Chapter 4
Theories and Lagrangians II: Introducing Gauge Fields

Gauge theories play a central role in our current understanding of the fundamental interactions. The weak, electromagnetic and strong interactions are well described by gauge theories. We introduce them in this chapter for the first time. Although we often talk about gauge invariance, or gauge symmetry, these terms are a bit misleading. The gauge symmetry is more a redundancy in the description of the physical degrees of freedom than a symmetry, as will be shown later on. The redundancy is of course very useful because it makes Lorentz invariance and locality explicit, but it is not a symmetry in the same sense as rotations or translations. Gauge theories have incredible richness and complexity. Many aspects of their dynamics are still poorly understood. In our presentation we just scratch the surface of a deep subject.

4.1 Classical Gauge Fields

In classical electrodynamics the basic physical quantities are the electric and magnetic fields **E** and **B**. They can be expressed in terms of the scalar and vector potentials φ and **A** as

$$\begin{aligned} \mathbf{E} &= -\nabla\varphi - \frac{\partial \mathbf{A}}{\partial t}, \\ \mathbf{B} &= \nabla \times \mathbf{A}. \end{aligned} \tag{4.1}$$

From these equations we see that specifying **E** and **B** does not uniquely determine the potentials, since the former do not change under the gauge transformations

$$\varphi(t,\mathbf{x}) \to \varphi(t,\mathbf{x}) + \frac{\partial}{\partial t}\varepsilon(t,\mathbf{x}), \quad \mathbf{A}(t,\mathbf{x}) \to \mathbf{A}(t,\mathbf{x}) - \nabla\varepsilon(t,\mathbf{x}). \tag{4.2}$$

From a classical point of view the introduction of φ and **A** is seen as a technicality that helps solving the Maxwell equations, but without physical relevance.

L. Álvarez-Gaumé and M. Á. Vázquez-Mozo, *An Invitation to Quantum Field Theory*, Lecture Notes in Physics 839, DOI: 10.1007/978-3-642-23728-7_4,

The equations of electrodynamics can be recast in a manifestly Lorentz invariant form using the four-vector gauge potential $A^\mu = (\varphi, \mathbf{A})$ and the antisymmetric field strength tensor defined by

$$F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu. \tag{4.3}$$

The four Maxwell equations

$$\begin{aligned} \nabla \cdot \mathbf{E} &= \rho, \\ \nabla \cdot \mathbf{B} &= 0, \\ \nabla \times \mathbf{E} &= -\frac{\partial}{\partial t}\mathbf{B}, \\ \nabla \times \mathbf{B} &= \mathbf{j} + \frac{\partial}{\partial t}\mathbf{E}, \end{aligned} \tag{4.4}$$

are written in the form

$$\begin{aligned} \partial_\mu F^{\mu\nu} &= j^\mu, \\ \varepsilon^{\mu\nu\sigma\eta}\partial_\nu F_{\sigma\eta} &= 0, \end{aligned} \tag{4.5}$$

where the four-current $j^\mu = (\rho, \mathbf{j})$ contains the charge density and the electric current. The second set of equations are called the Bianchi identities and are satisfied by any field strength (4.3). Notice that $F_{\mu\nu}$, and therefore the Maxwell equations, are invariant under the gauge transformations (4.2), which in covariant form read

$$A_\mu \longrightarrow A_\mu + \partial_\mu \varepsilon. \tag{4.6}$$

Finally, the equations of motion of a particle with mass m and charge q

$$m\ddot{\mathbf{x}} = q\Big(\mathbf{E} + \dot{\mathbf{x}} \times \mathbf{B}\Big) \tag{4.7}$$

take the form

$$m\frac{du^\mu}{d\tau} = q\,F^{\mu\nu}u_\nu, \tag{4.8}$$

where $u^\mu(\tau)$ is the particle four-velocity as a function of the proper time τ. These equations of motion, depending only on the field strength $F_{\mu\nu}$, are also gauge invariant.

The physical role of the vector potential becomes manifest only in quantum mechanics. Using the prescription of minimal substitution $\mathbf{p} \rightarrow \mathbf{p} - q\mathbf{A}$, the Schrödinger equation describing a particle with charge q moving in an electromagnetic field is

$$i\frac{\partial}{\partial t}\Psi = \left[-\frac{1}{2m}\left(\nabla - iq\mathbf{A}\right)^2 + q\varphi\right]\Psi. \tag{4.9}$$

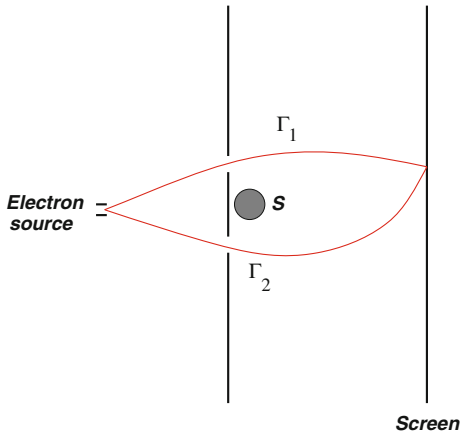

Fig. 4.1 Illustration of an interference experiment to show the Aharonov–Bohm effect. *S* represents the solenoid where the magnetic field is confined

Due to the explicit dependence on the electromagnetic potentials φ and $\mathbf{A}$, this equation seems to change under the gauge transformations (4.2). This is physically acceptable only if the ambiguity does not affect the probability density given by $|\Psi(t,\mathbf{x})|^2$. Therefore, a gauge transformation of the electromagnetic potential should amount to a change in the (unobservable) global phase of the wave function. This is indeed what happens: the Schrödinger equation (4.9) is invariant under the gauge transformations (4.2) provided the phase of the wave function is transformed at the same time according to

$$\Psi(t,\mathbf{x}) \longrightarrow e^{-iq\varepsilon(t,\mathbf{x})}\Psi(t,\mathbf{x}). \tag{4.10}$$

The Aharonov–Bohm Effect

This interplay between gauge transformations and the phase of the wave function gives rise to surprising phenomena. A first evidence of the role played by the electromagnetic potentials at the quantum level was pointed out by Yakir Aharonov and David Bohm [1]. Let us consider a double slit experiment as shown in Fig. 4.1, where we have placed a shielded solenoid just behind the first screen. Although the magnetic field is confined to the interior of the solenoid, the vector potential is nonvanishing also outside. The value of $\mathbf{A}$ outside the solenoid is locally a pure gauge, i.e., $\nabla \times \mathbf{A} = \mathbf{0}$, however since the region outside the solenoid is not simply connected the vector potential cannot be gauged to zero everywhere.

The dependence of the interference pattern with the magnetic field inside the solenoid can be calculated very easily using the path integral formalism introduced in Sect. 2.4. The probability amplitude for an electron emitted at $t = 0$ to be detected at some given position $\mathbf{x}$ on the screen at a later time τ is given by the propagator $K(\mathbf{x},\mathbf{x}_0;\tau)$, where $\mathbf{x}_0$ is the point where the electron is emitted. This propagator

admits a path integral representation, where the integration has to be done taking into account that there are two classes of paths that are topologically non-equivalent: those passing through the upper and the lower slits.

The classical action of a nonrelativistic particle of mass m and charge q in the presence of a vector potencial $\mathbf{A}$ is given by

$$S = \int dt \left(\frac{1}{2} m \dot{\mathbf{x}}^2 + q \dot{\mathbf{x}} \cdot \mathbf{A} \right) = \frac{1}{2} \int dt m \dot{\mathbf{x}}^2 + q \int_\gamma d\mathbf{x} \cdot \mathbf{A}, \tag{4.11}$$

where the second term in the last equation is a line integral along the particle trajectory γ. Using Stokes' theorem and $\nabla \times \mathbf{A} = \mathbf{0}$ we find that the value of this term only depends on the topological class of γ, but not in the particular curve within each class. Denoting by $K_1(\mathbf{x}, \mathbf{x}_0; \tau)$ and $K_2(\mathbf{x}, \mathbf{x}_0; \tau)$ the propagators of the electron going through each of the two slits in the absence of a magnetic field, the total propagator with the magnetic field switched on can be written as

$$\begin{aligned} K(\mathbf{x}, \mathbf{x}_0; \tau) &= e^{iq \int_{\Gamma_1} \mathbf{A} \cdot d\mathbf{x}} K_1(\mathbf{x}, \mathbf{x}_0; \tau) + e^{iq \int_{\Gamma_2} \mathbf{A} \cdot d\mathbf{x}} K_2(\mathbf{x}, \mathbf{x}_0; \tau) \\ &= e^{iq \int_{\Gamma_1} \mathbf{A} \cdot d\mathbf{x}} \left[K_1(\mathbf{x}, \mathbf{x}_0; \tau) + e^{iq \oint_\Gamma \mathbf{A} \cdot d\mathbf{x}} K_2(\mathbf{x}, \mathbf{x}_0; \tau) \right]. \end{aligned} \tag{4.12}$$

Here Γ_1 and Γ_2 are two arbitrary curves going through each of the two slits and joining $\mathbf{x}_0$ with $\mathbf{x}$ (see Fig. 4.1). Γ is the closed curve surrounding the solenoid defined by the union of Γ_1^{-1} and Γ_2.

The interference pattern on the screen is determined by the relative phase between the two terms in (4.12). The presence of the magnetic field confined to the solenoid introduces an extra term depending on the value of the vector potential outside the solenoid

$$U = \exp \left(iq \oint_\Gamma \mathbf{A} \cdot d\mathbf{x} \right). \tag{4.13}$$

Due again to Stokes' theorem and $\nabla \times \mathbf{A} = \mathbf{0}$ the value of the phase does not depend on the particular curve Γ chosen, so far as it surrounds the solenoid. The conclusion of this analysis is that the presence of the vector potential becomes observable even if the electrons do not feel the magnetic field directly. Performing the double-slit experiment when the magnetic field inside the solenoid is switched off we will observe the usual interference pattern on the second screen. Switching on the magnetic field a change in the interference pattern will appear due to the phase (4.13). This is the Aharonov–Bohm effect (see also [2] for an early prediction of the effect).

The first question that comes up is what happens with gauge invariance. Since $\mathbf{A}$ can be changed by a gauge transformation it seems that the resulting interference patters might depend on the gauge used. In fact the phase factor (4.13) is gauge invariant: the gauge variation of $\mathbf{A}$ is $-\nabla \varepsilon$ that, being a total derivative, gives zero upon integration over the close contour Γ.

The lesson we have learned is that in the quantum theory there are, apart from the electric and magnetic fields, other gauge invariant quantities giving observable

effects. An important difference with respect to **E** and **B** is that these gauge invariant observables are non-local, as can be seen from the definition of the phase U.

