TASI Lectures on The Strong CP Problem

Michael Dine

Santa Cruz Institute for Particle Physics, Santa Cruz CA 95064

Abstract

These lectures discuss the θ parameter of QCD. After an introduction to anomalies in four and two dimensions, the parameter is introduced. That such topological parameters can have physical effects is illustrated with two dimensional models, and then explained in QCD using instantons and current algebra. Possible solutions including axions, a massless up quark, and spontaneous CP violation are discussed.

1 Introduction

Originally, one thought of QCD as being described a gauge coupling at a particular scale and the quark masses. But it soon came to be recognized that the theory has another parameter, the θ parameter, associated with an additional term in the lagrangian:

$$\mathcal{L} = \theta \frac{1}{16\pi^2} F^a_{\mu\nu} \tilde{F}^{\mu\nu a} \tag{1}$$

where

$$\tilde{F}^a_{\mu\nu\rho\sigma} = \frac{1}{2}\epsilon_{\mu\nu\rho\sigma} F^{\rho\sigma a}. \tag{2}$$

This term, as we will discuss, is a total divergence, and one might imagine that it is irrelevant to physics, but this is not the case. Because the operator violates CP, it can contribute to the neutron electric dipole moment, d_n. The current experimental limit sets a strong limit on θ, $\theta \ll 10^{-9}$. The problem of why θ is so small is known as the strong CP problem. Understanding the problem and its possible solutions is the subject of this lectures.

In thinking about CP violation in the Standard Model, one usually starts by counting the parameters of the unitary matrices which diagonalize the quark and lepton masses, and then counting the number of possible redefinitions of the quark and lepton fields. In doing this counting, however, there is a subtlety. One of the field redefinitions induces a new term in the lagrangian, precisely the θ term above. The obstruction to making this transformation is known as an anomaly, and it is not difficult to understand.

Before considering real QCD, consider a simpler theory, with only a single flavor of quark. Before making any field redefinitions, the lagrangian takes the form:

$$\mathcal{L} = -\frac{1}{4g^2}F^2_{\mu\nu} + \bar{q}\,\slashed{q}^* + q\,\slashed{q}^* m\bar{q}q + m^*\bar{q}^*q^*. \tag{3}$$

Here, I have written the lagrangian in terms of two-component fermions, and noted that a priori, the mass need not be real,

$$m = |m|e^{i\theta}. \tag{4}$$

In terms of four-component fermions,

$$\mathcal{L} = \text{Re } m \ \bar{q}q + \text{Im } m \ q\bar{q}\gamma_5 q. \tag{5}$$

In order to bring the mass term to the conventional form, with no γ_5's, one would naively let

$$q \to e^{-i\theta/2}q \qquad \bar{a} \to e^{-i\theta/2}\bar{q}. \tag{6}$$

However, a simple calculation shows that there is a difficulty associated with the anomaly. Suppose, first, that M is very large. In that case we want to integrate out the quarks and obtain a low energy effective theory. To do this, we study the path integral:

$$Z = \int [dA_\mu] \int [dq][d\bar{q}]e^{iS} \tag{7}$$

2

Again suppose $m = e^{i\theta}M$, where θ is small and M is real. 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 $q \to^{-i\theta/2\gamma_5} q$).) 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:

$$Z = \int [dA_\mu] \int e^{iS_{eff}} \tag{8}$$

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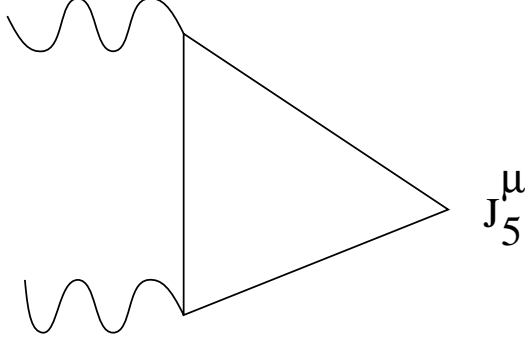

Figure 1: The triangle diagram associated with the four dimensional anomaly.

Because the field redefinition which eliminates θ is 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 making the transformation, treat the θ term as a small perturbation, and expand the exponential. Now consider a term in the effective action with two external gauge bosons. This is given by the Feynman diagram in fig. 1. The corresponding term in the action is given by

$$\delta\mathcal{L}_{eff} = -i\frac{\theta}{2}g^2 M Tr(T^a T^b) \int \frac{d^4k}{(2\pi)^4} \text{Tr}\gamma_5 \frac{1}{\slashed{k} + \slashed{q}_1 - M} \slashed{\epsilon}_1 \frac{1}{\slashed{k} - M} \slashed{\epsilon}_2 \frac{1}{\slashed{k} - \slashed{q}_2 - M} \tag{9}$$

Here, as in the figure, the q_i's are the momenta of the two photons, while the ϵ's are their polarizations and a and b are the color indices of the gluons. To perform the integral, it is convenient to introduce Feynman parameters and shift the k integral, giving:

$$\delta\mathcal{L}_{eff} = -i\theta g^2 M \text{Tr}(T^a T^b) \int d\alpha_1 d\alpha_2 \int \frac{d^4k}{(2\pi)^4} \text{Tr}\gamma_5 (\slashed{k} - \alpha_1 \slashed{q}_1 + \alpha_2 \slashed{q}_2 + \slashed{q}_1 + M) \slashed{\epsilon}_1 \tag{10}$$

$$\frac{(\slashed{k} - \alpha_1 \slashed{q}_1 + \alpha_2 \slashed{q}_2 + M) \slashed{\epsilon}_2 (\slashed{k} - \alpha_1 \slashed{q}_1 + \alpha_2 \slashed{q}_2 - \slashed{q}_2 + M)}{(k^2 - M^2 + O(q_i^2))^3} \tag{11}$$

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

$$\delta\mathcal{L}_{eff} = g^2 M^2 \theta \text{Tr}(T^a T^b) \epsilon_{\mu\nu\rho\sigma} q_1^\mu q_2^\nu \epsilon_1^\rho \epsilon_2^\sigma \int \frac{d^4k}{(2\pi)^4} \frac{1}{(k^2 - M^2)^3}. \tag{12}$$

This corresponds to a term in the effective action, after doing the integral over k and including a combinatoric factor of two from the different ways to contract the gauge bosons:

$$\delta\mathcal{L}_{eff} = \frac{1}{32\pi^2}\theta \text{Tr}(F\tilde{F}). \tag{13}$$

3

Now why does this happen? At the level of the path integral, the transformation would seem to be 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. 1, one sees that it contains terms which are linearly divergent, and thus it should be regulated. A simple way to regulate the diagram is to introduce a Pauli-Villars regulator, which means that one subtracts off a corresponding amplitude with some very large mass Λ. However, we have just seen that the result is independent of Λ! This sort of behavior is characteristic of an anomaly.

