On the Origins of Gauge Theory

Callum Quigley

April 14, 2003

1 Introduction

We know the Universe to be governed by four fundamental interactions: namely, the strong and the weak nuclear forces, electromagnetism and gravitation. It is a driving concept to unify these forces into a single, comprehensive theory. Though this task is far from its completion, there has been much progress.

The first great landmark in its development is attributable to James Maxwell, who in 1864 brought together the seemingly unrelated concepts of electricity, magnetism and optics into the now well known theory of electromagnetism. Nearly a hundred years later, the weak force too, was combined with the electromagnetic by Glashow, Weinberg and Salam, giving rise to the electroweak theory. Currently, attempts are being made to find a Grand Unified Theory which would explain all the forces, except gravity, as manifestations of the same fundamental interaction. We believe such a theory is plausible because these forces are all governed by the same principle: the gauge principle. In fact, we shall see that gravity also obeys this rule, which begs the question, "Are the four known forces all aspects of some single unified force?" Well, nobody knows, nor is it the purpose of this paper to chase that dream. Rather, we will investigate some of the historical developments which transformed this fundamental notion from a triviality into a cornerstone of physics.

Between the times of Maxwell and Salam, there were a number of other advances in unification. Many theories, such as Hermann Weyl's attempt to unify gravity and electricity in 1918, had to be abandoned. However, in Weyl's case, we shall see that a slight modification of his original proposal

forms the foundation of what is now known as gauge theory. The generalization of this concept, discovered by Yang and Mills, is the framework which explains both nuclear forces. We will restrict our attention, to the gravitational and electromagnetic (EM) forces. In particular, how they are both derivable from the gauge principle.

2 First Notions of Gauge Invariance

Roughly speaking, the gauge principle states:

> *If a physical system is invariant with respect to some global (space-time independent) group of continuous transformations, G, then it remains invariant when that group is considered locally (space-time dependant), that is $G \to G(x)$.*

Although this formulation is incomplete, we shall see under what conditions local invariance is possible. Throughout the paper, we'll use the following notation.

Notation Indices run from 0 to 3, unless otherwise noted. Square brackets [ ] denote anti-symmetrization. Repeated indices are summed over. Tensor indices are raised and lowered in the usual manner by the metric. Finally, we set $c = 1$.

2.1 Electromagnetism

The gauge principle was first recognized in electromagnetism, but in a rather trivial sense. We require the following definitions:

Definition 1a). The *electromagnetic 4-current density* $j^\mu = \{\rho, \mathbf{j}\}$, is a 4-vector where ρ is the electric charge density and $\mathbf{j}$ is the 3-dimensional electric current.
b) The *electromagnetic 4-potential* $A_\mu(x) = \{\phi(x), \mathbf{A}(x)\}$, is 1-form where ϕ and $\mathbf{A}$ are the electric and magnetic potentials, respectively.
c) Finally, the *electromagnetic field tensor* $F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu$, is a 2-form which encodes the EM information, via Maxwell's equations.

With this notation, and appropriate units, Maxwell's equations are compactly written:

$$\partial_{[\alpha}F_{\mu\nu]} = 0 \quad and \quad \partial_\mu F^{\mu\nu} = j^\nu. \tag{1}$$

Because of the equality of mixed partials, it follows that the transformation

$$A_\mu(x) \to A'_\mu(x) = A_\mu(x) + \partial_\mu \alpha(x), \tag{2}$$

leaves $F_{\mu\nu}$ unchanged, for any differentiable scalar function $\alpha(x)$. Thus, Maxwell's equations are unaltered by adding a gradient. Such a transformation is now termed a *gauge transformation*, for reasons that will become more clear. Furthermore, for fixed x it is easy to show that the transformations (2) form a commutative (abelian) group, with a single continuous parameter $\alpha(x)$.

For a long time, the EM potential A_μ was thought of only a mathematical tool for simplifying calculations, only the field $F_{\mu\nu}$ had any physical reality. So this gauge freedom, that is the ability to add a gradient onto the potential, was originally considered useful, but unphysical.

