\gamma_5). \tag{5.38}$$

This trace must be regularized. A simple procedure is to replace the determinant by

$$\mathrm{Tr}\;\left(i\theta\gamma_5 e^{-\lambda_n^2/M^2}\right). \tag{5.39}$$

At the end of the calculation we take $M \to \infty$. We can replace λ_n^2 by

$$\not{D}\,\not{D} = D^2 + \frac{1}{2}\sigma_{\mu\nu}F^{\mu\nu}. \tag{5.40}$$

Expanding in powers of $F^{\mu\nu}$, it is only necessary to work to first order (in the analogous calculation in four dimensions, it is necessary to work to second order). In other words, we expand the exponent to first order in $F^{\mu\nu}$ and make the replacement $D^2 \to p^2$. The required trace is given by

$$i\theta\int\frac{d^2p}{p^2}\,\mathrm{Tr}(\gamma_5\sigma_{\mu\nu})\,\frac{F^{\mu\nu}}{M}e^{-p^2/M^2}. \tag{5.41}$$

The trace in this expression now just refers to a trace over the Dirac indices. The momentum integral is elementary, and we obtain

$$\int\Pi\,da_n da_n^* \to \int\Pi\,da_n da_n^*\exp\left(i\frac{\theta}{2\pi}\int d^2x\,\epsilon_{\mu\nu}F^{\mu\nu}\right). \tag{5.42}$$

Interpreting the divergence of the current as the variation of the effective Lagrangian, we see that we have recovered the anomaly equation (5.15). The anomaly in four and other dimensions can also be calculated in this way. The exercises at the end of the chapter provide more details of these computations.

5.2.3 The CP^N model: an asymptotically free theory

The model we have considered so far is not quite like QCD in at least two ways. First, there are no instantons; second, the coupling e is dimensionful. We can obtain a theory closer to QCD by considering a class of theories with dimensionless couplings, the *non-linear sigma models*. These are models whose fields are the coordinates of some smooth manifold. They can be, for example, the coordinates of an n-dimensional sphere. An interesting case is the CP^N model; here the CP stands for "complex projective" space. This space is described by a set of coordinates z_i, $i = 1, \ldots, N+1$, where z_i is identified with αz_i and α is any complex constant. Alternatively, we can define the space through the constraint

$$\sum_i |z_i|^2 = 1, \tag{5.43}$$

where the point z_i is equivalent to $e^{i\alpha}z_i$. In the field theory, the z_is become two-dimensional fields $z_i(x)$. To implement the first constraint, we can add to the action a Lagrange multiplier field $\lambda(x)$. For the second, we observe that the identification of points in the "target space"

https://doi.org/10.1017/9781009290883.010 Published online by Cambridge University Press

CP^N must hold at every point in *ordinary space–time*, so this is a $U(1)$ gauge symmetry. Introducing a gauge field A_μ and the corresponding covariant derivative, we want to study the Lagrangian

$$\mathcal{L} = \frac{1}{g^2}\left[|D_\mu z_i|^2 - \lambda(x)(|z_i|^2 - 1)\right]. \tag{5.44}$$

Note that there is no kinetic term for A_μ, so we can simply eliminate it from the action using its equations of motion. This yields

$$\mathcal{L} = \frac{1}{g^2}\left(|\partial_\mu z_j|^2 + |z_j^* \partial_\mu z_j|^2\right). \tag{5.45}$$

It is easier, however, to proceed keeping A_μ in the action. In this case the action is quadratic in z, and we can integrate out the z fields:

$$\begin{aligned} Z &= \int [dA][d\lambda][dz_j] \exp(-S) = \int [dA][d\lambda] \exp\left(-\int d^2x\, \Gamma_{\text{eff}}[A, \lambda]\right) \\ &= \int [dA][d\lambda] \exp\left(-N \operatorname{Tr} \log(-D^2 - \lambda) - \frac{1}{g^2}\int d^2x\, \lambda\right). \end{aligned} \tag{5.46}$$

5.2.4 The large-N limit

By itself, the result in Eq. (5.46) is still rather complicated. The fields A_μ and λ have non-linear and non-local interactions. Things become much simpler if one takes the large-N limit, $N \to \infty$ with g^2N fixed. In this case the interactions of λ and A_μ are suppressed by powers of N. For large N the path integral is dominated by a single field configuration, which solves

$$\frac{\delta \Gamma_{\text{eff}}}{\delta \lambda} = 0, \tag{5.47}$$

or, setting the gauge field to zero,

$$N \int \frac{d^2k}{(2\pi)^2} \frac{1}{k^2 + \lambda} = \frac{1}{g^2}. \tag{5.48}$$

The integral on the left-hand side is ultraviolet divergent. We will simply cut it off at scale M. This gives

$$\lambda = m^2 = M \exp\left(-\frac{2\pi}{g^2 N}\right). \tag{5.49}$$

Here, a theory which is classically scale invariant exhibits a mass gap. This is the phenomenon of *dimensional transmutation*. These masses are related in a renormalization-group-invariant fashion to the cutoff. So the theory is quite analogous to QCD. We can read off the leading term in the beta function from the familiar formula

$$m = M \exp\left(-\int \frac{dg}{\beta(g)}\right). \tag{5.50}$$

So, with

$$\beta(g) = -\frac{1}{2\pi} g^3 b_0, \tag{5.51}$$

we have $b_0 = 1$.

Most important for our purposes is the question of θ-dependence. Just as in $(1 + 1)$-dimensional electrodynamics we can introduce a θ term,

$$S_\theta = \frac{\theta}{2\pi} \int d^2x\, \epsilon_{\mu\nu} F^{\mu\nu}. \tag{5.52}$$

Here $F_{\mu\nu}$ can be expressed in terms of the fundamental fields z_j. As usual, this is the integral of a total divergence. But, precisely as in the case of $(1 + 1)$-dimensional electrodynamics discussed above, this term is physically important. In a perturbation theory approach to the model, this is not entirely obvious; however, using our reorganization of the theory at large N, it is. The lowest-order action for A_μ is trivial, but at one loop (order $1/N$) a kinetic term for A is generated through the vacuum polarization loop:

$$\mathcal{L}_{\text{kin}} = \frac{N}{2\pi m^2} F_{\mu\nu}^2. \tag{5.53}$$

At this order, then, the effective theory consists of the gauge field with coupling $e^2 = 2\pi m^2/N$ and some coupling to a set of charged massive fields z. As we have already argued, θ corresponds to a non-zero background electric field due to charges at infinity, and the theory clearly has a non-trivial θ-dependence.

To this model one can add massless fermions. In this case one has an anomalous $U(1)$ symmetry, as in QCD. There is then no θ-dependence; by redefining the fermions according to $\psi \to e^{ia\theta}\psi$ one can eliminate θ. In this model the absence of a θ-dependence can be understood more physically: θ represents a charge at ∞, and it is possible to shield any such charge with massless fermions. But there is a non-trivial breaking of the $U(1)$ symmetry. At low energies, one has now a theory with a fermion coupled to a dynamical $U(1)$ gauge field. The breaking of the associated $U(1)$ symmetry in such a theory is a well-studied phenomenon, which we will not pursue here.