Consider now the case that $m \ll \Lambda_{QCD}$. In this case, we shouldn't integrate out the quarks, but we still need to take into account the regulator diagrams. For small m, the classical theory has an approximate symmetry under which

$$q \to e^{i\alpha} q \qquad \bar{q} \to e^{i\alpha} \bar{q} \tag{14}$$

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

$$j_5^\mu = \bar{q}\gamma_5\gamma_\mu q, \tag{15}$$

and classically,

$$\partial_\mu j_5^\mu = m\bar{q}\gamma_5 q. \tag{16}$$

Under a transformation by an infinitesmal angle α one would expect

$$\delta L = \alpha \partial_\mu j_5^\mu = m\alpha \bar{q}\gamma_5 q. \tag{17}$$

But what we have just discovered is that the divergence of the current contains another, m-independent, term:

$$\partial_\mu j_5^\mu = m\bar{q}\gamma_5 q + \frac{1}{32\pi^2} F\tilde{F}. \tag{18}$$

This anomaly can be derived in a number of other ways. One can define, for example, the current by "point splitting,"

$$j_5^\mu = \bar{q}(x + i\epsilon)e^{i \int_x^{x+\epsilon} dx^\mu A_\mu} q(x) \tag{19}$$

Because 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. A beautiful derivation, closely related to that we have performed above, is due to Fujikawa, described in [1]. 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[1]. We will do a calculation along these lines in a two dimensional model shortly.

The anomaly has important consequences in physics which will not be the subject of the lecture today, but it is worth at least listing a few before we proceed:

- π^o decay: the divergence of the axial isospin current,

$$(j_5^3)^\mu = \bar{u}\gamma_5\gamma^\mu\bar{u} - \bar{d}\gamma_5\gamma^\mu d \tag{20}$$

  has an anomaly due to electromagnetism. This gives rise to a coupling of the π^o to two photons, and the correct computation of the lifetime was one of the early triumphs of the theory of quarks with color.

- Anomalies in gauge currents signal an inconsistency in a theory. They mean that the 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 gauge anomalies is one of the striking ingredients of the standard model, and it is also crucial in extensions such as string theory.

Our focus in these lectures will be on another aspect of this anomaly: the appearance of an additional parameter in the standard model, and the associated "Strong CP problem."

What we have just learned is that, if in our simple model above, we require that the quark masses are real, we must allow for the possible appearance in the lagrangian of the standard model, of the θ-terms of eqn. 1. [a] This term, however, can be removed by a $B + L$ transformation. What are the consequences of these terms? We will focus on the strong interactions, for which these terms are most important. At first sight, one might guess that these terms are in fact of no importance. Consider, first, the case of QED. Then

$$\int d^4 x F\tilde{F} \tag{21}$$

is a the integral of a total divergence,

$$F\tilde{F} = \vec{E} \cdot \vec{B} = \frac{1}{2}\partial_\mu \epsilon^\mu_{\nu\rho\sigma} A^\nu F^{\rho\sigma}. \tag{22}$$

As a result, this term does not contribute to the classical equations of motion. One might expect that it does not contribute quantum mechanically either. If we think of the Euclidean path integral, configurations of finite action have field strengths, $F_{\mu\nu}$ which fall off faster than $1/r^2$ (where r is the Euclidean distance), and A which falls off faster than $1/r$, so one can neglect surface terms in $\mathcal{L}_\theta$.

(A parenthetical remark: This is almost correct. However, if there are magnetic monopoles there is a subtlety, first pointed out by Witten[2]. Monopoles can carry electric charge. In the presence of the θ term, there is an extra source for the electric field at long distances, proportional to θ and the monopole charge. So the electric charges are given by:

$$Q = n_e e - \frac{e\theta n_m}{2\pi} \tag{23}$$

[a]In principle we must allow a similar term, for the weak interactions. However, $B + L$ is a classical symmetry of the renormalizable interactions of the standard model. This symmetry is anomalous, and can be used to remove the weak θ term. In the presence of higher dimension $B + L$-violating terms, this is no longer true, but any effects of θ will be extremely small, suppressed by $e^{-2\pi/\alpha_w}$ as well as by powers of some large mass scale.

where n_m is the monopole charge in units of the Dirac quantum.)

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:

$$F\tilde{F} = \partial^\mu K_\mu \qquad K_\mu = \epsilon_{\mu\nu\rho\sigma}(A_\nu^a F_{\rho\sigma}^a - \frac{2}{3}f^{abc}A_\nu^a A_\rho^b A_\sigma^c). \tag{24}$$

But now the statement that F falls faster than $1/r^2$ does not permit an equally strong statement about A. We will see shortly that there are finite action configurations – finite action classical solutions – where $F \sim \frac{1}{r^4}$, but $A \to \frac{1}{r}$, so that the surface term cannot be neglected. These terms are called instantons. It is because of this that θ can have real physical effects.

2   A Two Dimensional Detour

Before considering four dimensions with all of its complications, it is helpful to consider two dimensions. Two dimensions are often 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.

