Internal symmetry and gauge interactions

16

- An introduction to the gauge theory is presented in this chapter. We start with a review of gauge invariance in electromagnetism. That is followed by a discussion of gauge symmetry in quantum mechanics, showing that the gauge transformation must involve a spacetime-dependent change of the phase of a charged particle's wavefunction (i.e. field). This can be viewed as a transformation in the (internal) charge space by changing the particle field's label.
- If we reverse the above procedure, instead of going from the change in EM potential then to the wavefunction transformation, we start first with the phase transformation in quantum mechanics. This initial step may be understood as changing a spacetime-independent symmetry of the quantum mechanics equation to a local symmetry—a procedure called "gauging a symmetry". EM potentials are viewed then as compensating factors needed to implement such a local symmetry—the presence of potentials (with appropriate transformation properties as gauge fields) is required so that the physics equations are covariant under such local symmetry transformations.
- We demonstrate how Maxwell's electrodynamics can be "derived" from the requirement of a local $U(1)$ symmetry in the internal charge space. In this way we understand the essence of Maxwell's theory as special relativity and gauge invariance. Much like the elevating by Einstein of the equality of gravitational and inertial masses to the equivalence principle of gravitation and inertia, we call the approach of finding dynamics by promoting a global symmetry to a local symmetry, the gauge principle. Using the gauge principle, we can then generalize this approach to electromagnetism to the investigation of other fundamental interactions.
- In 1919 Hermann Weyl first attempted to derive electromagnetism from a local scale invariance. He was inspired by the success of general relativity, Einstein's new theory of gravity formulated as a local spacetime symmetry. Weyl was ultimately successful in this endeavor; this came after the advent of modern quantum mechanics (QM) in 1926 when Vladimir Fock discovered that QM wave equations with electromagnetic coupling are invariant under local phase transformations. It was pointed out that Weyl's scale change in spacetime should

be understood as a spacetime-dependent $U(1)$ phase change in the charge space. However, Weyl's original terminology of gauge (i.e. scale) transformation has been retained in common usage.

- A $U(1)$ phase change, being commutative, is an abelian transformation. This was extended by C. N. Yang and R. L. Mills to the case of noncommuntative symmetries. The resultant equations are nonlinear—the gauge fields themselves are charged (unlike the abelian case of the electromagnetic field being electrically neutral, but like gravity where the gravitational field is itself a source of gravity). This richness is one of the key ingredients that allowed nonabelian gauge theories (also called Yang–Mills theories) to be the framework for modern particle theory.
- We describe briefly the steps of going from quantum electrodynamics to the formulation of the new theory of fundamental strong interaction, quantum chromodynamics. The gauge theory of electroweak interactions has a more complicated structure because its local symmetry must be spontaneously broken (via the Higgs mechanism) to account for the short-range nature of weak interactions. In sum, the successful formulation of the Standard Model shows that fundamental particle interactions are all gauge interactions. This is a mighty generalization of Einstein's symmetry principle, from spacetime to internal charge spaces. It allowed us first to have a deeper understanding of electromagnetism, which was crucial to our finding new theories for the strong and weak interactions.
- Abelian gauge symmetry is discussed in detail (Sections 16.1–16.3)—up to the point of seeing how Maxwell's equations follow from gauge symmetry. Quantum field theories of QED, QCD, and the Standard Model are described qualitatively in the subsequent sections.

16.1 Einstein and the symmetry principle

One of Einstein's greatest legacies in physics has been his bringing about of our realization of the importance of symmetry in physics. His theory of relativity was built on the foundation of invariance principles. Before Einstein, symmetries were generally regarded as mathematical curiosities of great value to crystallographers, but hardly worthy to be included among the fundamental laws of physics. We now understand that a symmetry principle is not only an organizational device, but also a method to discover new dynamics. Einstein's relativity theories based on coordinate symmetries have given us a deeper appreciation of the structure of physics. His formulation of the symmetry among inertial frames of reference showed us the true meaning of the Lorentz transformation; this allowed us to deduce all the (special) relativistic effects in a compact way and to discover new equations for other branches of physics (relativistic mechanics, etc.) so they could be compatible with the symmetry principle of relativity. The extension of this principle from a special

class of coordinates to all reference frames, his creation of general relativity, showed us the way of using spacetime-dependent (local) symmetry to generate dynamics—in the case of general relativity (GR), its gravitational interaction.

Ever since Einstein, a symmetry principle has been an essential guiding light in our effort to make new discoveries in theoretical physics. The topic of symmetry in physics is a rich one, especially in quantum physics. In this chapter we shall concentrate[1] on the gauge symmetry. It is one of the most important principles in fundamental physics. Local symmetry in some "internal" (or "charge") space has been the key to our discoveries of new basic physics, leading to the formulation of the Standard Model of particle physics. Starting from the work of Hermann Weyl (inspired by Einstein's GR discovery), we gradually learnt that electromagnetism could be understood as arising from a spacetime-dependent local symmetry (gauge symmetry) in the charge space. Namely, we discovered another profound lesson contained in Maxwell's equations: Besides teaching us the proper relation among inertial frames of reference (as given by the Lorentz transformation), these equations have such a structure as showing that electromagnetism is a gauge interaction. This simple $U(1)$ local symmetry associated with electromagnetism was later generalized to the noncommuting (Yang–Mills) gauge symmetry, which is a key element in the foundations of modern particle physics.

[1] We omit other important topics such as the relation between symmetry and conservation laws and degeneracy in the particle spectrum, and discrete symmetries (parity and time reversal invariance, etc.) and their violation, etc.

16.2 Gauge invariance in classical electromagnetism

We present in this chapter a pedagogical introduction to gauge theory. Since most students have their first exposure to gauge invariance in classical electrodynamics, this is where we will start—with a review of electromagnetic (EM) potentials and their gauge transformation. We then discuss gauge transformation in quantum mechanics. Because a quantum mechanical description is through a Hamiltonian (or through other energy quantities such as a Lagrangian), which can include a system's coupling to electromagnetism only through EM potentials, gauge symmetry plays an integral role in the QM description.

Classical electromagnetism[2] Any field theoretical description of the interaction between two particles involves a "two-step description". Call one the "source particle", giving rise to a field everywhere, which in turn acts locally on the "test particle". This two-step description can be represented schematically as follows:

$$\boxed{\text{Source particle}} \xrightarrow{\text{field eqns}} \boxed{\text{Field}} \xrightarrow{\text{eqns of motion}} \boxed{\text{Test particle}}$$

The "field equations" tell us how a source particle gives rise to the field everywhere. For the case of electromagnetism, they are Maxwell's equations. The "equations of motion" tell us the effects of the field on the motion of a test particle: how does the field cause the particle to accelerate. For the case of electromagnetism, they form the Lorentz force law.

[2] Here we repeat the essential elements of Maxwell's equations, first in familiar 3D vector notation (as already discussed in Sections A.1 and 3.1 as well as in Chapters 9 and 10) and then in the 4D spacetime formalism (in Section 16.4 as we have already done in Section 11.2.3).

- Equations of motion (Lorentz force law):

$$\mathbf{F} = e\left(\mathbf{E} + \frac{1}{c}\mathbf{v}\times\mathbf{B}\right) \tag{16.1}$$

  we note that this equation has a "double duty": It gives the definition of the electric and magnetic fields as well as acting as the equation of motion for a test charge placed in the electromagnetic field.
- Field equations (Maxwell equations):
  - Inhomogeneous Maxwell equations:

$$\nabla\cdot\mathbf{E} = \rho \qquad \text{Gauss's law} \tag{16.2}$$

$$\nabla\times\mathbf{B} - \frac{1}{c}\frac{\partial\mathbf{E}}{\partial t} = \frac{1}{c}\mathbf{j} \qquad \text{Ampere's law} \tag{16.3}$$

  - Homogeneous Maxwell equations:

$$\nabla\cdot\mathbf{B} = 0 \quad \text{Gauss's law for magnetism} \tag{16.4}$$

$$\nabla\times\mathbf{E} + \frac{1}{c}\frac{\partial\mathbf{B}}{\partial t} = 0 \quad \text{Faraday's law.} \tag{16.5}$$

16.2.1 Electromagnetic potentials and gauge transformation

It is easy to solve the homogeneous Maxwell equations (16.4) and (16.5) by noting that the divergence of any curl, as well as the curl of any gradient, must vanish:[3] Eq. (16.4) can be solved if the **B** field is the curl of a **vector potential A**:

$$\mathbf{B} = \nabla\times\mathbf{A}. \tag{16.6}$$

Substituting this into (16.5), the vanishing curl $\nabla\times\left(\mathbf{E} + \frac{1}{c}\frac{\partial\mathbf{A}}{\partial t}\right) = 0$ implies that the term in parentheses can be written as the gradient of a scalar potential Φ:

$$\mathbf{E} = -\nabla\Phi - \frac{1}{c}\frac{\partial\mathbf{A}}{\partial t}. \tag{16.7}$$

Thus we can replace $(\mathbf{E},\mathbf{B})$ fields by scalar and vector potentials $(\Phi,\mathbf{A})$ through the relations (16.6) and (16.7). Substituting these expressions into the inhomogeneous Maxwell equations of (16.2) and (16.3), we obtain the dynamics of the potentials once the source distribution $(\rho,\mathbf{j})$ is given. In other words, one can regard the homogeneous parts of Maxwell's equations as the "boundary conditions" telling us that fields can be expressed in terms of the potentials, and the true dynamics is contained in the inhomogeneous Maxwell equations.

[3] See the discussion leading up to Eq. (A.20) in Appendix A1.

Gauge invariance in classical electromagnetism

As outlined above we can simplify the description of the EM interactions by using four components of potentials $(\Phi,\mathbf{A})$ instead of six components of $(\mathbf{E},\mathbf{B})$. However this replacement of $(\mathbf{E},\mathbf{B})$ by $(\Phi,\mathbf{A})$ is not unique as the fields $(\mathbf{E},\mathbf{B})$,

hence also Maxwells equations, are invariant under the following change of potentials (called a **gauge transformation**):

$$\Phi \longrightarrow \Phi' = \Phi - \frac{1}{c}\frac{\partial \chi}{\partial t} \tag{16.8}$$

$$\mathbf{A} \longrightarrow \mathbf{A}' = \mathbf{A} + \nabla\chi \tag{16.9}$$

where $\chi = \chi(t, \mathbf{r})$ (called a **gauge function**) is an arbitrary scalar function of position and time. This invariance statement (**gauge symmetry**) can be easily verified:

$$\mathbf{B} = \nabla \times \mathbf{A} \;\longrightarrow\; \mathbf{B}' = \nabla \times \mathbf{A}' = \nabla \times \mathbf{A} + \nabla \times (\nabla\chi) = \nabla \times \mathbf{A} = \mathbf{B}$$

because the curl of a gradient must vanish. Similarly,

$$\mathbf{E} = -\nabla\Phi - \frac{1}{c}\frac{\partial \mathbf{A}}{\partial t} \;\longrightarrow\; \mathbf{E}' = -\nabla\Phi' - \frac{1}{c}\frac{\partial \mathbf{A}'}{\partial t}$$

$$= -\nabla\Phi + \frac{1}{c}\frac{\partial \nabla\chi}{\partial t} - \frac{1}{c}\frac{\partial \mathbf{A}}{\partial t} - \frac{1}{c}\frac{\partial \nabla\chi}{\partial t} = -\nabla\Phi - \frac{1}{c}\frac{\partial \mathbf{A}}{\partial t} = \mathbf{E}.$$

Gauge symmetry in classical electromagnetism does not seem to be very profound. It is merely the freedom to choose potentials (Coulomb gauge, radiation gauge, Lorentz gauge, etc.) to simply calculations. One can in principle avoid using potentials and stick with the $(\mathbf{E}, \mathbf{B})$ fields throughout, with no arbitrariness. On the other hand, the situation in quantum mechanics is different. As we shall see, the QM description of the electromagnetic interaction must necessarily involve potentials. Gauge symmetry must be taken into account in the QM description. As a consequence, it acquires a deeper significance.

