11

Instantons, fermions, and physical consequences

The most important physical consequences of the Yang–Mills instantons are associated with the presence of fermions in the theory. Because these turn out to be closely related to the axial anomaly, I begin with a brief review of that topic.

11.1 Anomalies

It sometimes happens that a transformation that is a symmetry of a classical field theory ceases to be a symmetry when the theory is quantized. Perhaps the best-known example of such an anomaly, and the one of relevance for us here, is that associated with the chiral symmetry of a theory with massless quarks. Classical analysis predicts a number of conserved vector and axial vector currents. However, it can happen that after quantization some of the axial currents are not conserved, but instead have anomalous divergences [255–257].

The simplest example of this axial anomaly occurs in a theory with massless fermion fields ψ_r, where the subscript $r = 1, 2, \ldots, N_f$ is a "flavor" index, and a Lagrangian density

$$\mathcal{L} = \bar{\psi}_r \left(i\gamma^\mu \partial_\mu + g\gamma^\mu A^a_\mu T^a \right) \psi_r + \cdots , \tag{11.1}$$

where the ellipsis denotes terms that do not contain the fermion fields. Here I will take the A^a_μ to be $\mathrm{SU}(N_c)$ gauge fields, with the T^a the corresponding generators. The "color" gauge indices on the fermions and the corresponding indices on the T^a have been suppressed. For simplicity, let us assume that all flavors of fermions transform under the fundamental representation of the gauge group.

Classically, this Lagrangian is invariant under both the $\mathrm{U}(1)\times\mathrm{U}(1)$ chiral transformations

$$\begin{aligned} \psi_r &\to e^{i\alpha_0} \psi_r \,, \\ \psi_r &\to e^{i\beta_0 \gamma^5} \psi_r \end{aligned} \tag{11.2}$$

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 17 Nov 2026 at 00:53:43, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9781139017787.012

and the $\mathrm{SU}(N_f)\times\mathrm{SU}(N_f)$ chiral transformations

$$\begin{aligned}\psi_r &\to \left(e^{i\alpha^a\mathcal{T}^a}\right)_{rs}\psi_s\,,\\ \psi_r &\to \left(e^{i\beta^a\mathcal{T}^a\gamma^5}\right)_{rs}\psi_s\,,\end{aligned}\tag{11.3}$$

where the $\mathcal{T}^a$ are generators of $\mathrm{SU}(N_f)$.

These symmetries imply conserved currents

$$\begin{aligned}j^\mu &= \bar\psi_r\gamma^\mu\psi_r\,,\\ j_5^\mu &= \bar\psi_r\gamma^\mu\gamma^5\psi_r\end{aligned}\tag{11.4}$$

and

$$\begin{aligned}j_a^\mu &= \bar\psi_r\mathcal{T}^a_{rs}\gamma^\mu\psi_s\,,\\ j_{5a}^\mu &= \bar\psi_r\mathcal{T}^a_{rs}\gamma^\mu\gamma^5\psi_s\,,\end{aligned}\tag{11.5}$$

with corresponding conserved charges. In particular, if we define right- and left-handed fields

$$\begin{aligned}\psi_{Rr} &= \frac{1+\gamma^5}{2}\psi_r\,,\\ \psi_{Lr} &= \frac{1-\gamma^5}{2}\psi_r\,,\end{aligned}\tag{11.6}$$

the charges corresponding to j^μ and j_5^μ are

$$\begin{aligned}Q &= \int d^3x\,\bar\psi_r^\dagger\psi_r = n_R + n_L\,,\\ Q_5 &= \int d^3x\,\bar\psi_r^\dagger\gamma^5\psi_r = n_R - n_L\,,\end{aligned}\tag{11.7}$$

where n_R is the number of right-handed particles minus the number of left-handed antiparticles, and similarly for n_L. Thus, Q is the total particle number and Q_5 the net chirality.

The problematic axial current is j_5^μ. If it were divergenceless, as predicted by the classical analysis, we would have the Ward identity

$$0 = k^\mu T_{\mu\alpha\beta}(k,q_1,q_2)\,,\tag{11.8}$$

where

$$T_{\mu\alpha\beta}(k,q_1,q_2) = i\int d^4x_1\,d^4x_2\langle 0|T[j_5^\mu(0)j_a^\alpha(x_1)j_a^\beta(x_2)]0\rangle e^{iq_1\cdot x_1+iq_2\cdot x_2}\,.\tag{11.9}$$

Here T denotes time-ordering, j_a^μ is the gauge current that couples to A_μ^a, and $k=q_1+q_2$. The leading contribution to the matrix element comes from the two

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 17 Nov 2026 at 00:53:43, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9781139017787.012

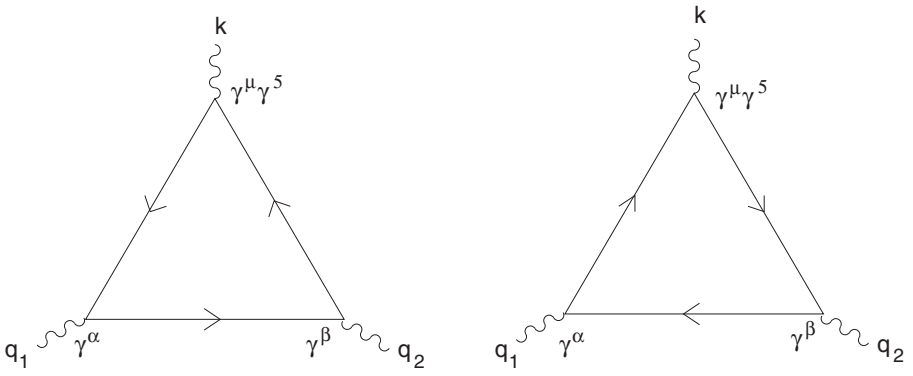


Fig. 11.1. The graphs that give the leading contribution to the matrix element in Eq. (11.9). The incoming momenta and the gamma matrix factors at each vertex are indicated.