2.1   The Anomaly In Two Dimensions

Consider, first, electrodynamics of a massless fermion in two dimensions. Let's investigate the anomaly. The point-splitting method is particularly convenient here. Just as in four dimensions, we write:

$$j_5^\mu = \bar{\psi}(x + i\epsilon)e^{i\int_x^{x+\epsilon} A_\rho dx^\rho}\gamma^\mu\psi(x) \tag{25}$$

For very small ϵ, we can pick up the leading singularity in the product of $\psi(x + \epsilon)\psi$ by using the operator product expansion, and noting that (using naive dimensional analysis) the leading operator is the unit operator, with coefficient proportional to $1/\epsilon$. We can read off this term by taking the vacuum expectation value, i.e. by simply evaluating the propagator. String theorists are particularly familiar with this Green's function:

$$\langle\bar{\psi}(x + \epsilon)\psi(x)\rangle = \frac{1}{2\pi}\frac{\not{\epsilon}}{\epsilon^2} \tag{26}$$

Expanding the factor in the exponential to order ϵ gives

$$\partial_\mu j_5^\mu = \text{naive piece} + \frac{i}{2\pi}\partial_\mu\epsilon_\rho A^\rho \text{tr}\frac{\not{\epsilon}}{\epsilon^2}\gamma^\mu\gamma^5. \tag{27}$$

Taking the trace gives $\epsilon_{\mu\nu}\epsilon^\nu$; averaging ϵ over angles ($< \epsilon_\mu\epsilon_\nu >= \frac{1}{2}\eta_{\mu\nu}\epsilon^2$), yields

$$\partial_\mu j_5^\mu = \frac{1}{4\pi}\epsilon_{\mu\nu}F^{\mu\nu}. \tag{28}$$

Exercise: Fill in the details of this computation, being careful about signs and factors of 2.

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

$$\partial_\mu j_5^\mu = \frac{1}{2\pi}\epsilon_{\mu\nu}\partial^\mu A^\nu. \tag{29}$$

So it appears possible to define a new current,

$$J^\mu = j_5^\mu - \frac{1}{2\pi}\epsilon_{\mu\nu}\partial^\mu A^\nu \tag{30}$$

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. Indeed, if one has charges, $\pm\theta$ at infinity, they give rise to a constant electric field, $F_{oi} = e\theta$. So θ has a very simple interpretation in this theory.

It is easy to see that 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 ∞.

2.2   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 the CP^N model (our treatment here will follow closely the treatment in Peskin and Schroeder's problem 13.3[1]). This model starts with a set of fields, z_i, $i = 1,\ldots N+1$. These fields live in the space CP^N. This space is defined by the constraint:

$$\sum_i |z_i|^2 = 1; \tag{31}$$

in addition, the point z_i is equivalent to $e^{i\alpha}z_i$. To implement the first of these constraints, 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," CP^N, must hold at every point in *ordinary space-time*, so this is a $U(1)$ gauge symmetry. So introducing a gauge field, A_μ, and the corresponding covariant derivative, we want to study the lagrangian:

$$\mathcal{L} = \frac{1}{g^2}[|D_\mu z_i|^2 - \lambda(x)(|z_i|^2 - 1)] \tag{32}$$

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

$$\mathcal{L} = \frac{1}{g^2}[|\partial_\mu z_j|^2 + |z_j^*\partial_\mu z_j|^2] \tag{33}$$

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

$$Z = \int [dA][d\lambda][dz_j]exp[-\mathcal{L}] \tag{34}$$

$$= \int [dA][d\lambda]e^{\int d^2x\Gamma_{eff}[A,\lambda]} \tag{35}$$

7

$$= \int [dA][d\lambda]exp[-N\mathrm{tr}\log(-D^2 - \lambda) - \frac{1}{g^2}\int d^2x\lambda]$$

2.3 The Large N Limit

By itself, the result of eqn. 35 is still rather complicated. The fields A_μ and λ have complicated, non-local interactions. Things become much simpler if one takes the "large N limit", a limit where one takes $N \to \infty$ with $g^2 N$ 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

$$\frac{\delta\Gamma_{eff}}{\delta\lambda} = 0 \tag{36}$$

or, setting the gauge field to zero,

$$N\int \frac{d^2k}{(2\pi)^2}\frac{1}{k^2 + \lambda} = \frac{1}{g^2}, \tag{37}$$

giving

$$\lambda = m^2 = M\exp[-\frac{2\pi}{g^2 N}]. \tag{38}$$

Here, M is a cutoff required because the integral in eqn. 37 is divergent. This result is remarkable. One has exhibited dimensional transmutation: a theory which is classically scale invariant contains non-trivial masses, related in a renormalization-group invariant fashion to the cutoff. This is the phenomenon which in QCD explains the masses of the proton, neutron, and other dimensionful quantities. So the theory is quite analogous to QCD. We can read off the leading term in the β-function from the familiar formula:

$$m = Me^{-\int\frac{dg}{\beta(g)}} \tag{39}$$

so, with

$$\beta(g) = -\frac{1}{2\pi}g^3 b_o \tag{40}$$

we have $b_o = 1$.

But most important for our purposes, it is interesting to explore the question of θ-dependence. Again, in this theory, we could have introduced a θ term:

$$\mathcal{L}_\theta = \frac{\theta}{2\pi}\int d^2\epsilon_{\mu\nu}F^{\mu\nu}, \tag{41}$$

where $F_{\mu\nu}$ 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 we discussed above, this term is physically important. In perturbation theory in the model, this is not entirely obvious. But 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 $1/N$), one generates a kinetic term for A through the usual vacuum polarization loop:

$$\mathcal{L}_{kin} = \frac{N}{2\pi m^2}F_{\mu\nu}^2. \tag{42}$$

8

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

There is, in addition, the possibility of including other light fields, for example massless fermions. In this case, one can again have an anomalous $U(1)$ symmetry. There is then no θ-dependence, since it is possible to shield any charge at infinity. But there is non-trivial breaking of the 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 $U(1)$ in such a theory is a well-studied phenomenon[3].

Exercise: Complete Peskin and Schroeder, Problem 13.3.

2.4   The Role of Instantons

There is another way to think about the breaking of the $U(1)$ symmetry and 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. In order that they be potentially important, it is necessary that these solutions have 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. Such solutions are not difficult to find in the CP^N model; we will describe them briefly below. These solutions carry non-zero values of the topological charge,

$$\frac{1}{2\pi} \int d^2 x \epsilon_{\mu\nu} F_{\mu\nu} = n \tag{43}$$

and have an action $2\pi n$. As a result, they contribute to the θ-dependence; they give a contribution to the functional integral:

$$Z_{inst} = e^{\frac{-2\pi n}{g^2}} e^{in\theta} \int [d\delta z_j] e^{-\delta z_i \frac{\delta^2 S}{\delta z_i \delta z_j} \delta z_j} + \dots \tag{44}$$

It follows that:

- Instantons generate θ-dependence

- In the large N limit, instanton effects are, formally, highly suppressed, much smaller that the effects we found in the large N limit

- Somewhat distressingly, the functional integral above can not be systematically evaluated. The problem is that the classical theory is scale invariant, as a result of which, instantons come in a variety of sizes. $\int [d\delta z]$ includes an integration over all instanton sizes, which diverges for large size (i.e. in the infrared). This prevents a systematic evaluation of the effects of instantons in this case. At high temperatures[4], it is possible to do the evaluation, and instanton effects are, indeed, systematically small.