16.2.2 Hamiltonian of a charged particle in an electromagnetic field

Before moving on to a discussion of gauge symmetry in QM, we undertake an exercise in classical EM of writing the Lorentz force law (16.1) in terms of the EM potentials. This form of the force law will be needed in the subsequent QM description of a charged particle in an EM field.

Lorentz force in terms of potentials

The ith component of the force law (16.1) may be written out in terms of the potentials via (16.7) and (16.6):

$$m\frac{dv_i}{dt} = -e\nabla_i\Phi - \frac{e}{c}\frac{\partial A_i}{\partial t} + \frac{e}{c}\epsilon_{ijk}v_j\epsilon_{klm}\nabla_l A_m. \tag{16.10}$$

For the last term, we shall use the identity $\epsilon_{ijk}\epsilon_{klm} = \delta_{il}\delta_{jm} - \delta_{im}\delta_{jl}$:

$$\frac{e}{c}(\mathbf{v} \times \mathbf{B})_i = \frac{e}{c}\epsilon_{ijk}v_j\epsilon_{klm}\nabla_l A_m = \frac{e}{c}\left[\mathbf{v} \cdot (\nabla_i \mathbf{A}) - (\mathbf{v} \cdot \nabla)\, A_i\right]. \tag{16.11}$$

The above expressions involve the differentiation of the vector potential $\mathbf{A}(\mathbf{r}, t)$ which depends on the time variables in two ways: through its explicit dependence on t, as well as implicitly through its dependence on position $\mathbf{r} = \mathbf{r}(t)$.

(The particle is moving!) Thus its full time derivative has a more complicated structure:

$$\frac{dA_i}{dt} = \frac{\partial A_i}{\partial t} + \nabla_j A_i \frac{dr_j}{dt} = \frac{\partial A_i}{\partial t} + (\mathbf{v}\cdot\boldsymbol{\nabla})A_i. \tag{16.12}$$

The two factors on the RHS of this equation are just those that appear on the RHS's of Eqs. (16.10) and (16.11); thus they can be combined into a $-\frac{e}{c}\frac{dA_i}{dt}$ term:

$$m\frac{dv_i}{dt} = -e\nabla_i\Phi - \frac{e}{c}\frac{dA_i}{dt} + \frac{e}{c}\mathbf{v}\cdot(\nabla_i\mathbf{A}) \tag{16.13}$$

which is then the expression of the Lorentz force in terms of $(\Phi, \mathbf{A})$ that we shall use in the discussion below, cf. Eq. (16.19).

Hamiltonian of a charged particle in an Electromagnetic field

Recall in QM that we do not use the concept of force directly in our description of particle interactions. Instead, the dynamics is governed by the Schrödinger equation,[4] which involves the Hamiltonian of the system. How do we introduce EM interactions in the Hamiltonian formalism? What is the Hamiltonian that represents the Lorentz force?

Recall that the Hamiltonian $H(\mathbf{r}, \mathbf{p})$ is a function of the position coordinate $\mathbf{r}$ and the canonical momentum $\mathbf{p}$, and the classical equations of motion are

Hamilton's equations

$$\frac{dr_i}{dt} = \frac{\partial H}{\partial p_i}, \qquad \frac{dp_i}{dt} = -\frac{\partial H}{\partial r_i}. \tag{16.14}$$

One can easily check that for $H = \frac{\mathbf{p}^2}{2m} + V(\mathbf{r})$, the first equation is just $\mathbf{p} = m\mathbf{v}$ (i.e. the canonical momentum is the same as the kinematic momentum) and the second, the usual $\mathbf{F} = m\mathbf{a}$. Now we claim that the Hamiltonian description of a charged particle (with mass m and charge e) in an EM field (represented by the potentials $\Phi, \mathbf{A}$), is given by

$$H = \frac{\left(\mathbf{p} - \frac{e}{c}\mathbf{A}\right)^2}{2m} + e\Phi. \tag{16.15}$$

To check this claim, let us work out the two Hamilton's equations:

1. The first equation in (16.14): Φ being a function of $\mathbf{r}$ only,

$$v_i \equiv \frac{dr_i}{dt} = \frac{\partial}{\partial p_i}\left[\frac{\left(\mathbf{p} - \frac{e}{c}\mathbf{A}\right)^2}{2m} + e\Phi\right] = \frac{1}{m}\left(p_i - \frac{e}{c}A_i\right).$$

Thus the canonical momentum ($\mathbf{p}$) differs from the kinematic momentum ($m\mathbf{v}$) by a factor related to the charge and vector potential,

$$\mathbf{p} = m\mathbf{v} + \frac{e}{c}\mathbf{A}, \tag{16.16}$$

and the first term in the Hamiltonian (16.15) remains the kinetic energy of $\frac{1}{2}mv^2$ (the second one, the electric potential energy).

[4] Here we start with elementary nonrelativistic quantum theory. However, all the results can be extended in a straightforward manner to relativistic Klein–Gordon and Dirac equations. For these cases, the simplest approach is (instead of the Hamiltonian) through the Lagrangian density as discussed in Section 16.4.2.

2. We expect the second equation in (16.14) to be the Lorentz force law. Let us verify this. Using the relation (16.16) between canonical and kinematic momenta, we have the LHS as

$$\frac{dp_i}{dt} = m\frac{dv_i}{dt} + \frac{e}{c}\frac{dA_i}{dt}. \tag{16.17}$$

The RHS may be written as

$$\begin{aligned} -\frac{\partial H}{\partial r_i} &= -\nabla_i \left[\frac{\left(p_j - \frac{e}{c}A_j\right)\left(p_j - \frac{e}{c}A_j\right)}{2m} + e\Phi \right] \\ &= \frac{\left(p_j - \frac{e}{c}A_j\right)}{m}\frac{e}{c}\nabla_i A_j - e\nabla_i \Phi \\ &= v_j \frac{e}{c}\nabla_i A_j - e\nabla_i \Phi \end{aligned} \tag{16.18}$$

where we have again used the relation (16.16). Equating (16.17) and (16.18) as in (16.14), we have

$$m\frac{dv_i}{dt} + \frac{e}{c}\frac{dA_i}{dt} = \frac{e}{c}\mathbf{v}\cdot(\nabla_i \mathbf{A}) - e\nabla_i \Phi, \tag{16.19}$$

which we recognize as the expression (16.13) of the Lorentz force in terms of the potentials, verifying our claim that Eq. (16.15) is the correct Hamiltonian for the Lorentz force law.

We have demonstrated that the Hamiltonian for a charged particle moving in an electromagnetic field can be compactly written in terms of the EM potentials $(\Phi, \mathbf{A})$ as (16.15). Perhaps the more important point is that there is no simple way to write the Hamiltonian, hence the QM description, in terms of the field strength $(\mathbf{E}, \mathbf{B})$ directly. As a consequence, we must study the invariance of the relevant QM equations under gauge transformation.

16.3 Gauge symmetry in quantum mechanics

Before launching into the study of gauge symmetry in QM, we shall take another look at the Hamiltonian (16.15) for a charged particle in the presence of an EM field. This prepares us for a new understanding of the theoretical significance of EM potentials $(\Phi, \mathbf{A})$.

16.3.1 The minimal substitution rule

Given the Hamiltonian (16.15), the Schrödinger equation $H\Psi = i\hbar\partial_t \Psi$, with the coordinate space representation of the canonical momentum $\mathbf{p} \doteq -i\hbar\nabla$, can be written out for a charged particle in an electromagnetic field as

$$\left[\frac{\left(\frac{\hbar}{i}\nabla - \frac{e}{c}\mathbf{A}\right)^2}{2m} + e\Phi \right] \Psi = i\hbar\frac{\partial \Psi}{\partial t}. \tag{16.20}$$

After a slight rearrangement of terms, it can be written as

$$-\frac{\hbar^2}{2m}\left(\nabla-\frac{ie}{\hbar c}\mathbf{A}\right)^2\Psi=i\hbar\left(\partial_t+\frac{ie}{\hbar}\Phi\right)\Psi. \tag{16.21}$$

When (16.21) is compared to the Schrödinger equation for **a free particle**,

$$-\frac{\hbar^2}{2m}\nabla^2\Psi=i\hbar\partial_t\Psi, \tag{16.22}$$

we see that the EM interaction (also referred to as the "EM coupling") can be introduced via the following replacement:

$$\nabla\longrightarrow\left(\nabla-\frac{ie}{\hbar c}\mathbf{A}\right)\equiv\mathbf{D}\quad\text{and}\quad\partial_t\longrightarrow\left(\partial_t+\frac{ie}{\hbar}\Phi\right)\equiv D_t. \tag{16.23}$$

This scheme of introducing the EM coupling is called **the minimal substitution rule**.[5] This procedure follows from the Hamiltonian (16.15) and is thus equivalent to the assumption of the Lorentz force law. While the procedure is simple, one is naturally curious for a deeper understanding: Is there a natural justification for this minimal coupling scheme? Namely, why does the EM coupling have the structure that it does?

[5] This is "minimal", because of the absence of other possible, but more complicated, couplings, e.g. those involving the spin operator and magnetic field $\boldsymbol{\sigma}\cdot\mathbf{B}$, etc.

Incidentally, the combinations $(D_t,\mathbf{D})$ of ordinary derivatives with EM potentials as defined in (16.23) are called **covariant derivatives**. As we shall discuss below they have the sane geometrical and physical significance as the covariant derivatives we encountered in our study of general relativity.