5.2.5 The role of instantons

There is another way to think about the breaking of the $U(1)$ symmetry and the θ-dependence in this theory. If one considers the Euclidean functional integral, it is natural to look for stationary points of the integration, i.e. for classical solutions of the Euclidean equations of motion. Since they are potentially important it is necessary that these solutions have a finite action, which means that they must be localized in Euclidean space and time. For this reason, such solutions were dubbed "instantons" by 't Hooft. Instantons are not difficult to find in the CP^N model; we will describe them below. These solutions carry non-zero values of the *topological charge*,

$$\frac{1}{2\pi} \int d^2x\, \epsilon_{\mu\nu} F_{\mu\nu} = n, \tag{5.54}$$

https://doi.org/10.1017/9781009290883.010 Published online by Cambridge University Press

and have an action $2\pi n$. If we write $z_i = z_{i\,\mathrm{cl}} + \delta z_i$ then the functional integral, in the presence of a θ term, has the form

$$Z_{\mathrm{inst}} = e^{\frac{-2\pi n}{g^2}} e^{in\theta} \int [d\delta z_j] \exp\left(-\delta z_i \frac{\delta^2 S}{\delta z_i \delta z_j} \delta z_j + \cdots\right). \tag{5.55}$$

It is easy to construct the instanton solution in the case of CP^1. Rather than write the theory in terms of a gauge field, as we have done above, it is convenient to parameterize it in terms of a single complex field Z. One can, for example, define Z as z_1/z_2 and let $\bar{Z}$ denote its complex conjugate. Then, with a bit of algebra, one can show that the action for Z which follows from Eq. (5.45) takes the following form (it is easiest to work backwards, starting with the equation below and deriving Eq. (5.45)):

$$\mathcal{L} = \frac{\partial_\mu Z \partial_\mu \bar{Z}}{(1 + \bar{Z}Z)^2}. \tag{5.56}$$

The function

$$g_{Z\bar{Z}} = \frac{1}{(1 + \bar{Z}Z)^2} \tag{5.57}$$

has an interesting significance. There is a well-known mapping of the unit sphere $x_1^2 + x_2^2 + x_3^2 = 1$ onto the complex plane:

$$z = \frac{x_1 + ix_2}{1 - x_3}. \tag{5.58}$$

The inverse is

$$x_1 = \frac{z + \bar{z}}{1 + |z|^2}, \quad x_2 = \frac{z - \bar{z}}{i(1 + |z|^2)}, \quad x_3 = \frac{|z|^2 - 1}{|z|^2 + 1}. \tag{5.59}$$

The line element on the sphere is mapped in a non-trivial way onto the plane:

$$ds^2 = dx_1^2 + dx_2^2 + dx_3^3 = g_{z\bar{z}} dz d\bar{z} = \frac{1}{(1 + \bar{z}z)^z} dz d\bar{z}. \tag{5.60}$$

So, the model describes a field that is constrained to move on a sphere; g is the metric of the sphere. In general, such a model is called a non-linear sigma model. This is an example of a Kahler geometry, a type of geometry which will figure significantly in our discussion of string compactification.

It is straightforward to write down the equations of motion:

$$\partial^2 Z g_{Z\bar{Z}} + \partial_\mu Z \left(\partial_\mu \bar{Z} \frac{\partial g}{\partial \bar{Z}} + \partial_\mu \phi \frac{\partial g}{\partial Z}\right) = 0, \tag{5.61}$$

or

$$\partial_z \partial_{\bar{z}} Z - \frac{2\partial_z Z \partial_{\bar{z}} \bar{Z}}{1 + \bar{Z}Z} = 0. \tag{5.62}$$

Now using space–time coordinates $z = x_1 + ix_2$, $\bar{z} = x_1 - ix_2$, we see that if Z is anti-analytic then the equations of motion are satisfied! So a simple solution, which, as you can check, has finite action, is

$$Z(\bar{z}) = \rho \bar{z}. \tag{5.63}$$

https://doi.org/10.1017/9781009290883.010 Published online by Cambridge University Press

In addition to evaluating the action you can evaluate the topological charge,

$$\frac{1}{2\pi}\int d^2x\,\epsilon_{\mu\nu}F^{\mu\nu} = 1, \tag{5.64}$$

for this solution. More generally, the topological charge measures the number of times that Z maps the complex plane into the complex plane; $Z = z^n$ has charge n.

We can generalize these solutions. The solution of Eq. (5.63) breaks several symmetries of the action: translation invariance, two-dimensional rotational invariance and the scale invariance of the classical equations. So we should be able to generate new solutions by translating, rotating and dilating the solution. You can check that

$$Z(z) = \frac{az+b}{cz+d} \tag{5.65}$$

is a solution with action 2π. The parameters $a, \ldots, d$ are called collective coordinates. They correspond to the symmetries of translations, dilations and rotations and special conformal transformations (forming the group $SL(2, C)$). In other words, any given finite-action solution breaks the symmetries. In the path integral the symmetry of Green's functions is recovered when one integrates over the collective coordinates. For translations this is particularly simple. Integrating over x_0, the instanton position,

$$\langle Z(x)Z(y)\rangle \approx \int d^2x_0\,\phi_{\rm cl}(x-x_0)\phi_{\rm cl}(y-x_0)e^{-S_0}. \tag{5.66}$$

(The precise measure is obtained by the Faddeev–Popov method.) Similarly, integration over the parameter ρ yields a factor

$$\int d\rho\,\rho^{-1}\exp\left(-\frac{2\pi}{g^2(\rho)}\right). \tag{5.67}$$

Here the first factor follows on dimensional grounds. The second follows from renormalization-group considerations. It can be found by explicit evaluation of the functional determinant. Note that, because of asymptotic freedom, this means that typical Green's functions will be divergent in the infrared.

There are many other features of this instanton that one can consider. For example, one can add massless fermions to the model; the resulting theory has a chiral $U(1)$ symmetry, which is anomalous. The instanton gives rise to non-zero Green's functions, which violate the $U(1)$ symmetry. We will leave investigation of fermions in this model to the exercises and turn to the theory of interest, which exhibits phenomena parallel to this simple theory.

5.3 Real QCD

The model of the previous section mimics many features of real QCD. Indeed, we will see that much of our discussion can be carried over, almost word for word, to the observed strong interactions. This analogy is helpful, given that in QCD we have no approximation which gives us control over the theory comparable with that which we found in the large-N limit of the CP^N model. As in that theory, we have the following.

- There is a θ parameter, which appears as an integral over the divergence of a non-gauge invariant current.
- There are instantons, which indicate that physical quantities should be θ-dependent. However, instanton effects cannot be considered in a controlled approximation, and there is no clear sense in which θ-dependence can be understood as arising from instantons.
- In QCD there is also a large-N expansion but, while it produces significant simplification, one cannot solve the theory even in the leading large-N approximation. Instead, an understanding of the underlying symmetries, and experimental information about chiral symmetry breaking, provides critical information about the behavior of the strongly coupled theory and allows computations of the physical effects of θ.