2.2 General Relativity

The idea of gauge invariance was first appreciated in Einstein's theory of General Relativity (GR). In fact, GR is derivable from the gauge principle, where the gauge transformations are rigid motions in spacetime. To see this, we must introduce the Christoffel connection,

$$\Gamma^\lambda_{\mu\nu} = \frac{1}{2} g^{\lambda\sigma}(\partial_\nu g_{\mu\sigma} + \partial_\mu g_{\nu\sigma} - \partial_\sigma g_{\mu\nu})$$

where the $g_{\mu\nu}$ are the spacetime metric coefficients. An infinitesimal spacetime interval's squared length is given by the 2-form

$$ds^2 = g_{\mu\nu} dx^\mu dx^\nu. \tag{3}$$

When a (co)vector is parallel transported an amount dx^ν, its components vary as

$$dv^\lambda = -v^\mu \Gamma^\lambda_{\mu\nu} dx^\nu \quad or \quad dv_\mu = v_\lambda \Gamma^\lambda_{\mu\nu} dx^\nu. \tag{4}$$

Since we write the scalar product between two vectors (at the same point) as $u \cdot v = g_{\mu\nu} u^\mu v^\nu$, then a vector's squared length is given by $|v|^2 = (g_{\mu\nu} v^\mu v^\nu) = (v_\nu v^\nu)$. We see that this length is invariant under parallel transport

$$\begin{aligned} dv^2 &= d(v_\nu v^\nu) \\ &= dv_\nu v^\nu + v_\nu dv^\nu \\ &= v_\lambda \Gamma^\lambda_{\mu\nu} dx^\mu v^\nu - v_\nu v^\mu \Gamma^\nu_{\mu\lambda} dx^\lambda \\ &= 0. \end{aligned} \tag{5}$$

In order for derivatives to remain co-ordinate invariant, that is gauge invariant, we must modify the partial derivative operator. Otherwise, a change of co-ordinates, from primed to unprimed, for a covector's partial derivative yields:

$$\partial_{\mu'} v^{\nu'} = (\partial_{\mu'} x^\mu \partial_\nu x^{\nu'} \partial_\mu + \partial_{\mu'} x^\mu \partial_{\mu\nu} v^{\nu'}) v^\nu. \tag{6}$$

The first term is fine, however, the second term is not a tensorial transformation. So, instead we use the covariant derivative, $\nabla_\mu(\Gamma)$, defined as

$$(\nabla_\mu)^\lambda_\nu = \delta^\lambda_\nu \partial_\mu + \Gamma^\lambda_{\mu\nu}$$

(where δ^α_β is the Kronecker delta) so that

$$\nabla_{\mu'} v^{\nu'} = \partial_{\mu'} x^\mu \partial_\nu x^{\nu'} \nabla_\mu v^\nu. \tag{7}$$

In other words, the covariant derivative transforms tensorialy. We will see that covariant derivatives are at the heart of gauge theory; through them, global invariance is preserved locally. The final essential geometric ingredient for GR is the Riemann curvature tensor, which can be expressed in terms of the connection, or the covariant derivative, as

$$\begin{aligned} R^\lambda_{\sigma\mu\nu} &= \partial_\mu \Gamma^\lambda_{\sigma\nu} - \partial_\nu \Gamma^\lambda_{\sigma\mu} + \Gamma^\lambda_{\alpha\mu} \Gamma^\alpha_{\sigma\nu} - \Gamma^\lambda_{\alpha\nu} \Gamma^\alpha_{\sigma\mu} \\ &= [\nabla_\sigma, \nabla_\mu]^\lambda_\nu. \end{aligned}$$

Note that the second definition highlights the non-commutivity of parallel transport, which tells us about the curvature of spacetime. We will write the contracted Riemann tensor $R^\lambda_{\mu\nu\lambda} \equiv R_{\mu\nu}$, and the Ricci scalar $R \equiv R^\nu_\nu$. Now we can write Einstein's field equations for gravitational interactions:

$$R_{\mu\nu} - \frac{1}{2} R g_{\mu\nu} = T_{\mu\nu} \tag{8}$$

where $T_{\mu\nu}$ is the (symmetric) stress-energy tensor.