triangle graphs shown in Fig. 11.1. Explicit calculation of these gives the nonzero result[1]

$$k^\mu T_{\mu\alpha\beta} = -\frac{N_f}{2\pi^2}\epsilon_{\alpha\beta\rho\sigma}q_1^\rho q_2^\sigma \,. \tag{11.10}$$

The failure of Eq. (11.8) can be understood by recalling that properly defining the quantum field theory requires specifying a regulator scheme. Any method for regulating the triangle graphs that respects the gauge symmetry, as is required for renormalizability, violates the chiral symmetry. Dimensional regularization has the problem that the γ^5 in the chiral transformation is an explicitly four-dimensional quantity that cannot be naturally continued to $4+\epsilon$ dimensions. Pauli–Villars regulation is gauge invariant and four-dimensional, but the massive Pauli–Villars regulator field explicitly breaks the chiral symmetry. This field gives a contribution to $\partial_\mu j_5^\mu$ that is proportional to the regulator mass M. This explicit factor of M multiplies a regulator graph proportional to $1/M$ to give a finite contribution in the $M \to \infty$ limit.

A useful way to interpret Eq. (11.10) is to include factors of A_α^a on the external gauge lines of the triangle graphs. We can then view $T_{\mu\alpha\beta}$ as a contribution to the expectation value of j_5^μ in the presence of a background gauge field. Equation (11.10) gives the divergence of this current in the background field as

$$\partial_\mu j_5^\mu = \frac{N_f g^2}{16\pi^2}\,\epsilon^{\mu\nu\rho\sigma}\mathrm{tr}\,(\partial_\mu A_\nu - \partial_\nu A_\mu)(\partial_\rho A_\sigma - \partial_\sigma A_\rho)\,. \tag{11.11}$$

[1] This assumes a regularization that keeps the two vector currents divergenceless.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 17 Nov 2026 at 00:53:43, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9781139017787.012

Including the effects of the analogous square and pentagon diagrams gives the additional terms needed to obtain the gauge-invariant result[2]

$$\partial_\mu j_5^\mu = \frac{N_f g^2}{8\pi^2} \operatorname{tr} F_{\mu\nu} \tilde{F}^{\mu\nu} \,. \tag{11.12}$$

11.2 Spectral flow and fermion zero modes

Notice that the anomalous divergence of the axial current, Eq. (11.12), is, up to a multiplicative constant, the same as the current j_A^μ that was defined in Eq. (10.17). This suggests that we combine them to form a divergenceless current

$$\mathcal{J}_5^\mu = j_5^\mu - 2N_f j_A^\mu \,. \tag{11.13}$$

Like j_A^μ, this current is gauge-variant, and so its analysis depends on the gauge in which we work. Measurable physical consequences must, of course, be the same in all gauges.

Let us start by working in $A_0 = 0$ gauge, where the instanton corresponds to tunneling between two vacua of different winding number. The charge associated with $\mathcal{J}_5^\mu$ is

$$\mathcal{Q}_5 = n_R - n_L - 2N_f \, n \,, \tag{11.14}$$

where n is the winding number of the gauge field. The divergenceless of $\mathcal{J}_5^\mu$ implies that $\mathcal{Q}_5$ is conserved. Hence, any change in winding number must be accompanied by a change in fermion chirality, with

$$\Delta(n_R - n_L) = 2N_f \, \Delta n \,. \tag{11.15}$$

To see how this comes about, let us consider the spectrum of the Dirac Hamiltonian in the presence of a background field A_μ. It is simplest to view this from the Dirac sea viewpoint, with both positive and negative energies in the spectrum, and antiparticles being unoccupied negative-energy states. For our massless fermions the Hamiltonian is

$$H = -i\alpha^j D_j \,, \tag{11.16}$$

where $\alpha^j = \gamma^0 \gamma^j$. In a basis with

$$\gamma^j = \begin{pmatrix} 0 & i\sigma^j \\ i\sigma^j & 0 \end{pmatrix} , \qquad \gamma^0 = \begin{pmatrix} 0 & -iI \\ iI & 0 \end{pmatrix} , \qquad \gamma^5 = \begin{pmatrix} I & 0 \\ 0 & -I \end{pmatrix} , \tag{11.17}$$

we have

$$\alpha^j = \begin{pmatrix} \sigma^j & 0 \\ 0 & -\sigma^j \end{pmatrix} . \tag{11.18}$$

[2] The anomaly can also be calculated from a careful examination of the behavior of the path integral measure under chiral transformations [258].

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 17 Nov 2026 at 00:53:43, subject to the Cambridge Core terms of use , available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9781139017787.012

The four-component fermions naturally split into a pair of two-component Weyl fermions. For the upper two components, corresponding to right-handed particles, the Hamiltonian becomes

$$H_R = -i\sigma^j D_j \,, \tag{11.19}$$

while for the lower two, left-handed, components we have

$$H_L = i\sigma^j D_j \,. \tag{11.20}$$

Because $H_L = -H_R$, each negative eigenvalue of one Hamiltonian corresponds to a positive eigenvalue of the other.

Let us start with the background field being a vacuum configuration with winding number n, and then consider the series of configurations along the tunneling path defined by an instanton. At intermediate stages along the way, A_μ is not in a vacuum state, there are nonzero field strengths, and the fermion spectrum is certainly different from the initial spectrum. Nevertheless, since the final configuration is also a vacuum configuration, albeit one with winding number $n+1$, the spectrum, which is gauge invariant, must be the same at the end as it was at the beginning.

This does not mean, however, that individual states must end up where they started. As shown schematically in Fig. 11.2, an individual energy level can have a net movement up or down the spectrum as the gauge field flows along the tunneling path. The only requirement is that its place be taken up by some other level. The relationship between the two Hamiltonians requires that for every level of H_R that moves up, an energy level of H_L with the opposite sign must move down, and vice versa.