16.3.2 The gauge transformation of wavefunctions

Since QM must necessarily involve the EM potentials, one wonders how gauge invariance is implemented here. A direct inspection of the effects of the gauge transformations $(\Phi,\mathbf{A})\longrightarrow\left(\Phi',\mathbf{A}'\right)$ would show that the Schrödinger equation (16.21) is **not** invariant under the transformation (16.8) and (16.9):

$$\begin{aligned}
\text{LHS}\ &\longrightarrow\ -\frac{\hbar^2}{2m}\left(\nabla-\frac{ie}{\hbar c}\mathbf{A}'\right)^2\Psi=-\frac{\hbar^2}{2m}\left(\nabla-\frac{ie}{\hbar c}\mathbf{A}\underbrace{-\frac{ie}{\hbar c}\nabla\chi}\right)^2\Psi,\\
\text{RHS}\ &\longrightarrow\ i\hbar\left(\partial_t+\frac{ie}{\hbar}\Phi'\right)\Psi=i\hbar\left(\partial_t+\frac{ie}{\hbar}\Phi\underbrace{-\frac{ie}{\hbar c}\partial_t\chi}\right)\Psi.
\end{aligned} \tag{16.24}$$

Namely, there are these extra terms $\underbrace{\ldots}$ involving the gauge function χ that do not match on two sides of the transformed equation; hence gauge invariance is lost under (16.8) and (16.9). However, as observed by Fock (1926), the invariance could be obtained if we supplement the transformations of (16.8) and (16.9) by an appropriate spacetime-dependent phase change of the wavefunction $\Psi(\mathbf{r},t)$,

$$\Psi(\mathbf{r},t)\longrightarrow\Psi'(\mathbf{r},t)=\exp\left[\frac{ie}{\hbar c}\chi(\mathbf{r},t)\right]\Psi(\mathbf{r},t), \tag{16.25}$$

so that the above-mentioned extra terms can be cancelled. The function χ in the exponent is the same gauge function that appears in (16.8) and (16.9). Let us see how the various terms in the Schrödinger equation (16.21) change under this combined transformation (16.8), (16.9), and (16.25). First consider the RHS:

$$\left(\partial_t + \frac{ie}{\hbar}\Phi\right)\Psi \longrightarrow \left(\partial_t + \frac{ie}{\hbar}\Phi'\right)\Psi' = \left(\partial_t + \frac{ie}{\hbar}\Phi - \frac{ie}{\hbar c}\partial_t\chi\right)\exp\left(\frac{ie}{\hbar c}\chi\right)\Psi.$$

When the time derivative ∂_t acts on the product $\left[\exp\left(\frac{ie}{\hbar c}\chi\right)\Psi(\mathbf{r},t)\right]$, two terms result: $\exp\left(\frac{ie}{\hbar c}\chi\right)\partial_t\Psi(\mathbf{r},t) + \left(\frac{ie}{\hbar c}\partial_t\chi\right)\exp\left(\frac{ie}{\hbar c}\chi\right)\Psi(\mathbf{r},t)$, thus the effect of "pulling the phase factor $\exp\left(\frac{ie}{\hbar c}\chi\right)$ to the left of the ∂_t operator" will result in another extra term which just cancels the unwanted term in (16.24):

$$\begin{aligned}\left(\partial_t + \frac{ie}{\hbar}\Phi\right)\Psi &\longrightarrow \exp\left(\frac{ie}{\hbar c}\chi\right)\left(\partial_t + \frac{ie}{\hbar}\Phi - \frac{ie}{\hbar c}\partial_t\chi + \frac{ie}{\hbar c}\partial_t\chi\right)\Psi \\ &= \exp\left(\frac{ie}{\hbar c}\chi\right)\left(\partial_t + \frac{ie}{\hbar}\Phi\right)\Psi. \qquad (16.26)\end{aligned}$$

Similarly we have

$$\left(\boldsymbol{\nabla} - \frac{ie}{\hbar c}\mathbf{A}\right)^2\Psi \longrightarrow \exp\left(\frac{ie}{\hbar c}\chi\right)\left(\boldsymbol{\nabla} - \frac{ie}{\hbar c}\mathbf{A}\right)^2\Psi. \qquad (16.27)$$

As a consequence, the transformed equation

$$-\frac{\hbar^2}{2m}\left(\boldsymbol{\nabla} - \frac{ie}{\hbar c}\mathbf{A}'\right)^2\Psi' = i\hbar\left(\partial_t + \frac{ie}{\hbar}\Phi'\right)\Psi' \qquad (16.28)$$

becomes

$$\exp\left(\frac{ie}{\hbar c}\chi\right)\left(\boldsymbol{\nabla} - \frac{ie}{\hbar c}\mathbf{A}\right)^2\Psi = \exp\left(\frac{ie}{\hbar c}\chi\right)\left(\partial_t + \frac{ie}{\hbar}\Phi\right)\Psi.$$

The same exponential factor $\exp\left(\frac{ie}{\hbar c}\chi\right)$ appears on both sides of the transformed Schrödinger equation; they cancel, showing that the validity the transformed equation (16.28) follows from the original equation (16.21), and we have gauge invariance restored. From now on, whenever we refer to gauge transformation it is understood to be the combined transformations of (16.8), (16.9), and (16.25).

16.3.3 The gauge principle

We will now turn the argument around and regard the transformation (16.25) of the wavefunction as being more fundamental, and from this we can "derive" the gauge transformation of the EM potentials, (16.8) and (16.9). The rationale for doing it this way will become clear as we proceed. Our wish is to generalize gauge symmetry beyond electromagnetism, and to use this symmetry as a tool to discover new physics. In such an endeavor it is much easier to start with the generalization of the gauge transformation of the wavefunction rather than that for the potentials. More importantly, as we shall see, this reversed procedure

will also explain why the electromagnetic couplings (16.23) have the structure that they have.

This is similar to what Einstein did when he elevated the empirically observed equality between gravitational and inertial masses to the principle of equivalence between gravitation and inertia. With this focus, he then applied EP to the physics beyond mechanics (cf. Sections 12.2 and 12.3). In the same way, once the approach is formulated as the gauge principle for electromagnetism, we can then apply it to the physics beyond, to strong and weak interactions, etc.

The Schrödinger equation for a free charged particle has global $U(1)$ symmetry

It is easier to start the generalization process by starting with the gauge transformation of the wavefunction because we can associate this part of the gauge transformation to a more familiar symmetry transformation. Consider the Schrödinger equation for a charged free particle,

$$-\frac{\hbar^2}{2m}\nabla^2\Psi = i\hbar\partial_t\Psi. \tag{16.29}$$

This equation is unchanged under the **global** phase change:

$$\Psi(\mathbf{r},t) \longrightarrow \Psi'(\mathbf{r},t) = \exp\left(\frac{ie}{\hbar c}\chi\right)\Psi(\mathbf{r},t). \tag{16.30}$$

In contrast to the transformation as given in (16.25), here the phase factor is a constant $\chi \neq \chi(\mathbf{r},t)$. Namely, we make the **same** phase change for the wavefunction at all space-time points! This simple phase change is a "unitary transformation in one dimension";[6] hence called a "global $U(1)$ transformation". Clearly Eq. (16.29) is invariant, as every term acquires the same phase that can be cancelled out, and this theory possesses global $U(1)$ symmetry. This symmetry has the associated electric charge conservation law, as expressed by the continuity equation

$$\partial_t\rho_e + \nabla\cdot\mathbf{j}_e = 0 \tag{16.31}$$

with $\rho_e = |\Psi|^2$ and $\mathbf{j}_e = \frac{-i\hbar}{2m}\left(\Psi^*\nabla\Psi - \Psi\nabla\Psi^*\right)$. We leave it as an elementary QM exercise to prove that this continuity equation follows from the free Schrödinger equation (16.29).

[6] The transformation $U = e^{i\chi}$ is "unitary" because it satisfies the condition $U^\dagger U = 1$; it is in one dimensional because it is specified by one parameter.

Gauging the symmetry

One may be dissatisfied with this **global** feature of the transformation: Why should the wavefunctions everywhere all undergo the same phase change? A more desirable form of symmetry would require the theory to be invariant under a **local** transformation. Namely, we replace the phase factor in the transformation (16.30) by a spacetime-dependent function

$$[\chi = \text{constant}] \longrightarrow [\chi = \chi(\mathbf{r},t)], \tag{16.32}$$

just as in (16.25). That is, we want the freedom of choosing the phase of the charge's wavefunction locally: a different one at each spacetime point.

Now the Schrödinger equation (16.29) is not invariant under such a local transformation as the derivative terms would bring down extra terms (because of the spacetime-dependent phase) that cannot be canceled, indicating that the equation is no longer symmetric. We can overcome this difficulty by replacing ordinary derivatives $(\partial_t, \boldsymbol{\nabla})$ by covariant derivatives $(D_t, \mathbf{D})$ as defined in (16.23). They have the desired property that covariant ("change in the same way") derivatives of the wavefunction transform in the same way as the wavefunction itself. Namely, just as $\Psi' = \exp\left(\frac{ie}{\hbar c}\chi\right)\Psi$, we have

$$(D_t\Psi)' = \exp\left(\frac{ie}{\hbar c}\chi\right)(D_t\Psi) \tag{16.33}$$

and similarly,

$$(\mathbf{D}\Psi)' = \exp\left(\frac{ie}{\hbar c}\chi\right)(\mathbf{D}\Psi) \;\text{ and also }\; \left(\mathbf{D}^2\Psi\right)' = \exp\left(\frac{ie}{\hbar c}\chi\right)\left(\mathbf{D}^2\Psi\right). \tag{16.34}$$

This replacement of derivatives

$$(\partial_t, \boldsymbol{\nabla}) \longrightarrow (D_t, \mathbf{D}) \tag{16.35}$$

calls to mind the principle of general covariance when going from special relativity to general relativity as discussed in Section 13.4. Similar to the situation here, when proceeding from SR to GR we go from a global symmetry to a local symmetry. This replacement (16.35) turns the Schrödinger equation (16.29) into

$$-\frac{\hbar^2}{2m}\mathbf{D}^2\Psi = i\hbar D_t\Psi. \tag{16.36}$$

Under the gauge transformations of (16.8), (16.9), and (16.25), we have

$$-\frac{\hbar^2}{2m}\left(\mathbf{D}^2\Psi\right)' = i\hbar(D_t\Psi)'. \tag{16.37}$$

The invariance of the equation can be checked because, through the relations in (16.33) and (16.34), it is

$$-\frac{\hbar^2}{2m}\left[\exp\left(\frac{ie}{\hbar c}\chi\right)\right]\left(\mathbf{D}^2\Psi\right) = i\hbar\left[\exp\left(\frac{ie}{\hbar c}\chi\right)\right](D_t\Psi). \tag{16.38}$$

With the exponential factors [...] canceled, this is just the original equation (16.36). This completes the proof of the equation's invariance[7] under such a local transformation.

The covariant derivatives (16.23) are constructed by an artful combination of the ordinary derivative with a set of newly introduced "compensating fields" $(\Phi, \mathbf{A})$ which themselves transform in such a way to compensate, to cancel, the unwanted extra factors that spoil the invariance. The replacement of ordinary derivatives by covariant derivatives as in (16.35) justifies the "principle of minimal substitution", used to introduce the EM coupling as done in (16.23). Equation (16.36) is just the Schrödinger equation (16.21) for a charged particle in an EM field that we discussed earlier. Thus we can understand this

[7] Properly speaking we should say "covariance of the equation", as the terms in an equation are not invariant, but they transform "in the same way" so that their relation is unchanged, and the same equation is obtained for the transformed quantitites.

coupling scheme as resulting from requiring the theory to have a **local** $U(1)$ symmetry in the charge space (i.e. with respect to a change of the wavefunction phase associated with the particle's charge). The general practice is that such local symmetry is called **gauge symmetry** and the process of turning a global symmetry into a local one as in (16.32) has come to be called "**gauging a symmetry**".

One can then understand the "origin" of the electromagnetic potentials $(\Phi, \mathbf{A})$ as the **gauge fields**[8] required to implement such a gauge symmetry. This procedure of understanding the origins of some dynamics (e.g. electromagnetism) through the process of turning a global symmetry into a local one is now called **the gauge principle**.

[8]From now on we shall often refer to $(\Phi, \mathbf{A})$ as "fields", rather than "potentials", and the equations they satisfy as field equations. This is compatible with the general practice in physics of calling any function of space and time "field".

We would like to emphasize the similarity of 'gauging a symmetry' and 'turning the global Lorentz symmetry in special relativity into the local coordinate symmetry of general relativity'. In both cases, we need to replace the usual derivatives by covariant derivatives. In the case of gauge theory, this introduces the gauge field; in the case of relativity this inserts a gravitational field intensity (in the form of Christoffel symbols) into the theory.