5.3.1 The theory and its symmetries

In order to understand the effects of θ it is sufficient to focus on the light quark sector of QCD. For simplicity in writing down some of the formulas, we will consider a simplified theory with two light quarks; it is not difficult to generalize the resulting analysis to the case of three. It is believed that the masses of the u and d quarks are of order 5 MeV and 10 MeV, respectively, much smaller than the scale of QCD. So we first consider an idealization of the theory in which these masses are set to zero. In this limit, the theory has a symmetry $SU(2)_\mathrm{L} \times SU(2)_\mathrm{R}$. Calling

$$q = \begin{pmatrix} u \\ d \end{pmatrix}, \quad \bar{q} = \begin{pmatrix} \bar{u} \\ \bar{d} \end{pmatrix}, \tag{5.68}$$

the two $SU(2)$ symmetries act separately on q and $\bar{q}$ (thought of as left-handed fermions),

$$q^T \to q^T U_\mathrm{L}, \quad \bar{q} \to U_\mathrm{R} \bar{q}. \tag{5.69}$$

This symmetry is spontaneously broken. The order parameter for the symmetry breaking is believed to be an expectation value for the quark bilinear product:

$$\mathcal{M} = \bar{q} q. \tag{5.70}$$

Under the original symmetry,

$$\mathcal{M} \to U_\mathrm{R} \mathcal{M} U_\mathrm{L}. \tag{5.71}$$

The expectation value (condensate) of $\mathcal{M}$ is

$$\langle \mathcal{M} \rangle = c \Lambda_\mathrm{QCD}^3 \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}. \tag{5.72}$$

This breaks some of the original symmetry but preserves the symmetry $U_\mathrm{L} = U_\mathrm{R}$. This symmetry is just the $SU(2)$ isospin symmetry. The Goldstone bosons associated with the three broken symmetry generators must transform in a representation of the unbroken symmetry: these are the pions, which an form isospin vector. One can think of the

https://doi.org/10.1017/9781009290883.010 Published online by Cambridge University Press

Goldstone bosons as being associated with a slow variation of the expectation value in space, so we can introduce

$$\mathcal{M} = \bar{q}q = M_0 \exp\left[i\frac{\pi_a(x)\tau_a}{f_\pi}\right]\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \tag{5.73}$$

The quark mass term in the Lagrangian is then (for simplicity taking $m_u = m_d = m_q$)

$$m_q\mathcal{M}. \tag{5.74}$$

Replacing $\mathcal{M}$ by the expression (5.73) gives a potential for the pion fields. Expanding $\mathcal{M}$ in powers of π/f_π, the minimum of the potential occurs for $\pi_a = 0$. Expanding to second order, one has

$$m_\pi^2 f_\pi^2 = m_q M_0. \tag{5.75}$$

We have been a bit cavalier about the symmetries. The theory also has two $U(1)$ symmetries:

$$q \to e^{i\alpha}q, \quad \bar{q} \to e^{i\alpha}\bar{q}, \tag{5.76}$$
$$q \to e^{i\alpha}q, \quad \bar{q} \to e^{-i\alpha}\bar{q}. \tag{5.77}$$

The first of these is baryon number symmetry and it is not chiral (and is not broken by the condensate). The second is the axial $U(1)_5$ symmetry; it is broken by the condensate. So, in addition to the pions there should be another approximate Goldstone boson. But there is no good candidate among the known hadrons. The η has the right quantum numbers but, as we will see below, it is too heavy to be interpreted in this way. The absence of this fourth (or, in the case of three light quarks, ninth) Goldstone boson is called the $U(1)$ problem.

The $U(1)_5$ symmetry suffers from an anomaly, however, and we might hope that this has something to do with the absence of a corresponding Goldstone boson. The anomaly is given by

$$\partial_\mu j_5^\mu = \frac{1}{16\pi^2}F\tilde{F}. \tag{5.78}$$

Again, we can write the right-hand side as a total divergence

$$F\tilde{F} = \partial_\mu K^\mu, \tag{5.79}$$

where

$$K_\mu = \epsilon_{\mu\nu\rho\sigma}\left(A_\nu^a F_{\rho\sigma}^a - \frac{2}{3}f^{abc}A_\nu^a A_\rho^b A_\sigma^c\right). \tag{5.80}$$

This accounts for the fact that in perturbation theory the axial $U(1)$ symmetry is conserved. Non-perturbatively, as we will now show, there are important configurations in the functional integral for which the right-hand side does not vanish rapidly at infinity.

https://doi.org/10.1017/9781009290883.010 Published online by Cambridge University Press

5.3.2 Instantons in QCD

In the Euclidean functional integral

$$Z = \int [dA][dq][d\bar{q}]e^{-S} \tag{5.81}$$

it is natural to look for stationary points of the effective action, i.e. finite-action classical solutions of the theory in imaginary time. The Yang–Mills equations are complicated nonlinear equations, but it turns out that, much as in the CP^N model, the instanton solutions can be found rather easily. The following tricks simplify the construction and turn out to yield the general solution. First, note that the Yang–Mills action satisfies an inequality, the Bogomol'nyi bound:

$$\int (F \pm \tilde{F})^2 = \int (F^2 + \tilde{F}^2 \pm 2F\tilde{F}) = \int (2F^2 \pm 2F\tilde{F}) \geq 0. \tag{5.82}$$

So, the action is bounded by $|\int F\tilde{F}|$, the bound being saturated when

$$F = \pm\tilde{F}, \tag{5.83}$$

i.e. if the gauge field is (anti-)self-dual.[1] This equation is a first-order equation, and it is easy to solve if one first restricts to an $SU(2)$ subgroup of the full gauge group. One makes the ansatz that the solution should be invariant under a combination of ordinary rotations and global $SU(2)$ gauge transformations. Take

$$g(x) = \frac{x_4 + i\vec{x} \cdot \vec{\tau}}{r} \tag{5.84}$$

and

$$A_\mu = f(r^2) g \partial_\mu g^{-1}. \tag{5.85}$$

Then, substituting in to the Yang–Mills equations yields

$$f = \frac{-ir^2}{r^2 + \rho^2}, \tag{5.86}$$

where ρ is an arbitrary quantity with dimensions of length. The choice of origin here is also arbitrary; this can be remedied by simply replacing x by $x - x_0$ everywhere in these expressions, where x_0 represents the location of the instanton.

From this solution, it is clear why $\int \partial_\mu K^\mu$ does not vanish for the solution: while A is a pure gauge at infinity, it falls only as $1/r$. Indeed, since $F = \tilde{F}$, for this solution we have

$$\int F^2 = \int d^4x 4\tilde{F}^2 = 32\pi^2. \tag{5.87}$$

[1] This is not an accident, nor was the analyticity condition in the CP^N case. In both cases we can add fermions so that the model becomes supersymmetric. Then one can show that if some supersymmetry generators Q_α annihilate a field configuration then the configuration is a solution. This is a first-order condition; in the Yang–Mills case it implies self-duality and in the CP^N case it requires analyticity.

https://doi.org/10.1017/9781009290883.010 Published online by Cambridge University Press

This result can also be understood topologically. Note that g defines a mapping from the “sphere at infinity” into the gauge group. It is straightforward to show that

$$\frac{1}{32\pi^2}\int d^4x\, F\tilde{F} \tag{5.88}$$

counts the number of times that g maps the sphere at infinity into the group (once for this specific example; n times more generally). In the exercises and suggested reading, features of the instanton are explored in more detail.

The expression in Eq. (5.85) is, by its nature, gauge-dependent and other presentations of the solution are sometimes convenient. For example, if one formally transforms by g^{-1}, one obtains a solution which falls more rapidly to zero but which is singular at the origin.