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 the theory in terms of a single complex field, ϕ. One can, for example, define $\phi = z_1/z_2$. Then, with a bit of algebra, one can show that the action for ϕ takes the form:

$$\mathcal{L} = (\partial_\mu \phi \partial_\mu \phi^*)\frac{1}{1 + \phi^* \phi} - \frac{\phi^* \phi}{(1 + \phi^* \phi)^2}. \tag{45}$$

One can think of the field ϕ as living on the space with metric given by the term in parenthesis, $g_{\phi\phi^*}$. One can show that this is the metric one obtains if one stereographically maps the sphere onto the complex plane. This mapping, which you may have seen in your math methods courses, is just:

$$z = \frac{x_1 + ix_2}{1 - x_3}; \tag{46}$$

The inverse is

$$x_1 = \frac{z + z^*}{1 + |z|^2} \qquad x_2 = \frac{z - z^*}{i(1 + |z|^2)} \tag{47}$$

$$x_3 = \frac{|z|^2 - 1}{|z|^2 + 1}.$$

It is straightforward to write down the equations of motion:

$$\partial^2 \phi g_{\phi\bar\phi} + \partial_\mu \phi (\partial_\mu \bar\phi \frac{\partial g}{\partial \bar\phi} + \partial_\mu \phi \frac{\partial g}{\partial \phi}) = 0 \tag{48}$$

Now calling the space time coordinates $z = x_1 + ix_2$, $z^* = x_1 - ix_2$, you can see that if ϕ is analytic, the equations of motion are satisfied! So a simple solution, which you can check has finite action, is

$$\phi(z) = \rho z. \tag{49}$$

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

$$\frac{1}{2\pi} \int d^2 x \epsilon_{\mu\nu} F^{\mu\nu} = 1 \tag{50}$$

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

Exercise: Verify that the action of eqn. 45 is equal to

$$\mathcal{L} = g_{\phi,\phi^*} \partial_\mu \phi \partial_\mu \phi^* \tag{51}$$

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 $g_{z,z} dz\ dz + g_{z,z^*} dz\ dz^* + g_{z^*z} dz^*\ dz + g_{dz^* dz^*} dz^* dz^*$. 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 eqns. 45,47.

More generally, $\phi = \frac{az+b}{cz+d}$ 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

10

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. If one studies a Green's function,

$$< \phi(x)\phi(y) > \approx \int d^2 x_o \phi_{cl}(x - x_o)\phi_{cl}(y - y_o)e^{-S_o} \tag{52}$$

The precise measure is obtained by the Fadeev-Popov method. Similarly, the integration over the parameter ρ yields a factor

$$\int d\rho \rho^{-1} e^{-\frac{2\pi}{g^2(\rho)}} \ldots \tag{53}$$

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[5]. 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 one can consider. For example, one can consider adding massless fermions to the model, by simply coupling them in a gauge-invariant way to A_μ. The resulting theory has a chiral $U(1)$ symmetry, which is anomalous. In the presence of an instanton, one can easily construct normalizable fermion zero modes (the Dirac equation just becomes the statement that ψ is analytic). As a result, Green's functions computed in the instanton background do not respect the axial $U(1)$ symmetry. But rather than get too carried away with this model (I urge you to get a little carried away and play with it a bit), let's proceed to four dimensions, where we will see very similar phenomena.

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 to that which we found in the large N limit of the CP^N model. As in that theory:

- 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 there should be real θ-dependence. 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.

- There is another approach to the theory, which shows that the θ-dependence is real, and allows computation of these effects. In QCD, this is related to the breaking of chiral symmetries.

While it is not in the spirit of much of this school, which is devoted to the physics of heavy quarks, it is sufficient, to understand the effects of θ, to focus on only the light quark sector of QCD. For simplicity in writing some of the formulas, we will consider 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 lighter 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 $SU(2)_L \times SU(2)_R$. This symmetry is spontaneously broken to a vector $SU(2)$. The three resulting Goldstone bosons are the π mesons. Calling

$$q = \begin{pmatrix} u \\ d \end{pmatrix} \qquad \bar{q} = \begin{pmatrix} \bar{u} \\ \bar{d} \end{pmatrix}, \tag{54}$$

the two $SU(2)$ symmetries act separately on q and $\bar{q}$ (thought of as left handed fermions). The order parameter for the symmetry breaking is believed to be the condensate:

$$\mathcal{M}_o = \langle \bar{q}q \rangle. \tag{55}$$

This indeed breaks the symmetry down to the vector sum. The associated Goldstone bosons are the π mesons. One can think of the Goldstone bosons as being associated with a slow variation of the expectation value in space, so we can introduce a composite operator

$$\mathcal{M} = \bar{q}q = M_o e^{i\frac{\pi_a(x)\tau_a}{f_\pi}} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \tag{56}$$

The quark mass term in the lagrangian is then (for simplicity writing $m_u = m_d = m_q$; more generally one should introduce a matrix)

$$m_q \mathcal{M}. \tag{57}$$

Expanding $\mathcal{M}$ in powers of π/f_π, it is clear that the minimum of the potential occurs for $\pi_a = 0$. Expanding to second order, one has

$$m_\pi^2 f_\pi^2 = m_q M_o. \tag{58}$$

But we have been a bit cavalier about the symmetries. The theory also has two $U(1)$'s;

$$q \to e^{i\alpha}q \qquad \bar{q} \to e^{i\alpha}\bar{q} \tag{59}$$

$$q \to e^{i\alpha}q \qquad \bar{q} \to e^{-i\alpha}\bar{q} \tag{60}$$

The first of these is baryon number and it is not chiral (and is not broken by the condensate). The second is the axial $U(1)_5$; It is also broken by the condensate. So, in addition to the pions, there should be another approximate Goldstone boson. The best candidate is the η, but, as we will see below (and as you will see further in Thomas's lectures), the η 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

$$\partial_\mu j_5^\mu = \frac{2}{32\pi^2} F\tilde{F} \tag{61}$$

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

$$F\tilde{F} = \partial_\mu K^\mu \tag{62}$$

where

$$K_\mu = \epsilon_{\mu\nu\rho\sigma}(A_\nu^a F_{\rho\sigma}^a - \frac{2}{3} f^{abc} A_\nu^a A_\rho^b A_\sigma^c). \tag{63}$$

So if it is true that this term accounts for the absence of the Goldstone boson, we need to show that there are important configurations in the functional integral for which the rhs does not vanish rapidly at infinity.