Magnetic Monopoles

It is very easy to check that the vacuum Maxwell equations

$$\begin{aligned} \nabla \cdot \mathbf{E} &= 0 \\ \nabla \cdot \mathbf{B} &= 0 \\ \nabla \times \mathbf{E} &= -\frac{\partial}{\partial t}\mathbf{B} \\ \nabla \times \mathbf{B} &= \frac{\partial}{\partial t}\mathbf{E} \end{aligned} \tag{4.14}$$

remain invariant under the transformation

$$\mathbf{E} - i\mathbf{B} \longrightarrow e^{i\theta}(\mathbf{E} - i\mathbf{B}), \quad \theta \in [0, 2\pi] \tag{4.15}$$

that for $\theta = \frac{\pi}{2}$ interchanges the electric and magnetic fields: $\mathbf{E} \to \mathbf{B}$, $\mathbf{B} \to -\mathbf{E}$. This duality symmetry is however broken in the presence of electric sources $(\rho, \mathbf{j})$. Nevertheless the Maxwell equations can be "completed" by introducing sources for the magnetic field $(\rho_m, \mathbf{j}_m)$ in such a way that the duality (4.15) is restored when supplemented by the transformation

$$\rho - i\rho_m \longrightarrow e^{i\theta}(\rho - i\rho_m), \quad \mathbf{j} - i\mathbf{j}_m \longrightarrow e^{i\theta}(\mathbf{j} - i\mathbf{j}_m). \tag{4.16}$$

In covariant language, this modification of the Maxwell equations implies adding sources on the right-hand side of the Bianchi identities

$$\partial_\mu \widetilde{F}^{\mu\nu} = j_m^\mu, \tag{4.17}$$

where $j_m^\mu = (\rho_m, \mathbf{j}_m)$ and

$$\widetilde{F}^{\mu\nu} = \frac{1}{2}\varepsilon^{\mu\nu\sigma\lambda}F_{\sigma\lambda} \tag{4.18}$$

is the dual electromagnetic tensor field. This means that, while electric charges act as sources for $F_{\mu\nu}$, magnetic charges are sources for $\widetilde{F}^{\mu\nu}$. The duality transformation (4.15, 4.16) is written now as

$$\begin{aligned} F_{\mu\nu} + i\widetilde{F}_{\mu\nu} &\longrightarrow e^{i\theta}\left(F_{\mu\nu} + i\widetilde{F}_{\mu\nu}\right), \\ j^\mu + ij_m^\mu &\longrightarrow e^{i\theta}\left(j^\mu + ij_m^\mu\right), \end{aligned} \tag{4.19}$$

keeping the extended Maxwell equations invariant. For $\theta = \frac{\pi}{2}$ electric and magnetic sources get interchanged and the field strength is replaced by its dual.

In 1931 Dirac [3] studied the possibility of finding solutions of the completed Maxwell equations with a magnetic monopoles of charge g as a source

$$\nabla \cdot \mathbf{B} = g\delta(\mathbf{x}). \tag{4.20}$$

Away from the position of the monopole $\nabla \cdot \mathbf{B} = 0$ and the magnetic field can still be derived locally from a vector potential $\mathbf{A}$ according to $\mathbf{B} = \nabla \times \mathbf{A}$. However, this potential cannot be regular everywhere since otherwise Gauss' theorem would imply that the magnetic flux threading a closed surface around the monopole should vanish, contradicting (4.20).

A solution to Eq. (4.20) in spherical coordinates is given by

$$B_r = \frac{1}{4\pi}\frac{g}{|\mathbf{x}|^2}, \quad B_\varphi = B_\theta = 0, \tag{4.21}$$

that for $\mathbf{x} \neq \mathbf{0}$ can be derived from the vector potential

$$A_\varphi = \frac{1}{4\pi}\frac{g}{|\mathbf{x}|}\tan\frac{\theta}{2}, \quad A_r = A_\theta = 0. \tag{4.22}$$

As expected, we find that this vector potential is singular at the half-line $\theta = \pi$ (see Fig. 4.2). This singular line starting at the position of the monopole is called the Dirac string and its position changes with a change of gauge but cannot be eliminated by any gauge transformation. Physically, we can see it as an infinitely thin solenoid confining a magnetic flux entering into the magnetic monopole from infinity that equals the outgoing magnetic flux from the monopole.

Since the position of the Dirac string depends on the gauge chosen it seems that we are facing a physical ambiguity. This would be rather strange since the Maxwell equations are gauge invariant also in the presence of magnetic sources. The solution to this apparent riddle lies in the fact that the presence of the Dirac string does not pose any consistency problem as far as it does not produce any physical effect, i.e., if its presence turns out to be undetectable. From our discussion of the Aharonov–Bohm effect we know that the wave function of charged particles picks up a phase (4.13) when surrounding a region where a magnetic flux is confined (such as the solenoid in the Aharonov–Bohm experiment). Since the Dirac string is like an infinitely thin solenoid, it will be unobservable if the phase picked up by the wave function of a charged particle surrounding it is equal to one. An evaluation of (4.13) in the field of the monopole shows that

$$e^{iqg} = 1 \quad \Longrightarrow \quad qg = 2\pi n \quad \text{with} \ \ n \in \mathbb{Z}. \tag{4.23}$$

Interestingly, we are led to the conclusion that the presence of a single magnetic monopole somewhere in the universe implies for consistency the quantization of the

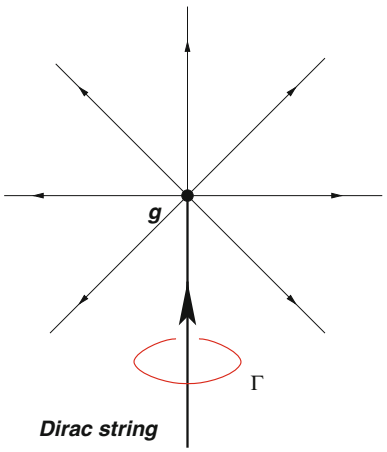

Fig. 4.2 The Dirac monopole

electric charge in units of $2\pi/g$, where g is the magnetic charge of the monopole.[1] This is called the Dirac charge quantization condition.

The idea of the magnetic monopole can be extended to dyons, particles having both electric and magnetic charge (q, g). The equations of motion for such particles in an electromagnetic field can be written remembering that magnetic charges couple to the dual field strength and requiring invariance under duality. This leads to

$$m\ddot{x}^{\mu} = \left(qF^{\mu\nu} + g\widetilde{F}^{\mu\nu}\right)\dot{x}_{\nu}, \tag{4.24}$$

where m is the mass of the dyon and the dot indicates differentiation with respect to the proper time. Writing the right-hand side of this equation in components in the nonrelativistic limit, we get the generalization of the Lorentz force acting on a dyon with charges (q, g)

$$\mathbf{F} = q\left(\mathbf{E} + \mathbf{v} \times \mathbf{B}\right) + g\left(\mathbf{B} - \mathbf{v} \times \mathbf{E}\right). \tag{4.25}$$

The invariance under duality is obvious noticing that the parentheses in the right-hand side of (4.24) can be written as $\mathrm{Im}[(q - ig)(F_{\mu\nu} - i\widetilde{F}_{\mu\nu})^*]$, which is manifestly invariant.

The Dirac quantization condition, valid for an electrically charged particle and a magnetic monopole, can be extended to two dyons with charges (q_1, g_1) and (q_2, g_2). To obtain this new condition one could proceed as in the case of the Dirac monopole

[1] The quantization of the electric charge has another consequence, which is that the gauge transformation of the wave function (4.10) is periodic. Using technical jargon one says that the U(1) gauge group gets compactified (see Appendix B). Although this might seem just a technical point, it has important physical consequences for the production of monopoles in gauge theories.

and impose that the corresponding singularities of the gauge potentials are unobservable. Here instead we are going to exploit the invariance of both the extended Maxwell equations and the equations of motion of the dyons under duality transformations.

These two facts imply that the proper quantization condition for the charge of the dyons should also be duality invariant and, moreover, reduce to the Dirac condition for the case $(q_1, g_1) = (q, 0)$, $(q_2, g_2) = (0, g)$. Taking into account the transformation of the electric and magnetic charges it is immediate to see that the following combination is duality invariant

$$\Big(q_1 - ig_1\Big)\Big(q_2 - ig_2\Big)^* = q_1q_2 + g_1g_2 + i\Big(q_1g_2 - q_2g_1\Big). \tag{4.26}$$

A look at the generalized Lorentz force shows $q_1g_2 - q_2g_1$ is the coupling constant of the velocity-dependent part of the force between the two dyons. The other duality-invariant combination, $q_1q_2 + g_1g_2$, gives the strength of the coupling of the velocity-independent part of this force, i.e., their "Coulomb" interaction. Since the imaginary part of Eq. (4.26) reduces to the Dirac quantization condition in the appropriate limit, we arrive at

$$q_1g_2 - q_2g_1 = 2\pi n, \quad \text{where } n \in \mathbb{Z}, \tag{4.27}$$

called the Dirac–Schwinger–Zwanziger quantization condition [4, 5].

There are some difficulties in considering quantum theories with *fundamental* magnetic monopoles. One of them is that they cannot be handled in perturbation theory, since the Dirac quantization condition implies that electric and magnetic coupling constants are inverse of each other and cannot be simultaneously small. This problem is avoided if monopoles are not fundamental objects but field configurations with finite size and energy. It was proved by 't Hooft and Polyakov [6, 7] that many gauge theories contain such monopoles as solitonic solutions. The 't Hooft-Polyakov monopoles have masses that scale with the inverse of the coupling constant, and therefore they are very heavy when the theory is weakly coupled. Only at large gauge couplings this objects become light and can be counted among the low-lying excitations of the system.

Monopoles are believed to have been produced copiously in the very early Universe. It is a generic prediction of grand unified theories that monopoles occur when a semisimple gauge group is spontaneously broken leaving a U(1) factor (spontaneous symmetry breaking will be explained in Chap. 7). The reason is that this U(1) is compact in the sense explained in the footnote of page 53, and therefore can "accommodate" monopole solutions. The fact that these monopoles are not observed today is believed to be the result of the dilution they underwent during the inflationary era that presumably followed their production.

4.2 Quantization of the Electromagnetic Field

We now proceed to the quantization of the electromagnetic field in the absence of sources $\rho = 0$, $\mathbf{j} = \mathbf{0}$. In this case the Maxwell equations (4.14) can be derived from the Lagrangian density

$$\mathscr{L}_{\text{Maxwell}} = -\frac{1}{4} F_{\mu\nu} F^{\mu\nu} = \frac{1}{2} \left(\mathbf{E}^2 - \mathbf{B}^2 \right). \tag{4.28}$$

Although in general the procedure to quantize the Maxwell Lagrangian is not very different from the one used for the Klein–Gordon or the Dirac field, here we need to deal with a new ingredient: gauge invariance. Unlike the cases studied so far, here the photon field A_μ is not unambiguously defined because the action and the equations of motion are insensitive to the gauge transformations $A_\mu \to A_\mu + \partial_\mu \varepsilon$. A first consequence of this symmetry is that the theory has less physical degrees of freedom than what would be expected for a vector field.

The way to tackle the problem of gauge invariance is to fix the freedom in choosing the electromagnetic potential before quantization. This can be done in several ways, for example by imposing the Lorentz gauge fixing condition

$$\partial_\mu A^\mu = 0. \tag{4.29}$$

Notice that this condition does not fix completely the gauge freedom since Eq. (4.29) is left invariant by gauge transformations satisfying $\partial_\mu \partial^\mu \varepsilon = 0$. One of the advantages of the Lorentz gauge is that it is covariant and therefore does not pose any danger to the Lorentz invariance of the quantum theory. Besides, applying it to the Maxwell equation $\partial_\mu F^{\mu\nu} = 0$ one finds

$$0 = \partial_\mu \partial^\mu A^\nu - \partial_\nu \left(\partial_\mu A^\mu \right) = \partial_\mu \partial^\mu A^\nu. \tag{4.30}$$

Since A_μ satisfies the massless Klein-Gordon equation the photon, the quantum of the electromagnetic interaction, has zero mass.