The instanton was presented by ’t Hooft in a fashion which is often more useful for actual computations. Defining the symbol η as follows,

$$\eta_{aij} = \epsilon_{aij},\ \eta_{a4i} = -\eta_{ai4} = -\delta_{ai}, \quad \bar{\eta}_{a\mu\nu} = (-1)^{\delta_{a\mu}+\delta_{a\nu}}\eta_{a\mu\nu}, \tag{5.89}$$

the instanton takes the simple form

$$A^a_\mu = \frac{2\eta_{a\mu\nu}x^\nu}{x^2+\rho^2} \tag{5.90}$$

while the field strength is given by

$$F^a_{\mu\nu} = \frac{4\eta_{a\mu\nu}\rho^2}{(x^2+\rho^2)^2}. \tag{5.91}$$

That this configuration solves the equations of motion follows from

$$\eta_{a\mu\nu} = \frac{1}{2}\epsilon_{\mu\nu\alpha\beta}\eta_{a\alpha\beta}. \tag{5.92}$$

The alert reader will note that the η symbols are connected to the embedding of $SU(2)$ of the gauge group into an $SU(2)$ subgroup of $O(4) = SU(2)\times SU(2)$. This can be understood by noting that

$$\eta_{a\mu\nu} = \frac{1}{2}\,\mathrm{Tr}(\sigma^a\sigma_{\mu\nu}), \quad \bar{\eta} = \mathrm{Tr}(\sigma^a\bar{\sigma}_{\mu\nu}). \tag{5.93}$$

In this form it is easy to check that $F = \tilde{F}$, so the equations are satisfied. Note the $1/r$ falloff of A^μ, as opposed to the $1/r^4$ falloff of $F_{\mu\nu}$.

So, we have exhibited potentially important contributions to the path integral which violate the $U(1)$ symmetry. How does this symmetry violation show up? Let us consider the path integral more carefully. Having found a classical solution, we want to integrate over small fluctuations over it. Including the θ term these have the form

$$\langle \bar{u}u\bar{d}d\rangle = e^{-8\pi^2/g^2}e^{i\theta}\int [d\delta A][dq][d\bar{q}]\exp\left(-\frac{\delta^2 S}{\delta A^2}\delta A^2 - S_{q,\bar{q}}\right)\bar{u}u\bar{d}d. \tag{5.94}$$

Now S contains an explicit factor $1/g^2$. As a result the fluctuations are formally suppressed by g^2 relative to the leading contribution. The one-loop functional integral yields a product of determinants for the fermions and a product of inverse square root determinants for the bosons.

Consider the integral over the fermions. It is straightforward, if challenging, to evaluate the determinants. However, if the quark masses are zero then the fermion functional

https://doi.org/10.1017/9781009290883.010 Published online by Cambridge University Press

integrals are also zero, because there is a zero mode for each of the fermions, i.e. for both q and $\bar{q}$ there is a normalizable solution of the equations

$$\not{D}u = 0, \quad \not{D}\bar{u} = 0 \tag{5.95}$$

and similarly for d and $\bar{d}$. It is straightforward to construct the solutions

$$u = \frac{\rho}{[\rho^2 + (x - x_0)^2]^{3/2}}\zeta, \tag{5.96}$$

where ζ is a constant spinor, and similarly for $\bar{u}$, etc.

Let's understand this a bit more precisely. Euclidean path integrals are conceptually simple. Consider some classical solution, $\Phi_{\rm cl}(x)$ (here Φ denotes collectively the various bosonic fields; we will treat, for now, the fermions as vanishing in the classical solutions). In the path integral, at small coupling we are interested in small fluctuations about the classical solution,

$$\Phi = \Phi_{\rm cl} + \delta\Phi. \tag{5.97}$$

Because the action is stationary at the classical solution,

$$S = S_{\rm cl} + \int d^4x\, \delta\Phi \frac{\partial^2 \mathcal{L}}{\partial \Phi^2} \delta\Phi + \cdots . \tag{5.98}$$

The second derivative here is a shorthand for a second-order differential operator, which we will simply denote by S'' and refer to as the *quadratic fluctuation operator*. We can expand $\delta\Phi$ in (normalizable) eigenfunctions of this operator Φ_n with eigenvalues λ_n, $\Phi = c_n\Phi_n$. The result of the functional integral is then $\prod \lambda_n^{-1/2}$. This is the leading correction to the classical limit. Higher-order corrections are suppressed by powers of g^2. This is most easily seen by working in the scaling where the action has a factor $1/g^2$. Then one can derive the perturbation theory from the path integral in the usual way; the main difference from the usual treatment with zero background fields is that the propagators are more complicated. The propagators for various fields in the instanton background are in fact known in closed form.

The form of the differential operator is familiar from our calculation of the beta function in the background field method (using the background field gauge). For the gauge bosons, in a suitable (background field) gauge it is

$$S'' = \mathcal{D}^2 + \mathcal{J}_{\mu\nu}F^{\mu\nu}. \tag{5.99}$$

Here $\mathcal{D}$ is just the covariant derivative, the vector potential corresponds to the classical solution (an instanton) and similarly for the field strength; $\mathcal{J}_{\mu\nu}$ is the generator of Lorentz transformations in the vector representation. The eigenvalue problem was completely solved by 't Hooft.

Both the bosonic and fermionic quadratic fluctuation operators have zero eigenvalues. For the bosons, these potentially give infinite contributions to the functional integral and they must be treated separately. The difficulty is that among the variations of the fields are symmetry transformations, which comprise changes in the location of the instanton (translations), rotations of the instanton and scale transformations. Consider translations. For every solution there corresponds an infinite set of other solutions obtained by shifting the origin (varying x_0). Thus, instead of integrating over a coefficient c_0, we integrate over the *collective coordinate* x_0 (one must also include a suitable Jacobian factor). The effect of this is to restore translational invariance in the Green's functions. We will see this explicitly shortly. Similarly, the instanton breaks the rotational invariance of the theory; correspondingly, we can find a three-parameter set of solutions and zero modes. Integrating over these rotational collective coordinates restores rotational invariance. (The instanton also breaks a global gauge symmetry, but a combination of rotations and gauge transformations is preserved.)

Finally, the classical theory is scale invariant; this is the origin of the parameter ρ in the solution. Again, one must treat ρ as a collective coordinate and integrate over ρ. There is a power of ρ arising from the Jacobian, which can be determined on dimensional grounds. For the Green's function Eq. (5.90), for example, which has dimension six, we have (if all the fields are evaluated at the same point),

$$\int d\rho\, \rho^{-7}. \tag{5.100}$$

However, there is additional ρ-dependence because the quantum theory violates scale symmetry. This can be understood by replacing g^2 by $g^2(\rho)$ in the functional integral and using

$$e^{-8\pi^2 g^2(\rho)} \approx (\rho M)^{b_0} \tag{5.101}$$

for small ρ. For three-flavor QCD, for example, $b_0 = 9$ and the ρ integral diverges for large ρ. This relation simply states that the integral is dominated by the infrared, where the QCD coupling becomes strong.

