where $\vec{\mathbf{E}}$ and $\vec{\mathbf{B}}$ are respectively the electric and magnetic fields; ρ and $\vec{\mathbf{J}}$ are the charge density and electric current. The continuity equation which dictates the conservation of charge,

$$\nabla \cdot \vec{\mathbf{J}} + \partial\rho/\partial t = 0, \tag{5}$$

indicates that Maxwell's equations describe a local theory, since you cannot destroy a charge locally and recreate it at a distant point instantaneously. The concept that the theory should be local is the cornerstone of the gauge theory used in quantum field theory, resulting in the Yang-Mills theory, the basis of the standard model.

Maxwell realized that since

$$\nabla \cdot \nabla \times \vec{\mathbf{B}} = 0, \tag{6}$$

equation (3) is inconsistent with (5); he altered (3) to read

$$\nabla \times \vec{\mathbf{B}} = \vec{\mathbf{J}} + \partial\vec{\mathbf{E}}/\partial t \tag{3'}$$

Thus a local conservation law mandated the addition of the $\partial\vec{\mathbf{E}}/\partial t$ term. Although equations (1), (2), (3′) and (4) are collectively known as Maxwell's equations, Maxwell himself was responsible only for (3′).

Maxwell calculated the speed of a wave propagated by the final set of equations and found its velocity very close to the speed of light. He thus hypothesized that light was an electromagnetic wave. Since the curl of a vector cannot be calculated in two dimensions, Maxwell's equations indicate that light, as we know it, cannot exist in a two-dimensional world. This is the first clue that electromagnetism is bound up with geometry. In fact, equation (6) is the vector analysis equivalent of the differential geometry result stating that if β is a p-form and $\mathbf{d}\beta$ is its exterior derivative, then $\mathbf{d(d}\beta\mathbf{)}$ or $\mathbf{d}^2\beta = 0$.

Unlike the laws of Newtonian mechanics, Maxwell's equations carry over to relativistic frames. The nonhomogeneous equations, (1) and (3′), become

$$\partial_\mu F^{\mu\nu} = J^\nu; \tag{7}$$

while the homogeneous equations, (2) and (4), become

$$\epsilon^{\alpha\beta\gamma\delta}\partial_\beta F_{\gamma\delta} = 0, \tag{8}$$

where $\epsilon^{\alpha\beta\gamma\delta}$ is the Levi-Civita symbol, $F^{\mu\nu} = \partial^\mu A^\nu - \partial^\nu A^\mu$, A^0 is the scalar potential, and A^i's (i = 1, 2, 3) the components of the vector potential $\vec{\mathbf{A}}$. Note that both equations (7) and (8) are manifestly covariant. In a later section we will show that equation (8) is due to the principle $d^2\omega = 0$, where ω is a p-form.

The following remark can be understood in differential geometry terms by using the generalized Stoke's theorem: $\int_M d\omega = \int_{\partial M}\omega$, where ω is a p-form and M is a $p+1$ dimension oriented manifold with boundary ∂M. For the purposes of this article a manifold is simply a space that is locally Euclidean. The remark is that equation (8), i.e., $d^2 = 0$, is related to the principle that the boundary of a region has no boundary, i.e., $\partial^2 M = \emptyset$. This last principle once noticed, is geometrically obvious. The fact that $d^2 = 0$ complements $\partial^2 M = \emptyset$ can be seen from $\int_{\partial^2 M}\omega = \int_{\partial M} d\omega = \int_M d^2\omega = 0$. Indeed Yang[1] notes that equation (8) complements $\partial^2 M = \emptyset$ which, as we show later, is in keeping with his championing of geometry's influence on field theory.

Gauge Invariance

In a 1918 article Hermann Weyl[2] tried to combine electromagnetism and gravity by requiring the theory to be invariant under a local scale change of the metric, i.e., $g_{\mu\nu} \to g_{\mu\nu}e^{\alpha(x)}$, where x is a 4-vector. This attempt was unsuccessful and was criticized by Einstein for being inconsistent with observed physical results. It predicted that a vector parallel transported from point p to q would have a length that was path dependent. Similarly, the time interval between ticks of a clock would also depend on the path on which the clock was transported. The article did, however, introduce

- the term "gauge invariance"; his term was *Eichinvarianz*. It refers to invariance under his scale change. The first use of "gauge invariance" in English[3] was in Weyl's translation[4] of his famous 1929 paper.
- the geometric interpretation of electromagnetism.
- the beginnings of nonabelian gauge theory. The similarity of Weyl's theory to nonabelian gauge theory is more striking in his 1929 paper.

By 1929 Maxwell's equations had been combined with quantum mechanics to produce the start of quantum electrodynamics. In his 1929 article[5] Weyl turned from trying to unify electromagnetism and gravity to following a suggestion originally thought to have been made by Fritz London in his 1927 article[6] and introduced as a phase factor an exponential in which the phase α is preceded by the imaginary unit i, e.g., $e^{+iq\alpha(\mathbf{x})}$, in the wave function for the wave equations (for instance, the Dirac equation is $(i\gamma^\mu\partial_\mu - m)\psi = 0$). It is here that Weyl correctly formulated gauge theory as a

[1] *Yang, C. N.,* Physics Today **6** *(1980), 42.*

[2] *Weyl, Hermann,* Sitzwingsber. Preuss. Akad., *Berlin, 1919, p. 465.*

[3] *See Jackson, J. D., and Okun, L. B.,* Rev. Mod. Phys. **73** *(2001), 663.*

[4] *Weyl, H.,* Proc. Natl. Acad. Sci. **15** *(1929), 32.*

[5] *Weyl, Hermann,* Z. f Phys. **330** *(1929), 56.*

[6] *London, Fritz,* Z. f Phys. **42** *(1927), 375.*

symmetry principle from which electromagnetism could be derived. It was to become the driving force in the development of quantum field theory. In their 2001 *Rev. Mod. Phys.* paper, Jackson and Okun point out that in a 1926 paper[7] predating London's, Fock showed that for a quantum theory of charged particles interacting with the electromagnetic field, invariance under a gauge transformation of the potentials required multiplication of the wave function by the now well-known phase factor. Many subsequent authors incorrectly cited the date of Fock's paper as 1927. Weyl's 1929 article—along with his 1918 one, and Fock's and London's, and other key articles—appears in translation in a work by O'Raifeartaigh[8] with his comments. Yang[9] discusses Weyl's gauge theory results as reported by Pauli[10] as a source for Yang-Mills gauge theory (although Yang did not find out until much later that these were Weyl's results):

> ... I was very much impressed with the idea that charge conservation was related to the invariance of the theory under phase changes and even more impressed with the fact the gauge-invariance determined all the electromagnetic interactions....