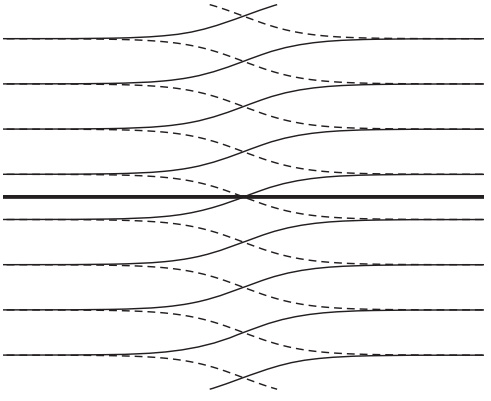


Fig. 11.2. Schematic illustration of the flow of fermion energy levels as the gauge field is varied from a vacuum configuration with winding number n (on the far left) to one with winding number $n+1$ (on the far right). The right-handed levels are indicated by solid lines and the left-handed ones by the dashed lines. The heavy horizontal line represents the zero of energy.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 17 Nov 2026 at 00:53:43, subject to the Cambridge Core terms of use , available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9781139017787.012

In particular, as indicated in the figure, some negative-energy states can become positive-energy states and some positive-energy states can become negative-energy states. Let us suppose that the fermions were originally in a vacuum state, with all negative-energy levels filled and all positive-energy levels empty. In the adiabatic approximation the occupation of the individual levels would not change, so the movement of a level from negative to positive energy would lead to the creation of a positive-energy particle, while a flow from positive to negative-energy would give a negative-energy hole, corresponding to a positive-energy antiparticle of the opposite chirality. If the adiabatic approximation is not applicable, particles may move between levels. However, because the original Hamiltonian only couples fermion fields of the same chirality, this movement must be between levels of the same chirality, and so won't affect the total chirality.

Thus, the relation between the changes in chirality and winding number in Eq. (11.15) can be understood if in the presence of a gauge field $A_\mu(x)$ with instanton number k there is a net flow upward of k right-handed levels from negative to positive energy, and an equal flow of left-handed levels from positive to negative energy. We can show that this is the case by means of an index theorem.

In order to follow the flow of the energy levels, let us write $x_4 = \tau$ and define

$$H_R(\tau) = -i\sigma^j \left[\partial_j - igA_j^a(\mathbf{x},\tau)T^a\right] \tag{11.21}$$

and label its instantaneous eigenvalues and two-component eigenfunctions as $\omega_n(\tau)$ and $\chi_n(\mathbf{x};\tau)$, respectively. Now consider the equation

$$0 = \mathcal{D}\chi = \left[-\frac{\partial}{\partial\tau} - H_R(\tau)\right]\chi\,. \tag{11.22}$$

For A_μ sufficiently slowly varying as a function of τ, this equation is solved by

$$\chi = e^{-\int_0^\tau d\tau'\,\omega_n(\tau')}\,\chi_n(\mathbf{x};\tau)\,. \tag{11.23}$$

If $\omega_n(\tau)$ has the same sign at $\tau = -\infty$ and $\tau = \infty$, this solution diverges as τ goes to either one limit or the other. On the other hand, if $\omega(-\infty) < 0$ and $\omega(\infty) > 0$, the solution goes to zero in both directions, and is a normalizable zero mode of $\mathcal{D}$. If instead $\omega(\tau)$ goes from positive to negative, a similar construction gives a normalizable zero mode of

$$\mathcal{D}^\dagger = \frac{\partial}{\partial\tau} - H_R(\tau)\,. \tag{11.24}$$

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 17 Nov 2026 at 00:53:43, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9781139017787.012

Thus, if k_+ levels of H_R move from negative energy to positive energy and k_- modes move from positive to negative, $\mathcal{D}$ will have k_+ zero modes, $\mathcal{D}^\dagger$ will have k_- zero modes, and the index of $\mathcal{D}$ will be[3]

$$\mathcal{I}(\mathcal{D}) = k_+ - k_- \,. \tag{11.25}$$

In fact, we have already calculated this index. The operator $\mathcal{D}$ defined above is the same as the one defined in Eq. (10.119), written in $A_0 = 0$ gauge. In Sec. 10.8 we showed that the index of $\mathcal{D}$ was given by Eq. (10.138). For fermions in the fundamental representation of SU(N), we have $T(R) = 1/2$, and hence

$$\mathcal{I}(\mathcal{D}) = k \,. \tag{11.26}$$

Thus, along the flow defined by a Euclidean gauge field with instanton number k there are k energy levels of H_R that move from negative to positive energy. Because $H_L = -H_R$, there are also k levels of H_L that move from positive to negative energy. This is repeated for each of the N_f flavors, so the net increase in chirality is $2N_f k$, just as required by Eq. (11.15).

The index calculation in Sec. 10.8 did not use the fact that the background field was self-dual, so our result applies even if $A_\mu(x)$ is not a solution of the Euclidean field equations. On the other hand, the vanishing theorem that showed that $\mathcal{D}^\dagger$ had no zero modes did use the self-duality of the gauge field, so for general background fields Eq. (11.26) only gives the difference between the numbers of zero modes of $\mathcal{D}$ and $\mathcal{D}^\dagger$, and thus a lower bound on the number of zero modes.

This analysis of the spectral flow was done in the $A_0 = 0$ gauge, where the instanton corresponds to a tunneling path between vacua of different winding number. In a gauge with a unique vacuum, the tunneling path represented by the instanton is gauge-transformed to one that begins and ends at the same point. This gauge transformation cannot change the physical consequences of the instanton. Hence, although the gauge field eventually returns to its initial value, the flow of the fermion eigenstates as the gauge field background evolves does not return them to their initial position. Instead, they are shifted, with some moving from negative to positive energy, or vice versa, just as in $A_0 = 0$ gauge. This leads to a net change in the fermion chirality, even though the winding number is unchanged. Thus, $\mathcal{Q}_5$ is not conserved. This is not in contradiction with the vanishing of $\partial_\mu \mathcal{J}_5^\mu$. The standard demonstration that a divergenceless current implies a conserved charge proceeds by integrating over the region between two spacelike surfaces and converting the integral to a sum of surface integrals. Conservation of the charge follows if the surface integrals at spatial infinity vanish, as is usually the case. However, although the gauge-variant current j_A^μ vanishes

[3] One might worry about the fact that Eq. (11.23) only solves Eq. (11.22) in the limit of slowly varying A_μ. However, the index is a topological invariant, and so the result for more rapidly varying fields must be the same as for a slowly varying field with the same instanton number.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 17 Nov 2026 at 00:53:43, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9781139017787.012

at large distance in $A_0 = 0$ gauge, it is nonvanishing and gives a finite surface integral in other gauges, with the result that conservation of $\mathcal{Q}_5$ is violated by instanton effects.