Fermion functional integrals introduce a new feature. In four-component language, it is necessary to treat q and $\bar{q}$ as independent fields. This rule gives the functional integral as a determinant rather than as, say, the square root of a determinant. (In two-component language, this corresponds to treating q and q^* as independent fields.) So, at the one-loop order, we need to study

$$\not{D} q_n = \lambda_n q_n, \quad \not{D} \bar{q}_n = \lambda_n \bar{q}_n. \tag{5.102}$$

For non-zero λ_n there is a pairing of solutions with opposite eigenvalues of γ_5. In four-component notation one can see this from

$$\not{D} q_n = \lambda_n q_n \;\rightarrow\; \not{D} \gamma_5 q_n = -\lambda_n \gamma_5 q_n. \tag{5.103}$$

Zero eigenvalues, however, are special. There is no corresponding pairing. This has implications for the fermion functional integral. Writing

$$q(x) = \sum a_n q_n(x), \tag{5.104}$$

$$S = \sum \lambda_n a_n^* a_n \tag{5.105}$$

we have

$$\int [dq][d\bar{q}] e^{-S} = \prod_{n=0}^{\infty} da_n da_n^* \exp\left(-\sum_{n\neq 0} \lambda_n a_n^* a_n\right). \tag{5.106}$$

https://doi.org/10.1017/9781009290883.010 Published online by Cambridge University Press

Because the zero modes do not contribute to the action, many Green's functions vanish. For example, $\langle 1 \rangle = 0$. In order to obtain a non-vanishing result, we need enough insertions of q to "soak up" all the zero modes.

We have seen that, in the instanton background, there are normalizable fermion zero modes, one for each left-handed field. This means that, in order for the path integral to be non-vanishing, we need to include insertions of enough qs and $\bar{q}$s to soak up all the zero modes. In other words, in two-flavor QCD, non-vanishing Green's functions have the form

$$\langle \bar{u}u\bar{d}d \rangle \tag{5.107}$$

and violate the symmetry. Note that the symmetry violation is just as predicted from the anomaly equation,

$$\Delta Q_5 = \frac{2}{16\pi^2} \int d^4x F\tilde{F} = 4. \tag{5.108}$$

This is a particular example of an important mathematical theorem known as the Atiyah–Singer index theorem.

We can put all this together to evaluate a Green's function which violates the classical $U(1)$ symmetry of the massless theory, $\langle \bar{u}(x)u(x)\bar{d}(x)d(x) \rangle$. Taking the gauge group to be $SU(2)$ there is one zero mode for each of u, $\bar{u}$, d and $\bar{d}$. The fields in this expectation value can soak up all these zero modes. The effect of the integration over x_0 is to give a result that is independent of x, since the zero modes are functions only of $x - x_0$. The integration over the rotational zero modes gives a non-zero result only if the Lorentz indices are contracted in a rotationally invariant manner (the same applies to the gauge indices). The integration over the instanton scale size – the conformal collective coordinate – is more problematic, exhibiting precisely the infrared divergence of Eq. (5.100).

So, we have provided some evidence that the $U(1)$ problem is solved in QCD, but no reliable calculation. What about the θ-dependence? Let us ask first about the θ-dependence of the vacuum energy. In order to get a non-zero result, we need to allow that the quarks are massive. Treating the mass as a perturbation, we obtain a result of the form

$$E(\theta) = C\Lambda_{\mathrm{QCD}}^9 m_u m_d \cos\theta \int d\rho\, \rho^{-3} \rho^9. \tag{5.109}$$

So, as in the CP^N model, we have evidence for θ-dependence but cannot do a reliable calculation. That we cannot do a calculation should not be a surprise. There is no small parameter in QCD to use as an expansion parameter. Fortunately, we can use other facts which we know about the strong interactions to get a better handle on both the $U(1)$ problem and the θ-dependence question.

Before continuing, however, let us consider the weak interactions. Here there is a small parameter and there are no infrared difficulties, so we might expect instanton effects to be small. The analog of the $U(1)_5$ symmetry in this case is baryon number. Baryon number has an anomaly in the standard model, since all the quark doublets have the same sign of the baryon number. 't Hooft showed that one could actually use instantons, in this case, to compute the violation of baryon number. Technically, there are no finite-action Euclidean solutions in this theory; this follows, as we will see in a moment, from a simple scaling argument. However, 't Hooft realized that one can construct important configurations

https://doi.org/10.1017/9781009290883.010 Published online by Cambridge University Press

having non-zero topological charge by starting with the instantons of the pure gauge theory and perturbing them. For the Higgs boson, one solves the equation

$$D^2\phi = V'(\phi). \tag{5.110}$$

For a light boson, one can neglect the right-hand side. Then this equation is solved by

$$\phi(x) = i\bar{\sigma}^\mu x^\mu \left(\frac{1}{x^2+\rho^2}\right)^{1/2} \langle\phi\rangle. \tag{5.111}$$

Note that at large x, this has the form $g(x)\langle\phi\rangle$. As a result, the action of the configuration is finite. One finds the following correction to the action:

$$\delta S = \frac{1}{g^2} v^2 \rho^2. \tag{5.112}$$

Including this in the exponential damps the ρ integral at large ρ, and leads to a convergent result.

Now including the fermions, there is a zero mode for each $SU(2)$ doublet. So, one obtains a non-zero expectation value for correlation functions of the form $\langle QQQLLL\rangle$, where the color and $SU(2)$ indices are contracted in a gauge-invariant way and the flavors for the Qs and Ls are all different. The coefficient is

$$\mathcal{A}_{\mathrm{bv}} = C\, e^{-2\pi/\alpha_{\mathrm{w}}}. \tag{5.113}$$

From this, one can see that baryon number violation occurs in the Standard Model but at an incredibly small rate. One can also calculate a term in the effective action, involving three quarks and three leptons, with a similar coefficient by studying Green's functions in which all the fields are widely separated. We will encounter this sort of computation later, when we discuss instantons in supersymmetric theories.

5.3.3 Physical interpretation of the instanton solution

We have derived dramatic physical effects from the instanton solution by direct calculation, but we have not provided a physical picture of the phenomena that the instanton describes. Already in quantum mechanics imaginary-time solutions of the classical equations of motion are familiar in the Wentzel–Kramers–Brillouin (WKB) analysis of tunneling, and the Yang–Mills instanton (and the CP^N instanton) also describe tunneling phenomena. In this subsection we will confine our attention to pure gauge theories. The generalization to theories with fermions and/or scalars is straightforward and interesting.

To understand the instanton in terms of tunneling, it is helpful to work in a non-covariant gauge, in which there is a Hamiltonian description. The gauge $A_0 = 0$ is particularly useful. In this gauge the canonical coordinates are the A_is and their conjugate momenta are the E_is (with a minus sign). This is too many degrees of freedom if all are treated as independent. The resolution lies in the need to enforce Gauss's law, which is now to be viewed as an operator constraint on states. For example, in a $U(1)$ theory,

$$G(\vec{x})|\Psi\rangle = (\vec{\nabla}\cdot\vec{E} - \rho)|\Psi\rangle = 0. \tag{5.114}$$

https://doi.org/10.1017/9781009290883.010 Published online by Cambridge University Press

The left-hand side is almost the generator of gauge transformations. On the gauge fields, for example,

$$\left[\int d^3x\,\omega(\vec{x})G(\vec{x}), A_i(\vec{y})\right] = -\int d^3x\,\partial_j\omega(\vec{x})[E(\vec{x})_j, A(\vec{y})_i] = \partial_i\omega(\vec{y}). \tag{5.115}$$

In the second step we have integrated by parts and dropped a possible surface term. This requires that $\omega \to 0$ fast enough at infinity. Such gauge transformations are called "small". We have learned that, in the $A_0 = 0$ gauge, states must be invariant under time-independent, small, gauge transformations.