16.4 Electromagnetism as a gauge interaction

The above discussion has allowed us to have a better understanding of the EM coupling as displayed in the Hamiltonian of (16.15), which is equivalent to the Lorentz force law. We will now show that gauge symmetry, together with special relativity, allows us to "derive" the electromagnetic field equations, the Maxwell equations. For this purpose we need a language that will simplify the expression of relativistic invariance. This is provided by Minkowski's four-dimensional spacetime formalism (cf. Chapter 11). We first provide a rapid review of this subject. Not only will this allow us to understand Maxwell's equations, it will also provide us, in a simple way, to infer the pattern of the $\pm$ signs and factors of c that have appeared in Eqs. (16.6)–(16.9), which were written in non relativistic notation.

16.4.1 The 4D spacetime formalism recalled

Gauging the $U(1)$ symmetry requires us to introduce the potentials $(\Phi, \mathbf{A})$ and thus the existence of electromagnetism. As we shall demonstrate, the requirements of gauge symmetry and Lorentz invariance (special relativity) can basically lead us to Maxwell's equations. For this we shall adopt the language of 4-vectors and 4-tensors in 4D Minkowski spacetime (as discussed in Chapter 11) so as to make it simpler to implement the condition of Lorentz symmetry.

The principal message of special relativity is that the arena for physics events is 4D Minkowski spacetime, with spatial and temporal coordinates being treated on an equal footing (cf. Section 11.3). In this 4D space, a position vector has four components x^μ with index $\mu = 0, 1, 2, 3$.

$$x^\mu = \left(x^0, x^1, x^2, x^3\right) = (ct, \mathbf{x}). \tag{16.39}$$

Noting that the contravariant and covariant tensor components are related, as shown in (11.19), by the Minkowski metric $\eta_{\mu\nu} = \eta^{\mu\nu} = \text{diag}(-1, 1, 1, 1)$, we list some of the 4-position and 4-derivatives:

$$\text{4D del operator} \qquad \partial_\mu = \left(\frac{1}{c}\partial_t, \boldsymbol{\nabla}\right) \text{ and } \partial^\mu = \left(-\frac{1}{c}\partial_t, \boldsymbol{\nabla}\right) \tag{16.40}$$

$$\text{momentum 4-vector} \qquad p^\mu = \left(\frac{E}{c}, \mathbf{p}\right) \text{ with } p_\mu p^\mu = -\frac{E^2}{c^2} + \mathbf{p}^2 = -m^2c^2. \tag{16.41}$$

Maxwell's equations

The six components of $(\mathbf{E}, \mathbf{B})$ are taken to be the elements of a 4×4 antisymmetric matrix: the "EM field tensor", $F_{\mu\nu} = -F_{\nu\mu}$ as displayed in Eq. (11.34). The inhomogeneous Maxwell equations (16.2)–(16.3) can then be written compactly as

$$\partial^\mu F_{\mu\nu} = -\frac{1}{c} j_\nu \qquad \text{Gauss + Ampere (inhomogeneous Maxwell)} \tag{16.42}$$

where $j^\mu \equiv (c\rho, \mathbf{j})$ is the "4-current density", and the homogeneous (16.4)–(16.5) as

$$\partial^\mu \tilde{F}_{\mu\nu} = 0 \qquad \text{Faraday + mag-Gauss (homogeneous Maxwell)} \tag{16.43}$$

where $\tilde{F}^{\mu\nu} = -\frac{1}{2}\varepsilon^{\mu\nu\lambda\rho} F_{\lambda\rho}$ is the dual field tensor.[9]

[9] The duality transformation (9.41) discussed in Section 9.5.1 corresponds to $F_{\mu\nu} \to \tilde{F}_{\mu\nu}$.

Electromagnetic potentials

It is easy to see that, with the "4-potential" being $A^\mu = (\Phi, \mathbf{A})$, namely, $\Phi = A^0 = -A_0$, the relation between potentials and the field tensor, (16.6) and (16.7), can be summarized as ($F_{\mu\nu}$ as the 4-curl of A_μ)

$$F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu, \tag{16.44}$$

while the gauge transformations (16.8) and (16.9) can be compactly written in the Minkowski notation as

$$A_\mu \longrightarrow A'_\mu = A_\mu + \partial_\mu \chi. \tag{16.45}$$

The electromagnetic field strength tensor $F_{\mu\nu}$, with components of $(\mathbf{E}, \mathbf{B})$, being related to the potentials A^μ as in (16.44), is clearly unchanged[10] under this transformation (16.45).

[10] $F'_{\mu\nu} = \partial_\mu A'_\nu - \partial_\nu A'_\mu = (\partial_\mu A_\nu - \partial_\nu A_\mu) + (\partial_\mu \partial_\nu \chi - \partial_\nu \partial_\mu \chi) = F_{\mu\nu}$ because $\partial_\mu \partial_\nu = \partial_\nu \partial_\mu$.

Such a notation also simplifies the steps when showing that (16.44) solves the homogeneous Maxwell equation (16.43):

$$\partial^\mu \tilde{F}_{\mu\nu} = \frac{1}{2}\epsilon_{\mu\nu\lambda\rho}\partial^\mu F^{\lambda\rho} = \frac{1}{2}\epsilon_{\mu\nu\lambda\rho}\partial^\mu\left(\partial^\lambda A^\rho - \partial^\rho A^\lambda\right) = \epsilon_{\mu\nu\lambda\rho}\partial^\mu\partial^\lambda A^\rho = 0.$$

The two RHS terms are combined when the dummy indices λ and ρ are relabeled. The final result vanishes because the indices $\mu\lambda$ are antisymmetric in 4D Levi-Civita symbols $\epsilon_{\nu\mu\lambda\rho}$ but symmetric in the double derivative

$\partial^\mu \partial^\lambda$. When the relation (16.44) is plugged into the inhomogeneous Maxwell equation (16.42), we have

$$\partial^\mu \left(\partial_\mu A_\nu - \partial_\nu A_\mu\right) = \Box A_\nu - \partial_\nu \left(\partial^\mu A_\mu\right) = -\frac{1}{c} j_\nu$$

which, after imposing the Lorentz gauge condition $\partial^\mu A_\mu = 0$, reduces to the simple wave equation $\Box A_\nu = -\frac{1}{c} j_\nu$ that we displayed in Section 3.1.

Similarly, the covariant derivatives D_t and $\mathbf{D}$ defined in (16.23) can be combined into a "4-covariant derivative" as $D_\mu = \left(\frac{1}{c} D_t, \mathbf{D}\right)$ so that

$$D_\mu \equiv \left(\partial_\mu - \frac{ie}{\hbar c} A_\mu\right), \tag{16.46}$$

and the minimal substitution rule is simply the replacement of $\partial_\mu \longrightarrow D_\mu$.

We also take note of a useful relation between the commutator of covariant derivatives and the field strength tensor [cf. Eq. (14.17)]

$$\left[D_\mu, D_\nu\right] = -\frac{ie}{\hbar c} F_{\mu\nu}. \tag{16.47}$$

This operator equation is understood that each term is an operator that acts, from the left, on some spacetime-dependent test function (such as a wavefunction). This can be verified by explicit calculation:

$$\begin{aligned}
&\left[D_\mu, D_\nu\right] \psi \\
&= \left(\partial_\mu - \frac{ie}{\hbar c} A_\mu\right)\left(\partial_\nu - \frac{ie}{\hbar c} A_\nu\right)\psi - \left(\partial_\nu - \frac{ie}{\hbar c} A_\nu\right)\left(\partial_\mu - \frac{ie}{\hbar c} g A_\mu\right)\psi \\
&= -\frac{ie}{\hbar c}\left\{\partial_\mu \left(A_\nu \psi\right) + A_\mu \left(\partial_\nu \psi\right) - \partial_\nu \left(A_\mu \psi\right) - A_\nu \left(\partial_\mu \psi\right)\right\} \qquad (16.48) \\
&= -\frac{ie}{\hbar c}\left(\partial_\mu A_\nu - \partial_\nu A_\mu\right)\psi = -\frac{ie}{\hbar c} F_{\mu\nu}\psi.
\end{aligned}$$

The relevance of this relation in a generalized gauge symmetry will be discussed below—see the displayed equation (16.68).

16.4.2 The Maxwell Lagrangian density

We now add further detail to the statement: "the electromagnetic interaction is a gauge interaction", or equivalently, "electrodynamics is a gauge theory". So far we have concentrated on the "equation of motion" part of the field description (the Lorentz force law). Now we discuss the "field equation" part. In the case of electromagnetism, it is Maxwell's equation. A field can be viewed as a system having an infinite number of degrees freedom with its generalized coordinate being the field itself $q = \phi(x)$, where $\phi(x)$ is some generic field. For such a continuum system, the field equation, as discussed in Section A.5.2, is the Euler–Lagrange equation (A.64) written in terms of the **Lagrangian density**

$\mathcal{L}$ (with a Lagrangian $L = \int d^3x\mathcal{L}$ and an action $S = \int d^4x\mathcal{L}(x)$) which is a function of the field and its derivatives $\mathcal{L} = \mathcal{L}\big(\phi, \partial_\mu\phi\big)$:

$$\partial_\mu \frac{\partial \mathcal{L}}{\partial\big(\partial_\mu\phi\big)} - \frac{\partial \mathcal{L}}{\partial\phi} = 0. \tag{16.49}$$

Knowledge of the (Lorentz invariant) Lagrangian density $\mathcal{L}$ is equivalent to knowing the (Lorentz covariant) field equation. Thus, knowledge of Maxwell's Lagrangian density:

$$\mathcal{L}(x) = -\frac{1}{4}F_{\mu\nu}F^{\mu\nu} + \frac{1}{c}j^\nu A_\nu \tag{16.50}$$

is tantamount to knowing Maxwell's field equations. The Euler–Lagrange equation for the $A_\mu(x)$ field is [namely, Eq. (16.49) with $\phi(x) = A_\nu(x)$]:

$$\partial_\mu \frac{\partial \mathcal{L}}{\partial\big(\partial_\mu A_\nu\big)} - \frac{\partial \mathcal{L}}{\partial A_\nu} = 0, \tag{16.51}$$

which is just the familiar Maxwell equation (16.42), as we have

$$\frac{\partial \mathcal{L}}{\partial A_\nu} = \frac{1}{c}j^\nu, \tag{16.52}$$

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial\big(\partial_\mu A_\nu\big)} &= \frac{\partial}{\partial\big(\partial_\mu A_\nu\big)}\left(-\frac{1}{4}\big(\partial_\alpha A_\beta - \partial_\beta A_\alpha\big)^2\right)\\ &= \frac{\partial}{\partial\big(\partial_\mu A_\nu\big)}\left(-\frac{1}{2}\partial_\alpha A_\beta\big(\partial^\alpha A^\beta - \partial^\beta A^\alpha\big)\right)\\ &= -F^{\mu\nu}.\end{aligned} \tag{16.53}$$

16.4.3 Maxwell equations from gauge and Lorentz symmetries

From the above discussion, we see that a derivation of Maxwell's Lagrangian density (16.50) is tantamount to a derivation of Maxwell's equations (16.42) themselves. Gauging a symmetry requires the introduction of the gauge field; in the case of $U(1)$ symmetry, it is the vector $A_\mu(x)$ field. To have a dynamical theory for the $A_\mu(x)$ field, we need to construct a Lagrangian density from this $A_\mu(x)$ field and its derivatives $\partial_\mu A_\nu$ (because the kinetic energy term must involve spacetime derivatives). The simplest gauge-invariant combination of $\partial_\mu A_\nu$ is

$$\partial_\mu A_\nu - \partial_\nu A_\mu = F_{\mu\nu}. \tag{16.54}$$

The Lagrangian density must also be a Lorentz scalar (i.e. all spacetime indices are contracted) so the resulting Euler–Lagrange equations are relativistic covariant. The simplest combination[11] is the expression

$$\mathcal{L}_A = -\frac{1}{4}F_{\mu\nu}F^{\mu\nu}. \tag{16.55}$$

[11] In principle, higher powers $(F_{\mu\nu}F^{\mu\nu})^n$ are also gauge and Lorentz symmetric. However such terms are "nonrenormalizable" and our current understanding of quantum field theory informs us that they should be highly suppressed (i.e. at the relevant energy scale we consider, they make negligible contribution).