For the wave equations to be gauge invariant, i.e., have the same form after the gauge transformation as before, the local phase transformation $\psi(\mathbf{x}) \to \psi(x)e^{+iq\alpha(\mathbf{x})}$ has to be accompanied by the local gauge transformation

$$\mathbf{A}_\mu \to \mathbf{A}_\mu - q^{-1}\partial_\mu \alpha(\mathbf{x}). \tag{9}$$

(The phase and gauge transformations are local because $\alpha(\mathbf{x})$ is a function of $\mathbf{x}$.) This dictates that the ∂_μ in the wave equations be replaced by $\partial_\mu + iqA_\mu$ in order for the $\partial_\mu \alpha(\mathbf{x})$ terms to cancel each other out. Thus gauge invariance determines the type of interaction; here, the inclusion of the vector potential. This is called the *gauge principle,* and A_μ is called the *gauge field* or *gauge potential.* Gauge invariance is also called *gauge symmetry.* In electromagnetism $\mathbf{A}$ is the space-time vector potential representing the photon field, while in electroweak theory $\mathbf{A}$ represents the intermediate vector bosons $W^\pm$ and Z^0 fields; in the strong interaction, $\mathbf{A}$ represents the colored gluon fields. The fact that the q in $\psi(x)e^{+iq\alpha(\mathbf{x})}$ must be the same as the q in $\partial_\mu + iqA_\mu$ to insure gauge invariance means that the charge q must be conserved.[11] Thus gauge invariance dictates charge conservation. By Noether's theorem, a conserved current is associated with a symmetry. Here the symmetry is the nonphysical rotation invariance in an internal space called a *fiber*. In electromagnetism the rotations form the group U(1), the group of unitary 1-dimensional matrices. U(1) is an example of a *gauge group,* and the fiber is S^1, the circle.

A *fiber bundle* is determined by two manifolds and the structure group G which acts on the fiber: the first manifold, called the total space E, consists of many copies of the fiber F—one for each point in the second manifold, the base manifold M which for our discussion is the space-time manifold. The fibers are said to project down to the base manifold. A *principal fiber bundle*[12] is a fiber bundle in which the structure group G acts on the total space E in such a way that each fiber is mapped onto itself and the action of an individual fiber looks like the action of the structure group on itself by left-translation. In particular, the fiber F is diffeomorphic to the structure group G.

The gauge principle shows how electromagnetism can be introduced into quantum mechanics. The transformation $\partial_\mu \to \partial_\mu + iqA_\mu$ is also called the *minimal principle* and the operation $\partial_\mu + iqA_\mu$ is the covariant derivative of differential geometry, $\mathbf{D} = \mathbf{d} + \mathbf{iqA}$, where $\mathbf{A}$ is the connection on a fiber bundle. A connection on a fiber bundle allows one to identify fibers over points $b_i \in M$ via parallel transport along a path γ from b_1 to b_2. In general, the particular identification is path dependent. It turns out that the parallel transport depends only on the homotopy class of the path if and only if the curvature of the connection vanishes identically. Recall that two paths are homotopic if one can be deformed continuously onto the other keeping the end points fixed.

In his 1929 paper Weyl also includes an expression for the curvature Ω of the connection $\mathbf{A}$, namely, Cartan's second structural equation, which in modern differential geometry notation is $\Omega = \mathbf{dA} + \mathbf{A} \wedge \mathbf{A}$. It is the same form as the equation used by Yang and Mills, which in modern notation is $\Omega = \mathbf{dA} + \frac{1}{2}[\mathbf{A}, \mathbf{A}]$. Since the transformations in (9) form an abelian group U(1), the space-time vector potential $\mathbf{A}$ commutes with itself. Thus in electromagnetism the curvature of the connection $\mathbf{A}$ is just

$$\Omega = \mathbf{dA} \tag{10}$$

[7] *Fock, V.,* Z. Phys. **39** *(1926), 226.*

[8] *O'Raifeartaigh, L.,* The Dawning of Gauge Theory, *Princeton University,* 1997.

[9] *Yang, C. N., "Hermann Weyl's contribution to physics",* Hermann Weyl *(ed. Chandrasekharan, K.), Springer-Verlag, 1980.*

[10] *Pauli, W.,* Rev. Mod. Phys. **13** *(1941), 203.*

[11] *See section 2.6 of Aitchison, I. J. R., and Hey, A. J. G.,* Gauge Theories in Particle Physics, *Adam Hilger, 1989.*

[12] *In giving these definitions, we restrict our attention to the smooth manifolds, which are adequate for our discussion.*

which, as we will see in the next section, is the field strength $\mathbf{F}$ defined as $\mathbf{F} = \mathbf{dA}$.

Differential Geometry

Differential geometry—principally developed by Levi-Civita, Cartan, Poincaré, de Rham, Whitney, Hodge, Chern, Steenrod and Ehresmann—led to the development of fiber bundle theory, which is used in explaining the geometric content of Maxwell's equations. It was later used to explain Yang-Mills theory and to develop string theory. The successes of gauge theory in physics sparked mathematicians' interest in it. In the 1970s Sir Michael Atiyah initiated the study of the mathematics of the Yang-Mills equations; and in 1983 his student Simon Donaldson, using Yang-Mills theory, discovered a unique property of smooth manifolds[13] in $\mathbb{R}^4$. Michael Freedman went on to prove that there exist multiple exotic differential structures only on $\mathbb{R}^4$. It is known that in other dimensions the standard differential structure on $\mathbb{R}^n$ is unique.