Once gauge invariance is fixed, $A_\mu(t, \mathbf{x})$ can be expanded in a complete basis of plane-wave solutions to Eq. (4.30)

$$\varepsilon_\mu(\mathbf{k}, \lambda) e^{-i|\mathbf{k}|t + i\mathbf{k}\cdot\mathbf{x}}, \tag{4.31}$$

where $\varepsilon_\mu(\mathbf{k}, \lambda)$ are the polarization vectors. In principle there are four independent polarizations for the photon, labelled by λ. The Lorentz gauge condition (4.29), however, forces the polarization vectors to be transverse

$$k^\mu \varepsilon_\mu(\mathbf{k}, \lambda) = k^\mu \varepsilon_\mu(\mathbf{k}, \lambda)^* = 0. \tag{4.32}$$

This condition can be used to eliminate one polarization. We can get rid of another one by using the on-shell condition $k^2 = 0$ and the residual gauge transformations mentioned after Eq. (4.29). Finally we are left with just two physical independent

transverse polarizations $\lambda = \pm 1$. They correspond to right and left circularly polarized photons.

Now, upon quantization, the gauge field operator $\hat{A}_\mu(t, \mathbf{x})$ can be written as the following expansion

$$\hat{A}_\mu(t, \mathbf{x}) = \sum_{\lambda=\pm 1} \int \frac{d^3k}{(2\pi)^3} \frac{1}{2|\mathbf{k}|} \Big[\varepsilon_\mu(\mathbf{k}, \lambda) \hat{a}(\mathbf{k}, \lambda) e^{-i|\mathbf{k}|t + i\mathbf{k}\cdot\mathbf{x}} + \varepsilon_\mu(\mathbf{k}, \lambda)^* \hat{a}^\dagger(\mathbf{k}, \lambda) e^{i|\mathbf{k}|t - i\mathbf{k}\cdot\mathbf{x}} \Big], \tag{4.33}$$

where the canonical commutation relations imply that

$$\begin{aligned} \Big[\hat{a}(\mathbf{k}, \lambda), \hat{a}^\dagger(\mathbf{k}', \lambda') \Big] &= (2\pi)^3 (2|\mathbf{k}|) \delta(\mathbf{k} - \mathbf{k}') \delta_{\lambda\lambda'} \\ \Big[\hat{a}(\mathbf{k}, \lambda), \hat{a}(\mathbf{k}', \lambda') \Big] &= \Big[\hat{a}^\dagger(\mathbf{k}, \lambda), \hat{a}^\dagger(\mathbf{k}', \lambda') \Big] = 0. \end{aligned} \tag{4.34}$$

Therefore $\hat{a}(\mathbf{k}, \lambda)$, $\hat{a}^\dagger(\mathbf{k}, \lambda)$ form a set of creation-annihilation operators for photons with momentum $\mathbf{k}$ and helicity λ.

Had we kept the unphysical degrees of freedom removed by the residual gauge transformations, the spectrum would contain states with negative norm. To decouple these states with negative probability is one of the main concerns in quantizing theories with gauge invariance. In these theories there is a redundancy in the way physical states are represented by rays in the Hilbert space $\mathscr{H}$: a physical state is represented by infinitely many rays in $\mathscr{H}$. Here we have dealt with this problem by eliminating this redundancy explicitly, i.e., keeping only those polarizations that are physical. Other strategies to handle this problem can be found in standard textbooks (see Ref. [1–15] in Chap. 1). In Sect. 4.6 we will return to the problem of fixing the gauge redundancy, this time using the path integral formalism.

From the previous discussion the reader might think that we have worked too hard unnecessarily. If the photon has only two physical degrees of freedom, perhaps we could describe it using two scalar degrees of freedom, instead of introducing a redundant four-component gauge field. The obstacle is Lorentz invariance: the only known way of describing the two photon polarizations in a Lorentz invariant way is through the gauge field A_μ. The gauge redundancy is the prize we pay for a Lorentz invariant and local description of massless photons.

4.3 Coupling Gauge Fields to Matter

Once we know how to quantize the electromagnetic field we can consider interacting theories containing electrically charged particles, for example electrons. To couple the Dirac Lagrangian to electromagnetism we use the analysis of the Schrödinger equation for a charged particle presented in pages 48–49. There we learned that the gauge ambiguity of the electromagnetic potential is compensated by a U(1) phase

shift in the wave function. The Lagrangian (3.36) is invariant under $\psi \to e^{-iq\varepsilon}\psi$, with ε a constant. This invariance is broken as soon as one identifies ε with the position-dependent gauge transformation parameter of the electromagnetic field.

To promote this global U(1) symmetry of the Dirac Lagrangian to a local one $\psi \to \psi' = e^{-iq\varepsilon(x)}\psi$ it is enough to replace ∂_μ by a covariant derivative D_μ, also transforming under a gauge transformation $D_\mu \to D'_\mu$, and satisfying

$$D'_\mu \psi' = D'_\mu \left[e^{-iq\varepsilon(x)}\psi\right] = e^{-iq\varepsilon(x)} D_\mu \psi. \tag{4.35}$$

Such a covariant derivative can be constructed in terms of the gauge potential A_μ as

$$D_\mu = \partial_\mu + iqA_\mu. \tag{4.36}$$

The gauge transformation of A_μ absorbs the derivative of the gauge parameter and Eq. (4.35) is satisfied. The electromagnetic field strength can be written in terms of the commutator of two covariant derivatives as

$$[D_\mu, D_\nu] = iqF_{\mu\nu}. \tag{4.37}$$

This identity will be useful in the construction of nonabelian gauge theories in the next section.

The Lagrangian of quantum electrodynamics (QED), i.e., a spin-$\frac{1}{2}$ field coupled to electromagnetism,

$$\mathscr{L}_{\text{QED}} = -\frac{1}{4}F_{\mu\nu}F^{\mu\nu} + \overline{\psi}(i\not{D} - m)\psi, \tag{4.38}$$

is invariant under the U(1) gauge transformations

$$\psi \longrightarrow e^{-iq\varepsilon(x)}\psi, \quad A_\mu \longrightarrow A_\mu + \partial_\mu \varepsilon(x). \tag{4.39}$$

Unlike the theories we encountered so far, QED is an interacting theory. By plugging (4.36) into the Lagrangian we find that the interaction term between fermions and photons has the form

$$\mathscr{L}^{(\text{int})}_{\text{QED}} = -\mathscr{H}^{(\text{int})}_{\text{QED}} = -qA_\mu \overline{\psi}\gamma^\mu \psi. \tag{4.40}$$

This shows that, as anticipated in the previous chapter (see page 43), the electric current four-vector is given by $j^\mu = q\overline{\psi}\gamma^\mu\psi$. In the following we stick to the general convention and denote the charge by e. In the case of electrons or muons, for example, e is negative and equal to the elementary charge.

The quantization of interacting field theories like QED poses new problems that we did not meet in the case of the free theories. In particular in most cases it is not possible to solve the theory exactly. When this happens the physical observables have to be computed in perturbation theory in powers of the coupling constant. An added problem appears in the computation of quantum corrections to the classical

result, which is plagued with infinities that should be taken care of. All these issues will be addressed in Chaps. 6 and 8.

Here we can connect with the comments made at the beginning of the chapter. The end result of our quantization procedure is to write the gauge field in terms of the two physical degrees of freedom appearing in (4.33). Out of the four components of A_μ only two represent physical degrees of freedom. It is clear that if we wrote the theory (after including interactions) only in terms of the transverse degrees of freedom the result would be a theory without explicit Lorentz symmetry and also with non-local interactions. The inclusion of longitudinal- and timelike photons makes these apparently lost, but fundamental properties, explicit. The basic problem in the quantization of gauge theories is to make sure that at the quantum level the additional components continue to be irrelevant. Unfortunately this is not always possible, in some cases there are quantum anomalies making the theory inconsistent (see Chap. 9).

4.4 Nonabelian Gauge Theories

QED is the simplest example of a gauge theory coupled to matter based on the abelian gauge symmetry of local U(1) phase rotations. Gauge theories based on nonabelian groups can also be constructed. Our knowledge of the strong and weak interactions is in fact based on the use of the nonabelian generalizations of QED, called *Yang–Mills theories*.

Let us consider a gauge group G with hermitian generators T^A, $A = 1, \ldots, \dim G$ satisfying the Lie algebra[2]

$$\left[T^A, T^B\right] = i f^{ABC} T^C. \tag{4.41}$$

We introduce a vector field $A_\mu \equiv A_\mu^A T^A$ taking values on the Lie algebra $\mathfrak{g}$ of the group G. Its gauge transformation is given by

$$A_\mu \longrightarrow A'_\mu = -\frac{1}{i g_{\rm YM}} U \partial_\mu U^{-1} + U A_\mu U^{-1}, \quad U = e^{i\chi(x)}, \tag{4.42}$$

where $\chi(x) = \chi^A(x) T^A$ and $g_{\rm YM}$ is the coupling constant. These gauge transformations are non-linear in the gauge function $\chi(x)$. Infinitesimally, the matrix-valued field A_μ transforms according to

$$\delta A_\mu = \frac{1}{g_{\rm YM}} \partial_\mu \chi - i[A_\mu, \chi], \tag{4.43}$$

which in components reads

[2] Some basics facts about Lie groups have been summarized in Appendix B.

$$\delta A_\mu^A = \frac{1}{g_{\rm YM}} \partial_\mu \chi^A + f^{ABC} A_\mu^B \chi^C. \tag{4.44}$$

As in the abelian case, the coupling of matter to a nonabelian gauge field is done by introducing a covariant derivative. Let Φ be a field (scalar or spinor) transforming in a representation $\mathbf{R}$ of the gauge group G

$$\Phi \longrightarrow \Phi' = U_{\mathbf{R}} \Phi. \tag{4.45}$$

The covariant derivative satisfying $D'_\mu \Phi' = U_{\mathbf{R}} D_\mu \Phi$ is defined by

$$D_\mu \Phi = \partial_\mu \Phi - i g_{\rm YM} A_\mu \Phi, \tag{4.46}$$

where $A_\mu = A_\mu^A T_{\mathbf{R}}^A$, with $T_{\mathbf{R}}^A$ the generators in the representation $\mathbf{R}$. In the particular case of the adjoint representation the generators can be written in terms of the structure constants

$$\left(T_{\rm adj}^A\right)_C^B = -i f^{ABC}, \tag{4.47}$$

and the covariant derivative takes the form

$$D_\mu \Phi = \partial_\mu \Phi - i g_{\rm YM} \left[A_\mu, \Phi\right] \qquad \text{(adjoint representation)}. \tag{4.48}$$

Comparing this expression with (4.43) we find that the infinitesimal transformation of the gauge field can be expressed as

$$\delta A_\mu = \frac{1}{g_{\rm YM}} D_\mu \chi. \tag{4.49}$$

Our last task is to find the kinetic term for the nonabelian gauge fields. Generalizing Eq. (4.37), we write

$$\left[D_\mu, D_\nu\right] = -i g_{\rm YM} F_{\mu\nu}, \tag{4.50}$$

where $F_{\mu\nu}$ is the nonabelian field strength

$$F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu - i g_{\rm YM} \left[A_\mu, A_\nu\right] \tag{4.51}$$

This expression reduces to (4.3) for abelian gauge groups, when the commutator of the gauge fields vanishes. The field strength tensor takes values in the Lie algebra, $F_{\mu\nu} = F_{\mu\nu}^A T^A$, where

$$F_{\mu\nu}^A = \partial_\mu A_\nu^A - \partial_\nu A_\mu^A + g_{\rm YM} f^{ABC} A_\mu^B A_\nu^C. \tag{4.52}$$

Unlike the case of the Maxwell theory the field strength for nonabelian gauge fields is not gauge invariant. Using (4.50) and the transformation of the covariant derivative it is easy to show that it transforms as

$$F_{\mu\nu} \longrightarrow U F_{\mu\nu} U^{-1}. \tag{4.53}$$

This gives the clue to constructing a gauge invariant Lagrangian for the nonabelian gauge field A_μ as

$$\mathscr{L} = -\frac{1}{2}\mathrm{Tr}\Big(F_{\mu\nu}F^{\mu\nu}\Big) = -\frac{1}{4}F^A_{\mu\nu}F^{A\mu\nu}, \tag{4.54}$$

where the normalization $\mathrm{Tr}(T^A T^B) = \frac{1}{2}\delta^{AB}$ has been used. A crucial difference between this and the Lagrangian of electromagnetism is the presence of cubic and quartic terms in the gauge field A_μ. This means that, unlike the photon, the nonabelian gauge bosons act themselves as sources of the field. The equations of motion derived from the Lagrangian (4.54) can be written as

$$D_\mu F^{\mu\nu} = 0, \tag{4.55}$$

where D_μ is the covariant derivative in the adjoint representation shown in Eq. (4.48).