One can then add $\frac{1}{c}j^{\nu}A_{\nu}$ as the source term to arrive at the result in (16.50). (Factors like $-\frac{1}{4}$ and c are unimportant—all a matter of system of units and convention.)

The Maxwell equations were discovered by experimentation and deep theoretical invention; but this derivation shows **why** the four equations,[12] (16.2)–(16.5), take on the form they take. Their essence is special relativity plus gauge invariance. With this insight of electromagnetism we can then generalize the approach to the investigation of other fundamental interactions.

[12]Of course these equations have already been greatly simplified with the vector notation embodying rotational symmetry, which is part of Lorentz invariance.

16.5 Gauge theories: A narrative history

The above discussion shows that we can understand electromagnetism as a "gauge interaction". From the requirement of a local $U(1)$ symmetry in charge space, the presence of a vector gauge field $A_{\mu}(x)$ is deduced. In the rather restrictive framework of special relativity, its dynamics can be fixed (to be that described by Maxwell's equation). This way of using symmetry to deduce the dynamics has been very fruitful in our attempts to understand (i.e. to construct theories of) other particle interactions as well.

One of the crowning achievement in the physics of the twentieth century is the establishment of the Standard Model (SM) of elementary particle interactions.[13] This gives a complete and correct description of all nongravitational physics. This theory is based on the principle of gauge symmetry. Strong, weak, and electromagnetic interactions are all gauge interactions. In this section we give a very brief account of this SM gauge theory of particle physics.

[13]The progress of physics depends both on theory and experiment. A proper account of the experimental accomplishments in the establishment of the Standard Model is however beyond the scope of this presentation. This omission should not be viewed in any way as the author's lack of appreication of their importance.

16.5.1 Einstein's inspiration, Weyl's program, and Fock's discovery

The rich and interesting history of gauge invariance and electromagnetic potentials in classical electromagnetism is beyond the scope of our presentation; we refer the curious reader to the authoritative and accessible account given by Jackson and Okun (2001). Here we shall present a narrative of the development of the gauge symmetry idea[14] as rooted in Einstein's general theory of relativity.

[14]For a more detailed gauge theory survey with extensive references, see Cheng and Li (1988).

What is the origin of the name "gauge symmetry"? The term **eichinvarianz** (gauge invariance) was coined in 1919 by Hermann Weyl (1885–1955) in the context of his attempt to "geometrize" the electromagnetic interaction and to construct in this way a unified geometrical theory of gravity and electromagnetism (Weyl 1918, 1919). He invoked the invariance under a local change of the scale, the "gauge", of the metric field $g_{\mu\nu}(x)$:

$$g_{\mu\nu}(x) \longrightarrow g'_{\mu\nu}(x) = \lambda(x) g_{\mu\nu}(x), \tag{16.56}$$

where $\lambda(x)$ is an arbitrary function of space and time. Weyl was inspired by Einstein's geometric theory of gravity, general relativity, which was published in 1916 (cf. Chapters 13 and 14). This was, of course, before the emergence of

modern quantum mechanics in 1925–26. A key QM concept was to identify the dynamical variables of energy and momentum with the operators $-i\hbar\partial_\mu$ and, in the presence of an electromagnetic field, with $-i\hbar\partial_\mu + \frac{e}{c}A_\mu$ as in the "minimal substitutional rule". In this context, Vladimir Fock (1898–1974) discovered in 1926 that the quantum mechanical wave equation was invariant under the combined transformation $A'_\mu = A_\mu + \partial_\mu\chi$ and

$$\Psi'(x) = \exp\left[\frac{ie}{\hbar c}\chi(x)\right]\Psi(x); \tag{16.57}$$

he called it the **gradient transformation**.[15] Given the central role played by (16.57) in showing that here one is discussing transformations in charge space, one would say that it was Fock who first discovered the modern notion of gauge invariance in physics.

Fritz London observed in 1927 that, if the i was dropped from Fock's exponent in (16.57), this phase transformation becomes a scale change (16.56) and the transformation of (16.45) and (16.57) was equivalent to Weyl's old eichtransformation (London 1927). However, when Weyl finally worked out this approach later on he retained his original terminology of "gauge invariance" because he believed that a deep understanding of the local transformation of gauge invariance could come about only through the benefit of general relativity.[16] Most importantly it was Weyl who first declared [especially in his famous book: *Theory of Groups and Quantum Mechanics* (Weyl 1928, 1931)] gauge invariance as a fundamental principle—the requirement of the matter wave equation being symmetric under the gauge transformation leading to the introduction of the electromagnetic field. Subsequently this principle has become the key pathway in the discovery of modern theories of fundamental particle interactions; this calling a local symmetry (in charge space) a "gauge symmetry" has become the standard practice in physics.

[15] This designation originates from the transformation $A'_\mu = A_\mu + \partial_\mu\chi$. From the title of the Fock (1926) paper, it is clear what Fock wanted to emphasize is that this new symmetry, involving the transformation of $\Psi(x)$ and $A_\mu(x)$, is a symmetry of charge space. Thus in the first 40 years or so after the invention of the gauge principle, people often followed the practice of designating global symmetry in charge space as '**gauge symmetry of the first kind**' and local symmetry in charge space as '**gauge symmetry of the second kind**'. But nowadays by gauge symmetry we mean gauge symmetry of the second kind and distinguish it from the first kind simply by calling the latter global symmetry.

[16] The connection between gauge symmetry and general relativity would be deepened further with the advent of the nonabelian theory of Yang and Mills in the 1950s.

16.5.2 Quantum electrodynamics

We have so far discussed gauge symmetry in the quantum description of a nonrelativistic charged particle interacting with an electromagnetic field. The proper framework for particle interaction should be quantum field theory,[17] which is a union of quantum mechanics with special relativity. We first comment on the prototype quantum field theory of quantum electrodynamics (QED), which is the quantum description of relativistic electrons and photons. However in this introductory presentation of gauge symmetry, we shall by and large stay with a classical field description.

[17] Some elementary aspects of quantum field theory were presented in Section 6.4. For an introduction to the Standard Model in the proper quantum field theoretical framework, see, for example, Cheng and Li (1984).

Dirac equation

QED is the theory of electrons interacting through the electromagnetic field. While the EM field equation is already relativistic, we must replace the Schrödinger equation by the relativistic wave equation for the electron, first discovered by Paul Dirac. Namely, instead of the Schrödinger equation (16.29), we should use the Dirac equation for a free electron,

$$\left(i\hbar\gamma^\mu\partial_\mu - mc\right)\psi = 0 \tag{16.58}$$

with ψ being the four-component electron spinor field and γ^μ is a set of four 4×4 "Dirac gamma matrices" obeying the anticommutation relation $\{\gamma^\mu, \gamma^\nu\} = -2g^{\mu\nu}$. In momentum space with $p^\mu = (E/c, \mathbf{p})$ being the 4-momentum vector this equation becomes

$$\left(\gamma^\mu p_\mu - mc\right)\psi = 0. \tag{16.59}$$

When operated on from the LHS by $(\gamma^\nu p_\nu + mc)$, this equation, after using the anticommutation relation, implies[18] $\left(p^\mu p_\mu + m^2c^2\right)\psi = 0$, which we recognize as the relativistic energy momentum relation of (16.41). To couple it to an EM field, we replace the derivative ∂_μ in (16.58) by the covariant derivative D_μ of (16.46): $\left(i\hbar\gamma^\mu D_\mu - mc\right)\psi = 0$—in just the same way as we obtained Eq. (16.21) from the free Schrödinger equation (16.22). We can display the role of the gauge field $\left(A_\mu\right)$/electron field (ψ) cross-term as the source factor by separating out and moving it to the RHS:

$$\left(i\hbar\gamma^\mu \partial_\mu - mc\right)\psi = \frac{e}{c}\gamma^\mu A_\mu \psi. \tag{16.60}$$

Lagrangian density for QED

Instead of field equations, we can equivalently work with the Lagrangian density. Thus instead of Eq. (16.58) we can concentrate on the equivalent quantity

$$\mathcal{L}_\psi = \bar{\psi}\left(i\hbar\gamma^\mu \partial_\mu - mc\right)\psi, \tag{16.61}$$

which is manifestly Lorentz invariant, with $\bar{\psi}$ being the conjugate $\psi^\dagger\gamma_0$. As discussed above, EM coupling can be introduced through the covariant derivative and, after adding the density $\mathcal{L}_A$ of the EM field (16.55), we have the full QED Lagrangian density

$$\mathcal{L}_{\text{QED}} = \mathcal{L}_\psi + \mathcal{L}_A + \mathcal{L}_{\text{int}}. \tag{16.62}$$

The interaction density,[19] which comes from part of the covariant derivative, is just the source density in (16.50)

$$\mathcal{L}_{\text{int}} = \frac{1}{c} j^\mu A_\mu = \frac{e}{c}\bar{\psi}\gamma^\mu \psi A_\mu \tag{16.63}$$

where j^μ is shown now as the 4-current density of electron. A graphical representation of a gauge boson coupled to a current is shown in Fig. 16.1(a).

[19] $\mathcal{L}_\psi$ and $\mathcal{L}_A$ are the Lagrangians for free electrons and photons as they are quadratic (harmonic) in their respective fields, while $\mathcal{L}_{\text{int}}$ represents the interaction as it has more than two field powers (hence anharmonic).

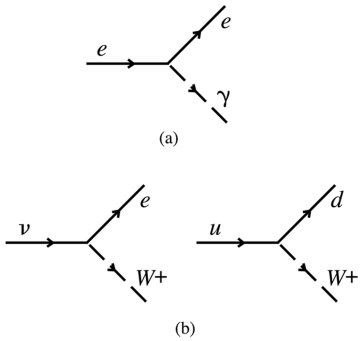

Fig. 16.1 (a) Trilinear coupling of a photon to an electron; (b) weak vector boson W coupled respectively to weak currents of leptons and quarks.

[18] We first find

$$\left(\gamma^\nu p_\nu + mc\right)\left(\gamma^\mu p_\mu - mc\right) = \gamma^\nu p_\nu \gamma^\mu p_\mu - m^2c^2.$$

Since $p_\nu p_\mu = p_\mu p_\nu$, we should symmetrize the gamma matrices as well:

$$\frac{1}{2}\left\{\gamma^\nu, \gamma^\mu\right\} p_\mu p_\nu - m^2c^2 = -p^\mu p_\mu - m^2c^2 = 0.$$

To reach the last expression, we have used the anticommutation relation of gamma matrices.