In 1959 Aharonov and Bohm[14] established the primacy of the vector potential by proposing an electron diffraction experiment to demonstrate a quantum mechanical effect: A long solenoid lies behind a wall with two slits and is positioned between the slits and parallel to them. An electron source in front of the wall emits electrons that follow two paths: one path through the upper slit and the other path through the lower slit. The first electron path flows above the solenoid, and the other path flows below it. The solenoid is small enough so that when no current flows through it, the solenoid does not interfere with the electrons' flow. The two paths converge and form a diffraction pattern on a screen behind the solenoid. When the current is turned on, there is no magnetic or electric field outside the solenoid, so the electrons cannot be affected by these fields. However, there is a vector potential $\vec{\mathbf{A}}$, and it affects the interference pattern on the screen. Thus Einstein's objection to Weyl's 1918 paper can be understood as saying that there is no Aharonov-Bohm effect for gravity. Because of the necessary presence of the solenoid, the upper path cannot be continuously deformed into the lower one. Therefore, the two paths are not homotopically equivalent.

The solution of $(1/2m)(-i\hbar\nabla - q\mathbf{A}/c)^2\psi + qV\psi = E\psi$, the time-independent Schrödinger's equation for a charged particle, is $\psi_0(\mathbf{x})e^{(iq/c\hbar)\int^{\mathbf{s(x)}} \mathbf{A(y)}d\mathbf{s}'(\mathbf{y})}$ where $\psi_0(\mathbf{x})$ is the solution of the equation for $\mathbf{A}$ equals zero and $\mathbf{s(x)}$ represents each of the two paths. Here c is the speed of light, and $\hbar$ is Plank's constant divided by 2π. The interference term in the superposition of the solution for the upper path and of that for the lower path produces a difference in the phase of the electrons' wave function called a *phase shift*. Here the phase shift is $(q/c\hbar)\oint \mathbf{A}(\mathbf{x})d\mathbf{x}$. By Stoke's theorem, the phase shift is $(q/c\hbar)\phi$, where ϕ is the magnetic flux in the solenoid, $\int \vec{\mathbf{B}} \cdot d\vec{\mathbf{S}}$. Mathematically, their proposal corresponds to the fact that even if the curvature (the electromagnetic field strength) of the connection vanishes (as it does outside the solenoid), parallel transport along nonhomotopic paths can still be path-dependent, producing a shift in the diffraction pattern.

Chambers[15] performed an experiment to test the Aharonov and Bohm (AB) effect. The experiment, however, was criticized because of leakage from the solenoid. Tonomura[16] et al. performed beautiful experiments that indeed verified the AB prediction. Wu and Yang[17] analyzed the prediction of Aharonov and Bohm and comment that different phase shifts $(q/c\hbar)\phi$ may describe the same interference pattern, whereas the phase factor $e^{(iq/c\hbar)\phi}$ provides a unique description as we now show. The equation $e^{2\pi Ni} = 1$, where N is an integer, means that

$$e^{(iq/c\hbar)(\phi+2\pi Nc\hbar/q)} = e^{(iq/c\hbar)\phi}e^{2\pi Ni} = e^{(iq/c\hbar)\phi}.$$

Thus flux of $\phi, \phi + 2\pi c\hbar/q, \phi + 4\pi c\hbar/q...$ all describe the same interference pattern. Moreover, they introduced a dictionary relating gauge theory terminology to bundle terminology. For instance, the gauge theory phase factor corresponds to the bundle parallel transport, and as we shall see, the Yang-Mills gauge potential corresponds to a connection on a principal fiber bundle.

Let's see how by using the primacy of the four-vector potential $\mathbf{A}$, we can derive the homogeneous Maxwell's equations from differential geometry simply by using the gauge transformation. Then we will get the nonhomogeneous Maxwell's equations using the fact that our world is a four-dimensional (space-time) world.

We will also show that Maxwell's equations are invariant under the transformations $\mathbf{A}_\mu \to \mathbf{A}_\mu + \partial_\mu\alpha(\mathbf{x})$, or expressed in differential geometry terms, $\mathbf{A} \to \mathbf{A} + \mathbf{d}\alpha(\mathbf{x})$. We want $\alpha(\mathbf{x})$ to vanish when a function of $\mathbf{A}$ is assigned to the $\vec{\mathbf{E}}$ and $\vec{\mathbf{B}}$ fields. Taking the exterior derivative of $\mathbf{A}$ will do this, since $\mathbf{d}^2\alpha(\mathbf{x}) = \mathbf{0}$. Set $\mathbf{A}$ to be the 1-form $A = -A_0\mathbf{d}t + A_x\mathbf{d}x + A_y\mathbf{d}y + A_z\mathbf{d}z$. Evaluating $\mathbf{dA}$ and realizing that the wedge product $\mathbf{d}x^i \wedge \mathbf{d}x^j = -\mathbf{d}x^j \wedge \mathbf{d}x^i$ and therefore $\mathbf{d}x^j \wedge \mathbf{d}x^j = 0$, where $\mathbf{d}x^0$ is $\mathbf{d}t$, $\mathbf{d}x^1$ is $\mathbf{d}x$, $\mathbf{d}x^2$ is $\mathbf{d}y$ and $\mathbf{d}x^3$ is $\mathbf{d}z$, produces a 2-form consisting of terms such as $(\partial_x A_0 +$