3.2 Instantons

It is easiest to study the Euclidean version of the theory. This is useful if we are interested in very low energy processes, which can be described by an effective action expanded about zero momentum. In the functional integral,

$$Z = \int [dA][dq][d\bar{q}] e^{-S} \tag{64}$$

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, non-linear equations, but it turns out that, much as in the CP^N model, the instanton solutions can be found rather easily[6]. The following tricks simplify the construction, and turn out to yield the general solution. First, note that the Yang-Mills action satisfies an inequality:

$$\int (F \pm \tilde{F})^2 = \int (F^2 + \tilde{F}^2 \pm 2F\tilde{F}) = \int (2F^2 + 2F\tilde{F}) \geq 0. \tag{65}$$

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

$$F = \pm\tilde{F} \tag{66}$$

i.e. if the gauge field is (anti) self dual.[b] 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:

$$A_\mu = f(r^2) + h(r^2)\vec{x} \cdot \vec{\tau} \tag{67}$$

---

[b]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 is supersymmetric. Then one can show that if some of the 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.

where we are using a matrix notation for the gauge fields. One can actually make a better guess: define the gauge transformation:

$$g(x) = \frac{x_4 + i\vec{x} \cdot \vec{\tau}}{r} \tag{68}$$

and take

$$A_\mu = f(r^2) g \partial_\mu g^{-1} \tag{69}$$

Then plugging in the Yang-Mills equations yields:

$$f = \frac{r^2}{r^2 + \rho^2} \tag{70}$$

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

Exercise Check that eqns. 69,70 solve 66.

¿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,

$$\int F^2 = \int \tilde{F}^2 = 32\pi^2 \tag{71}$$

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

$$\frac{1}{32\pi^2} \int d^4x F\tilde{F} \tag{72}$$

counts the number of times g maps the sphere at infinity into the group (one for this specific example; n more generally). We do not have time to explore all of this in detail; I urge you to look at Sidney Coleman's lecture, "The Uses of Instantons"[7]. To actually do calculations, 't Hooft developed some notations which are often more efficient than those described above[5].

So we have exhibited potentially important contributions to the path integral which violate the $U(1)$ symmetry. How does this violation of the symmetry show up? Let's think about the path integral in a bit more detail. Having found a classical solution, we want to integrate about small fluctuations about it:

$$Z = e^{-\frac{8\pi^2}{g^2}} e^{i\theta} \int [d\delta A][dq][d\bar{q}] e^{i\delta^2 S} \tag{73}$$

Now S contains an explicit factor of $1/g^2$. 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 of inverse square root determinants for the bosons. Consider, first, the integral over the fermions. It is straightforward, if challenging, to evaluate the determinants[5]. But if the quark

14

masses are zero, the fermion functional integrals are 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 equation:

$$\not{D}u = 0 \qquad \not{D}\bar{u} = 0 \tag{74}$$

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

$$u = \frac{\rho}{(\rho^2 + (x - x_o)^2)^{3/2}}\zeta \tag{75}$$

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

This means that in order for the path integral to be non-vanishing, we need to include insertions of enough q's and $\bar{q}$'s to soak up all of the zero modes. In other words, non-vanishing Green's functions have the form

$$\langle \bar{u}u\bar{d}d \rangle \tag{76}$$

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

$$\Delta Q_5 = 4\frac{1}{\pi^2} \int d^4x F\tilde{F} = 4 \tag{77}$$

However, the calculation we have described here is not self consistent. The difficulty is that among the variations of the fields we need to integrate over are changes in the location of the instanton (translations), rotations of the instanton, and scale transformations. The translations are easy to deal with; one has simply to integrate over x_o (one must also include a suitable Jacobian factor[7]). Similarly, one must integrate over ρ. There is a power of ρ arising from the Jacobian, which can be determined on dimensional grounds. For our Green's function above, for example, which has dimension 6, we have (if all of the fields are evaluated at the same point),

$$\int d\rho \rho^{-7}. \tag{78}$$

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

$$e^{-8\pi^2 g^2(\rho)} \approx (\rho M)^{b_o} \tag{79}$$

for small ρ. For 3 flavor QCD, for example, $b_o = 9$, and the ρ integral diverges for large ρ. This is just the statement that the integral is dominated by the infrared, where the QCD coupling becomes strong.

So we have provided some evidence that the $U(1)$ problem is solved in QCD, but no reliable calculation. What about θ-dependence? Let us ask first about 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

$$E(\theta) = C\Lambda_{QCD}^9 m_u m_d \cos(\theta) \int d\rho \rho^{-3}\rho^9. \tag{80}$$

15

So again, 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 question of θ-dependence.

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 of the quark doublets have the same sign of the baryon number. 't Hooft realized 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 of non-zero topological charge by starting with the instantons of the pure gauge theory and perturbing them. If one simply takes such an instanton, and plugs it into the action, one necessarily finds a correction to the action of the form

$$\delta S = \frac{1}{g^2} v^2 \rho^2. \tag{81}$$

This damps the ρ integral at large ρ, and leads to a convergent result. Affleck showed how to develop this into a systematic computation[9]. Note that from this, one can see that baryon number violation occurs in the standard model, and that the rate is incredibly small, proportional to $e^{-2\frac{\pi}{\alpha_W}}$.