QED as a $U(1)$ gauge theory

The discussion carried out in the previous sections of this chapter demonstrates that one can "derive" $\mathcal{L}_{\text{QED}}$ from the requirement of Lorentz and gauge symmetry. Namely, the theory can be understood as following from "gauging the $U(1)$ symmetry" of the free Dirac equation. The original global $U(1)$ symmetry is directly linked to the familiar electric charge conservation. The quantization (and renormalization) procedure based on $\mathcal{L}_{\text{QED}}$ is rather complicated and is beyond the scope of this presentation. Still, what we need to know is that the full QED theory can be worked out on the basis of this Lagrangian density (16.62). We also note that quanta of the electromagnetic field are photons. They can now be viewed as the gauge particles (spin-1 bosons) of QED theory. Parenthetically, the common practice in quantum field theory of describing electrons interacting through the electromagnetic field is "interaction through the exchange of photons". This language is particularly convenient when, as we shall see, describing the strong and weak interactions. An important feature of the QED Lagrangian is the absence of a term of the form of $A^\mu A_\mu$ because it is forbidden by gauge invariance. Such a term would correspond to a gauge boson (photon) mass. Thus gauge invariance automatically predicts a massless photon, which accounts for the long-range nature of the (electromagnetism) interaction that it transmits.[20]

Because of its many redundant degrees of freedom, quantization of gauge theory is rather intricate. The necessary renormalization program for QED was successfully formulated through the work of Julian Schwinger (1918–94), Richard Feynman (1918–88), Sin-Itiro Tomonaga (1906–79), and Freeman Dyson (1923–). The close interplay of high-precision experimental measurement and theoretical prediction brought about this notable milestone in the history of physics.

[20] The relation between the range of interaction and the mass of the mediating particle is discussed in Section 6.4.2.

16.5.3 QCD as a prototype Yang–Mills theory

Here we shall discuss a highly nontrivial extension of the gauge symmetry of electromagnetism. This extension makes it even clearer that the transformation in the charge space involves the change of particle/field labels of the theory.

Abelian versus nonabelian gauge symmetries

The gauge symmetry for electromagnetism is based on the $U(1)$ symmetry group; its transformation involves the multiplications of phase factors to the wavefunction $\Psi \to \Psi' = U\Psi$ with $U(x) = e^{i\theta(x)}$ and the wavefunction $\Psi(x)$ is itself a simple function. A $U(1)$ phase transformation is equivalent to a rotation around a fixed axis in a 2D plane (in charge space). Hence $U(1)$ is isomorphic to the 2D rotation group (also called the 2D special orthogonal group): $U(1) = SO(2)$. Clearly such transformations are commutative $U_1U_2 = U_2U_1$, and the symmetry is said to be an **abelian symmetry**. On the other hand, general rotations in 3D space are represented by noncommutative matrices. The symmetry based on such rotations, called $SO(3) = SU(2)$, is **nonabelian symmetry**. The corresponding wavefunction (i.e. field) Ψ is a multiplet in some multidimensional charge space; its components would correspond to different

particle states. As it turns out, we can understand other elementary particle interactions due to strong and weak forces as gauge interactions also, but their gauge symmetries are nonabelian—their respective symmetry transformations are noncommutative. Gauge symmetry with such noncommuntative transformations was first studied in 1954 in particle physics by C.N. Yang (1922–) and Robert L. Mills (1927–99), hence nonabelian gauge theory is often referred to as **Yang–Mills theory**.

Quarks and gluons From the heroic experimental and deep phenomenological studies of the strong and weak interactions, it was discovered that strongly interacting particles (called **hadrons**[21]) are composed of even more elementary constituents. These spin-1/2 particles were invented and named **quarks** by Murray Gell-Mann (1929–). There are six '**quark flavors**' (*up, down, strange, charm, bottom* and *top*); each has three hidden degrees of freedom called '**color**'.[22] Namely, each quark flavor can be in three different color states—they form an $SU(3)$ triplet representation in the color charge space:

$$q(x) = \begin{pmatrix} q_1(x) \\ q_2(x) \\ q_3(x) \end{pmatrix}. \tag{16.64}$$

If it is an 'up-quark', we can call them, for example, 'red', 'blue', and 'white' up-quarks. This triplet undergoes the transformation, $q' = Uq$, with U being a 3×3 unitary matrix having unit determinant (hence called special unitary). Namely, the particle fields can not only change their phases but the particle labels as well. When this symmetry is "gauged" (i.e. turned into a local symmetry) we have the $SU(3)$ gauge theory, called **quantum chromodynamics** (QCD). Just as QED is the theory of electrons interacting through the exchange of abelian gauge fields of photons, QCD is the fundamental strong interaction of quarks through the exchange of a set (8) of nonabelian gauge particles called **gluons**.

Yang–Mills gauge particles To implement such a local symmetry, we need to introduce a covariant derivative involving gauge fields:

$$D_\mu = \partial_\mu - ig_s G_\mu. \tag{16.65}$$

This is the same as (16.46) of the abelian case (for simplicity of notation we have suppressed, or absorbed, the factor $\hbar c$). In the strong interacting QCD case, in place of the electromagnetic coupling strength e we have the strong coupling g_s. Instead of the $U(1)$ gauge field A_μ, we have Yang–Mills fields G_μ. But now the D_μ and G_μ are matrices in the color charge space; the gluon field matrix G_μ has eight independent components (being a traceless 3×3 Hermitian matrix) corresponding to the eight gluons of the strong interaction.

The basic property for covariant derivatives in Yang–Mills theory is still the same as that for the abelian case: the covariant derivative of a wavefunction $(D_\mu q)$ transforms in the same way as the wavefunction $q(x)$ itself:

$$q' = Uq \quad \text{and} \quad (D'_\mu q') = U(D_\mu q). \tag{16.66}$$

[21] Examples of hadrons are the proton, neutron, pion and omega, etc.

[22] 'Color' is the whimsical name given to the strong interaction charge and has nothing to do with the common understanding of different frequencies of visible EM waves. Here we give an example of the type of phenomenology from which the color degrees of freedom were deduced. The omega baryon is composed of three strange quarks. Being a system of identical fermions, its wavefunction should be antisymmetric with respect to the interchange of any two quarks. Yet both its spin (3/2) and orbital angular momentum (S-wave) wavefunctions are symmetric. Spin-statistics is restored only when its antisymmetric color (singlet) wavefunction is taken into account.

We then have $D'_\mu = UD_\mu U^{-1}$ and more explicitly

$$\partial_\mu - ig_s G'_\mu = U\partial_\mu U^{-1} - ig_s U G_\mu U^{-1}.$$

This means that the gauge fields must transform as

$$G'_\mu = UG_\mu U^{-1} - \frac{1}{ig_s} U\big(\partial_\mu U^{-1}\big). \tag{16.67}$$

One can easily check that in the abelian case with $U = \exp(ig_s\chi)$ this reduces to Eq. (16.45). The factor $UG_\mu U^{-1}$ indicates that the gauge field itself transform nontrivially under the gauge group. Namely, the gauge fields (or gauge particles) themselves carry gauge charges.

The QCD Lagrangian Another place where one can see that nonabelian gauge fields themselves carry gauge charges is in a property of the nonabelian field tensor $F_{\mu\nu}$, which is similarly related to the covariant derivatives as (16.47), $\big[D_\mu, D_\nu\big] = -ig_s F_{\mu\nu}$. Working it out as in (16.48), we find

$$F_{\mu\nu} = \partial_\mu G_\nu - \partial_\nu G_\mu - \frac{1}{ig_s}\big[G_\mu, G_\nu\big] \tag{16.68}$$

showing a nonvanishing commutator because now G_μ is also a matrix in the charge space. This nonabelian $F_{\mu\nu}$ is now quadratic in G_μ. Thus when we construct[23] the QCD Lagrangian density for the gauge field, like we did for the abelian case of (16.55),

$$\mathcal{L}_{\rm A} = -\frac{1}{4}{\rm tr}F_{\mu\nu}F^{\mu\nu}, \tag{16.69}$$

we get, besides quadratic $\big(G^2\big)$ terms, also cubic $\big(G^3\big)$ and quartic $\big(G^4\big)$ gauge field terms. While quadratic terms in $\mathcal{L}$ correspond to the free-particle Lagrangian, higher powers represent interactions. Again, these cubic and quartic couplings reflect the fact that nonabelian gauge fields, the gluons, must now be charged fields. See Fig. 16.2. Very much like Eq. (16.62) for QED, the Lagrangian density for QCD can be written down:

$$\mathcal{L}_{\rm QCD} = \mathcal{L}_{\rm q} + \mathcal{L}_{\rm A} + \mathcal{L}_{\rm qA}. \tag{16.70}$$

$\mathcal{L}_{\rm q} = \bar{q}(i\hbar\gamma^\mu\partial_\mu - mc)q$ is the Lagrangian density for free quarks, much like $\mathcal{L}_\psi$ in Eq. (16.61) for free electrons. Since the quark field is a triplet, a sum of three terms (one for each color) is understood in $\mathcal{L}_{\rm q}$. The Euler Lagrange equation for the gluon field based on $\mathcal{L}_{\rm A}$ is nonlinear,[24] reflecting the fact that gluons carry color charges themselves. The quark (q)/gauge field (G_μ) coupling

[23] The symbol "tr" in (16.69) stands for the operation "trace", taken in charge space, (i.e. all charge space indices of the two $F_{\mu\nu}s$ are to be summed over in order to get a gauge symmetric quantity).

[24] This is entirely similar to the nonlinearity of the Einstein field equation in GR. A gravitational field carries energy ("gravity charge"), thus is itself a source of a gravitational field.

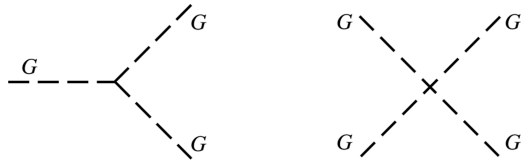

Fig. 16.2 Cubic and quartic self-couplings of charged gauge bosons. The similarity to the trilinear couplings shown in Fig.16.1 should be noted.

$\mathcal{L}_{\rm qA}$ comes from the covariant derivative and is of a form entirely similar to the QED interaction term (16.63)

$$\mathcal{L}_{\rm qA} = \frac{1}{c} j^\mu G_\mu = \frac{g_s}{c} \bar{q} \gamma^\mu G_\mu q. \tag{16.71}$$

However now the quark field q is a triplet and G_μ is a 3×3 Hermitian matrix in the color charge space. With nonvanishing off-diagonal terms in the color matrix G_μ, a quark's color charges can be changed by such a quark/gluon coupling.

Asymptotic freedom and quark confinement The Yang–Mills gauge particles as transmitters of interactions being charged, this feature leads to the important physical consequence that the effective interaction strength[24a] (the so-called "running coupling") grows, logarithmically, as the distance between quarks increases. Namely, we have an antiscreen effect on the color charge—as the color charge is probed further away from the source, the effective charge is seen to increase! The increase in coupling strength means that it would take more and more energy to separate colored charges. Thus a colored particle must be confined to short subnuclear distances. This explains why no free quarks have ever been seen. All the observed strong-interaction particles (hadrons) are colorless compounds of quarks and gluons. This short-range confinement effect explains why even though gluons, like photons, are massless, the strong interaction they transmit, unlike the EM interaction, is nevertheless short-ranged. The other side of the same property (called **asymptotic freedom**[25]) is that the effective coupling becomes small at short distances and a perturbation approach can be used to solve the QCD equations in the high-energy and large-momentum-transfer regime, leading to precise QCD predictions that have been verified to high accuracy by experiments.