[13] *Donaldson, S. K.,* Bull. Amer. Math. Soc. **8** *(1983), 81.*

[14] *Aharonov, Y., and Bohm, D.,* Phys. Rev. **115** *(1959), 485.*

[15] *Chambers, R. G.,* Phys. Rev. Lett. **5** *(1960), 3.*

[16] *Tonomura, Akira, et. al.,* Phys. Rev. Lett. **48** *(1982), 1443; and* Phys. Rev. Lett. **56** *(1986), 792.*

[17] *Wu, T. T., and Yang, C. N.,* Phys. Rev. D **12***(1975), 3845.*

$\partial_t A_x)dtdx$ and $(\partial_x A_y - \partial_y A_x)dxdy$. When all the components are evaluated, these terms become respectively $\nabla \mathbf{A_0} + \partial\vec{\mathbf{A}}/\partial_t$ and $\nabla \times \vec{\mathbf{A}}$. The analysis up to now has been purely mathematical. To give it physical significance we associate these terms with the field strengths $\vec{\mathbf{B}}$ and $\vec{\mathbf{E}}$. In electromagnetic theory, two fundamental principles are $\nabla \cdot \vec{\mathbf{B}} = 0$ (no magnetic monopoles) and for time-independent fields $\vec{\mathbf{E}} = -\nabla A_0$ (the electromagnetic field is the gradient of the scalar potential), so consistency dictates that in the time-dependent case, we assign the two terms to $\vec{\mathbf{B}}$ and $\vec{\mathbf{E}}$ respectively:

$$\vec{\mathbf{B}} = \nabla \times \vec{\mathbf{A}} \text{ and } \vec{\mathbf{E}} = -\nabla A_0 - \partial_t \vec{\mathbf{A}}. \tag{11}$$

The gradient, curl, and divergence are spatial operators—they involve the differentials dx, dy, and dz. The exterior derivative of a scalar is the gradient, the exterior derivative of a spatial 1-form is the curl, and the exterior derivative of a spatial two-form is the divergence. In the 1-form $\mathbf{A}$, the $-A_0\mathbf{d}t$ is a spatial scalar and when the exterior derivative is applied gives rise to ∇A_0. The remaining terms in $\mathbf{A}$ are the coefficients of dx^i constituting a spatial 1-form and thus produce $\nabla \times \vec{\mathbf{A}}$.

We define the field strength $\mathbf{F}$ as $\mathbf{F} = \mathbf{dA}$, and from equation (10) we see that the field strength is the curvature of the connection $\mathbf{A}$. Using the equations in (11) and the 2-form $\mathbf{dA}$, we get

$$\begin{aligned} F = {} & E_x dx\mathbf{d}t + E_y \mathbf{d}y\mathbf{d}t + E_z \mathbf{d}z\mathbf{d}t \\ & + B_x \mathbf{d}y\mathbf{d}z + B_y \mathbf{d}z\mathbf{d}x + B_{\mathbf{d}} x\mathbf{d}y \end{aligned} \tag{12}$$

where, for example, $\mathbf{d}x\mathbf{d}t$ is the wedge product $\mathbf{d}x \wedge \mathbf{d}t$. Since $\mathbf{d}^2\mathbf{A} = \mathbf{0}$

$$\mathbf{dF} = \mathbf{0}. \tag{13}$$

Evaluating $\mathbf{dF}$ gives the homogeneous Maxwell's equations. In equation (12), since the $\mathbf{E}$ part is a spatial 1-form, when the exterior derivative is applied it produces the $\nabla \times \vec{\mathbf{E}}$ part of Maxwell's homogeneous equations. Since the $\mathbf{B}$ part of equation (12) is a spatial 2-form, it results in the $\nabla \cdot \vec{\mathbf{B}}$ part. Since $\mathbf{dF} = \mathbf{0}$, $\mathbf{F}$ is said to be a closed 2-form.

To get the expression for the nonhomogeneous Maxwell's equations, i.e., the equivalent of equation (7), we use

$$\mathbf{J} = \rho\mathbf{d}t + J_x\mathbf{d}x + J_y\mathbf{d}y + J_z\mathbf{d}z \tag{14}$$

and calculate the Hodge dual using the Hodge star operator. The Hodge duals are defined[18] by $*F_{\alpha\beta} = 1/2\epsilon_{\alpha\beta\gamma\delta}F^{\gamma\delta}$ and $*J_{\alpha\beta\gamma} = \epsilon_{\alpha\beta\gamma\delta}J^{\delta}$. The Hodge star[19] operates on the differentials in equations (12) and (14) using $*(dx^i dt) = dx^j dx^k$ and $*(dx^j dx^k) = -dx^i dt$, where i, j, and k refer to x, y, and z and are taken in cyclic order. The metric used is (-+++). Thus the Hodge star takes a spatial 1-form $dx^i dt$ into a spatial 2-form and vice versa with a sign change.

The nonhomogeneous Maxwell's equations are then expressed by

$$\mathbf{d}{*}\mathbf{F} = 0 \text{ (source-free)} \tag{15}$$

$$\mathbf{d}{*}\mathbf{F} = {*}\mathbf{J} \text{ (non-source-free)} \tag{15$'$}$$

where the 2-form $*\mathbf{F}$ and the 3-form $*\mathbf{J}$ are respectively the Hodge duals of $\mathbf{F}$ and $\mathbf{J}$. $*\mathbf{F}$ and $*\mathbf{J}$ are defined as

$$\begin{aligned} *\mathbf{F} = {} & -B_x\mathbf{d}x\mathbf{d}t - B_y\mathbf{d}y\mathbf{d}t - B_z\mathbf{d}z\mathbf{d}t \\ & + E_x\mathbf{d}y\mathbf{d}z + E_y\mathbf{d}z\mathbf{d}x + E_z\mathbf{d}x\mathbf{d}y \end{aligned} \tag{16}$$

$$\begin{aligned} *\mathbf{J} = {} & \rho\mathbf{d}x\mathbf{d}y\mathbf{d}z - J_x\mathbf{d}t\mathbf{d}y\mathbf{d}z \\ & - J_y\mathbf{d}t\mathbf{d}z\mathbf{d}x - J_z\mathbf{d}t\mathbf{d}x\mathbf{d}y \end{aligned} \tag{17}$$