3.3 Real QCD and the U(1) Problem

In real QCD, it is difficult to do a reliable calculation which shows that there is not an extra Goldstone boson, but the instanton analysis we have described makes clear that there is no reason to expect one. Actually, while perturbative and semiclassical (instanton) techniques have no reason to give reliable results, there are two approximation methods techniques which are available. The first is large N, where one now allows the N of $SU(N)$ to be large, with $g^2 N$ fixed. In contrast to the case of CP^N, this does not permit enough simplification to do explicit computations, but it does allow one to make qualitative statements about the theory. available in QCD. Witten has pointed out a way in which one can at 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 point is to note that the anomaly is an effect suppressed by a power of N, in the large N limit. This is because the loop diagram contains a factor of g^2 but not of N. So, in large N, it can be treated as a perturbation, and the the η is massless. $\partial_\mu j_5^\mu$ is like a creation operator for the η, so (just like $\partial_\mu j_5^{\mu\,3}$ is a creation operator for the π meson), so one can compute the mass if one knows the correlation function, at zero momentum, of

$$\langle \partial_\mu j_5^\mu(x) \partial_\mu j_5^\mu(y) \rangle \propto \frac{1}{N^2} \langle F(x)\tilde{F}(x) F(y)\tilde{F}(y) \rangle \tag{82}$$

To leading order in the $1/N$ expansion, this correlation function can be computed in the theory without quarks. Witten argued that while this vanishes order by order in perturbation theory, there is no reason

16

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 AdS-CFT correspondence recently discovered in string theory. Both methods give promising results.

So the U(1) problem should be viewed as solved, in the sense that absent 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 to 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.

3.4   Other Uses of Instantons: A Survey

In the early days of QCD, it was hoped that instantons, being a reasonably well understood non-perturbative effect, might give insight into many aspects of the strong interactions. Because of the infrared divergences discussed earlier, this program proved to be a disappointment. There was simply no well-controlled approximation to QCD in which instantons were important. Indeed, Witten[10] stressed the successes of the large N limit in understanding the strong interactions, and argued that in this limit, anomalies could be important but instantons would be suppressed exponentially. This reasoning (which I urge you to read) underlay much of our earlier discussion, which borrowed heavily on this work.

In the years since Coleman's "Uses of Instantons" was published[7], many uses of instantons in controlled approximations have been found. What follows is an incomplete list; I hope this will inspire some of you to read Coleman's lectures and develop a deeper understanding of the subject.

- θ-dependence at finite temperature: Within QCD, instanton calculations are reliable at high temperatures. So, for example, one can calculate the θ-dependence of the vacuum energy in the early universe, and other quantities to which instantons give the leading contribution[11].

- Baryon number violation in the standard model: We have remarked that this can be reliably calculated, though it is extremely small. However, as explained in [7], instanton effects are associated with tunneling, and in the standard model, they describe tunneling between states with different baryon number. It is reasonable to expect that baryon number violation is enhanced at high temperature, where one has plenty of energy to pass over the barrier without tunneling. This is indeed the case. This baryon number violation might be responsible for the matter-antimatter asymmetry which we observe[12].

- Instanton effects in supersymmetric theories: this has turned out to be a rich topic. Instantons, in many instances, are the leading effects which violate non-renormalization theorems in perturbation theory, and they can give rise to superpotentials, supersymmetry breaking, and other phenomena.

17

More generally, they have provided insight into a whole range of field theory and string theory phenomena.

4 The Strong CP Problem

4.1 θ-dependence of the Vacuum Energy

The fact that the anomaly resolves the $U(1)$ problem in QCD, however, raises another issue. Given that $\int d^4x F\tilde{F}$ has physical effects, the theta 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 masses, 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 energy. In QCD, the quark mass term in the lagrangian has the form:

$$\mathcal{L}_m = m_u \bar{u}u + m_d \bar{d}d + \text{h.c.} \tag{83}$$

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 so that there is no $\theta F\tilde{F}$ term in the action, but there is a phase in m_u and m_d. Clearly, we have some freedom in making this choice. In the case that 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). We can, by convention, take θ to be the phase of the overall lagrangian:

$$\mathcal{L}_m = (m_u \bar{u}u + m_d \bar{d}d)\cos(\theta/2) + \text{h.c.} \tag{84}$$

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 bilinears have non-zero vacuum expectation values, so

$$E(\theta) = (m_u + m_d)e^{i\theta}\langle\bar{q}q\rangle. \tag{85}$$

While, without a difficult non-perturbative calculation, we can't calculate the separate quantities on the right hand side of this expression, we can, using current algebra, relate them to measured quantities. A simple way to do this is to use the effective lagrangian method (which will be described in more detail in Thomas's lectures). The basic idea is that at low energies, the only degrees of freedom which can readily be excited in QCD are the pions. So parameterize $\bar{q}q$ as

$$\bar{q}q = \Sigma = <\bar{q}q> e^{i\frac{\pi^a(x)\sigma^a}{2f_\pi}} \tag{86}$$

We can then write the quark mass term as

$$\mathcal{L}_m = e^{i\theta}TrM_q\Sigma. \tag{87}$$

18

Ignoring the θ term at first, we can see, plugging in the explicit form for Σ, that

$$m_\pi^2 f_\pi^2 = (m_u + m_d) < \bar{q}q > . \tag{88}$$

So the vacuum energy, as a function of θ, is:

$$E(\theta) = m_\pi^2 f_\pi^2 \cos(\theta). \tag{89}$$

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

4.2   The Neutron Electric Dipole Moment

As Scott Thomas will explain in much greater detail in his lectures, the most interesting physical quantities to study in connection with CP violation are electric dipole moments, particularly that of the neutron, d_n. It has been possible to set strong experimental limits on this quantity. Using current algebra, the leading contribution to the neutron electric dipole moment due to θ can be calculated, and one obtains a limit $\theta < 10^{-98}$. The original paper on the subject is quite readable. Here we outline the main steps in the calculation; I urge you to work out the details following the reference. 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 [84] as a perturbation. We can also 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 of as a small chiral transformation from the vacuum. Then, e.g., for the τ_3 direction, $q \to (1 + i\pi_3\tau_3)q$ (the π field parameterizes the transformation), so the action becomes:

$$\frac{m}{f_\pi}\pi_3(\bar{q}\gamma_5 q + \theta\bar{q}q) \tag{90}$$

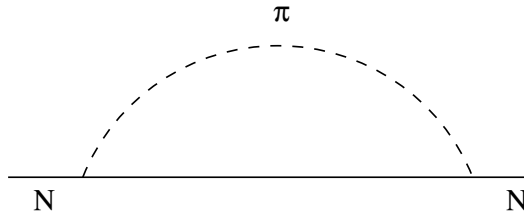In other words, we have calculated a CP violating coupling of the mesons to the pions.