Just as in the Maxwell theory, the components of the nonabelian field strength tensor $F^A_{\mu\nu}$ in four dimensions can be decomposed into electric and magnetic fields $\mathbf{E}^A$ and $\mathbf{B}^A$

$$E^A_i = F^A_{0i}, \quad B^A_i = -\frac{1}{2}\varepsilon_{ijk}F^A_{jk}. \tag{4.56}$$

From (4.53) it follows that the nonabelian electric and magnetic fields are gauge dependent. In terms of them the Lagrangian (4.54) becomes

$$\mathscr{L} = \frac{1}{2}\Big(\mathbf{E}^A\cdot\mathbf{E}^A - \mathbf{B}^A\cdot\mathbf{B}^A\Big). \tag{4.57}$$

In QCD $\mathbf{E}^A$ and $\mathbf{B}^A$ are respectively known as chromoelectric and chromomagnetic fields.

With all this information we can write a generic Lagrangian for a nonabelian gauge field coupled to scalars ϕ and spinors ψ as

$$\begin{aligned}\mathscr{L} = &-\frac{1}{2}\mathrm{Tr}\Big(F_{\mu\nu}F^{\mu\nu}\Big) + i\overline{\psi}\not{D}\psi + (D_\mu\phi)^\dagger D^\mu\phi \\ &- \overline{\psi}\Big[M_1(\phi) + i\gamma_5 M_2(\phi)\Big]\psi - V(\phi),\end{aligned} \tag{4.58}$$

where the covariant derivatives are in the representation of the field involved. The Lagrangian of the standard model is of this form, with $M_1(\phi)$ and $M_2(\phi)$ linear in ϕ and $V(\phi)$ of quartic order. This particular form of the functions appearing in (4.58) is related to the good properties of the standad model at high energies.

4.5 Understanding Gauge Symmetry

In classical mechanics the application of the Hamiltonian formalism starts with the replacement of generalized velocities by momenta

$$p_i \equiv \frac{\partial L}{\partial \dot{q}_i} \qquad \Longrightarrow \qquad \dot{q}_i = \dot{q}_i(q, p). \tag{4.59}$$

Most of the time there is no problem in inverting the relations $p_i = p_i(q, \dot{q})$. However in some systems these relations might not be invertible and result in a number of constraints of the type

$$f_a(q, p) = 0, \quad a = 1, \ldots, N_1. \tag{4.60}$$

These systems are called degenerate or constrained [8, 9].

The presence of constraints of the type (4.60) makes the formulation of the Hamiltonian formalism more involved. The first problem is related to the ambiguity in defining the Hamiltonian, since the addition of any linear combination of the constraints does not modify its value. Secondly, one has to make sure that the constraints are consistent with the time evolution in the system. In the language of Poisson brackets this means that further constraints have to be imposed in the form

$$\{f_a, H\}_{PB} \approx 0. \tag{4.61}$$

Following [8], we use the symbol $\approx$ to indicate a "weak" equality holding when the constraints $f_a(q, p) = 0$ are satisfied. Notice however that since the computation of the Poisson brackets involves derivatives, the constraints can be used only after the bracket is computed. In principle, the conditions (4.61) can give rise to a new set of constraints $g_b(q, p) = 0,\ b = 1, \ldots, N_2$. Again these constraints have to be consistent with time evolution and we have to repeat the procedure. Eventually this finishes when a set of constraints is found that do not require any further constraint to be preserved in time.[3]

Once all the constraints of a degenerate system have been found we consider the so-called first class constraints $\phi_a(q, p) = 0, a = 1, \ldots, M$, those whose mutual Poisson bracket vanishes weakly

$$\{\phi_a, \phi_b\}_{\text{PB}} = c_{abc}\phi_c \approx 0. \tag{4.62}$$

The constraints that do not satisfy this condition, called second class constraints, can be eliminated by modifying the Poisson bracket [8], so for all practical purposes we can forget about them. The total Hamiltonian of the theory is defined as the canonical Hamiltonian plus a linear combination of all first-class constraints with arbitrary coefficients

[3] In principle it is also possible that the procedure finishes because some kind of inconsistent identity is found. In this case the system itself is inconsistent as it happens with the Lagrangian $L(q, \dot{q}) = q$.

$$H_T = p_i \dot{q}_i - L + \sum_{a=1}^{M} \lambda_a(t)\phi_a. \tag{4.63}$$

The total Hamiltonian and the canonical one coincide on the submanifold of phase space defined by the first class constraints, where the dynamical evolution of the system takes place.

What is the relation with gauge invariance? The answer lies in the fact that for a singular system the first class constraints ϕ_a generate gauge transformations. Indeed, the time evolution generated by the Hamiltonian (4.63) is ambiguous due to the presence of the arbitrary functions $\lambda_a(t)$. Specifying the state of the system by the values of the canonical variables at some reference time t_0, the ambiguity in the time evolution translates into a redundancy in the description of the state of the system in terms of the values of the canonical variables at a later time t: the phase space trajectories related by the infinitesimal transformations

$$\begin{aligned} q_i &\longrightarrow q_i + \sum_{a=1}^{M} \varepsilon_a(t)\{q_i, \phi_a\}_{\mathrm{PB}}, \\ p_i &\longrightarrow p_i + \sum_{a=1}^{M} \varepsilon_a(t)\{p_i, \phi_a\}_{\mathrm{PB}} \end{aligned} \tag{4.64}$$

describe one and the same state.

This ambiguity in the description of the system in terms of the generalized coordinates and momenta can be traced back to the equations of motion in Lagrangian language. Writing them in the form

$$\frac{\partial^2 L}{\partial \dot{q}_i \partial \dot{q}_j} \ddot{q}_j = -\frac{\partial^2 L}{\partial \dot{q}_i \partial q_j} \dot{q}_j + \frac{\partial L}{\partial q_i}, \tag{4.65}$$

we find that in order to determine the accelerations in terms of the positions and velocities, the matrix $\frac{\partial^2 L}{\partial \dot{q}_i \partial \dot{q}_j}$ has to be invertible. However, the existence of constraints (4.60) precisely implies that the determinant of this matrix vanishes and therefore the time evolution is not uniquely determined in terms of the initial conditions.

Applications to Electrodynamics

After a general discussion we particularize the analysis to the Maxwell Lagrangian

$$L = -\frac{1}{4} \int d^3x F_{\mu\nu} F^{\mu\nu}. \tag{4.66}$$

The generalized momenta conjugate to A_μ is defined by

$$\pi^\mu = \frac{\delta L}{\delta(\partial_0 A_\mu)} = F^{\mu 0}, \tag{4.67}$$

hence, $\pi^0 = 0$ and $\pi^i = E^i$. The Hamiltonian is given by

$$H = \int d^3x \left(\pi^\mu \partial_0 A_\mu - \mathscr{L}\right) = \int d^3x \left[\frac{1}{2}\left(\mathbf{E}^2 + \mathbf{B}^2\right) + A_0 \nabla \cdot \mathbf{E}\right], \tag{4.68}$$

where we have used $\partial_0 \mathbf{A} = \nabla A_0 - \pi = \nabla A_0 - \mathbf{E}$ and integrated by parts the second term in the last integral.

The Hamiltonian (4.68) shows that $A_0(x)$ plays the role of a Lagrange multiplier implementing Gauss' law $\nabla \cdot \mathbf{E} = 0$ as a constraint.[4] Thus $\pi^0 = 0$ and $\nabla \cdot \pi = 0$ form a set of two first class constraints generating gauge transformations. The ones generated by π^0 can be used to fix the value of $A_0(x)$, thus defining a temporal gauge. This does not completely fix the gauge freedom, since there are the gauge transformations generated by Gauss' law. Using the canonical Poisson brackets

$$\{A_i(t, \mathbf{x}), E_j\left(t, \mathbf{x}'\right)\}_{\mathrm{PB}} = \delta_{ij}\delta\left(\mathbf{x} - \mathbf{x}'\right) \tag{4.69}$$

we find these to be

$$\delta A_i(t, \mathbf{x}) = \{A_i(t, \mathbf{x}), \int d^3x' \varepsilon(t, \mathbf{x}') \nabla \cdot \mathbf{E}(t, \mathbf{x}')\}_{\mathrm{PB}} = \partial_i \varepsilon(t, \mathbf{x}), \tag{4.70}$$

while $A_0(t, \mathbf{x})$ is left invariant. This is equivalent to a general gauge transformation generated by a time-independent gauge function $\varepsilon(\mathbf{x})$. Thus, for consistency, we take $\varepsilon(t, \mathbf{x})$ in (4.70) to depend only on the spatial coordinates. The constraint $\nabla \cdot \mathbf{E} = 0$ can be implemented by demanding $\nabla \cdot \mathbf{A} = 0$, reducing the three degrees of freedom of $\mathbf{A}$ to the two physical degrees of freedom of the photon.

So much for the classical analysis. In the quantum theory the constraint $\nabla \cdot \mathbf{E} = 0$ has to be imposed on the physical states $|\mathrm{phys}\rangle$. This is done by defining the following unitary operator in the Hilbert space

$$\mathscr{U}(\varepsilon) \equiv \exp\left[i \int d^3x \varepsilon(\mathbf{x}) \nabla \cdot \mathbf{E}\right]. \tag{4.71}$$

By definition, physical states should not change when a gauge transformations is performed. This is implemented by requiring the operator $\mathscr{U}(\varepsilon)$ to act trivially on them

$$\mathscr{U}(\varepsilon)|\mathrm{phys}\rangle = |\mathrm{phys}\rangle \qquad \Longrightarrow \qquad (\nabla \cdot \mathbf{E})|\mathrm{phys}\rangle = 0. \tag{4.72}$$

---

[4] This constraint can also be obtained from the requirement that $\pi^0 = 0$ be preserved by the time evolution, $\{\pi^0, H\}_{\mathrm{PB}} = 0$. A detailed analysis of Maxwell electrodynamics using the general formalism for constrained systems can be found in [9].