Thus the Hodge star reverses the roles of $\vec{\mathbf{E}}$ and $\vec{\mathbf{B}}$ from what they were in $\mathbf{F}$. In $*\mathbf{F}$ the coefficient of the spatial 1-form is now $-\vec{\mathbf{B}}$, which will produce the curl in the nonhomogeneous Maxwell's equations; and the coefficient of the spatial 2-form is $\vec{\mathbf{E}}$, which will produce the divergence. In $\mathbb{R}^n$ the Hodge star operation on a p-form produces an (n-p)-form. Thus the form of Maxwell's equations is dictated by the fact that we live in a four-dimensional world. When the 1-form $\mathbf{A}$ undergoes the local gauge transformation $\mathbf{A} \to \mathbf{A} + \mathbf{d}\alpha(\mathbf{x})$, $\mathbf{dA}$ remains the same, since $\mathbf{d}^2\alpha = 0$. Since $\vec{\mathbf{B}}$ and $\vec{\mathbf{E}}$ are unchanged, Maxwell's theory is gauge invariant.

The Dirac and Electromagnetism Lagrangians

To prepare for the discussion of the Yang-Mills equations, let's investigate the Dirac and Electromagnetism Lagrangians. The Dirac equation is

$$(i\gamma^{\mu}\partial_{\mu} - m)\psi = 0, \tag{18}$$

where the speed of light, c, and Plank's constant $\hbar$ are set to one. Its Lagrangian density is

$$\mathcal{L} = \bar{\psi}(i\gamma^{\mu}\partial_{\mu} - m)\psi. \tag{19}$$

The Euler-Lagrange equations minimize the action $\mathbf{S}$ where $S = \int \mathcal{L}\mathbf{dx}$. Using the Euler-Lagrange equation where the differentiation is with respect to $\bar{\psi}$, i.e.,

$$\partial_{\mu}\left(\tfrac{\partial\mathcal{L}}{\partial(\partial_{\mu}\bar{\psi})}\right) - \tfrac{\partial\mathcal{L}}{\partial\bar{\psi}} = 0, \tag{20}$$

yields equation (18).

The same gauge invariant argument used in the "Gauge Invariance" section applies here. In order for the Lagrangian to be invariant under the phase transformation $\psi(\mathbf{x}) \to \psi(x)e^{+iq\alpha(\mathbf{x})}$, this transformation has to be accompanied by the local gauge transformation $\mathbf{A}_{\mu} \to \mathbf{A}_{\mu} - q^{-1}\partial_{\mu}\alpha(\mathbf{x})$ and ∂_{μ} has

[18] *Misner, C. W., Thorne, K. S., and Wheeler, J. A.,* Gravitation, *Freeman, San Francisco 1973.*

[19] *Flanders, H.,* Differential Forms, *Academic Press, 1963.*

to be replaced by $\partial_\mu + ieA_\mu$. The Lagrangian density becomes

$$\mathcal{L} = \bar{\psi}(i\gamma^\mu \partial_\mu - m)\psi - e\bar{\psi}\gamma^\mu \psi A_\mu. \quad (21)$$

The last term is the equivalent of the interaction energy with the electromagnetic field, $j^\mu A_\mu$. In order for the Euler-Lagrangian equation differentiated with respect to A_μ to yield the inhomogeneous Maxwell equation (7), we must add $-(\frac{1}{4})(F_{\mu\nu})^2$, getting

$$\mathcal{L} = \bar{\psi}(i\gamma^\mu \partial_\mu - m)\psi - e\bar{\psi}\gamma^\mu \psi A_\mu - (\tfrac{1}{4})(F_{\mu\nu})^2. \quad (22)$$

The Euler-Lagrange equation yields

$$\partial_\mu F^{\mu\nu} = e\bar{\psi}\gamma^\nu \psi, \quad (23)$$

which equals J^ν. Note that the gauge field A_μ does not carry a charge and that there is no gauge field self-coupling, which would be indicated by an $[A_\mu, A_\nu]$ term in (23). The Lagrangian density does not yield the homogeneous Maxwell equations. They are satisfied trivially, because the definition of $F^{\mu\nu}$ satisfies the homogeneous equations automatically.[20]

From this it is apparent that the Lagrangian density for the electromagnetic field alone,

$$\mathcal{L} = -J^\mu A_\mu - (\tfrac{1}{4})(F_{\mu\nu})^2, \quad (24)$$

yields all of Maxwell's equations.

In differential geometry, if $j = 0$, this Lagrangian density becomes

$$\mathcal{L} = -\tfrac{1}{2} F \wedge *F. \quad (25)$$

The Yang-Mills Theory

The Yang-Mills theory incorporates isotopic spin symmetry introduced in 1932 by Heisenberg, who observed that the proton and neutron masses are almost the same (938.272 MeV versus 939.566 MeV respectively). He hypothesized that if the electromagnetic field were turned off, the masses would be equal and the proton and neutron would react identically to the strong force, the force that binds the nucleus together and is responsible for the formation of new particles and the rapid (typical lifetimes are about 10^{-20} seconds) decay of others. In a nonphysical space (also known as an internal space) called *isospin space*, the proton would have isospin up, for instance, and the neutron, isospin down but other than that they would be identical. The wave function for each particle could be transformed to that for the other by a rotation using the spin matrices of the nonabelian group SU(2). Because of charge independence, the strong interactions are invariant under rotations in isospin space. Since the ratio of the electromagnetic to strong force is approximately α, where $\alpha = e^2/4\pi\hbar c = 1/137$, for a good approximation we can neglect the fact that the electromagnetic forces break this symmetry. By Noether's theorem, if there is a rotational symmetry in isospin space, the total isotopic spin is conserved. This hypothesis enables us to estimate relative rates of the strong interactions in which the final state has a given isospin. The spin matrices turn out to be the Pauli matrices σ_i. The theory just described is a global one; i.e., the isotopic spin rotation is independent of the space-time coordinates, and thus no connection is used. We will see that Yang and Mills[21] elevated this global theory to a local one. In 1954 they proposed applying the isospin matrices to electromagnetic theory in order to describe the strong interactions. Ultimately their theory was used to describe the interaction of quarks in the electroweak theory[22] and the gluons fields of the strong force. In the next section we will give an example using the up quark u, which has a charge of $\frac{2}{3}e$, and the down quark d, which has a charge of -$\frac{1}{3}e$.