In electrodynamics this is not particularly interesting. But the same manipulations hold in non-Abelian theories, and in this case there are interesting large gauge transformations. An example is

$$g(\vec{x}) = \exp\left(i\pi\frac{\vec{x}\cdot\vec{\sigma}}{\sqrt{\vec{x}^2 + a^2}}\right). \tag{5.116}$$

We can also consider powers g^n of g. We can think of g as mapping three-dimensional space into the group $SU(2)$. The number of times that the mapping wraps around the gauge group is known as the winding number, and it can be written as

$$n = \frac{1}{24\pi^2}\int d^3x\,\epsilon_{ijk}\,\mathrm{Tr}(\partial_i g\partial_j g\partial_k g). \tag{5.117}$$

However, g_n is not unique; we can multiply by any small gauge transformation without changing n. The zero-energy states consist of $A_i = ig^{-n}\partial_i g^n$ averaged over the small gauge transformations in such a way as to make them invariant.

With just a little algebra one can show that $n = \int d^3x K_0$, where K^μ is the topological current encountered in Eq. (5.80). So an instanton, in $A_0 = 0$ gauge, corresponds to a tunneling between states of different n. More precisely, there is a non-zero matrix element of the Hamiltonian between states of different n,

$$\langle n|H|n \pm 1\rangle = \epsilon. \tag{5.118}$$

This is analogous to the situation in crystals, and the energy eigenstates are similar to Bloch waves,

$$|\theta\rangle = \sum_n e^{in\theta}|n\rangle, \tag{5.119}$$

with energy $\epsilon\cos\theta$. This θ is precisely the quantity which entered as a parameter in the Lagrangian.

5.3.4 QCD and the $U(1)$ problem

In real QCD we have seen that, on the one hand, instanton configurations violate the axial $U(1)$ symmetry. In general, there is no small parameter which governs the size of this breaking, so there is no reason to expect a light (pseudo)Goldstone. Consistent with this, explicit calculations are infrared divergent. Again, this is not a surprise; there is no small parameter which would justify the use of a semiclassical approximation, but

https://doi.org/10.1017/9781009290883.010 Published online by Cambridge University Press

the instanton analysis we have described makes clear that there is no reason to expect that there is a light Goldstone boson. Actually, while there is no obvious reason why perturbative and semiclassical (instanton) techniques should give reliable results, there are two approximation method techniques available. The first is for large N, where one now allows the N of $SU(N)$ to be large, with g^2N fixed. In contrast with the case of CP^N, this does not give enough simplification to permit explicit computations, but it does allow one to make qualitative statements about the theory. Witten has pointed out a way in which one can relate the mass of the η (or η' if one is thinking in terms of $SU(3) \times SU(3)$ current algebra) to quantities in a theory without quarks. The anomaly is then an effect suppressed by a power of N, in the large-N limit, because the loop diagram contains a factor g^2 but not a factor N. So, for large N it can be treated as a perturbation and the η is almost massless. The quantity $\partial_\mu j_5^\mu$ acts as a creation operator for η (just as $\partial_\mu j_5^{\mu\,3}$ is a creation operator for the π meson), so one can compute the mass if one knows the correlation function at zero momentum,

$$\left\langle \partial_\mu j_5^\mu(x) \partial_\mu j_5^\mu(y) \right\rangle \propto \frac{1}{N^2} \langle F(x)\tilde{F}(x)F(y)\tilde{F}(y)\rangle. \tag{5.120}$$

To leading order in the $1/N$ expansion, the $F\tilde{F}$ correlation function can be computed in the theory without quarks. Witten argued that, while it vanishes order by order in perturbation theory, there is no reason that this correlation function need vanish in the full theory. Attempts have been made to compute this quantity both in lattice gauge theory and using the anti-de Sitter–conformal-theory (AdS–CFT) correspondence discovered in string theory and discussed later in this text. Both methods give promising results.

So, the $U(1)$ problem should be viewed as solved, in the sense that in the absence of any argument to the contrary, there is no reason to think that there should be an extra Goldstone boson in QCD.

The second approximation scheme which gives some control of QCD is known as chiral perturbation theory. The masses of the u, d and s quarks are small compared with the QCD scale, and the mass terms for these quarks in the Lagrangian can be treated as perturbations. This will figure in our discussion in the next section.

5.4 The strong CP problem

5.4.1 The θ-dependence of the vacuum energy

The assumption that the anomaly resolves the $U(1)$ problem in QCD raises another issue. Given that $\int d^4x\, F\tilde{F}$ has physical effects, a θ term in the action has physical effects as well. Since this term is CP odd, this means that there is the potential for strong CP-violating effects. These effects should vanish in the limit of zero quark mass since, in this case, by a field redefinition we can remove θ from the Lagrangian. In the presence of quark masses, the θ-dependence of many quantities can be computed. Consider, for example, the vacuum

https://doi.org/10.1017/9781009290883.010 Published online by Cambridge University Press

energy. In QCD, the quark mass term in the Lagrangian has the form

$$\mathcal{L}_{\rm m} = m_u \bar{u}u + m_d \bar{d}d + \text{h.c.} \tag{5.121}$$

Were it not for the anomaly we could, by redefining the quark fields, take m_u and m_d to be real. Instead, we can define these fields in such a way that there is no $\theta F\tilde{F}$ term in the action but a phase in m_u and m_d. Clearly, we have some freedom in making this choice. In the case where m_u and m_d are equal, it is natural to choose these phases to be the same. We will explain shortly how one proceeds when the masses are different (as they are in nature). So

$$\mathcal{L}_{\rm m} = (m_u \bar{u}u + m_d \bar{d}d)e^{i\theta} + \text{h.c.} \tag{5.122}$$

Now we want to treat this term as a perturbation. At first order, it makes a contribution to the ground-state energy proportional to its expectation value. We have already argued that the quark bilinear forms have non-zero vacuum expectation values, so

$$E(\theta) = (m_u + m_d)\cos\theta \langle \bar{q}q\rangle. \tag{5.123}$$

While without a difficult non-perturbative calculation we cannot calculate the separate quantities on the right-hand side of this expression, we can, using current algebra, relate them to measured quantities. It is shown in Appendix B that

$$m_{\pi^2} f_{\pi^2} = \text{Tr}\,(M_q\langle\mathcal{M}\rangle) = (m_u + m_d)\langle\bar{q}q\rangle. \tag{5.124}$$

Replacing the quark mass terms in the Lagrangian by their expectation values, we can immediately read off the energy of the vacuum as a function of θ:

$$E(\theta) = m_\pi^2 f_\pi^2 \cos\theta. \tag{5.125}$$

This expression can readily be generalized to the case of three light quarks, by similar methods. So, we see that there is real physics in θ even if we do not understand how to do an instanton calculation. In the next section we will calculate a more interesting quantity: the neutron electric dipole moment as a function of θ.