In the presence of a charge density ρ, this condition becomes $(\nabla \cdot \mathbf{E} - \rho)|\text{phys}\rangle = 0$.

The action of the gauge transformations in the quantum theory is very illuminating in understanding the real role of gauge invariance [10–12]. We have learned that the presence of a gauge symmetry in a theory reflects a degree of redundancy in the description of physical states in terms of the degrees of freedom appearing in the Lagrangian. In classical mechanics, for example, the state of a system is determined by the value of the canonical coordinates (q_i, p_i). We know, however, that this is not the case for constrained Hamiltonian systems, where the transformations generated by the first class constraints change the value of q_i and p_i without actually changing the physical state. Physical (i.e., measurable) quantities have to be free from such ambiguity and therefore be represented by gauge invariant objects. The same happens in classical field theory: in the Maxwell theory for every physical configuration determined by the gauge invariant quantities **E** and **B** there is an infinite number of possible values of A_μ related by gauge transformations $\delta A_\mu = \partial_\mu \varepsilon$.

In the quantum theory this means that one should identify into a single physical state all rays in the Hilbert space related by the operator $\mathscr{U}(\varepsilon)$ with any gauge function $\varepsilon(x)$. In other words, each physical state corresponds to a whole orbit of states transforming among themselves by gauge transformations.

This explains the necessity of gauge fixing. In order to avoid the redundancy in the states a further condition should be given selecting one single state on each orbit. Once again, we connect with the opening comments in this chapter. In the Hamiltonian quantization we see very clearly described how the gauge symmetry is more a redundancy than a symmetry. In going to the timelike gauge, i.e., imposing $A_0 = 0$, we eliminate one of the components of the gauge field. In the initial value surface we need to impose Gauss' law (by requiring for example $\nabla \cdot \mathbf{A} = 0$) to eliminate yet one more degree of freedom, reducing the number of physical degrees of freedom to two per gauge group generator.

4.6 Gauge Fields and Path Integrals

The redundancy in the Hilbert space is a source of complications when quantizing gauge theories. This we have seen already in Sect. 4.2: the photon had two unphysical polarizations removed using the Lorentz gauge fixing condition and the residual gauge invariance.

In the path integral formalism the problem of gauge invariance reflects in the necessity of carrying out the integration over gauge fields in a way that avoids over-counting. This means that two field configurations related by a gauge transformation should be considered as physically equivalent and included only once. For example, a naive evaluation of the vacuum-to-vacuum amplitude (partition function)

$$\mathscr{Z} = \int \mathscr{D} A_\mu e^{-\frac{i}{2}\int d^4x \mathrm{Tr}(F_{\mu\nu}F^{\mu\nu})} \tag{4.73}$$

would include together with each gauge field configuration A_μ all others obtained from it by an arbitrary gauge transformation, thus overcounting the result by an infinite factor equal to the volume of the gauge group.

The correct evaluation of the integral (4.73) requires restricting the integration to fields not related by gauge transformations. A practical way to do this is to notice that the computation of observables in quantum field theory generically involves quotients of path integrals [see Chap. 6 and in particular Eq. (6.35)]. Then it suffices to cancel the (infinite) volume factor in the numerator and denominator.

To carry out this program we follow ideas due to Faddeev and Popov [13] and begin by imposing a set of gauge fixing conditions of the form

$$\mathscr{F}^A(A_\mu) = 0. \tag{4.74}$$

They can be visualized as a "slice" in the space of all gauge field configurations. Each A_μ falls into a gauge orbit generated by the gauge transformations acting on it. Two gauge field configurations are nonequivalent if they lie on different orbits. The condition (4.74) selects a representative on each orbit and has to satisfy a number of requirements: it has to be reachable from any A_μ, i.e., each gauge orbit should have a representative satisfying (4.74), and this representative should be unique. To keep expressions simple in the following we drop the group theory index in Eq. (4.74) and denote the gauge conditions collectively by $\mathscr{F}(A_\mu) = 0$.

The next step is to split the functional integral (4.73) into an integration over the orbit representatives and an integral over each gauge orbit. This last integration results in a common factor equal to the volume of the gauge group. This is done by introducing the functional $\Delta_{\rm FP}[A_\mu]$ through the following definition

$$1 = \Delta_{\rm FP}[A_\mu] \int \mathscr{D}U \delta\Big[\mathscr{F}(A_\mu^U)\Big], \tag{4.75}$$

where we are integrating over all gauge transformations and by A_μ^U we denote the gauge potential transformed by U. For reasons that will be explained soon, $\Delta_{\rm FP}[A_\mu]$ is called the Faddeev–Popov determinant. It is not difficult to show that it is gauge invariant. Indeed, for any gauge transformation U' we have

$$\begin{aligned} \Delta_{\rm FP}[A_\mu^{U'}]^{-1} &= \int \mathscr{D}U \delta\Big[\mathscr{F}\left(A_\mu^{UU'}\right)\Big] \\ &= \int \mathscr{D}U'' \delta\Big[\mathscr{F}\left(A_\mu^{U''}\right)\Big] = \Delta_{\rm FP}[A_\mu]^{-1}, \end{aligned} \tag{4.76}$$

where we have made the change of variables $U'' = UU'$ and used the gauge invariance of the integration measure over the gauge group, $\mathscr{D}U'' = \mathscr{D}U$.

We insert now the identity (4.75) into the function integral (4.73)

$$\mathscr{Z} = \int \mathscr{D}A_\mu \mathscr{D}U \, \Delta_{\rm FP}[A_\mu] \delta\Big[\mathscr{F}\left(A_\mu^U\right)\Big] e^{-\frac{i}{2}\int d^4x {\rm Tr}(F_{\mu\nu}F^{\mu\nu})}. \tag{4.77}$$

Doing the change of variables $A_\mu \to A_\mu^{U^{-1}}$ and using the gauge invariance of both the action and $\Delta_{\rm FP}[A_\mu]$, we remove all dependence on U from the integrand. If the integration measure over the gauge fields $\mathscr{D}A_\mu$ is gauge invariant, this change of variables does not induce any Jacobian and the integration over the gauge group can be factored out

$$\mathscr{Z} = \left(\int \mathscr{D}U\right) \int \mathscr{D}A_\mu \Delta_{\rm FP}[A_\mu] \delta\Big[\mathscr{F}(A_\mu)\Big] e^{-\frac{i}{2}\int d^4x {\rm Tr}(F_{\mu\nu}F^{\mu\nu})}. \tag{4.78}$$

We can ignore the divergent prefactor and replace (4.73) by the gauge-fixed functional integral

$$\mathscr{Z} = \int \mathscr{D}A_\mu \Delta_{\rm FP}[A_\mu] \delta\Big[\mathscr{F}(A_\mu)\Big] e^{-\frac{i}{2}\int d^4x {\rm Tr}(F_{\mu\nu}F^{\mu\nu})}. \tag{4.79}$$

The delta function restricts the integration to gauge configurations lying on the slice $\mathscr{F}(A_\mu) = 0$, i.e., the integral only includes the contributions of the representatives of each gauge orbit.

To find an explicit expression for $\Delta_{\rm FP}[A_\mu]$ we use a functional version of the delta-function identity (2.10), namely

$$\delta\Big[\mathscr{F}\left(A_\mu^U\right)\Big] = \left|\det\left[\left.\frac{\delta\mathscr{F}(A_\mu^U)}{\delta U}\right|_{U=U'}\right]\right|^{-1} \delta(U-U'), \tag{4.80}$$

where U' is a gauge transformation such that $\mathscr{F}\left(A_\mu^{U'}\right) = 0$ for a given A_μ. Going back to Eq. (4.75) and integrating over U using the delta function, we find that $\Delta_{\rm FP}\left[A_\mu\right]$ can be expressed as the following functional determinant

$$\Delta_{\rm FP}[A_\mu] = \det\left[\left.\frac{\delta\mathscr{F}(A_\mu^U)}{\delta U}\right]\right|_{U=\mathbf{1}}. \tag{4.81}$$

In writing this expression we have used that $\Delta_{\rm FP}[A_\mu] = \Delta_{\rm FP}[A_\mu^{U'^{-1}}]$. This means that in the computation of the Faddeev–Popov determinant we have to impose that the gauge field lies on the gauge slice $\mathscr{F}(A_\mu) = 0$.

It should be clear that the value of the path integral (4.79) is not modified by changing the position of the slice defined by (4.74). That is, the value of $\mathscr{Z}$ does not change if we replace $\mathscr{F}(A_\mu)$ by $\mathscr{F}(A_\mu) = f(x)$, where $f(x)$ is an arbitrary Lie algebra valued function of the coordinates,

$$\mathscr{Z} = \int \mathscr{D}A_\mu \Delta_{\rm FP}[A_\mu] \delta\Big[\mathscr{F}(A_\mu) - f(x)\Big] e^{-\frac{i}{2}\int d^4x {\rm Tr}(F_{\mu\nu}F^{\mu\nu})}. \tag{4.82}$$

Since the previous expression is independent of $f(x)$ we can insert the constant term

$$\int \mathscr{D} f e^{-\frac{i}{\xi}\int d^4x \mathrm{Tr}[f(x)^2]} = \text{constant}, \tag{4.83}$$

and carry out the integration over $f(x)$ using the delta function. Modulo a global normalization, this gives

$$\mathscr{Z} = \int \mathscr{D} A_\mu \Delta_{\mathrm{FP}}[A_\mu] e^{i\int d^4x \mathrm{Tr}\left[-\frac{1}{2}F_{\mu\nu}F^{\mu\nu} - \frac{1}{\xi}\mathscr{F}(A_\mu)^2\right]}, \tag{4.84}$$

where ξ is an arbitrary real parameter. The new term added to the action is called the gauge fixing term.

We illustrate the previous discussion with two examples. We begin with QED and impose the Lorentz gauge $\mathscr{F}(A_\mu) = \partial_\mu A^\mu$. Using $U(x) = e^{ie\varepsilon(x)}$ we find

$$\mathscr{F}(A^U_\mu) = \partial_\mu A^\mu + \partial_\mu \partial^\mu \varepsilon \qquad \Longrightarrow \qquad \left.\frac{\delta \mathscr{F}(A^U_\mu)}{\delta U}\right|_{U=\mathbf{1}} = -\frac{1}{ie}\partial_\mu \partial^\mu. \tag{4.85}$$

Hence $\Delta_{\mathrm{FP}}[A_\mu] = |\det(-\frac{1}{ie}\partial_\mu\partial^\mu)|$ is independent of the gauge field. This means that we do not have to bother computing the determinant because it goes out of the path integral as an irrelevant global normalization constant. The typical functional integral for QED can be written as

$$\mathscr{Z}_{\mathrm{QED}} = \int \mathscr{D}\overline{\psi}\mathscr{D}\psi\mathscr{D}A_\mu e^{i(S_{\mathrm{QED}}+S_{\mathrm{gf}})}, \tag{4.86}$$

where the action and the gauge-fixing term read

$$S_{\mathrm{QED}} + S_{\mathrm{gf}} = \int d^4x \left[\overline{\psi}(i\not{D} - m)\psi - \frac{1}{4}F_{\mu\nu}F^{\mu\nu} - \frac{1}{2\xi}\left(\partial_\mu A^\mu\right)^2\right]. \tag{4.87}$$

The conclusion is that the problem of gauge invariance in the path integral quantization of QED is handled in a Lorentz-invariant way by adding a gauge fixing term to the action. The constant ξ is arbitrary and can be chosen to make some expressions simpler. In Chap. 6 we will learn how to compute observables in QED.