We have seen that the field strength (which is also the curvature of the connection on the fiber) is given by $\mathbf{F} = \mathbf{dA} + \mathbf{A} \wedge \mathbf{A}$. In electromagnetism $\mathbf{A}$ is a 1-form with scalar coefficients for $\mathbf{d}x^i$, so $\mathbf{A} \wedge \mathbf{A}$ vanishes. If, however, the coefficients are noncommuting matrices, $\mathbf{A} \wedge \mathbf{A}$ does not vanish and provides for gauge field self-coupling. Yang and Mills formulated the field strength using the letter B instead of A, so we will follow suit. The field is

$$F_{\mu\nu} = (\partial_\mu B_\nu - \partial_\nu B_\mu) + i\epsilon(B_\mu B_\nu - B_\nu B_\mu) \quad (26)$$

or equivalently $F_{\mu\nu} = (\partial_\mu B_\nu - \partial_\nu B_\mu) + i\epsilon[B_\mu, B_\nu]$, where B is the connection on a principal fiber bundle, i.e., the gauge potential and where ϵ is the coupling constant analagous to q in (9). Therefore, as opposed to the electromagnetic field strength which is linear, their field strength is nonlinear. They proposed using a local phase,

$$\psi(\mathbf{x}) \to \psi(x)e^{-i\epsilon\alpha_j(\mathbf{x})\sigma^j}, \quad (27)$$

where σ^j are the Pauli matrices and j goes from 1 to 3. Thus the isotopic spin rotation is space-time dependent, i.e., local. The Pauli matrices do not commute:

$$[\tfrac{\sigma^i}{2}, \tfrac{\sigma^j}{2}] = i\tfrac{\sigma^k}{2}\epsilon^{ijk}.$$

Since $B_\mu = b^i_\mu \sigma_i$ or $B_\mu = \vec{\mathbf{b}} \cdot \vec{\sigma}$ (where b^i_μ is called the isotopic spin vector gauge field), the four-vectors B_μ and B_ν in (26) do not commute leading to nonabelian theory. The purpose of the Pauli spin matrices in the connection B is to rotate the particles in isospin space so that they retain their identities at different points in $\mathbb{R}^4$. Equation (26) can be rewritten so that the curvature is defined

[20] *Jackson, J. D.,* Classical Electrodynamics, *3rd ed., John Wiley and Sons, 1998.*

[21] *Yang, C. N., and Mills, R. L.,* Phys. Rev. **96** *(1954), 91.*

[22] *The weak and electromagnetic forces are the two manifestations of the electroweak force.*

as $\mathbf{F} = \mathbf{dB} + \frac{1}{2}\, \mathbf{i}\epsilon[\mathbf{B}, \mathbf{B}]$. As opposed to the Maxwell's equations case, the exterior derivative of the curvature $\mathbf{dF}$ does not equal zero because of the commutator in the expression for the curvature. Thus the exterior derivative for the 2-form $\mathbf{F}$ has to be altered to include the connection $\mathbf{B}$ in order to generalize the homogeneous Maxwell's equations.

The Yang-Mills Lagrangian

$$\mathcal{L} = \bar{\psi}(i\gamma^{\mu}(D_{\mu} - m)\psi - (\tfrac{1}{4})Tr(F_{\mu\nu}F^{\mu\nu}) \tag{28}$$

is invariant under the gauge transformation for the covariant derivative given as

$$D_{\mu} = \partial_{\mu} - i\epsilon B_{\mu}. \tag{29}$$

The connection B_{μ} transforms as

$$B_{\mu} \to B_{\mu} + \epsilon^{-1}\partial_{\mu}\alpha - i[\alpha, B_{\mu}], \tag{30}$$

the fiber is the sphere S^2, and the gauge group is SU(2).

Since there are three components of the vector gauge field b^i_{μ}, there are three vector gauge fields representing three gauge particles having spin one. They were later identified as the intermediate vector bosons $W^{\pm}$ and Z^0 which mediate the electroweak interactions. The fact that there are three gauge particles is dictated by the fact that the gauge field is coupled with the three Pauli spin matrices. Also, since the charges of the up quark and down quark differ by one, the gauge field particles that are absorbed and emitted by them in quark-quark interactions can have charges of ± 1 or zero. It is astonishing that Yang and Mills in their 1954 paper predicted the existence of the three intermediate vector bosons.

Independently of Yang and Mills and slightly after them, Shaw and Utiyama separately developed nonabelian gauge theories similar to the Yang-Mills one. Utiyama's theory also included a gauge theory for gravity and electromagnetism. Their papers are included in O'Raifeartaigh's *The Dawning of Gauge Theory.*

The gauge particles predicted by the Lagrangian (28) have zero mass, since any mass term added to (28) would make the Lagrangian noninvariant under a gauge transformation. So the force associated with the particles would have infinite range, as the photons of the electromagnetic interaction do. Of course the weak force (the force responsible for particles decaying slowly; typically their lifetimes are about 10^{-10} seconds or much less) and strong nuclear force are short range. This discrepency was corrected some years later by the introduction of spontaneous symmetry breaking in the electroweak $SU(2) \times U(1)$ theory of Weinberg, Salam and Glashow (WSG) using the Higgs mechanism. The WSG theory, which explains the electromagnetic and weak forces, predicts the mass of the $W^{\pm}$ (80.37 $\pm$0.03 GeV) and Z^0 (92 $\pm$ 2 GeV) intermediate vector bosons, where GeV represents a billion electron volts. The $W^{\pm}$ was discovered[23] in 1983 (its mass is now reported at 80.425 Gev $\pm$ 0.033 GeV), and later that year the Z^0 was discovered[24] (its mass is now reported at a mass of 91.187 $\pm$ 0.002 GeV).