3 Weyl's Unified Theory

As powerful and profound as Einstein's gravitational theory was, many felt it was only the beginning. To describe both known forces (the nuclear forces were not yet discovered), the EM field tensor had to be put in by hand. Many, including Einstein himself, sought a unified theory to explain both phenomena, preferably in a geometric fashion like GR.

The first attempt to generalize GR to encompass EM, was proposed by Weyl three years later. Unhappy with Riemannian geometry, Weyl developed his *purely infinitesimal geometry* which did not allow comparison at a distance.

3.1 Scale Invariance

As is well known, in Euclidean geometry, translation of a vector preserves its length and direction. In Riemann's geometry, the Christoffel connection guarantees length preservation, however, a vector's orientation is path dependant. However, the angle between to vectors, following the same path, *is* preserved under translation. Weyl wondered why the remnant of planar geometry, length preservation, persisted. After all, our measuring standards (rigid rods and clocks) are known only at one point in spacetime. To measure lengths at another point, we must bring our measuring tools along with us. According to Weyl, only the *relative* lengths of any two vectors (at the same point), and the angle between them, are preserved under parallel transport; the length of any single vector is arbitrary. To encode this mathematically, Weyl made the following substitution

$$g_{\mu\nu}(x) \to \lambda(x) g_{\mu\nu}(x). \tag{9}$$

Where the *conformal factor* $\lambda(x)$ is an arbitrary, positive, smooth function of position. Weyl required that in addition to GR's co-ordinate invariance, formulas must remain invariant under the substitution (9). Weyl called this a *gauge transformation.*

Remark 1. This was the first deliberate application of the gauge principle. In Riemann's geometry, the metric is fixed up to a global scale factor. Weyl's idea was to make that scale a local property of the metric.

Remark 2. The term *gauge* was introduced into mathematics and physics by Weyl during this period. Until now, its use in this paper has been purely from a modern perspective.

In this setting, if a vector, v^α, at a point $P = (x_1, x_2, \ldots x_n)$ is parallel transported to the point $P' = (x_1 + dx^1, x_2 + dx^2, \ldots x_n + dx^n)$, then

$$v^\alpha \to v^\alpha + dv^\alpha \quad where \quad dv^\alpha = -v^\mu \{^{\alpha}_{\mu\nu}\} dx^\nu. \tag{10}$$

This transformation is identical to the Riemannian case, except the Christoffel connection Γ has been replaced by a similar object, the *conformal connection* $\{\ \}$, which is also symmetric in the lower indices.

3.2 The Conformal Connection

To find the conformal connection explicitly, consider two vectors, u^α and v^α, at P. Under parallel transport to P', they become $u^\alpha + du^\alpha$ and $v^\alpha + dv^\alpha$. By Weyl's hypothesis, the relationship between the vector's scalar products at each point is given by:

$$(g_{\alpha\beta} + dg_{\alpha\beta})(u^\alpha + du^\alpha)(v^\beta + dv^\beta) = (1 + d\phi)(g_{\alpha\beta} u^\alpha v^\beta). \tag{11}$$

That is, the scalar products at P and P' are *not* equal, rather they are proportional. The factor of proportionality $1 + d\phi$, which is infinitesimally close to unity, distinguishes this geometry from Riemann's. By expanding (11) up to linear differential terms, we have

$$\begin{aligned} g_{\alpha\beta} u^\alpha v^\beta d\phi &= dg_{\alpha\beta} u^\alpha v^\beta + g_{\alpha\beta}(u^\alpha dv^\beta + du^\alpha v^\beta) \\ &= dg_{\alpha\beta} u^\alpha v^\beta - g_{\alpha\beta}(u^\alpha \{^{\beta}_{\mu\nu}\} v^\mu dx^\nu + v^\beta \{^{\alpha}_{\mu\nu}\} u^\mu dx^\nu) \\ &= dg_{\alpha\beta} u^\alpha v^\beta - \{_{\alpha\mu\nu}\} u^\alpha v^\mu dx^\nu - \{_{\beta\mu\nu}\} u^\mu v^\beta dx^\nu, \end{aligned} \tag{12}$$