The case of nonabelian Yang–Mills theories is more complicated and here we only outline the procedure. Using the Lorentz condition $\mathscr{F}(A_\mu) = \partial_\mu A^\mu$ and the gauge transformation $\delta A_\mu = \frac{1}{g_{\mathrm{YM}}} D_\mu \chi$ we find

$$\left.\frac{\delta \mathscr{F}(A^U_\mu)}{\delta U}\right|_{U=\mathbf{1}} = \frac{1}{ig_{\mathrm{YM}}}\partial_\mu D^\mu, \tag{4.88}$$

where D_μ is the covariant derivative in the adjoint representation, given by (4.48). Unlike the case of QED, now the Faddeev–Popov determinant depends on the gauge

field, even after imposing the Lorentz condition $\partial_\mu A^\mu = 0$. This has to be taken into account when carrying out the integration over A_μ. The standard way to proceed now is to write $\Delta_{\rm FP}[A_\mu]$ as a path integral over some unphysical fields called the Faddeev–Popov ghosts. The details can be found in most of the textbooks listed in Ref. [1–15] of Chap. 1.

The use of Faddeev–Popov ghosts in nonabelian gauge theories can be avoided, for example, in the axial gauge $n^\mu A_\mu = 0$, with $n_\mu n^\mu < 0$. In this case

$$\left.\frac{\delta \mathscr{F}(A^U_\mu)}{\delta U}\right|_{U=\mathbf{1}} = \frac{1}{ig_{\rm YM}} n^\mu D_\mu. \tag{4.89}$$

Imposing the gauge condition $n^\mu A_\mu = 0$, we find that $n^\mu D_\mu = n^\mu \partial_\mu$ and $\Delta_{\rm FP}[A_\mu]$ is independent of the gauge field. It can be absorbed in the global normalization of the path integral, and the partition function (4.79) becomes

$$\begin{aligned} \mathscr{Z} &= \int \mathscr{D}A_\mu \delta[n^\nu A_\nu] e^{-\frac{i}{2}\int d^4x \mathrm{Tr}(F_{\mu\nu}F^{\mu\nu})} \\ &= \int \mathscr{D}A_\mu e^{i\int d^4x \mathrm{Tr}(-\frac{1}{2}F_{\mu\nu}F^{\mu\nu} - \frac{1}{\xi}n^\mu n^\nu A_\mu A_\nu)}. \end{aligned} \tag{4.90}$$

4.7 The Structure of the Gauge Theory Vacuum

The topology of the gauge group plays an important physical role in Yang–Mills theories. To illustrate the issue, we first look at a toy model: a U(1) gauge theory in $1+1$ dimensions. Later we will be more general. We will also point out a number of subtleties involved in the definition of the topology of the gauge field making the arguments presented more semiclassical rather than nonperturbative.

In the Hamiltonian formalism, gauge transformations $g(\mathbf{x})$ are functions defined on $\mathbb{R}$ with values on the gauge group U(1)

$$g : \mathbb{R} \longrightarrow U(1). \tag{4.91}$$

We assume that $g(x)$ is regular at infinity. In this case we can add to the real line $\mathbb{R}$ the point at infinity and compactify it to the circle S^1 (see Fig. 4.3). Once this is done, the $g(x)$'s are functions defined on S^1 with values on $\mathrm{U}(1) = S^1$ that can be parametrized as

$$g : S^1 \longrightarrow U(1), \quad g(x) = e^{i\alpha(x)}, \tag{4.92}$$

with $x \in [0, 2\pi]$.

Since S^1 does have a nontrivial topology, $g(x)$ is divided into topological sectors. They are labelled by an integer number $n \in \mathbb{Z}$ and defined by

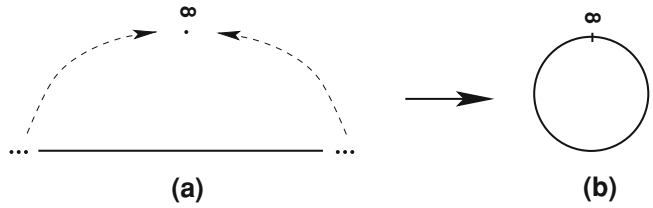

Fig. 4.3 Compactification of the real line (a) into the circumference S^1 (b) by adding the point at infinity

$$\alpha(2\pi) = \alpha(0) + 2\pi n. \tag{4.93}$$

Geometrically, n is the number of times that the spatial S^1 winds around the gauge group U(1). This winding number can be written equivalently as

$$\oint_{S^1} g(x)^{-1} dg(x) = 2\pi n, \tag{4.94}$$

where the integral is along the spatial S^1.

Something similar happens in the case of a SU(2) gauge theory in $3+1$ dimensions.[5] Demanding $g(\mathbf{x}) \in SU(2)$ to be regular at spatial infinity, $|\mathbf{x}| \to \infty$, we can compactify $\mathbb{R}^3$ into a three-dimensional sphere S^3, exactly as we did in $1+1$ dimensions. The matrices $g(\mathbf{x})$ can be parameterized as

$$g(\mathbf{x}) = a^0(\mathbf{x})\mathbf{1} + i\mathbf{a}(\mathbf{x}) \cdot \sigma, \tag{4.95}$$

with σ_i the Pauli matrices. The conditions $g(\mathbf{x})^\dagger g(\mathbf{x}) = \mathbf{1}$, $\det g = 1$ imply $(a^0)^2 + \mathbf{a}^2 = 1$. Hence SU(2) is a three-dimensional sphere and $g(\mathbf{x})$ defines a map from the spatial S^3 to the S^3 defined by the gauge group

$$g : S^3 \longrightarrow S^3. \tag{4.96}$$

As in the $(1+1)$-dimensional case, the gauge transformations $g(\mathbf{x})$ are divided into topological sectors labelled this time by the integer winding number

$$n = \frac{1}{24\pi^2} \int_{S^3} d^3x \varepsilon_{ijk} \mathrm{Tr}\Big[\left(g^{-1}\partial_i g\right) \left(g^{-1}\partial_i g\right) \left(g^{-1}\partial_i g\right) \Big]. \tag{4.97}$$

In U(1) and SU(2), gauge transformations split into different sectors labelled by an integer. Since this winding number is a continuous function of the gauge transformation $g(\mathbf{x})$, two transformations with different values of n cannot be smoothly

[5] Although we present for simplicity only the case of SU(2), similar arguments apply to any simple group.

deformed into each other. The sector with $n = 0$ corresponds to those transformations that can be continuously connected with the identity.

Now we will be a bit more formal. Let us consider a gauge theory in $3+1$ dimensions with gauge group G and let us denote by $\mathscr{G}$ the set of all gauge transformations $g(\mathbf{x})$ approaching the identity at spatial infinity, $\mathscr{G} = \{g : S^3 \to G\}$. At the same time we introduce the subgroup $\mathscr{G}_0 \subset \mathscr{G}$ containing all transformations in $\mathscr{G}$ that can be smoothly deformed into the identity. Our theory will have topological sectors if

$$\mathscr{G}/\mathscr{G}_0 \neq \mathbf{1}. \tag{4.98}$$

The existence of these topological sectors in $(3+1)$-dimensional gauge theories is controlled by a mathematical object called the third homotopy group of the gauge group that is denoted by $\pi_3(G)$. For example, it can be proved [14] that $\pi_3(S^1) = \mathbf{1}$, i.e., the third homotopy group of U(1) is trivial and therefore no topological sectors appear in $(3+1)$-dimensional electrodynamics. On the other hand, $\pi_3(S^3) = \mathbb{Z}$ and as a consequence the topological sectors of the SU(2) gauge theory are labelled by a single integer, the winding number[6] (4.97).

In the case of electromagnetism, we have seen that Gauss' law annihilates physical states. For a nonabelian theory the analysis is similar and leads to the condition

$$\mathscr{U}(g_0)|\text{phys}\rangle \equiv \exp\left[i \int d^3x \chi^A(\mathbf{x})(\mathbf{D}\cdot\mathbf{E})^A\right]|\text{phys}\rangle = |\text{phys}\rangle, \tag{4.99}$$

where $g_0(\mathbf{x}) = e^{i\chi^A(\mathbf{x})T^A}$ is in the connected component of the identity $\mathscr{G}_0$, and D_i is the covariant derivative in the adjoint representation. The important point here is that only the elements of $\mathscr{G}_0$ can be written as exponentials of the infinitesimal generators. Since these generators annihilate the physical states, this implies $\mathscr{U}(g_0)|\text{phys}\rangle = |\text{phys}\rangle$ only when $g_0(\mathbf{x}) \in \mathscr{G}_0$.

What happens with gauge transformations in the other topological sectors? If $g \in \mathscr{G}/\mathscr{G}_0$ there is still a unitary operator $\mathscr{U}(g)$ implementing gauge transformations on the Hilbert space of the theory. However since g is not in the connected component of the identity, it cannot be written as the exponential of Gauss' law. Still, gauge invariance is preserved if $\mathscr{U}(g)$ only changes the overall global phase of the physical states. For example, if $g_1(\mathbf{x})$ is a gauge transformation with winding number $n = 1$

$$\mathscr{U}(g_1)|\text{phys}\rangle = e^{i\theta}|\text{phys}\rangle. \tag{4.100}$$

It is easy to convince oneself that all transformations with winding number $n = 1$ have the same value of θ modulo 2π. This can be shown by noticing that if $g(\mathbf{x})$ has $n = 1$ then $g(\mathbf{x})^{-1}$ has opposite winding number $n = -1$. It is a simple exercise to prove that the winding number is additive: given two transformations g_1, g_2 with winding number 1, $g_1^{-1}g_2$ has winding number $n = 0$. This leads to

---

[6] The existence of topological sectors in $(1+1)$-dimensional electrodynamics is a consequence of the nontrivial character of the *first* homotopy group of S^1, namely $\pi_1(S^1) = \mathbb{Z}$.

$$\begin{aligned} |\text{phys}\rangle &= \mathscr{U}(g_1^{-1}g_2)|\text{phys}\rangle = \mathscr{U}(g_1)^\dagger \mathscr{U}(g_2)|\text{phys}\rangle \\ &= e^{i(\theta_2-\theta_1)}|\text{phys}\rangle, \end{aligned} \tag{4.101}$$

thus $\theta_1 = \theta_2 \bmod 2\pi$. Therefore a gauge transformation $g_n(\mathbf{x})$ with winding number n acts on physical states according to

$$\mathscr{U}(g_n)|\text{phys}\rangle = e^{in\theta}|\text{phys}\rangle, \quad n \in \mathbb{Z}. \tag{4.102}$$

To find a physical interpretation of this result, we look for a similar situation in a more familiar setup, for example the quantum states of electrons in the periodic potential produced by the ion lattice in a solid. For simplicity, we discuss the one-dimensional case where the minima of the potential are separated by a distance a. When the barrier between consecutive degenerate vacua is high enough, we can neglect tunneling between different vacua and consider the ground states $|na\rangle$ of the potential near the minimum located at $x = na$ ($n \in \mathbb{Z}$) as possible vacua of the theory. These ground states are not invariant under lattice translations

$$e^{ia\hat{P}}|na\rangle = |(n+1)a\rangle. \tag{4.103}$$

It is nevertheless possible to define a new vacuum state

$$|k\rangle = \sum_{n\in\mathbb{Z}} e^{-ikna}|na\rangle, \tag{4.104}$$

which under $e^{ia\hat{P}}$ transforms just by a global phase

$$e^{ia\hat{P}}|k\rangle = \sum_{n\in\mathbb{Z}} e^{-ikna}|(n+1)a\rangle = e^{ika}|k\rangle. \tag{4.105}$$

This ground state is labelled by the momentum k and corresponds to the Bloch wave function.