Let us find the explicit form of the fermion zero mode in the background of a single SU(2) instanton centered at the origin. Rather than working in $A_0 = 0$ gauge, it is simpler to work in a gauge where the instanton is given by Eq. (10.87). In terms of the four e_p defined in Eq. (10.63), the zero-mode equation takes the form

$$0 = (e_p^\dagger)_{\alpha\beta} \left[\delta_{ab}\partial_p - ig(A_p)_{ab}\right] \Phi_{b\beta} \,. \tag{11.27}$$

Here Greek subscripts denote spinor indices 1 or 2 while Latin subscripts from the beginning of the alphabet denote SU(2) indices, which also take values 1 or 2. The distinction between the two types of indices can be ignored if we view $\Phi_{b\beta}$ as a 2×2 matrix and rewrite the zero-mode equation as the matrix equation

$$0 = i \left[\partial_p - igA_p\right] \Phi \, (e_p^\dagger)^t \,. \tag{11.28}$$

Now recall that the Pauli matrices satisfy

$$\sigma_j^t = (i\sigma_2)\sigma_j(i\sigma_2) \,, \qquad j = 1, 2, 3, \tag{11.29}$$

which implies that

$$(e_p^\dagger)^t = e_p^* = -(i\sigma_2)e_p(i\sigma_2) \,, \qquad p = 1, 2, 3, 4, \tag{11.30}$$

so that our zero-mode equation becomes

$$0 = i \left[\partial_p - igA_p\right] \Phi(i\sigma_2) \, e_p \,. \tag{11.31}$$

The instanton is invariant under the combination of a rotation and an SU(2) global gauge transformation. Since there is only one fermion zero mode, it must be a singlet under such a transformation. Because $\epsilon_{b\beta} = (i\sigma_2)_{b\beta}$ is an SU(2) invariant tensor, a natural ansatz is

$$\Phi_{b\beta} = i\epsilon_{b\beta} \, h(x^2) \,. \tag{11.32}$$

Substituting this into Eq. (11.31) leads to

$$0 = \left[2x_p \, h'(x^2) - igA_p h(x^2)\right] e_p \,, \tag{11.33}$$

where the prime indicates differentiation with respect to x^2. Using the explicit form of the instanton,

$$A_p = \frac{1}{g} \, \frac{\eta_{pq} x_q}{x^2 + \lambda^2} \,, \tag{11.34}$$

and the easily verified identity

$$\eta_{pq} e_p = i \left(\delta_{pq} - e_p e_q^\dagger\right) e_p = 3i e_q \,, \tag{11.35}$$

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 17 Nov 2026 at 00:53:43, subject to the Cambridge Core terms of use , available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9781139017787.012

we obtain

$$0 = h' + \frac{3}{2}\,\frac{h}{x^2+\lambda^2}\,, \tag{11.36}$$

whose solution is

$$h = \frac{B}{(x^2+\lambda^2)^{3/2}}\,, \tag{11.37}$$

with B an arbitrary integration constant. The normalized fermion zero mode can therefore be written as

$$\Psi = \frac{\sqrt{2}}{\pi}\,\frac{\lambda}{(x^2+\lambda^2)^{3/2}}\,\chi\,, \tag{11.38}$$

where χ is a fixed isospinor four-component Dirac spinor. In the basis in which we have been working, the lower components of χ vanish and the upper ones are given by the two-component Weyl spinor $\Phi_{a\alpha} = \epsilon_{a\alpha}$.

This zero mode is for a fermion in the doublet representation of SU(2). Because the unit instanton in any larger gauge group is an embedding of the SU(2) instanton, the generalization to other groups is straightforward. In particular, the unit instanton for SU(N_c) is the embedding of the SU(2) instanton in a 2×2 block. The fundamental representation zero mode is obtained by inserting the zero mode of Eq. (11.38) into the corresponding two components of the fermion field, and then setting the remaining $N_c - 2$ components equal to zero.

It is instructive to summarize the chirality-violating processes associated with an instanton by a nonlocal effective Lagrangian density. This must contain the product of $2N_f$ fermion fields, one ψ_L and one $\bar{\psi}_R$ for each flavor. It can be written in the form [238]

$$\mathcal{L}_{\text{eff}} = Ce^{-8\pi^2/g^2(\lambda)} \prod_{s=1}^{N_f} (\bar{\psi}_R^s \omega)(\bar{\omega}\psi_L^s)\,, \tag{11.39}$$

where an integration over the positions of the fermion fields is understood, ω is a fixed Dirac spinor transforming under the fundamental representation of the gauge group, and the constant C is obtained from the one-loop corrections to the instanton. The nontrivial contribution from $\mathcal{L}_{\text{eff}}$ comes from the terms in the fermion fields corresponding to the zero mode; i.e., the product of the zero mode and the corresponding creation or annihilation operator. The term shown here is not Hermitian; one must add its Hermitian conjugate, which gives the effects of an anti-instanton. One must also, of course, integrate over all instanton positions and scales.[4]

[4] This effective Lagrangian contains ω, whose form depends on the gauge orientation of the instanton. However, physical results must be gauge-independent. A gauge-invariant effective Lagrangian can be obtained by integrating over the gauge orientations of ω [238, 259].

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 17 Nov 2026 at 00:53:43, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9781139017787.012

11.3 QCD and the U(1) problem

Even before QCD was discovered, it was realized that the eight light pseudoscalar mesons could be understood as approximate Goldstone bosons arising from the spontaneous breaking of an approximate SU(3)×SU(3) chiral symmetry, with the especially low masses of the pions indicating that a chiral SU(2)×SU(2) was even closer to being an exact symmetry of the Lagrangian.