The Euler-Lagrange equations for equation (28) give the Dirac equation

$$\gamma^{\mu}(\partial_{\mu} - ieB_{\mu})\psi + m\psi = 0, \tag{31}$$

and also the vector equation for the vector field $\mathbf{F}$, namely

$$\partial^{\mu}\mathbf{F}_{\mu\nu} - i\epsilon[\mathbf{B}^{\mu}, \mathbf{F}_{\mu\nu}] = -i\epsilon\bar{\psi}\gamma_{\nu}\psi = -\mathbf{J}_{\nu}, \tag{32}$$

which, if it weren't for the commutator, is the same form as the nonhomogeneous four-vector Maxwell equation. The commutator causes the gauge particles to interact with themselves.

The effect of these equations is explained by 't Hooft,[25] who with Veltman proved the renormalizability of Yang-Mills theories:

> ...The B quanta would be expected to be exchanged between any pair of particles carrying isospin, generating not only a force much like the electromagnetic force, but also a force that rotates these particles in isospin space, which means that elementary reactions envolving the transmutation of particles into their isospin partners will result. A novelty in the Yang-Mills theory was that the B quanta are predicted to interact directly with one another. These interactions originate from the commutator term in the $\mathbf{F}_{\mu\nu}$ field [equation (32)], but one can understand physically why such interactions have to occur: in contrast with ordinary photons, the Yang-Mills quanta also carry isospin, so they will undergo isospin transitions themselves, and furthermore, some of them are charged, so the neutral components of the Yang-Mills fields cause Coloumb-like interactions between these charged particles.

So the Yang-Mills equations indicate that, for instance, for the up quark-down quark doublet, the W^- generates a force that rotates the d into the u in isospin space exhibited by the transition $d \to u + W^-$. The commutator in equation (32) is responsible for interactions like $W \to W + Z$

[23] *Arnison, G., et al.,* Phys. Lett. **122B** *(1983), 103.*

[24] *Arnison, G., et. al.,* Phys. Lett. **126B** *(1983), 398.*

[25] *'t Hooft, Gerardus, ed.,* 50 Years of Yang-Mills Theory, *World Scientific, 2005.*

occurring,[26] and the W can radiate, producing a photon in $W \to W + \gamma$.

The Yang-Mills field strength contribution (also called the kinetic term) to the differential geometry Lagrangian density, where k is a constant is:

$$\mathcal{L} = -kTr(F \wedge *F). \tag{33}$$

The Euler-Lagrange equations produce $d_B F = 0$ (it is called the Bianchi Identity, which is purely geometric), and in the absence of matter fields the field equation $d_B * F = 0$, where d_B is the exterior covariant derivative. These are the Yang-Mills equations in compact form.

The Feynman Stamp

In QED after Schwinger, Tomonaga and Feynman addressed the singularities produced by the self-energy of the electron by renormalizing the theory. They were then exceedingly successful in predicting phenomena such as the Lamb shift and anomalous magnetic moment of the electron.

Feynman introduced[27] schematic diagrams, today called *Feynman diagrams*, to facilitate calculations of particle interaction parameters. External particles, represented by lines (edges) connected to only one vertex are real, i.e., observable. They are said to be on the mass shell, meaning their four-momentum squared equals their actual mass, i.e., $m^2 = E^2 - p^2$. Internal particles are represented by lines that connect vertices and are therefore intermediate states—that is why they are said to *mediate* the interaction. They are virtual and are considered to be off the mass shell. This means their four-momentum squared differs from the value of their actual mass. This is done so that four-momentum is conserved at each vertex. The rationale for this difference is the application of the uncertainty principle $\Delta E \cdot \Delta t = \hbar$. Since Δt, the time spent between external states is very small for that short time period, ΔE, and thus the difference between the actual and calculated mass can be large. In the following Feynman diagrams, the time axis is vertical upwards.

The diagram on the upper-left of the stamp (Figure 1) is a vertex diagram and as such represents a component of a Feynman diagram. It illustrates the creation of an electron-positron pair from a photon, γ; it is called *pair production.* The γ is represented by a wavy line. The Feynman-Stuckelberg interpretation of negative-energy solutions indicates that here the positron, the electron's antiparticle, which is propagating forward in time, is in all ways equivalent to an electron going backwards in time. If all the particles here were external, the process

would not conserve energy and momentum. To see this you must first remember that since the photon has zero mass due to the gauge invariance of electromagnetic theory, its energy and momentum are equal. Thus β, which equals $\frac{p}{E}$, has the value 1; but $\beta = \frac{v}{c}$, so that the photon's velocity is always c, the speed of light. In the electron-positron center of mass frame (more aptly called the center of momentum frame, since the net momentum of all the particles is zero there), the electron and positron momenta are equal and are in opposite directions. The photon travels at the speed of light, and therefore its momentum cannot be zero; but there is no particle to cancel its momentum, so the interaction cannot occur. (For it to occur requires a Coloumb field from a nearby nucleus to provide a virtual photon that transfers momentum, producing a nuclear recoil.) Therefore the γ in the diagram is internal. Its mass is off the mass shell and cannot equal its normal value, i.e., zero.

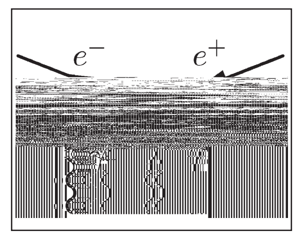

Figure 1. A pair production vertex.

The diagram on the lower-left of the stamp (Figure 2) is also a vertex diagram and represents an electron-positron pair annihilation producing a γ. Again, if all the particles are external, conservation energy and momentum prohibit the reaction from occurring; therefore the γ must be virtual.