[24a] In quantum field theory the interaction strength is always modified by the quantum fluctuations represented by the production and reabsorption of virtual particles. The effect of such a quantum cloud depends how closely in distance is the coupling being probed.

[25] This fundamental property of Yang–Mills theory was discovered in 1973 by David Gross (1941–), Frank Wilczek (1951–), and David Politzer (1949–).

[26] Fermi's theory was invented to describe the then only known weak interaction—the neutron's beta decay: $n \to p + e + \bar{\nu}$. The neutron/proton transition can be interpreted at the quark level as a down/up transition because a neutron has valence quarks of (ddu) and a proton has (udu). The quark decay of $d \to u + e + \bar{\nu}$ is directly related to the scattering $\nu + d \to e + u$ when the final antineutrino is turned into a neutrino in the initial state as depicted in Fig. 16.3(b).

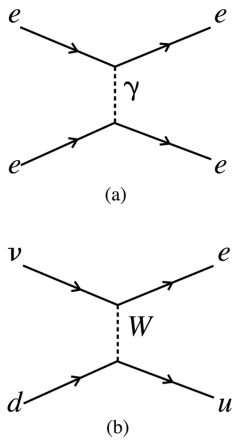

Fig. 16.3 (a) QED description of $e + e \to e + e$ through the exchange of a vector photon; (b) weak scattering $\nu + d \to e + u$ as due to the exchange of a heavy vector boson W. In this sense Fermi's weak interaction theory was based on the analog of QED.

16.5.4 Hidden gauge symmetry and the electroweak interaction

The Standard Model of particle interactions describes the strong, weak, and electroweak interactions. We have already discussed the gauge theories of the electromagnetic and strong interactions. We now discuss the gauge theory of the weak interaction.

The electroweak $SU(2) \times U(1)$ gauge symmetry

In the early 1930s Enrico Fermi proposed a quantum field description of weak interactions. It was modeled on QED. His proposal[26] can be translated and updated in the language of quarks and weak vector bosons as follows. Just as electron–electron scattering $e + e \to e + e$ is described in quantum field theory as due to the exchange of a photon, the weak process of neutrino scattering off a down-quark producing an electron and an up-quark $\nu + d \to e + u$ is due to the exchange of a heavy vector boson W (see Fig. 16.3). Thus instead of the trilinear coupling of $e\bar{e}A_\mu$ we need $\bar{\nu} e W_\mu^+$ and $\bar{u} d W_\mu^+$, etc. (Here, except for the photon, we use particle names for their respective fields.) Namely, unlike

electrodynamics, weak interaction couplings change particle labels, or we can say, "changes the weak interaction charge of a particle" (i.e. we can regard, for example, a neutrino and an electron as different weak interaction states of a leptonic particle).

This feature can be easily accommodated by a nonabelian gauge symmetry. For example, in the weak charge space the electron and electron–neutrino can be placed in an $SU(2)$ doublet of leptons, and in the same way, the up- and down-quarks form a weak doublet:[27]

$$l = \begin{pmatrix} \nu_e \\ e \end{pmatrix} \quad \text{and} \quad q = \begin{pmatrix} u \\ d \end{pmatrix}. \tag{16.72}$$

[27] An $SU(3)$ quark color triplet has components of quarks with different '**colors**', while a quark weak doublet has different '**flavors**'.

The lepton/gauge-boson coupling, much like the electron/photon coupling shown in (16.71), can have the weak charge structure:[28]

$$g_2 \bar{l} \gamma^\mu W_\mu l = g_2 (\bar{\nu}_e,\ \bar{e}) \gamma^\mu \begin{pmatrix} W_\mu^0 & W_\mu^+ \\ W_\mu^- & -W_\mu^0 \end{pmatrix} \begin{pmatrix} \nu_e \\ e \end{pmatrix} \tag{16.73}$$

which (when only particle labels are displayed) contains a trilinear vertex $\bar{\nu} e W_\mu^+$ of an (electrically) charged gauge boson W^+ coupled to an electron and an antineutrino. Similarly, if we replace the lepton by the quark doublet we can have a flavor-changing quark and gauge boson coupling like $\bar{u} d W_\mu^+$. See Fig. 16.1(b).

[28] The discovery of parity violation in weak interactions, through the work of T.D. Lee (1926–), C.N. Yang, and many others in the mid-1950s, stimulated a great deal of progress in particle physics. This symmetry violation can be accommodated elegantly by the stipulation that the above displayed weak doublets involve only the left-handed helicity state of each particle. Parity violation comes about because the left-handed states and right-handed states have different weak charges (i.e. they belong to different types of weak multiplets). See footnote 36 in this chapter for further comments.

One would naturally try to identify the neutral gauge boson W_μ^0 with the photon. However, this would not be feasible because, as can be seen in (16.73), W_μ^0 must couple oppositely to the neutrino and to the electron: $\left(\bar{\nu}\nu W_\mu^0 \text{ and } -\bar{e}e W_\mu^0\right)$; similarly, oppositely to up- and to down-quarks $\left(\bar{u}u W_\mu^0 \text{ and } -\bar{d}d W_\mu^0\right)$, but all these fermions do not have electric charges opposite to their doublet partners. While such a unification of the weak and electromagnetic interactions, involving only a symmetry group of $SU(2)$ with gauge bosons W_μ^0 and $W_\mu^\pm$, does not work out, we can nevertheless achieve a partial unification by the simple addition of another $U(1)$ gauge factor, having an abelian gauge boson $B_{\mu.}$ Namely, we have a unified theory of electromagnetic and weak interactions (electroweak theory) based on the gauge symmetry of $SU(2) \times U(1)$. While neither W_μ^0 nor B_μ can be the photon field, we can assign leptons and quarks with new $U(1)$ charges (called "weak hypercharges") so that one of their linear combination has just the correct coupling property for a photon field A_μ:

$$\begin{aligned} A_\mu &= \cos\theta_{\rm w} B_\mu + \sin\theta_{\rm w} W_\mu^0 \\ Z_\mu^0 &= -\sin\theta_{\rm w} B_\mu + \cos\theta_{\rm w} W_\mu^0 \end{aligned} \tag{16.74}$$

where the mixing angle $\theta_{\rm w}$ is called the Weinberg angle. The combination Z_μ^0, orthogonal to A_μ, is another physical neutral vector boson mediating yet another set of weak interaction processes.[29] We have only a "partial unification" because, to describe two interactions, we still have two independent coupling strengths as each gauge factor comes with an independent gauge coupling:[30] g_1 for $U(1)$ and g_2 for $SU(2)$.

[29] Historically the weak interactions that were first studied are those mediated by the charged vector bosons $W_\mu^\pm$; they are called charged current reactions. Thus one of the firm predictions of this electroweak unification is the existence of Z_μ^0-mediated "neutral current proccesses".

[30] These gauge coupling constants are directly related to the experimentally more accessible constants of electric charge and Weinberg angle

$$e = \frac{g_1 g_2}{\left(g_1^2 + g_2^2\right)^{1/2}} \quad \text{and} \quad \cos\theta_{\rm w} = \frac{g_2}{\left(g_1^2 + g_2^2\right)^{1/2}}.$$

Spontaneous symmetry breaking

Mass problems in electroweak gauge theory Gauge boson mass terms are forbidden by gauge invariance. While gluons are massless, the QCD gauge interaction is still effectively short-ranged because color particles are confined within short distances. Now if the weak interaction is to be formulated as a gauge interaction with the weak vector bosons $W^{\pm}$ and Z^0 identified as gauge bosons, they would also be required to be massless. But observationally the weak interaction is very short ranged (even shorter than the strong interaction range) hence the interaction transmitters must be massive.[31] This had been a major obstacle in the formulation of weak interaction as a gauge force. If we simply insert vector-boson mass terms (hence breaking the weak gauge symmetry), one would end up with uncontrollable ultraviolet divergences (technically speaking, making the theory unrenormalizable). This is the gauge boson mass problem. There is also a fermion mass problem. Symmetry (whether global or local) implies mass degeneracy of particles belonging to the same symmetry multiplet. Thus all three color states of the quark triplet (16.64) have identical masses. But to have a gauge theory of the weak interaction with the weak doublets as shown in (16.72), such fermion mass degeneracy would contradict observation, as the electron and the neutrino have different masses $m_e \neq m_\nu$, so have the up- and down-quarks $m_u \neq m_d$.

[31]The relation between the interaction range and the mass of the mediating particle is discussed in Section 6.4.2.

Symmetry is hidden The mass problems discussed above were solved eventually by spontaneous symmetry breaking (SSB).[32] This is the possibility that physics equations with symmetry may have asymmetric solutions. A ferromagnet is a familiar example: above the critical temperature ($T > T_c$) it is a system of randomly oriented magnetic dipoles, reflecting the rotational symmetry of the physics equation describing such a system. But, below the critical temperature, all the dipoles are aligned in one particular direction—breaking the rotational symmetry even though the underlying physics equation is rotational invariant. This comes about because in a certain parameter space the theory would yield a ground state, instead being symmetric (i.e. a symmetry singlet state as shown in Fig. 16.4a), being a set of degenerate states related to each other through the symmetry transformation (as shown in Fig. 16.4b). Since the physical ground state (the vacuum state in a quantum field system) has to be unique, its selection, out of the degenerate set, must necessarily break the symmetry. Thus, the ground state, a solution to the symmetrical equation, is itself asymmetric. The rest of the physics (built on this vacuum state) will also have asymmetric features such as nondegenerate masses in a multiplet, etc. Since the underlying equations are symmetric while the outward appearance is not, SSB can best be described as the case of a "symmetry being hidden".

[32]Important contribtutions were made by P.W. Anderson (1923–), Y. Nambu (1921–), J. Goldstone (1933–), S. Weinberg (1933–), J. Schwinger, P.W. Higgs (1929–), and many others.

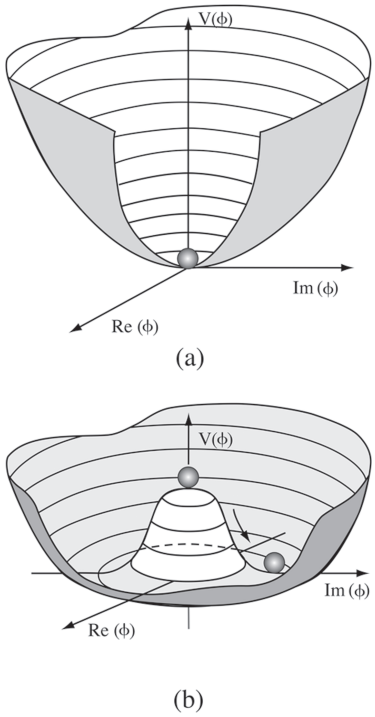

Fig. 16.4 The potential energy function $V(\phi)$ illustrates the occurrence of spontaneous symmetry breaking. (a) Normal symmetry realization: the ground state is a symmetry singlet. (b) A case of hidden symmetry: the ground state is a set of degenerate states—the circle at the bottom of the wine-bottle shaped energy surface; the selection of the true vacuum as one point in this circle breaks the symmetry. The small ball indicates the location of the physical ground state.