Figure 2: Diagram in which CP-violating coupling of the pion contributes to d_n.

This coupling is difficult to measure directly, but it was observed in [8] that this coupling gives rise, in a calculable fashion, to a neutron electric dipole moment. Consider the graph of fig. 2. This graph generates a neutron electric dipole moment, if we take one coupling to be the standard pion-nucleon coupling, and the second the coupling we have computed above. The resulting Feynman graph is infrared

19

divergent; we cut this off at m_π, while cutting off the integral in the ultraviolet at the QCD scale. Because of this infrared sensitivity, the low energy calculation is reliable. The exact result is:

$$d_n = g_{\pi NN} \frac{-\theta m_u m_d}{f_\pi (m_u + m_d)} \langle N_f | \bar{q} \tau^a q | N_i \rangle \ln(M_N/m_\pi) \frac{1}{4\pi^2} M_N.$$ (91)

The matrix element can be estimated using ordinary $SU(3)$, yielding $d_n = 5.2 \times 10^{-16}\theta$ cm. The experimental bound gives $\theta < 10^{-9-10}$. Understanding why CP violation is so small in the strong interactions is the "strong CP problem."

5   Possible Solutions

What should our attitude towards this problem be? We might argue that, after all, 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 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.

5.1   Massless u Quark

Suppose the bare mass of the u quark was zero (i.e. at some high scale, the u quark mass were zero). Then, by a redefinition of the u quark field, we could eliminate θ from the lagrangian. Moreover, as we integrated out physics from this high scale to a lower scale, instanton effects would generate a small u quark mass. In fact, a crude estimate suggests that this mass will be comparable to the estimates usually made from current algebra. Suppose that we construct a Wilsonian action at a scale, say, of order twice the QCD scale. Call this scale Λ_o. Then we would expect, on dimensional grounds, that the u quark mass would be of order:

$$m_u = \frac{m_d m_s}{\Lambda_o}$$ (92)

Now everything depends on what you take Λ_o to be, and there is much learned discussion about this. The general belief seems to be that the coefficient of this expression needs to be of order three to explain the known facts of the hadron spectrum. There is contentious debate about how plausible this possibility is.

Note, even if one does accept this possibility, one would still like to understand why the u quark mass at the high scale is exactly zero (or extremely small). It is interesting that in string theory, one knows of discrete symmetries which are anomalous, i.e. one has a fundamental theory where there are discrete symmetries which can be broken by very tiny effects. Perhaps this could be the resolution of the strong CP problem?

20

CP as a Spontaneously Broken Symmetry

A second possible solution comes from the observation that if the underlying theory is CP conserving, a "bare" θ parameter is forbidden. In such a theory, the observed CP violation must arise spontaneously, and the challenge is to understand why this spontaneous CP violation does not lead to a θ parameter. For example, if the low energy theory contains just the standard model fields, then some high energy breaking of CP must generate the standard model CP violating phase. This must not generate a phase in $\det m_q$, which would be a θ parameter. Various schemes have been devised to accomplish this[13,14]. Without supersymmetry, they are generally invoked in the context of grand unification. There, it is easy to arrange that the θ parameter vanishes at the tree level, including only renormalizable operators. It is then necessary to understand suppression of loop effects and of the contributions of higher dimension operators. In the context of supersymmetry, it turns out that understanding the smallness of θ in such a framework, requires that the squark mass matrix have certain special properties (there must be a high degree of squark degeneracy, and the left right terms in the squark mass matrices must be nearly proportional to the quark mass matrix.)[15].

Again, it is interesting that string theory is a theory in which CP is a fundamental (gauge) symmetry; its breaking is necessarily spontaneous. Some simple string models possess some of the ingredients required to implement the ideas of [13,14].

The Axion

Perhaps the most popular 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[16].

Consider the vacuum energy as a function of θ, eqn. [85]. 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 particle, a, with coupling to QCD:

$$\mathcal{L}_{axion} = (\partial_\mu a)^2 + \frac{(a/f_a + \theta)}{32\pi^2} F\tilde{F} \tag{93}$$

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

$$a \rightarrow a + \alpha \tag{94}$$

for constant α. Then by a shift in a, one can eliminate 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 $E(\theta)$.

$$m_a^2 \approx \frac{m_\pi^2 f_{\pi^2}}{f_a^2}. \tag{95}$$

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. As for the θ-dependence of the vacuum energy, it is not difficult to get the factors straight using current algebra methods. A collection of formulae, with great care about factors of 2, appears in [17]

Actually, there are several questions one can raise about this proposal:

- Should the axion already have been observed? In fact, as Scott Thomas will explain in greater detail in his lectures, the couplings of the axion to matter can be worked out in a straightforward way, using the methods of current algebra (in particular of non-linear lagrangians). All of the couplings of the axion are suppressed by powers of f_a. So if f_a is large enough, 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 only is suppressed 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 a bit more detail below, the axion can be copiously produced in the early universe. As a result, there is an *upper bound* on the axion decay constant. In this case, as we will explain below, the axion could constitute the dark matter.

- Can one search for the axion experimentally[18]? 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 a group at Livermore (the collaboration contains members from MIT, University of Florida) and Kyoto. The basic idea is to suppose that the dark matter in the halo 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. 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 postulates a symmetry, preserved to some high degree of approximation, but which is not a symmetry; it is at least broken by tiny QCD effects. Is this reasonable? To understand the nature of the problem, consider one of the ways an axion can arise. In some approximation, we can suppose 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, with an associated (pseudo)-Goldstone boson, a. We can parameterize ϕ as:

$$\phi = f_a e^{ia/f_a} \qquad |\langle\phi\rangle| = f_a \tag{96}$$

If this field couples to fermions, so that they gain mass from its expectation value, then at one loop we generate a coupling $aF\tilde{F}$ from integrating out the fermions. This calculation is identical

22

to the corresponding calculation for pions we 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 dimension 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. For example, suppose one has a symmetry breaking operator

$$\frac{\phi^{n+4}}{M_p^n} \tag{97}$$

Such a term gives a linear contribution to the axion potential of order $\frac{f_a^{n+3}}{M_p^n}$. If $f_a \sim 10^{11}$, this swamps the would-be QCD contribution $(\frac{m_\pi^2 f_\pi^2}{f_a})$ unless $n > 12^{19}$!