The form of the QCD Lagrangian clarifies the origin of these symmetries. The part containing the quark fields q_r (with the subscript labeling quark flavor) is

$$\mathcal{L} = \sum_r \left[\bar{q}_r \left(i\gamma^\mu \partial_\mu + g\gamma^\mu A^a_\mu T^a\right) q_r + m_r \bar{q}_r q_r\right] . \tag{11.40}$$

The quark "masses" m_r are neither the positions of poles in Green's functions nor effective constituent masses in hadrons, but simply parameters in the Lagrangian. Current-algebra arguments lead to the conclusion that the up and down quark masses are only a few MeV, while the strange quark mass is roughly 100 MeV. All three are small compared to a typical QCD scale (e.g., constituent up and down quark masses of about 300 MeV) suggesting that this Lagrangian can be viewed as an approximation to one with either two or three massless quarks. The SU(2)×SU(2) and SU(3)×SU(3) chiral symmetries would then correspond to the transformations in Eq. (11.3).

However, a Lagrangian of this form is also invariant under the transformations of Eq. (11.2), making the symmetry either U(2)×U(2) or U(3)×U(3) and predicting either a fourth or a ninth approximate Goldstone boson. In the former case, with two quarks considered to be light, the only plausible candidate for the extra Goldstone boson is the η, but its mass of 548 MeV seems to be far too high compared to that of the pions. (Indeed, one can show that the mass of the fourth Goldstone boson cannot be more than $\sqrt{3}$ times that of the other three [260].) With three quarks considered to be light, the Goldstone bosons from the breaking of SU(3)×SU(3) are the three pions (135 and 140 MeV), the four kaons (494 and 498 MeV), and the η. Again, the only candidate for the ninth Goldstone boson, the η' at 958 MeV, is much too massive. The absence of a satisfactory candidate for the ninth (or fourth) Goldstone boson was known as the U(1) problem.

Note that this U(1) problem only arises after the form of the Lagrangian is determined. It is quite possible to write down theories (the sigma model is an example) that are invariant under SU(N_f)×SU(N_f) symmetry but do not have a U(N_f)×U(N_f) symmetry.

The effective Lagrangian of Eq. (11.39), generalized to an SU(3) gauge group, provides a resolution of the U(1) problem. If we define

$$\mathcal{M}_{rs} = (\bar{q}_{Rr}\omega)(\bar{\omega} q_{Ls}) , \tag{11.41}$$

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 17 Nov 2026 at 00:53:43, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9781139017787.012

the anticommutivity and Grassmann nature of the fermion fields implies that

$$\prod_{s=1}^{N_f}(\bar{q}_{Rs}\omega)(\bar{\omega}q_{Ls}) = \frac{1}{(N_f)!}\det\mathcal{M}\,. \tag{11.42}$$

The transformations in Eqs. (11.2) and (11.3) can all be written in the form

$$q_L \to U_L q_L\,, \qquad q_R \to U_R q_R\,, \tag{11.43}$$

where U_L and U_R are $N_f \times N_f$ unitary matrices. Under such transformations

$$\mathcal{M} \to U_R^\dagger \mathcal{M} U_L\,. \tag{11.44}$$

The determinant of $\mathcal{M}$, and hence $\mathcal{L}_{\text{eff}}$, is unchanged if U_L and U_R are both SU(N_f) matrices, with unit determinant, or if $U_L = U_R$. However, $\mathcal{L}_{\text{eff}}$ is not invariant under U(1) transformations with $U_L = U_R^\dagger$, which are precisely those transformations corresponding to the anomalous current j_5^μ. Thus, the U(1) that appeared to be a spontaneously broken symmetry in the massless quark limit is not a symmetry at all once instanton effects are included, and so there is no longer any prediction of an extra Goldstone boson.

It is important to recognize that even though $\mathcal{L}_{\text{eff}}$ arises from instanton effects, it is not a small correction to the Lagrangian. We saw in Sec. 10.12 that the integration over instanton scales diverges and the dilute-gas approximation breaks down for large instantons. Although this prevents us from obtaining reliable quantitative results for the instanton effects, we can expect them to be large and comparable to other strong interaction effects.

11.4 Baryon number violation by electroweak processes

We have seen that instanton effects in QCD can be large, but cannot be reliably calculated because of the divergence of the integration over instanton size. By contrast, in the SU(2)×U(1) electroweak theory there are instanton effects that are calculable although, at first sight, they seem to be negligibly small. These are associated with the weak isospin SU(2) factor, with the U(1) playing no role.

The essential new factor here is the presence of the Higgs doublet, which must approach its nonzero vacuum value $\langle\phi\rangle$ as $x^2 \to \infty$ but vanish at the center of the instanton. The Higgs vacuum expectation value breaks the classical scale invariance. Its effects favor smaller instanton size, so that the fields deviate from the vacuum over a smaller region, thus reducing the classical action. On the other hand, we have seen that the renormalization of the gauge coupling by the one-loop quantum effects favors large instantons. The net result is that the integration over instanton size peaks around $\lambda \sim 1/\langle\phi\rangle$.

The left-handed fermions of the standard model fall into SU(2) doublets. For each generation there are three quark doublets (because of the three colors) and

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 17 Nov 2026 at 00:53:43, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9781139017787.012

one lepton doublet. The former each carry baryon number $B = 1/3$, while the latter have lepton number $L = 1$. With three generations, each instanton leads to violations of baryon and lepton numbers [261]

$$\Delta B = \Delta L = 3\,. \tag{11.45}$$

Although B and L are not separately conserved, their difference $B - L$ is conserved. This corresponds to the fact that the $B - L$ current (unlike the currents of B and L separately) does not have an anomalous divergence in the Standard Model.

This is a truly striking result. Even though all perturbative Standard Model processes conserve baryon number, nonperturbative instanton effects allow baryon number violating processes such as the annihilation of a proton and neutron to yield an antinucleon and three leptons.[5] However, the prospects for experimentally observing such a process are rather dim, since the rate is suppressed by a factor of

$$\left(e^{-8\pi^2/g^2}\right)^2 = e^{-16\pi^2 \sin^2\theta_W/e^2}. \tag{11.46}$$

Because the size of the instantons responsible for this process is given by the electroweak scale, let us evaluate the quantities on the right-hand side of this equation at the Z mass. This gives 10^{-161}. The observable universe contains about 10^{78} protons and has an age of approximately 10^{10} years, or 10^{40} times a typical strong interaction time scale of 10^{-23} seconds. Thus, the probability that baryon number violation by such a process ever happened in our past light cone would seem to be vanishingly small.