5.4.2 The neutron electric dipole moment

The most interesting physical quantities to study in connection with CP violation are electric dipole moments, particularly that of the neutron, d_n. If CP were badly violated in strong interactions, one might expect $d_n \approx e\,\text{fm} \approx 10^{-14}$ cm (here e is the electron charge). But the experimental limit on the dipole moment is striking,

$$d_n < 10^{-25}\, e\,\text{cm}. \tag{5.126}$$

Using current algebra the leading contribution to the neutron electric dipole moment due to θ can be calculated, and one obtains a limit $\theta < 10^{-9}$. Here we outline the main steps in the calculation; I urge you to work out the details following the reference in the suggested reading. We will simplify the analysis by working in an exact $SU(2)$-symmetric limit, i.e. by taking $m_u = m_d = m$. We again treat the Lagrangian of Eq. (5.122) as a perturbation. We can understand how this term depends on the π fields by making an axial $SU(2)$ transformation on the quark fields. In other words, a background π field can be thought

https://doi.org/10.1017/9781009290883.010 Published online by Cambridge University Press

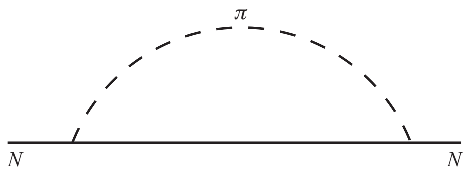

Fig. 5.2 Diagram in which CP-violating coupling of the pion contributes a newtron electric dipole moment d_n.

of as a small chiral transformation on the vacuum. Then, for example, for the τ_3 direction, $q \to (1 + i\pi_3\tau_3)q$ (the π field parameterizes the transformation), so the action becomes

$$\frac{m}{f_\pi}\pi_3(\bar{q}\gamma_5 q + \theta\bar{q}q). \tag{5.127}$$

The second term gives rise to a CP-violating coupling, $\bar{g}_{\pi NN}\pi^a\bar{N}\tau^a N$, of the pions and nucleons N. This is related to the matrix elements of $\bar{q}\tau^a q$ between nucleons. These, in turn, can be estimated by noting that at zero moment they are the matrix elements of an isospin charge operator between nucleons. The latter matrix elements can be estimated using the Gell-Mann and Ne'eman $SU(3)$ symmetry (a similar operator with coefficient m_s is responsible for the splitting between the members of the baryon octet). One obtains, in this way,

$$\bar{g}_{\pi NN} \approx -\theta\frac{(m_\Xi - m_N)m_u m_d}{2f_\pi(m_u + m_d)m_s} \approx 0.38. \tag{5.128}$$

This coupling is difficult to measure directly, but it gives rise, in a calculable fashion, to a neutron electric dipole moment. Consider the graph of Fig. 5.2. This graph generates a neutron electric dipole moment, if we take one coupling to be the standard pion–nucleon coupling and the other the coupling we have computed above. The resulting Feynman graph is infrared divergent; we cut this off at m_π while cutting off the integral in the ultraviolet at the QCD scale. The low-energy calculation is reliable in the limit that m_π is small, so that $\ln(m_\pi/\Lambda_{\rm QCD})$ is large compared to unity. The result is

$$d_{\rm n} = \frac{g_{\pi NN}\bar{g}_{\pi NN}}{4\pi^2 m_N}\ln\frac{M_N}{m_\pi}. \tag{5.129}$$

The matrix element can be estimated using the $SU(3)$ symmetry of Gell-Mann and Ne'eman, as mentioned above, yielding $d_{\rm n} = 5.2 \times 10^{-16}\theta$ cm. The experimental bound gives $\theta < 10^{-9}$–10^{-10}. Understanding why CP violation is so small in strong interactions is known as the strong CP problem.

5.5 Possible solutions of the strong CP problem

What should our attitude towards this problem be? We might argue that, on the one hand, some Yukawa couplings are as small as 10^{-5}, so why is 10^{-9} so bad? On the other hand, we suspect that the smallness of the Yukawa couplings is related to approximate

https://doi.org/10.1017/9781009290883.010 Published online by Cambridge University Press

symmetries, and that these Yukawa couplings are telling us something. Perhaps there is some explanation of the smallness of θ, and perhaps this is a clue to new physics. In this section we review some of the solutions which have been proposed to understand the smallness of θ.

5.5.1 Zero *u* quark mass

Suppose that the mass of the u quark were zero. In this case, by a field redefinition of the u quark

$$u \to e^{-i\theta} u, \tag{5.130}$$

one could make the θ term vanish as a consequence of the anomaly. This would be a simple enough explanation, but there are two issues. First, why should we make this redefinition? We might imagine that it is the result of a symmetry, but this symmetry cannot be a real symmetry of the underlying theory since it is violated by QCD (through the anomaly). We will see later in this book that discrete symmetries, with anomalies of the kind required to understand a vanishing u quark mass, do in fact frequently arise in string theory. So, perhaps this sort of explanation is plausible. We would not, then, expect that the u quark mass should be *exactly* zero but, instead, examining our formula for the neutron electric dipole moment, we would require that the ratio m_u/m_d should be less than about 10^{-10}.

As we described in Chapter 3, however, lattice gauge theory computations establish a non-zero value of the u quark mass with large statistical significance. It is worth noting why researchers in the past contemplated this possibility. Examining the mass spectrum of the pseudoscalar mesons, using the methods of current algebra or chiral Lagrangians (we will discuss these further in Chapter 8), one obtains $m_u/m_d \approx 0.5$. The question, however, is which mass values should actually appear in this formula? In particular, in a theory in which $m_u = 0$ at some high scale, instantons will generate a non-zero mass for m_u at lower scales. The resulting expression is infrared divergent, but we take as the main lesson that it is proportional to $m_d m_s$. Because m_s is not so different from the characteristic scales of QCD, one might imagine that an effective mass of the needed size could be found. It is this possibility which has been excluded by modern lattice computations.

5.5.2 Spontaneous CP violation

Suppose that the underlying theory respects CP and that the observed CP violation is spontaneous. Because θ is CP odd, the underlying theory has $\theta = 0$. One might hope that this feature would be preserved when the symmetry is spontaneously broken. Satisfying this condition and simultaneously generating an order-one CP-violating angle in the CKM matrix is a model-building challenge which we will not review here. Suffice it to say that this can be achieved at tree level. However, existing realizations rely on model-building cleverness and do not have a clear conceptual basis. So, one must ask how plausible is this possibility, and does it survive quantum corrections.

There are a number of ways in which θ might be generated in the low-energy theory. First, suppose that CP is broken by the expectation value of a complex field Φ. There might well be direct couplings such as

$$\frac{1}{16\pi^2}\,(\mathrm{Im}\,\Phi)\,F\tilde{F}. \tag{5.131}$$

Note that Φ might also couple to fermions, giving them a large mass through its expectation value. When these fermions are integrated out this would also generate an effective θ. This is likely, simply because of the anomalous field redefinitions which may be required to make the masses of these fields real. There do exist, however, models which, while complicated, meet the requirements of small θ.

5.5.3 The axion

Perhaps the most compelling explanation of the smallness of θ involves a hypothetical particle called the axion. We present here a slightly updated version of the original idea of Peccei and Quinn.

Consider the vacuum energy as a function of θ (Eq. (5.123)). This energy has a minimum at $\theta = 0$, i.e. at the CP-conserving point. As Weinberg noted long ago, this is almost automatic: points of higher symmetry are necessarily stationary points. As it stands this observation is not particularly useful, since θ is a parameter, not a dynamical variable. But, suppose that one has a field a with coupling to QCD:

$$\mathcal{L}_{\mathrm{axion}} = (\partial_\mu a)^2 + \frac{a/f_a + \theta}{32\pi^2} F\tilde{F}, \tag{5.132}$$

where f_a is known as the axion decay constant. Suppose, in addition, that the rest of the theory possesses a symmetry, called the Peccei–Quinn symmetry,

$$a \to a + \alpha \tag{5.133}$$

for constant α. Then, by a shift in a one can eliminate θ. What we have previously called the vacuum energy as a function of θ, $E(\theta)$, is now $V(a/f_a)$, the potential energy of the axion. It has a minimum at $\theta = 0$. The strong CP problem is solved.