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 $M_{GUT} - M_p$. Whether this can be made compatible with cosmology, or whether one can obtain a lower scale, is an open question.

6 The Axion in Cosmology

Despite the fact that it is so weakly coupled, it be copiously produced in the early universe[20]. The point is the weak coupling itself. In the early universe, we know the temperature was once at lease 1 MeV, and if the temperature was above a GeV, the potential of the axion was irrelevant. Indeed, if the universe is radiation dominated, the equation of motion for the axion is:

$$\frac{d^2}{dt^2}\phi + \frac{3}{2t}\frac{d}{dt}\phi + V'(\phi) = 0. \tag{98}$$

For $t^{-1} \gg m_a$, the system is overdamped, and the axion does not move. There is no obvious reason that the axion should sit at its minimum in this early era. So one can imagine that the axion sits at its minimum until $t \sim m_a^{-1}$, and then begins to roll. For $f_a \sim M_p$, this occurs at the QCD temperature; for smaller f_a it occurs earlier. After this, the axion starts to oscillate in its potential; it looks like a coherent state of zero momentum particles. At large times,

$$\phi = \frac{c}{R^{3/2}(t)}\cos(m_a t). \tag{99}$$

so the density is simply diluted by the expansion. The energy density in radiation dilutes like T^4, so eventually the axion comes to dominate the energy density. If $f_a \sim M_p$, the axion energy density is comparable when oscillations start. If f_a is smaller, oscillation starts earlier and there is more damping. Detailed study (including the finite temperature behavior of the axion potential) gives a limit $f_a < 10^{11}$ GeV.

7 Conclusions: Outlook

The strong CP problem, on the one hand, seems very subtle, but on the other hand it is in many ways similar to the other problems of flavor which we confront when we examine the standard model. θ is one more parameter which is surprisingly small. The smallness of the Yukawa couplings may well be the result of approximate symmetries. Similarly, all of the suggestions we have discussed above to understand the smallness of θ involve approximate symmetries of one sort or another.

In the case of other ideas about flavor, there is often no compelling argument for the scale of breaking of the symmetries. The scale could be so high as to be unobservable, and there is little hope for testing the hypothesis. What is perhaps most exciting about the axion is that if we accept the cosmological bound, the axion might well be observable.

That said, one should recognize that there are reasons to think that the axion scale might be higher. In particular, as mentioned earlier, string theory provides one of the most compelling settings for axion physics, and one might well expect the Peccei-Quinn scale to be of order the GUT scale or Planck scale. There have been a number of suggestions in the literature as to how the cosmological bound might be evaded in this context.

At a theoretical level, there are other areas in which the axion is of interest. Such particles inevitably appear in string theory and in supersymmetric field theories. (Indeed, it is in this context that Peccei-Quinn symmetries of the required type *for* QCD most naturally appear). These symmetries and the associated axions are a powerful tool for understanding these theories.

Acknowledgements:

1. The text, M.E. Peskin and D.V. Schroeder, *An Introduction to Quantum Field Theory*, Addison Wesley (1995) Menlo Park, has an excellent introduction to anomalies (chapter 19).
2. E. Witten, "Dyons of Charge $\frac{e\theta}{2\pi}$", Phys. Lett. **86B** (1979) 283. For a pedagogical introduction to monopoles, I strongly recommend J.A. Harvey, "Magnetic Monopoles, Duality and Supersymmetry", hep-th/9603086.
3. J. Kogut and L. Susskind, Phys. Rev. **D11** (1976) 3594.
4. I. Affleck, Phys.Lett. **B92** (1980) 149.
5. G. 't Hooft, Phys. Rev. **D14** (1976) 3432.
6. A.A. Belavin, A.M. Polyakov, A.S. Schwartz and Y.S. Tyupkin, Phys. Lett. **59B** (1975) 85.
7. S. Coleman, "The Uses Of Instantons," in *Aspects of Symmetry*, Cambridge University Press, Cambridge, 1985.
8. R.J. Crewther, P. Di Vecchia, G. Veneziano and E. Witten, Phys. Lett. **88B** (1979) 123.
9. I. Affleck, "On Constrained Instantons," Nucl. Phys. **B191** (1981) 429.

10. E. Witten, Nucl.Phys. **B149** (1979) 285.

11. D. Gross, R. Pisarski and L.G. Yaffe, "QCD and Instantons at Finite Temperature," Rev. Mod. Phys. **53** (1981) 43.

12. A.G. Cohen, D.B. Kaplan and A.E. Nelson, "Progress in Electroweak Baryogenesis," Ann. Rev. Nucl. Part. Sci. **43** (1993) 27, hep-ph/9302210.

13. M. A. B. Beg and H. Tsao, Phys. Rev. Lett. **41** (1978) 281; R. N. M. and Senjanović, Physics Letters **B79** (1978) 283; H. Georgi, Had. Journ. **1** (1978) 155.

14. A. Nelson, Phys. Lett. **136B** (1984) 387; S.M. Barr, Phys. Rev. Lett. **53** (1984) 329.

15. M. Dine, R. Leigh and A. Kagan, Phys. Rev. **D48** (1993) 4269, hep-ph/9304299.

16. R.D. Peccei and H.R. Quinn, Phys. Rev. Lett. **38** (1977) 1440; Phys. Rev. **D16** (1977) 1791.

17. M. Srednicki, "Axion Couplings to Matter: CP Conserving Parts," Nucl. Phys. **B260** (1985) 689.

18. P. Sikivie, "Axion Searches," Nucl.Phys.Proc.Suppl. **87** (2000) 41, hep-ph/0002154.

19. M. Kamionkowski and J. March-Russell, "Planck Scale Physics and the Peccei-Quinn Mechanism," Phys. Lett. **B282** (1992) 137, hep-th/9202003.

20. M. Turner, "Windows on the Axion," Phys.Rept. **197** (1990) 67.