This is very similar to what we found for nonabelian gauge theories. The vacuum state labelled by θ plays a role similar to the Bloch wave function for the periodic potential with the identification of θ with the momentum k. To make this analogy more precise, let us write the Hamiltonian for nonabelian gauge theories

$$H = \frac{1}{2}\int d^3x \Big(\pi^A \cdot \pi^A + \mathbf{B}^A \cdot \mathbf{B}^A\Big) = \frac{1}{2}\int d^3x \Big(\mathbf{E}^A \cdot \mathbf{E}^A + \mathbf{B}^A \cdot \mathbf{B}^A\Big), \tag{4.106}$$

where we have used the expression of the canonical momenta π^{iA}. Moreover, we work in the gauge $A_0 = 0$ and assume that the Gauss law constraint is satisfied. The first term in the integral is the kinetic energy, $T = \frac{1}{2}\pi^A \cdot \pi^A$, and the second the potential energy, $V = \frac{1}{2}\mathbf{B}^A \cdot \mathbf{B}^A$. Since $V \geqslant 0$, the vacua of the theory can be identified with those gauge field configurations for which $V = 0$, modulo gauge transformations. This happens when $\mathbf{A}(0, \mathbf{x})$ is a pure gauge. Since gauge transformations are classified by their winding number, there are infinitely many ground

states. Indeed, taking a representative gauge transformation $g_n(\mathbf{x})$ in the sector with winding number n, these vacua will be associated with the gauge potentials

$$\mathbf{A}(0,\mathbf{x}) = -\frac{1}{ig_{\rm YM}} g_n(\mathbf{x})\nabla g_n(\mathbf{x})^{-1}, \tag{4.107}$$

modulo topologically trivial gauge transformations. Thus the theory is characterized by an infinite number of ground states $|n\rangle$ labelled by the winding number.

These vacua are not gauge invariant. A gauge transformation with $n = 1$ changes the winding number of the vacuum by one unit

$$\mathscr{U}(g_1)|n\rangle = |n+1\rangle. \tag{4.108}$$

As with Bloch waves, a gauge invariant vacuum can be defined

$$|\theta\rangle = \sum_{n\in\mathbb{Z}} e^{-in\theta}|n\rangle \quad \text{with } \theta \in \mathbb{R}, \tag{4.109}$$

transforming under a gauge transformation by a global phase

$$\mathscr{U}(g_1)|\theta\rangle = e^{i\theta}|\theta\rangle. \tag{4.110}$$

We have concluded that the nontrivial topology of the gauge group has very important physical consequences for the quantum theory. In particular, it implies an ambiguity in the definition of the vacuum. This can also be seen in a Lagrangian analysis. In constructing the Lagrangian for the nonabelian version of the Maxwell theory we only considered the term $F^A_{\mu\nu}F^{\mu\nu A}$. However this is not the only Lorentz and gauge invariant term containing just two derivatives. We can write the more general action

$$S = -\frac{1}{2}\int d^4x \mathrm{Tr}\Big(F_{\mu\nu}F^{\mu\nu}\Big) - \frac{\theta g^2_{\rm YM}}{16\pi^2}\int d^4x \mathrm{Tr}\Big(F_{\mu\nu}\widetilde{F}^{\mu\nu}\Big), \tag{4.111}$$

where $\widetilde{F}_{\mu\nu}$ is the dual of the field strength defined by

$$\widetilde{F}_{\mu\nu} = \frac{1}{2}\varepsilon_{\mu\nu\sigma\lambda}F^{\sigma\lambda}. \tag{4.112}$$

The constant θ is dimensionless in natural units. The extra term in (4.111), proportional to $\mathbf{E}^A \cdot \mathbf{B}^A$, is a total derivative and does not change the equations of motion or the quantum perturbation theory.

This, however, does not mean that the addition of the second piece in the action (4.111) does not change the physics. It can be directly checked that

$$\frac{g^2_{\rm YM}}{16\pi^2}\mathrm{Tr}\Big(F_{\mu\nu}\widetilde{F}^{\mu\nu}\Big) = \partial_\mu \mathscr{J}^\mu \tag{4.113}$$

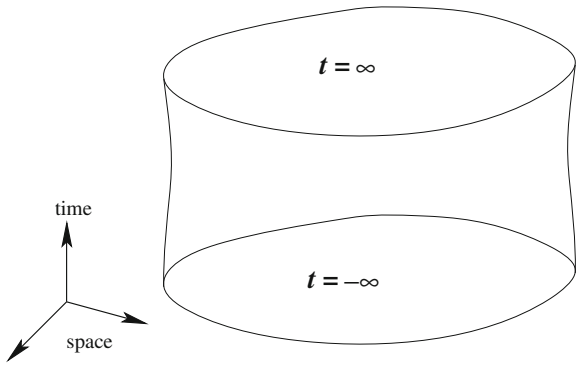

Fig. 4.4 Region of integration to compute the contribution of the θ-term to the gauge theory action. The gauge field $\mathbf{A}(t, \mathbf{x})$ tends to pure gauge configurations both at early and late times $t \to \pm\infty$ and at spatial infinity $|\mathbf{x}| \to \infty$ (the side of the cylinder)

with

$$\mathscr{J}^{\mu} = \frac{g_{\mathrm{YM}}^2}{16\pi^2}\varepsilon^{\mu\nu\sigma\lambda}\mathrm{Tr}\left(F_{\nu\sigma}A_{\lambda} - \frac{2ig_{\mathrm{YM}}}{3}A_{\nu}A_{\sigma}A_{\lambda}\right). \tag{4.114}$$

Thus, the contribution of the second term in (4.111) can be computed using Gauss' theorem. To ensure the convergence of the integral we assume that $\mathbf{A}(t, \mathbf{x})$ approaches a pure gauge configuration both at spatial infinity and at late and early times $t \to \pm\infty$. To be more precise we assume that

$$\mathbf{A}(t \to \infty, \mathbf{x}) \longrightarrow -\frac{1}{ig_{\mathrm{YM}}}g(\mathbf{x})\nabla g(\mathbf{x})^{-1}, \tag{4.115}$$

while $\mathbf{A}(t, \mathbf{x})$ is taken to vanish at $t \to -\infty$. This last condition implies no loss of generality, since it can always be achieved by an appropriate gauge transformation.

In the gauge $A^0 = 0$ it is easy to check that $\mathscr{J}^i \to 0$ at spatial infinity. Hence, the integral of the topological term in the action only receives contributions from the boundaries at $t \to \pm\infty$ (see Fig. 4.4). This yields

$$\begin{aligned}&\frac{g_{\mathrm{YM}}^2}{16\pi^2}\int d^4x\mathrm{Tr}\Big(F_{\mu\nu}\widetilde{F}^{\mu\nu}\Big)\\ &\quad = \frac{1}{24\pi^2}\int d^3x\varepsilon_{ijk}\mathrm{Tr}\Big[\left(g\partial_i g^{-1}\right)\left(g\partial_j g^{-1}\right)\left(g\partial_k g^{-1}\right)\Big].\end{aligned} \tag{4.116}$$

Comparing this expression with Eq. (4.97) we obtain

$$\frac{\theta g_{\mathrm{YM}}^2}{16\pi^2}\int d^4x\mathrm{Tr}\Big(F_{\mu\nu}\widetilde{F}^{\mu\nu}\Big) = \theta n[g] \equiv \theta n[\mathbf{A}^A]. \tag{4.117}$$

This term distinguishes gauge fields according to topological sectors: two gauge fields are in the same sector if the corresponding gauge transformations g giving their asymptotic behavior at late times have the same winding numbers. This is very important in the quantum theory, because it means that one must sum over all

topological sectors when performing the functional integration, each one weighted by a θ-dependent phase. Symbolically,

$$\mathscr{D}\mathbf{A}_\mu^A = \sum_{n\in\mathbb{Z}} e^{-i\theta n}[\mathscr{D}\mathbf{A}^A]_n \tag{4.118}$$

where $[\mathscr{D}\mathbf{A}^A]_n$ indicates that the integration is performed over gauge fields in the topological class $n[\mathbf{A}^A] = n$. We have reobtained, in the Lagrangian language, the vacuum degeneracy found above in the canonical formalism.

The presence of the θ-term in the gauge theory action has several important physical consequences. One of them is that it violates both parity P and the combination of charge conjugation and parity CP. This will be further studied in Chap. 11.

Subtleties and Technicalities

Before closing this section we would like to mention a number of subtleties in the arguments presented concerning the structure of the gauge group. We have used the fact that $\pi_3(S^3) = \mathbb{Z}$ to characterize the number of components of the gauge group SU(2). In the argument it is crucial that the spatial topology is S^3. If this is not the case, the treatment should be refined. For instance, in *noncompact* three-dimensional Euclidean space the type of gauge transformations described by the elements of $\mathscr{G} = \{g : S^3 \to G\}$ are those approaching the identity at infinity fast enough and in a way that does not depend on angles. An equivalent way to describe them is to consider those gauge transformations that, outside a compact set surrounding the origin of coordinates, go to the identity very fast. The classes generated by these gauge transformations can be characterized by an integer number, but this may not exhaust the topological characterization of all possible nontrivial transformations.

Working on a three-dimensional box with periodic boundary conditions results in a spatial topology that is that of a three-dimensional torus T^3, and the topological structure of the mappings $\mathscr{G} = \{g : T^3 \to G\}$ is in general richer than the one described by the single winding number appearing in S^3. In this case we also have other gauge transformations not included in $\mathscr{G}_0$ and associated to the fact that the space is not simply connected. These additional transformations are physically relevant, and play an important role in 't Hooft's theory of confinement in nonabelian gauge theories. From this point of view, the topology of the space of gauge transformation often depends on the type of physical questions asked. Hence, apart from the θ angle, there may be other angles or quantum number characterizing the physical states (or the vacuum) of the theory (see for instance [15] and references therein).

To summarize, the implementation of the Gauss' law constrain and the set of physical parameter that characterize it depends on the physics and topology of the problem at hand. In the case of the θ-angle, we can introduce it by either refining our arguments on the structure of the space of nontrivial gauge transformations, or simply by arguing that the second term in (4.111) should be included because it is

local, gauge and Lorentz invariant and with the same canonical dimension as the kinetic term. Extracting the dependence of physical quantities on the vacuum angle is in general a highly non-trivial problem that is not fully understood.