One can estimate the axion mass by simply examining $E(\theta)$, (Eq. 5.125):

$$m_a^2 \approx \frac{m_\pi^2 f_{\pi^2}}{f_a^2}. \tag{5.134}$$

If $f_a \sim$ TeV, this yields a mass of order keV. If $f_a \sim 10^{16}$ GeV, this gives a mass of order 10^{-9} eV.

There are several questions one can raise about this proposal.

- Should the axion already have been observed? The couplings of the axion to matter can be worked out in a given model in a straightforward way, using the methods of current algebra (in particular non-linear Lagrangians). All the couplings of the axion are suppressed by powers of f_a. This is characteristic of a Goldstone boson. At zero momentum a change in the field is like a symmetry transformation so, before including the QCD effects which explicitly break the symmetry, axion couplings are suppressed by powers of momentum over f_a; QCD effects are suppressed by $\Lambda_{\mathrm{QCD}}/f_a$. Thus if f_a is

https://doi.org/10.1017/9781009290883.010 Published online by Cambridge University Press

large enough then the axion is difficult to see. The strongest limit turns out to come from red giant stars. The production of axions is "semiweak", i.e. it is suppressed only by one power of f_a rather than two powers of m_W; as a result, axion emission is competitive with neutrino emission until $f_a > 10^{10}$ GeV or so.

- As we will describe in more detail in Chapter 18, the axion could have been copiously produced in the early universe. As a result there is an *upper bound* on the axion decay constant, of about 10^{11} GeV. If this bound is saturated, the axion constitutes the dark matter. We will discuss this bound in detail in Chapter 19.
- Can one search for the axion experimentally? Typically, the axion couples not only to the $F\tilde{F}$ of QCD but also to the same object in QED. This means that in a strong magnetic field an axion can convert to a photon. Precisely this effect is being searched for by the ADMX experiment at the University of Washington. The basic idea is to suppose that the dark matter in the halo of our galaxy consists principally of axions. Using a (superconducting) resonant cavity with a high Q value in a large magnetic field, one searches for the conversion of these axions into excitations of the cavity due to the coupling of the axion to the electromagnetic field, $F\tilde{F} = \vec{E} \cdot \vec{B}$. The experiments have already reached a level where they set interesting limits; the next generation of experiments will cut a significant swath in the presently allowed parameter space.
- The coupling of the axion to $F\tilde{F}$ violates the shift symmetry; this is why the axion can develop a potential. But this seems rather paradoxical: one is postulating a symmetry, preserved to some high degree of approximation but which is not a symmetry: it is at the least broken by tiny QCD effects. Is this reasonable? To understand the nature of the problem, consider one of the ways in which an axion can arise. In some approximation we can suppose that we have a global symmetry under which a scalar field ϕ transforms as $\phi \to e^{i\alpha}\phi$. Suppose, further, that ϕ has an expectation value. This could arise due to a potential, $V(\phi) = -\mu^2|\phi|^2 + \lambda|\phi|^4$. Associated with the symmetry breaking would be a (pseudo)-Goldstone boson, a. We can parameterize ϕ as follows:

$$\phi = f_a e^{ia/f_a}, \quad |\langle\phi\rangle| = f_a. \tag{5.135}$$

If this field couples to fermions, they gain mass from its expectation value. At one loop, the same diagrams as those discussed in our anomaly analysis generate a coupling $aF\tilde{F}$, from integrating out the fermions. This calculation is identical to the corresponding calculation for pions discussed earlier. But we usually assume that global symmetries in nature are accidents. For example, baryon number is conserved in the Standard Model simply because there are no gauge-invariant renormalizable operators which violate the symmetry. We believe it is violated by higher-dimensional terms. The global symmetry we postulate here is presumably an accident of the same sort. But for the axion, the symmetry must be *extremely* good. We can introduce an axion quality Q_a,

$$Q_a = \frac{1}{m_a^2 f_a}\frac{\partial V}{\partial a}, \tag{5.136}$$

which must be less than 10^{-10}. Suppose, for example, one has a symmetry breaking operator $\phi^{n+4}/M_{\mathrm{p}}^n$. Such a term gives a linear contribution to the axion potential of

https://doi.org/10.1017/9781009290883.010 Published online by Cambridge University Press

order $f_a^{n+3}/M_{\mathrm{p}}^n$. If $f_a \sim 10^{11}$, this swamps the would-be QCD contribution $m_\pi^2 f_\pi^2/f_a$ unless $n > 12$!

This last objection finds an answer in string theory. In this theory there are axions with just the right properties, i.e. there are symmetries in the theory which are *exact* in perturbation theory, but which are broken by exponentially small non-perturbative effects. The most natural value for f_a would appear to be of order M_{GUT} or M_{p}. Whether this can be made compatible with cosmology, or whether one can obtain a lower scale, is an open question to which we will return.

Suggested reading

There are a number of excellent books and reviews on anomalies, as well as good treatments in quantum field theory textbooks. The texts of Peskin and Schroeder (1995), Pokorski (2000) and Weinberg (1995) have excellent treatments of different aspects of anomalies. The string textbook of Green *et al.* (1987) provides a good introduction to anomalies in higher dimensions. One of the best introductions to the physics of instantons is provided in the article of Coleman (1985). The $U(1)$ problem in two-dimensional electrodynamics, and its role as a model for confinement, was discussed by Casher *et al.* (1974). The serious reader should study 't Hooft's instanton paper from 1976, in which he both uncovers much of the physical significance of the instanton solution and also performs a detailed evaluation of the determinant. The propagators in the instanton background are given in Brown *et al.* (1978). Instantons in CP^N models were studied by Affleck (1980). The dependence of d_{n} on θ was calculated by Crewther *et al.* (1979) in a short and quite readable paper.

Exercises

(1) Derive Eq. (5.15).

(2) Calculate the decay rate of the π^0 to two photons. You will need the matrix element

$$\left\langle \pi(q) \middle| \partial_\mu j_5^{\mu 3} \middle| 0 \right\rangle = f_\pi q^\mu e^{iq\cdot x}, \tag{5.137}$$

where $f_\pi = 93\mathrm{MeV}$. You will need also to compute the anomaly in the third component of the axial isospin current.

(3) Fill in the details of the anomaly computation in two dimensions, being careful about signs and factors of 2.

(4) Fill in the details of the Fujikawa computation of the anomaly, in the CP^N model, again being careful about factors of 2. Make sure that you understand why one is calculating

https://doi.org/10.1017/9781009290883.010 Published online by Cambridge University Press

a determinant and why the factors appear in the exponential. Verify that the action of Eq. (5.56) is equal to

$$\mathcal{L} = g_{\phi,\phi^*}\partial_\mu\phi\partial_\mu\phi^*, \tag{5.138}$$

where g is the metric of the sphere in complex coordinates, i.e. it is the line element $dx_1^2 + dx_2^2 + dx_3^2$ expressed as $g_{z,z}dzdz + g_{z,z^*}dzdz^* + g_{z^*z}dz^*dz + g_{dz^*dz^*}dz^*dz^*$. A model with an action of this form is called a non-linear sigma model; the idea is that the fields live on some "target" space, with metric g. Verify Eqs. (5.56) and (5.59).

(5) Check that Eqs. (5.85) and (5.86) solve (5.83).

https://doi.org/10.1017/9781009290883.010 Published online by Cambridge University Press