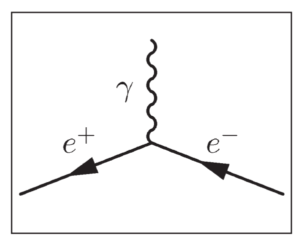

Figure 2. A pair annihilation vertex.

The diagram on the bottom to the right of Feynman (Figure 3) was meant to represent an electron-electron scattering with a single photon exchange. This is called Møller scattering. (It can, however, represent any number of interactions exchanging a photon.) The diagram represents the t-channel of Møller scattering; there is another diagram not shown here representing the u-channel contribution, where u, t and

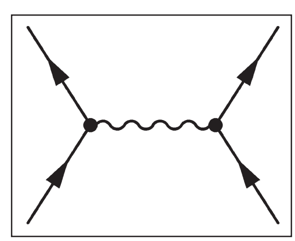

Figure 3. Electron-electron (Møller) scattering.

[26] *This is indicated in Figure 1 of the Yang-Mills paper. See also F. Halzen and A. D. Martin,* Quarks and Leptons, *John Wiley & Sons, 1984, p. 343.*

[27] *Feynman, R. P.,* Phys. Rev. **76** *(1949), 769.*

another variable s are called the *Mandelstam variables*. They are used in general to describe 2-body → 2-body interactions. If you rotate the diagram in Figure 3 by 90°, you have the s-channel diagram for electron-positron scattering, called Bhabha scattering, shown in Figure 4 but not on the stamp. Here an electron and positron annihilate, producing a virtual photon which in turn produces an electron-positron pair. There is also a t-channel contribution to Bhabha scattering. The cross-section for Bhabha scattering can be easily obtained from the one for Møller scattering by interchanging the s and u in the cross-section expression in a process called *crossing*. Small angle Bhabha scattering is used to test the luminosity in e^+-e^- colliding beam accelerators.

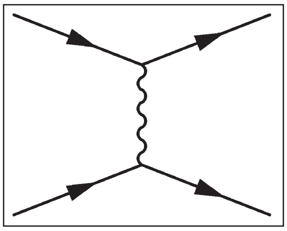

Figure 4. Electron-positron (Bhabha) scattering.

To the right of the Møller scattering diagram is a vertex correction to electron scattering, shown in Figure 5, where the extra photon forms a loop. It is used to calculate both the anomalous magnetic moment of the electron and muon, also the anomalous magnetic moment contribution to the Lamb shift.[28] The other two contributions to the Lamb shift are the vacuum polarization and the electron mass renormalization. The Lamb shift explains the splitting in the spectrum of the $2S_{\frac{1}{2}}$ and $2P_{\frac{1}{2}}$ levels of hydrogen, whereas Dirac theory alone incorrectly predicted that these two levels should be degenerate.

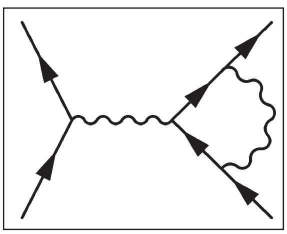

Figure 5. Radiative correction.

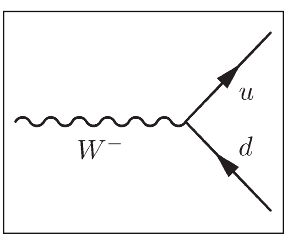

Figure 6. A flavor non-conserving transition vertex.

The low-order solution of the Dirac equation predicts a value of 2 for the g-factor used in the expression for the magnetic moment of the electron. The vertex correction shown in Figure 5, however, alters the g-factor, producing an anomalous magnetic moment contribution, written as $\frac{g-2}{2}$. When this and higher-order contributions are included, the calculated value of $\frac{g-2}{2}$ for the electron is $1159\,652\,460(127)(75) \times 10^{-12}$ and the experimental value is $1159\,652\,193(10) \times 10^{-12}$, where the numbers in parenthesis are the errors. This seven-significant figure agreement is a spectacular triumph for QED. We need not emphasize that the calculations for all these diagrams use the gauge principal for quantum electrodynamics.

The other diagrams on the stamp are all vertex diagrams and show how Feynman's work, originally applicable to QED, was then later used to elucidate the electroweak force. This is exemplified on the stamp by flavor-changing transitions, e.g., $d \to W^- + u$ shown in Figure 6, and flavor-conserving transitions, e.g., $d \to Z^0 + d$ of the electroweak force—the u and d quarks have different values of flavor. The process in Figure 6 occurs for instance in β decay, where a neutron (udd) decays into a proton (udu) and electron and an antineutrino. What happens is that the transition $d \to u + W^-$ corresponds to a rotation in isospin space. This rotation is caused by the virtual W^- which mediates the decay. It in turn decays into an electron and an antineutrino. The calculations for these transitions all use the Yang-Mills theory. Although quarks are confined in hadrons (particles that undergo strong interactions like the proton and neutron), they are free to interact with the intermediate vector bosons.

[28] *See, for instance, Griffith, David,* Introduction to Elementary Particles, *John Wiley and Sons, 1987, p. 156.*

Who Designed the Stamp?

Feynman's daughter, Michelle, was sent a provisional version of the stamp by the United States Postal Service and advised on the design of it by, among others, Ralph Leighton, coauthor with Richard Feynman of two popular books, and CalTech's Steven Frautschi and Kip Thorne. Frautschi and Leighton edited the Feynman diagrams, and Frautschi rearranged them and composed the final design. The person who chose the original Feynman diagrams that form the basis for the stamp remains a mystery.

Acknowledgements

The author thanks Jeff Cheeger, Bob Ehrlich, J. D. Jackson, and C. N. Yang for their suggestions on the article. Their contributions improved the paper. The author especially thanks J. D. Jackson for generously giving of his time, and periodically providing advice since the author first corresponded with him many years ago.