where we have used the substitution (10) for du^α and dv^β, and then used the metric to lower the indices of $\{\ \}$. For the above relation to hold, for any vectors u and v, we require that

$$\begin{aligned} g_{\alpha\beta} d\phi &= dg_{\alpha\beta} - \{_{\alpha\beta\nu}\} dx^\nu - \{_{\beta\alpha\nu}\} dx^\nu \\ &= \partial_\nu g_{\alpha\beta} dx^\nu - \{_{\alpha\beta\nu}\} dx^\nu - \{_{\beta\alpha\nu}\} dx^\nu, \end{aligned} \tag{13}$$

where we have used the chain rule in the second line. Thus, $d\phi$ is a linear differential form: $d\phi = \phi_\nu dx^\nu$. Plugging this into the above, we find

$$g_{\alpha\beta}\phi_\nu = \partial_\nu g_{\alpha\beta} - \{_{\alpha\beta\nu}\} - \{_{\beta\alpha\nu}\}. \tag{14}$$

By cyclicly permuting α, β and ν, then subtracting the first arrangement from the sum of the other two, and finally applying the inverse metric $g^{\nu\lambda}$, yields the *conformal connection*

$$\{^\lambda_{\alpha\beta}\} = \Gamma^\lambda_{\alpha\beta} + \frac{1}{2}(\delta^\lambda_\alpha\phi_\beta + \delta^\lambda_\beta\phi_\alpha + g_{\alpha\beta}g^{\nu\lambda}\phi_\nu),$$

where $\Gamma^\lambda_{\alpha\beta}$ is the usual Christoffel connection, derived from the metric. Thus, in Weyl's geometry, the affine connection is doubly dependant; it is determined by a) the metric tensor $g_{\alpha\beta}$ and b) the covector ϕ_ν. We will call this covector the *length connection*, as it relates the scales between two points on a manifold.

Remark 3. Unless the ϕ_ν are known, the conformal connection is not uniquely determined. Instead, there exists an equivalence class of connections which preserve relative lengths and angles under parallel transport.

3.3 Electromagnetic Interpretation

Now, if we perform the gauge transformation (9), since $\{_{\alpha\beta\nu}\} = g_{\alpha\lambda}\{^\lambda_{\beta\nu}\}$, then (14) becomes (suppressing the x dependance),

$$\begin{aligned} \lambda g_{\alpha\beta}\phi'_\nu &= \lambda\partial_\nu g_{\alpha\beta} + g_{\alpha\beta}\partial_\nu\lambda - \lambda\{_{\alpha\beta\nu}\} - \lambda\{_{\beta\alpha\nu}\} \\ &= \lambda g_{\alpha\beta}\phi_\nu + g_{\alpha\beta}\partial_\nu\lambda. \end{aligned} \tag{15}$$

From which we conclude, that the transformation

$$g_{\mu\nu} \to \lambda g_{\mu\nu} \quad \Leftrightarrow \quad \phi_\nu \to \phi'_\nu = \phi_\nu + \frac{\partial_\nu\lambda}{\lambda}. \tag{16}$$

Since λ is an arbitrary smooth positive function, we can just as well set $\lambda(x) = e^{\alpha(x)}$, for some function $\alpha(x)$. Now the gauge transformation reads

$$g_{\mu\nu}(x) \to e^{\alpha(x)}g_{\mu\nu}(x) \quad \Leftrightarrow \quad \phi_\nu(x) \to \phi_\nu(x) + \partial_\nu\alpha(x). \tag{17}$$

In this highly suggestive form, Weyl's gauge transformation of the length connection bears a striking resemblance to the gauge transformations of the

EM potential, A_μ. Weyl postulated that of all the possible conformal connections in the equivalence class, only one had any relation to physics. Specifically, the connection where $\phi_\nu = \frac{e}{\gamma} A_\nu$, where e is the electron charge and γ is an undetermined constant.