However, matters are not quite so simple. This instanton-mediated process involves tunneling through the potential energy barrier separating two vacua with different winding number. An alternative possibility is to pass over the top of the barrier via a thermal fluctuation. Although unfeasible today, such a process might have been possible at the much higher temperatures that were present in the very early universe.

The crucial quantity here is the height of the barrier that must be traversed. On any path over the barrier there is a high point of maximum energy. If there is a lowest such maximum, it will dominate the rate for thermal fluctuations; its energy is the minimum energy needed to be able to cross the barrier without tunneling. Because this lowest maximum is a stationary point of the potential energy, it must correspond to a static solution (in $A_0 = 0$ gauge) of the field equations. Since it is a saddle point, rather than a local minimum, it is an unstable solution. This solution is known as a sphaleron [262].

[5] Note that the spatial location of this process gives a physical interpretation to the instanton position, which did not have a directly observable meaning in the effects considered previously.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 17 Nov 2026 at 00:53:43, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9781139017787.012

The size of the sphaleron is set by the Higgs vacuum expectation value v. This leads to a rough estimate of its energy,

$$E_{\rm sph} \sim \frac{4\pi v}{g}\,, \tag{11.47}$$

where g is the SU(2) gauge coupling. [The value of the U(1) coupling g' plays a lesser role, because it is the non-Abelian part of the theory that gives rise to the effect.] Using a spherically symmetric ansatz, Manton and Klinkhamer [263] found a sphaleron solution with an energy that ranged between 7.5 TeV and 13.3 TeV, depending on the Higgs boson mass; current experimental bounds on the Higgs mass put it in the upper part of this range.

A naïve estimate of the rate for sphaleron processes would be $\Gamma \sim \exp(-E_{\rm sph}/T)$. However at finite temperature it is the free energy that is the relevant quantity, so we expect the somewhat larger rate

$$\Gamma \sim e^{-F_{\rm sph}/T}\,, \tag{11.48}$$

where the calculation of $F_{\rm sph}$ must take into account the finite temperature corrections to the effective potential and the fact that these reduce the expectation value of the Higgs field. Even at this level of approximation it is clear that at temperatures near (and above) that of the electroweak phase transition, baryon number violation via electroweak processes can proceed at significant rates. This has the potential of washing out, or at least significantly diluting, any pre-existing baryon asymmetry. For a detailed review of this and related processes, see [264].

11.5 CP violation and the $\theta F\tilde{F}$ term

The possibility of vacuum tunneling led to the realization that a term

$$\Delta\mathcal{L} = \frac{\theta g^2}{16\pi^2}\mathrm{tr}\, F_{\mu\nu}\tilde{F}^{\mu\nu} = \frac{\theta g^2}{32\pi^2}\epsilon^{\mu\nu\alpha\beta}\mathrm{tr}\, F_{\mu\nu}F_{\alpha\beta} \tag{11.49}$$

can be added to the Yang–Mills Lagrangian. Because this term is a total divergence, it has no effect classically. Even in the quantum theory, it has no effect on Feynman diagrams, as was already noted below Eq. (10.32).

A clear demonstration that terms such as this can, nevertheless, have an effect in the full quantum theory is provided by the one-particle Lagrangian of Eq. (10.41) with the total time derivative term of Eq. (10.42) included. If we drop the potential energy term, we have

$$L = \frac{1}{2}B\dot{\alpha}^2 + \frac{\theta}{2\pi}\dot{\alpha}\,. \tag{11.50}$$

If α is taken to be an angle with period 2π, this can be interpreted as the Lagrangian for a rigid rotor. The momentum conjugate to α, $J = B\dot{\alpha} + \theta/2\pi$, is

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 17 Nov 2026 at 00:53:43, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9781139017787.012

therefore quantized and takes on integer values (in units where $\hbar = 1$). Converting from the Lagrangian to the Hamiltonian and setting $J = n$ gives the energy eigenvalues

$$E_n = \frac{1}{2B}\left(n - \frac{\theta}{2\pi}\right)^2 . \tag{11.51}$$

The shift $\theta \to \theta + 2\pi$ leaves the spectrum invariant, although with a relabeling of states, reflecting the periodicity of the θ parameter.

For $\theta = 0$ all levels but the ground state are degenerate, with $E_n = E_{-n}$. For $\theta = \pi$ all levels are paired, with $E_n = E_{1-n}$. In both cases the degeneracy is a consequence of the invariance of the Hamiltonian under time reversal. For all other values of θ this time-reversal invariance is broken and the degeneracy is absent.

Now suppose that we add a potential energy $K(1-\cos\alpha)$, with $K \gg 1/B$; the rotor is now perhaps better viewed as a rigid pendulum. There is now an energy barrier against motions in which the pendulum makes a full rotation. For energies less than the height of this barrier the classical pendulum only oscillates back and forth, but the quantum pendulum can also tunnel through the energy barrier and make a full rotation. In the path integral, it is only the latter type of path that is sensitive to θ. Because these paths are associated with tunneling processes, the θ-dependence of the low-energy eigenvalues is exponentially suppressed, as are the T-violating energy splittings.

Let us now return to the Yang–Mills theory, and ask what observable effects the θ term might have. There could be a θ-dependence of the vacuum energy, but since θ is fixed there would be no way of observing this.[6] A more promising possibility is to look for signals of the symmetry-breaking properties of the θ term. The presence of $\epsilon^{\mu\nu\alpha\beta}$ means that this term violates both parity and time reversal; because of CPT invariance, the latter implies CP violation.

In QCD, the apparent divergence of the integral over instanton sizes means that, although these are instanton effects, there is no exponential suppression. Could a QCD θ term, then, provide an explanation for the experimentally observed CP violation that would be an alternative to that based on a phase in the CKM matrix? Just two pieces of data are sufficient to show that it cannot.