The Higgs sector A hidden symmetry scenario can take place in both global and local symmetries. For global symmetry, one has the interesting consequence that such a hidden symmetry scheme leads to the existence of massless scalar bosons (called Nambu–Goldstone bosons). One would then be concerned with the following unpalatable prospect of a theory with local

symmetry: not only do we have the unwanted massless vector gauge bosons, we also have these unwanted massless scalar bosons. As it turns out, in the realization of SSB in gauge theories (**the Higgs mechanism**) each of these two ills is the cure of the other. The massless Goldstone scalar combines with the (two states of) a massless vector boson to form (the three states of) a massive vector boson. In the end we have a hidden gauge symmetry without any unwanted massless states. The $SU(2) \times U(1)$ electroweak gauge theory starts out in the symmetric limit with all particles (gauge bosons, leptons, and quarks) being massless. The explicit realization of the Higgs mechanism involves the introduction of a doublet of elementary (complex) scalar boson fields $\left(\phi^+, \phi^0\right)$, and their dynamics is such that the ground state value of ϕ^0 is a nonvanishing constant. The vacuum is permeated with this constant scalar field. All the particles, originally massless, gain their respective masses while propagating in this vacuum. The electroweak theory has a structure such that the photon gauge particle (as well as the neutrino states) remains massless. However this scalar sector, often referred to as the Higgs sector, is less constrained by the symmetry of the theory. In particular we are free to adjust the couplings of the scalars to leptons and quarks in order to obtain their respective observed masses. (Namely the Standard Model does not predict the lepton and quark masses.) Of the complex doublet $\left(\phi^+, \phi^0\right)$ we have four independent scalar bosons; while $\phi^\pm$ and one of $\phi^{0\prime}s$ are "eaten" by the gauge bosons to make three massive vector bosons of $W^\pm$and Z^0, the remaining scalar boson is a real massive spin-0 particle. This 'Higgs boson' with characteristic couplings (to leptons, quarks, photons, and other intermediate vector bosons) should be an observable signature of the SSB feature of the electroweak theory.

The development of electroweak gauge theory

The Glashow–Weinberg–Salam model Many have contributed to the development of the gauge theory of electroweak interactions. We mention some milestones. Sheldon Glashow (1932–) was the first one in 1957 to write down an $SU(2) \times U(1)$ gauge theory and also made major contributions later on in building a consistent quark sector of the theory. However in the original Glashow model, the vector boson masses were introduced by hand, hence it was not a self-consistent quantum field theory. In 1967 Steven Weinberg formulated an electroweak gauge theory of leptons with gauge bosons and electron masses generated by the Higgs mechanism. At about the same time Abdus Salam (1926–96) presented an electroweak gauge theory with spontaneous symmetry breaking as well, although not in a formal journal publication. Their results did not generate great enthusiasm right away in the physics community because the quantization[33] and renormalization[34] of nonabelian gauge theories were still being worked out in those years.

Yang–Mill theories are renormalizable with or without SSB For almost two decades (1950s and 1960s), there was in fact a great deal of pessimism in the physics community that quantum field theory could be the proper framework for the study of strong and weak interactions. The strong interaction did not appear to have a small coupling and its field equation could not be solved

[33]The quantization of Yang–Mills theory, because of its many redundant degrees of freedom, is highly nontrivial. Its consistent program was finally achieved through the work of many, by Bryce DeWitt (1923–2004), R.P. Feynman, Ludvig Faddeev (1934–) and Victor Popov (1937–94), *et al.*

[34]One of the important steps in the renormalization is the implementation of the regularization procedure that renders the divergent integrals finite so that the calculation is well defined for further mathematical manipulations. The renormalizability of a theory with symmetry depends critically on the cancellation of divergences as enforced by symmetry relations. The **dimensional regularization scheme**, by going to a lower spacetime dimension, makes theory finite without violating its symmetry properties. This elegant procedure was invented independently by several groups, among them G. 't Hooft (1946–) and M.J.G. Veltman (1931–).

by the only known method: perturbation theory. Without knowing their solutions, one did not know how to test such theories. While the weak interaction had features of a gauge interaction and the perturbation should be applicable, it was generally thought that quantum theory with massive vector bosons was not renormalizable. It is in this light that one must appreciate the result obtained in 1971 by Gerard 't Hooft, a student of Martinus Veltman, proving that Yang–Mills theory was renormalizable, with or without spontaneous symmetry breaking. The significance of this achievement was appreciated very quickly by their worldwide physics colleagues. This transformed the whole field of theoretical particle physics and brought about the renaissance of quantum field theory in the 1970s.

We discussed QCD before electroweak theory, because QCD, without the need of a hidden symmetry, is a simpler gauge theory to present. Historically the nonabelian gauge theory for the weak interaction was successfully developed first. Politzer, Gross, and Wilczek then proved that Yang–Mills theory, and only Yang–Mills theory, has the property of asymptotic freedom. That allowed the QCD quantum field theory of strong interactions at the short-distance regime to be solved perturbatively and tested experimentally.

16.5.5 The Standard Model and beyond

The Standard Model of particle interactions (Table 16.1) is a gauge theory based on the symmetry group of $SU(3) \times SU(2) \times U(1)$. QCD is the $SU(3)$ gauge theory for the strong interaction. The $SU(2) \times U(1)$ gauge theory with spontaneous symmetry breaking describes the electroweak interaction. Even though their coupling strengths are the same, the weak interaction appears to be much weaker than the electromagnetic force because its effects are usually suppressed by the large masses of the $W^{\pm}$ and Z bosons.

Grand unified gauge theories

The Standard Model has been remarkably successful in its confrontation with experiment tests. Nevertheless it does not explain why the three generations of leptons and quarks have the same charge and representation assignments.[35] Furthermore, the theory must be specified by 18 parameters: three gauge couplings, the SSB energy scale (which fixes the vector boson masses), three lepton masses, as well as six masses and four angles of a complex mixing matrix of the quarks, and, finally, the Higgs boson mass. The consensus is that the Standard

[35] The theoretical structures for each of the three generations (e, ν_e, u, d), (μ, ν_μ, c, s), and (τ, ν_τ, t, b) in the Standard Model are identical.

Table 16.1 Gauge symmetry, gauge bosons, and gauge couplings of the Standard Model.

interactions	sym group	vector gauge fields	partial unification
Electromagnetic	$U(1)$	photon A_μ	$SU(2) \times U(1)$ electroweak e and $\theta_{\rm w}$
Weak	$SU(2)$	weak vector bosons $W^{\pm}_\mu$, Z_μ	
Strong	$SU(3)$	gluons G^a_μ, $a = 1, 2, .. 8$	g_s

Model is only a low-energy effective theory of some more fundamental theory with an intrinsic energy scale much higher than the electroweak scale.

As a first step going beyond the Standard Model, people have explored the possibility of 'grand unification' of strong, weak, and electromagnetic interactions in the framework of larger groups that are 'simple' (with only one gauge coupling) that contain $SU(3) \times SU(2) \times U(1)$ as their subgroup. Namely, in the Standard Model the three interactions are still described by three separate gauge groups with distinctive coupling strengths. In a more unified simple gauge group there is only one coupling strength—truly one gauge interaction. This unified strength at some very high 'grand unified' energy scale $\Lambda_{\rm GU}$ can evolve into the distinctive couplings of strong, weak, and electromagnetic couplings at a lower energy scale if there is another spontaneous symmetry breaking taking place at $\Lambda_{\rm GU}$ with all gauge bosons other than those belonging to the $SU(3) \times SU(2) \times U(1)$ group gaining masses $O(\Lambda_{\rm GU})$. The decoupling of these heavy particles implies that the subgroup couplings g_1, g_2, and g_3 will evolve differently below the $\Lambda_{\rm GU}$ scale, giving rise to the observed different interaction strengths for the strong, weak, and electromagnetic forces observed in our more familiar low-energy scales (see Fig. 16.5).

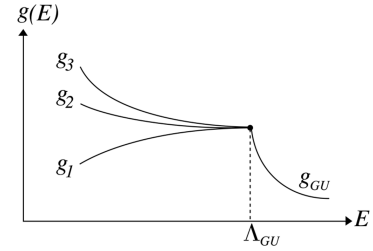

Fig. 16.5 Running coupling strengths as a function of energy. Coupling constant unification occurs at some super-high energy ($\Lambda_{\rm GU} = 10^{16}$ GeV?). Spontaneous symmetry breaking of the unification gauge group with a single gauge coupling $g_{\rm GU}$ will cause the gauge couplings of subgroup $SU(3) \times SU(2) \times U(1)$ to evolve differently towards lower-energy regimes, giving rise to the different interaction strengths as observed of the strong, weak, and electromagnetic forces.

Successful GUTs such as the gauge theory based on $SU(5)$ have been constructed; they can explain (as the coupling unification discussed above) why the strong interaction is strong and why the weak interaction is weak. Moreover all the quark lepton gauge charges can be understood based on a simple assignment of GUT charges for these fermions.[36] Their description can in fact be improved upon with the introduction of supersymmetry; in particular the precise coupling unification discussed above can come about only by the inclusion of supersymmetric particles. This program is still a work in progress; it very much needs guidance from experimental discoveries. In this connection, we comment below on the distinction that Einstein made between constructive theories versus theories of principle.

[36]That two-component fermions form the fundamental representations of the Lorentz group provides us with a natural explanation of parity violation by fundamental interactions. That QCD turns out to be parity conserving is explained by the GUT charge assignment which just leads to the same $SU(3)$ color charges for the left-handed and right-handed quarks.

The Standard Model as a constructive theory and as a theory of principle

Abraham Pais in his Einstein biography wrote[37]

. . . a distinction that Einstein liked to make between two kinds of physical theories. Most theories, according to Einstein, are constructive, they interpret complex phenomena in terms of relatively simple propositions. An example is the kinetic theory of gases, in which the mechanical, thermal, and diffusional properties of gases are reduced to molecular interactions and motions . . . then there are the theories of principle, which use the analytic rather than the synthetic method . . . An example is the impossibility of a perpetuum mobile *in thermodynamics. Then Einstein went on to say, 'The theory of relativity is a theory of principle'.*

[37]See Pais (1982, p. 27), based on Einstein's letter to his gymnasium teacher H. Friedmann, March 18, 1929.

We would like to suggest that the Standard Model of elementary particle interactions is a good example of a theory that is **both** a constructive theory **and** a theory of principle.

The discovery of the quark and lepton as the basic constituents of matter, and that of the symmetry groups of $SU(3)$ and $SU(2) \times U(1)$, followed the

practice of a constructive theory with its trial-and-error theoretical propositions followed by the experimental checks.

Einstein and the Standard Model Einstein did not participate directly in the construction of the Standard Model as described above. Also, the Standard Model is an example of a quantum field theory, which Einstein never accepted as an acceptable theoretical framework. However the influence of his idea has been of paramount importance in the successful creation of the Standard Model. Besides the fundamental importance of special relativity, photons, and Bose–Einstein statistics to particle physics, the use of local symmetry to generate dynamics (the gauge principle) is very much in the spirit of Einstein's theory of principle as represented in particular by his masterful deployment of the invariance principle, and the equivalence principle. This approach of utilizing an overarching principle in the search of the new patterns in Nature will become even more relevant as we explore physical realms that are ever more inaccessible to direct experimentation.