4.8 Instantons in Gauge Theories

The existence of multiple vacua in nonabelian gauge theories makes natural to study the possibility of tunneling between them. As explained in Sect. 2.5, in semiclassical tunneling this is described by solutions to the Euclidean field equations with finite action. For nonabelian gauge theories, the analytical continuation to imaginary times $t \to -it$, $A_0 \to iA_0$, leads to the Euclidean action

$$S_E[A_\mu] = \frac{1}{2}\int d^4x \mathrm{Tr}\Big(F_{\mu\nu}F^{\mu\nu}\Big), \tag{4.119}$$

where the indices now are lowered and raised using $\delta_{\mu\nu}$. Since we are interested in solutions to the Euclidean field equations with finite action, the gauge field $A_\mu(t, \mathbf{x})$ has to approach a pure gauge configuration both at spatial infinity $|\mathbf{x}| \to \infty$ as well as at "early" and "late" Euclidean times, $t \to \pm\infty$.

In the semiclassical evaluation of the path integral, the contribution of each saddle point comes weighted by the exponential factor $\exp\{-S_E[A_\mu]\}$, so the leading contribution is the one with the lowest value of the Euclidean action. To identify the dominant field configurations we use the following inequality valid in Euclidean space

$$0 \leqslant \mathrm{Tr}\Big[\left(F_{\mu\nu} \mp \widetilde{F}_{\mu\nu}\right)\left(F^{\mu\nu} \mp \widetilde{F}^{\mu\nu}\right)\Big] = 2\mathrm{Tr}\left(F_{\mu\nu}F^{\mu\nu}\right) \mp 2\mathrm{Tr}\left(F_{\mu\nu}\widetilde{F}^{\mu\nu}\right). \tag{4.120}$$

The combination of the inequalities for the two signs leads to the bound

$$S_E[A_\mu] \geqslant \frac{1}{2}\left|\int d^4x \mathrm{Tr}\Big(F_{\mu\nu}\widetilde{F}^{\mu\nu}\Big)\right|. \tag{4.121}$$

The right-hand side of this expression we already encountered in its Minkowskian version in Eq. (4.113). In the present setup, it is defined in four-dimensional Euclidean space and, being a total derivative, it can be written as an integral over the three-dimensional sphere at infinity, $|x| \equiv \sqrt{x^\mu x_\mu} \to \infty$. Notice that this term is independent of the metric and therefore it does not change when continued to Euclidean space, unlike the Yang–Mills action that picks up an imaginary unit in front, $S[A_\mu] \to iS_E[A_\mu]$.

Since the gauge field approaches a pure gauge when $|x| \to \infty$

$$A_\mu(x) \overset{|x|\to\infty}{\longrightarrow} -\frac{1}{ig_{\mathrm{YM}}} g\partial_\mu g^{-1}, \tag{4.122}$$

the integral in (4.121) is given in terms of the instanton charge Q defined by

$$Q = \frac{1}{24\pi^2} \int\limits_{S^3_\infty} d\sigma_\mu \varepsilon^{\mu\nu\sigma\lambda} \mathrm{Tr}\Big(g\partial_\nu g^{-1} g\partial_\sigma g^{-1} g\partial_\lambda g^{-1}\Big), \tag{4.123}$$

where the integration is performed over the three-dimensional sphere at infinity. In terms of it, the action bound reads

$$\frac{1}{2}\int d^4x \mathrm{Tr}\Big(F_{\mu\nu}F^{\mu\nu}\Big) \geqslant \frac{8\pi^2}{g^2_{\mathrm{YM}}}|Q|. \tag{4.124}$$

A look at (4.120) shows that the previous inequality is saturated if an only if the Euclidean gauge field is either selfdual or anti-selfdual, namely if its field strength tensor satisfies

$$F_{\mu\nu} = \pm\widetilde{F}_{\mu\nu}. \tag{4.125}$$

Euclidean solutions satisfying these conditions are called respectively *instantons* (+sign) and *anti-instantons* (− sign). These are the configurations dominating the Euclidean amplitudes in the semiclassical limit within each topological sector. It is important to notice that any (anti-)selfdual gauge field is automatically a solution of the Euclidean field equations: the (anti)-selfduality condition reduces the equations of motion $D_\mu F^{\mu\nu} = 0$ to the Bianchi identities,

$$\varepsilon^{\mu\nu\sigma\lambda} D_\nu F_{\sigma\lambda} = 0, \tag{4.126}$$

that are identically satisfied by any field strength tensor (the reader is invited to prove it as an exercise). Finally, it is easy to see that instantons and anti-instantons have positive and negative topological charges respectively.

We study the solutions to the selfduality equation with instanton charge $Q = 1$. To keep things simple, we consider the case of a SU(2) gauge theory. In fact this does not mean a big loss of generality: the instanton solutions for other gauge groups can be constructed in terms of their SU(2) factors. The calculation of the instanton solution is rather long and its details can be found, for example, in [16]. The result for the gauge potential in a generic gauge is

$$A^a_\mu(x) = \frac{2}{g_{\mathrm{YM}}} \frac{\eta^a_{\mu\nu}\left(x^\nu - x^\nu_0\right)}{(x-x_0)^2 + \rho^2}, \tag{4.127}$$

where $a = 1, 2, 3$ is the SU(2) index and $\eta^a_{\mu\nu}$ are the 't Hooft symbols introduced in Chap. 3 (see page 35). The field strength

$$F^a_{\mu\nu}(x) = \frac{4}{g_{\mathrm{YM}}} \frac{\eta^a_{\mu\nu}\rho^2}{[(x-x_0)^2 + \rho^2]^2} \tag{4.128}$$

is selfdual and the Euclidean action saturates the bound (4.121) with unit instanton charge

$$S_E[A_\mu] = \frac{8\pi^2}{g^2_{\rm YM}}. \tag{4.129}$$

The solution (4.127) depends on a number of arbitrary parameters: the coordinates of its center x_0^μ and the size ρ. These are part of the so-called *collective coordinates* of the instanton. They are generated by applying to a given solution the invariances of the Euclidean action, in our case translations and dilatations

$$A^a_\mu(x) \longrightarrow A^a_\mu(x+\xi), \quad A^a_\mu(x) \longrightarrow \lambda A^a_\mu(\lambda x) \tag{4.130}$$

respectively. In addition to (x_0^μ, ρ) the general instanton solution have three additional collective coordinates associated with its orientation in SU(2) space, making a total of eight collective coordinates. This number might seem smaller than expected, since the Euclidean gauge action is invariant under the full conformal group that includes the Euclidean group (rotations and translations), dilatations and special conformal transformations. The reason why rotations and special conformal transformations do not generate collective coordinates is that the two can be combined with translations, dilatations and SU(2) rotations to leave the instanton solutions invariant up to a gauge transformation. As a result only 8 of the total 18 generators [15 of the Euclidean group plus 3 of SU(2)] give rise to collective coordinates.

Finite action classical solutions to the Euclidean field equations of motion represent tunneling between different vacua of the theory (see Sect. 2.5). Next we want to show how the instanton solutions (4.127) describe indeed the semiclassical tunneling between gauge field configurations with topological numbers differing by one unit (the topological charge of the instanton). In order to make the connection with the analysis of the gauge theory vacua presented in the previous section, we have to change from the generic gauge used in writing (4.127) to the gauge $A^a_0 = 0$. This is accomplished by a gauge transformation $U(t, \mathbf{x})$ satisfying

$$U(t,\mathbf{x})^{-1}\partial_0 U(t,\mathbf{x}) = -ig_{\rm YM}A_0(t,\mathbf{x}), \tag{4.131}$$

such that in this new gauge

$$\begin{aligned} A'_0(t,\mathbf{x}) &= 0 \\ \mathbf{A}'(t,\mathbf{x}) &= -\frac{1}{ig_{\rm YM}} U\nabla U^{-1} + U\mathbf{A}(t,\mathbf{x})U^{-1}. \end{aligned} \tag{4.132}$$

The general solution to the differential equation (4.131) depends on an arbitrary function of $\mathbf{x}$. This is fixed by demanding that the spatial components of the instanton in the new gauge, $\mathbf{A}'(t,\mathbf{x})$, tend to zero at early Euclidean times, $t \to -\infty$. With this condition, the gauge transformation $U(t,\mathbf{x}) = \exp(i\chi^a T^a)$ is determined to be

$$\chi^a(t,\mathbf{x}) = \frac{2\left(x^a - x_0^a\right)}{\sqrt{(\mathbf{x}-\mathbf{x}_0)^2+\rho^2}}\left[\frac{\boldsymbol{\pi}}{2} + \arctan\left(\frac{t-t_0}{\sqrt{(\mathbf{x}-\mathbf{x}_0)^2+\rho^2}}\right)\right]. \tag{4.133}$$

Since the spatial components of the instanton solution (4.127) vanish as $t \to \pm\infty$, the Euclidean gauge field (4.132) approaches a pure gauge configuration both at early and late Euclidean times. Therefore it can be interpreted as interpolating between two vacua of the SU(2) gauge theory. As $t \to -\infty$ the gauge field is identically zero, whereas when $t \to \infty$ the instanton solution approach the vacuum configuration

$$A_i'(t, \mathbf{x}) \longrightarrow -\frac{1}{i g_{\rm YM}} g(\mathbf{x}) \partial_i g(\mathbf{x})^{-1} \tag{4.134}$$

with

$$g(\mathbf{x}) \equiv \lim_{t \to +\infty} U(t, \mathbf{x}) = \exp\left[\frac{2\pi i (x^a - x_0^a)}{\sqrt{(\mathbf{x} - \mathbf{x}_0)^2 + \rho^2}} T^a \right]. \tag{4.135}$$

This, unlike the $\mathbf{A}(t, \mathbf{x}) = \mathbf{0}$ vacuum in the asymptotic Euclidean "past", is a gauge configuration with nonvanishing topological number, namely [cf. equation (4.116)]

$$n[\mathbf{A}'] = \frac{1}{24\pi^2} \int d^3x \varepsilon_{ijk} \mathrm{Tr}\Big[\left(g \partial_i g^{-1}\right) \left(g \partial_i g^{-1}\right) \left(g \partial_i g^{-1}\right) \Big] = 1. \tag{4.136}$$

The final conclusion of our analysis is that the instanton solution (4.127) describes the tunneling from a gauge theory vacuum with vanishing winding number to a nontrivial vacuum with winding number equal to one, the difference being equal to the topological charge of the instanton. A similar analysis can be repeated for anti-instanton solutions, obtained from (4.127) by replacing the 't Hooft symbols by their duals $\overline{\eta}^a_{\mu\nu}$ [see Eq. (3.11)]. They have instanton charge $Q = -1$ and interpolate between gauge theory vacua with winding numbers that differ by this amount.

(Anti-)Instanton contributions to physical quantities are weighted by

$$\exp\left(-\frac{8\pi^2}{g_{\rm YM}^2} |Q| \right). \tag{4.137}$$

This factor is nonanalytic around $g_{\rm YM} = 0$, showing that the effect of tunneling between different gauge theory vacua is truly nonperturbative. We see that at weak coupling instanton effects are exponentially suppressed and therefore overshadowed by any perturbative contribution to the same process, that necessarily scales with a positive power of $g_{\rm YM}$. This is the reason why instantons are mostly relevant in physical situations where perturbative terms are known to be zero.

References

1. Aharonov, Y., Bohm, D.: Significance of the electromagnetic potentials in the quantum theory. Phys. Rev. **115**, 485 (1955)
2. Ehrenberg, W., Siday, R.E.: The refractive index in electron optics and the principles of dynamics. Proc. Phys. Soc. B **62**, 8 (1949)