The first observations of CP violation were in kaon decays. For example, the ratio of the amplitudes for the CP-violating decay $K_L^0 \to \pi^+\pi^-$ and the CP-conserving decay $K_S^0 \to \pi^+\pi^-$ is [265]

$$|\eta_{+-}| = \left|\frac{A(K_L^0 \to \pi^+\pi^-)}{A(K_S^0 \to \pi^+\pi^-)}\right| = 2.2 \times 10^{-3} . \tag{11.52}$$

On the other hand, a neutron electric dipole moment d_n, which would be T-violating (and therefore CP-violating) has not been observed. A natural scale

[6] However, in axion theories the axion field effectively plays the role of a spacetime-dependent θ and has an effect on the energy density that is in principle observable.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 17 Nov 2026 at 00:53:43, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9781139017787.012

for such a moment would be e times 10^{-13} cm, the characteristic length associated with the nucleon. Experimentally, however, [265]

$$d_n < 2.9 \times 10^{-26}\, e\text{–cm}\,, \tag{11.53}$$

thirteen orders of magnitude below the natural scale.

In the standard model, this extra 10 orders of magnitude suppression of the neutron electric dipole moment is attributable to the fact that CP violation is a weak interaction effect. There is no such suppression in the kaon decays, because the decays themselves are already weak interaction processes. On the other hand, if the θ term were the origin of the CP violation, there would be no need to invoke the weak interactions and therefore no reason to expect such a large discrepancy in the magnitudes of the two effects. Any value of θ large enough to explain CP violation in the kaon system would produce a neutron dipole moment far in excess of the experimental bounds. Hence, a θ term cannot be the explanation of the observed CP violation.

Rather than being a solution, the possibility of such a term is actually a serious problem. Because there is no weak interaction suppression, the smallness of the neutron electric dipole moment places a stringent limit on θ. An estimate using current-algebra methods gives [266]

$$d_n \approx 5 \times 10^{-16}\, \theta\, e\text{–cm}\,. \tag{11.54}$$

Comparing this with Eq. (11.53), we see that θ must be less than 10^{-10} or so. Understanding why this parameter in the Lagrangian should take on such an unnaturally small value has become known as the strong CP problem.

This problem is exacerbated by the presence of fermions in the theory. To understand this, consider the mass term for a single fermion field. This is usually written in the form

$$\mathcal{L}_m = -m\bar{\psi}\psi = -m\bar{\psi}_L\psi_R + \text{h.c.} \tag{11.55}$$

with m real and h.c. denoting the Hermitian conjugate. However, by redefining fields we can convert this to a complex mass parameter. With $\psi' = e^{i\alpha\gamma^5}\psi$, Eq. (11.55) becomes

$$\begin{aligned} \mathcal{L}_m &= -me^{-2i\alpha}\bar{\psi}'_L\psi'_R + \text{h.c.} \\ &\equiv -m'\bar{\psi}'_L\psi'_R + \text{h.c.} \end{aligned} \tag{11.56}$$

This redefinition can be viewed as a chiral transformation of the Lagrangian. Classically, this would be a symmetry if the mass term were absent and would imply the conservation of the Noether current j_5^μ. Now recall that in the presence of a symmetry-breaking term the divergence of a Noether current is related to the infinitesimal change in the Lagrangian $\Delta\mathcal{L}$ by

$$\partial_\mu J^\mu_{\text{Noether}} = \Delta\mathcal{L}\,. \tag{11.57}$$

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 17 Nov 2026 at 00:53:43, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9781139017787.012

Taking α in Eq. (11.56) to be infinitesimal, we would then conclude that

$$\alpha\partial_\mu j_5^\mu = 2i\alpha\left(m\bar{\psi}_L\psi_R - \text{h.c.}\right) . \tag{11.58}$$

However, we know that this classical result is not the whole story, since even in the absence of a fermion mass j_5^μ has a nonzero anomalous divergence given by Eq. (11.12). This, in turn, tells us that the change in the Lagrangian is not given just by the right-hand side of Eq. (11.58), but rather by

$$\Delta\mathcal{L} = 2i\left(m\bar{\psi}_L\psi_R - \text{h.c.}\right) + \frac{g^2}{8\pi^2}\text{tr}\, F_{\mu\nu}\tilde{F}^{\mu\nu} . \tag{11.59}$$

Thus, the same transformation that rotates the phase of m also adds an additional $F\tilde{F}$ term; i.e., it shifts the value of θ. The invariant quantity is $\bar{\theta} = \theta - \arg m$. If there are several flavors of fermions the mass term becomes

$$\mathcal{L}_m = M_{rs}\bar{\psi}_{rR}\psi_{sL} + \text{h.c.} , \tag{11.60}$$

where the possibility of flavor mixing means that the mass matrix M may have nonzero off-diagonal terms. Generalizing the one-flavor calculation shows that the invariant CP-violating parameter is

$$\bar{\theta} = \theta - \arg\det M . \tag{11.61}$$

Before considering the effects of fermions it was hard to understand why θ should be so small. This becomes even harder to understand once we realize that the effective value of θ has a contribution from a fermion mass matrix that arises from the coupling to a Higgs field whose vacuum expectation value has an arbitrary phase.

One possible solution is for one of the fermions to be massless. The phase of its mass (and, more generally, of the determinant of the mass matrix) would then be ambiguous and θ could be shifted at will. Hence, all values of θ would be equivalent and there would be no θ-dependent physical quantities. It has therefore been suggested that the strong CP problem would be resolved if the up quark were massless. However, current algebra and other evidence suggests that it is not.

Perhaps a more plausible resolution of the puzzle is a Peccei–Quinn [267, 268] type mechanism, in which there is a coupling to a complex scalar field that dynamically sets $\bar{\theta}$ to zero. There is then a spontaneously broken approximate U(1) symmetry whose pseudo-Goldstone boson is the axion [269, 270].

Another consequence of a nonzero θ, first pointed out by Witten, is seen in the electric charges of magnetically charged objects [69]. Recall that the electric charge plays a dual role in theories that include charged fields. On the one hand, it is a quantity that is dynamically coupled to electric fields and that can be measured by examining the Coulomb tail of these fields. On the other hand, it

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 17 Nov 2026 at 00:53:43, subject to the Cambridge Core terms of use , available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9781139017787.012

is proportional to the conserved Noether charge that arises from the invariance of the theory under phase rotations of the complex charged fields.

The Noether charge is the generator of the symmetry transformation and is constructed from products of the variation of the fields and the corresponding conjugate momenta. Consider a transformation of the form

$$\delta \mathbf{A}_\mu = \frac{1}{e}\,\mathbf{D}_\mu(\boldsymbol{\phi}/|\boldsymbol{\phi}|)\,,$$
$$\delta \boldsymbol{\phi} = 0 \tag{11.62}$$

in a theory with gauge coupling e where SU(2) is broken to U(1) by a triplet field $\boldsymbol{\phi}$. [As in Chap. 5, boldface vector notation refers to SU(2) group indices.] The corresponding Noether charge is

$$Q_{\text{Noether}} = \int d^3x\, \delta \mathbf{A}_j \cdot \boldsymbol{\Pi}^j\,, \tag{11.63}$$

where $\boldsymbol{\Pi}^j$ is the momentum conjugate to $\mathbf{A}_j$. If $\theta = 0$,

$$\boldsymbol{\Pi}_j = \frac{\partial \mathcal{L}}{\partial_0 \mathbf{A}_j} = \mathbf{F}^{j0}\,, \tag{11.64}$$

so

$$eQ_{\text{Noether}} = \int d^3x\, \mathbf{D}_j \hat{\boldsymbol{\phi}} \cdot \mathbf{F}^{j0} = \int d^3x\, \left[\partial_j \left(\hat{\boldsymbol{\phi}} \cdot \mathbf{F}^{j0}\right) - \hat{\boldsymbol{\phi}} \cdot \mathbf{D}_j \mathbf{F}^{j0}\right]. \tag{11.65}$$

The field equation $\mathbf{D}_j \mathbf{F}^{j0} = \boldsymbol{\phi} \times \mathbf{D}^0 \boldsymbol{\phi}$ shows that the last term in the second integral vanishes, so

$$eQ_{\text{Noether}} = \int d^2 S_j \hat{\boldsymbol{\phi}} \cdot \mathbf{F}^{j0} = Q_E\,, \tag{11.66}$$

where the integral is over the surface at spatial infinity. In a gauge with constant $\hat{\boldsymbol{\phi}}$, Eqs. (11.65) and (11.66) reduce to the expression for the electric charge in Eq. (5.84). Because Q_{Noether} is an integer, we recover the familiar result $Q_E = ne$.

Adding the θ term of Eq. (10.32) changes the conjugate momenta, so that now

$$\boldsymbol{\Pi}^j = \frac{\partial \mathcal{L}}{\partial_0 A_j} = \mathbf{F}^{j0} - \frac{\theta e^2}{16\pi^2} \epsilon^{jkl} \mathbf{F}_{kl}\,. \tag{11.67}$$

Repeating the above steps using this modified $\boldsymbol{\Pi}^j$ and using the Bianchi identity $\epsilon^{jkl} \mathbf{D}_j \mathbf{F}_{kl} = 0$ leads to

$$\begin{aligned} eQ_{\text{Noether}} &= \int d^2 S_j \left(\hat{\boldsymbol{\phi}} \cdot \mathbf{F}^{j0} - \frac{\theta e^2}{16\pi^2} \epsilon^{jkl} \hat{\boldsymbol{\phi}} \cdot \mathbf{F}_{kl}\right) \\ &= Q_E - \frac{e\theta}{2\pi}\left(\frac{e}{4\pi} Q_M\right). \end{aligned} \tag{11.68}$$

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 17 Nov 2026 at 00:53:43, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9781139017787.012

The Noether charge, being conjugate to a periodic variable, remains quantized in integer units. Hence, a monopole with magnetic charge $Q_M = 4\pi/e$ must actually be a dyon with electric charge

$$Q_E = ne + \frac{e\theta}{2\pi} \tag{11.69}$$

for some integer n. The periodicity of θ can be seen here by noting that the replacement $\theta \to \theta + 2\pi$ leaves the spectrum of allowed electric charges invariant. Because all magnetically charged objects have electric charges of this form, these noninteger charges are consistent with the generalized charge quantization condition of Eq. (5.16).

Looked at from a distance, the monopole appears like a point object and the charged fields whose transformation led to the Noether charge are not evident. The anomalous θ-dependent charge can then be understood by noting that we effectively have ordinary electromagnetism supplemented by a term

$$\Delta\mathcal{L}_{\rm em} = -\frac{\theta e^2}{8\pi^2}\mathbf{E}\cdot\mathbf{B}\,. \tag{11.70}$$

The Abelian Gauss's law then becomes

$$\boldsymbol{\nabla}\cdot\mathbf{E} - \frac{\theta e^2}{8\pi^2}\boldsymbol{\nabla}\cdot\mathbf{B} = \rho\,, \tag{11.71}$$

where ρ represents any purely electric sources.[7] It immediately follows that any magnetic charge must be accompanied by an electric charge at the same point.

It should be stressed that this is not an instanton effect. The connection with instantons (and the reason for including it here rather than in Chap. 5) is merely the historical accident that it was the discovery of the instanton solutions that led to the consideration of the effects of a θ-term.

Finally, there is an apparent puzzle if the theory contains massless fermions. As was noted already, all values of θ would then be equivalent, which would seem to be in conflict with the presence of θ in the dyon charge quantization condition. The resolution can be understood in light of the discussion of fermions and monopoles in Sec. 5.7. We saw there that the charged fermions create a condensate cloud about the monopole core, with the radius of the cloud inversely proportional to the fermion mass. As the fermion mass goes to zero, the charge density of the fermion–dyon system moves out to spatial infinity, and so the charge as measured at any finite radius goes to zero, regardless of the value of θ [93, 118].

[7] The presence of the θ-term here might seem to contradict the statement that adding a total divergence to the Lagrangian density should not affect the equations of motion. The explanation lies in the singularities of the Abelian gauge potential when a monopole is present.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 17 Nov 2026 at 00:53:43, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9781139017787.012