This is a most remarkable result. By allowing scales to vary between points on a manifold, Weyl could to formulate a unified theory of EM and gravity. The gravitational fields are encoded in the metric tensor, and EM fields are derived from the length connection form. What is perhaps even more remarkable is this theory sprung forth from a purely mathematical concept: *Weyl's infinitesimal geometry*. GR, however, was derived from an entirely physical fact: the equivalence of inertial and gravitational mass.

With this choice of the length connection ϕ, the *Weyl connection* is written

$$\binom{\lambda}{\alpha\beta} = \Gamma^\lambda_{\alpha\beta} + \frac{e}{2\gamma}(\delta^\lambda_\alpha A_\beta + \delta^\lambda_\beta A_\alpha + g_{\alpha\beta} g^{\nu\lambda} A_\nu).$$

It depends on both the gravitational fields, $g_{\alpha\beta}$, and the EM potential A_μ. We can form all the objects from Riemannian geometry, such as the covariant derivative and curvature tensors, by replacing the Christoffel connection with Weyl's.

Rather than develop Weyl's geometry further, let's return to equation (11). Suppose a vector, v^μ at P has the (squared) length $l = g_{\mu\nu} v^\mu v^\nu$, then under parallel transport to P', this length changes by

$$dl = l d\phi = l \frac{e}{\gamma} A_\mu dx^\mu. \tag{18}$$

If instead, this vector is (parallel) transported along the path C to some distant point Q, then upon integrating (18), we find

$$l = l_o e^{\frac{e}{\gamma} \int_C A_\mu(x) dx^\mu}, \tag{19}$$

where l_o is l at P. So, a vector's length is, in general, path dependant. At this point, a clever reader might raise the following objection. Suppose we have two identical clocks, at P. Let l be the length of a time-like vector corresponding to some unit of time. Now, transport the two clocks on different paths C_1 and C_2, which both end at Q. Let l_1 and l_2 denote the new values of l given by each clock, at the point Q. Then, by Stokes theorem, $l_1 = l_2$ if and only if

$$\oint_{C_1 - C_2} A_\mu(x) dx^\mu = \int_D F_{\mu\nu}(x) dx^\mu \wedge dx^\nu = 0 \tag{20}$$

where $D = int(C_1 - C_2)$ and $F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu$ is the EM field tensor. Thus, in the presence of an EM field, the two clock rates will differ. As Einstein pointed out, the frequency of the spectral lines of atomic clocks would depend on the location, both past and present, of the atom. However, we know the atomic spectral lines to be quite definite, and independent of position.

Einstein was able to refute Weyl's theory, with this simple physical argument. Weyl probably wished he had not sent his paper to Einstein to be published, since Einstein included this negation as a postscript. He, nonetheless, admired Weyl as a brilliant mathematician, and was greatly impressed by his novel geometric ideas. Weyl, on the other hand, was not convinced and continued to develop his true infinitesimal geometry. He thought,

> It would be remarkable if in Nature there was realized instead an illogical quasi-infinitesimal geometry, with an electromagnetic field attached to it.

Weyl's gauge theory was paid little heed during the next decade. With the coming of the Quantum era, attention moved to the microscopic regime.

4 Gauge Theory Revived

As pointed out in the introduction, Weyl's unified theory, though flawed, was not utterly doomed. It was to find salvation during the development of Quantum Theory.

4.1 Quantum Mechanics

In classical mechanics, a system, composed of a particle of mass m in a potential V, is said to be *conservative* if the total energy E is a constant of motion. That is to say, $\frac{d}{dt}E = 0$. Conservative systems are described by the *Hamiltonian function*

$$H(\mathbf{x}, \mathbf{p}) = E = \frac{\mathbf{p}^2(t)}{2m} + V(t, \mathbf{x}) \tag{21}$$

which is just the sum of kinetic and potential energies (bold variables are now 3-vectors). In Quantum Mechanics (QM), a particle is described by a complex-valued *wave function* $\psi(t, \mathbf{x})$. For simplicity, we shall only consider non-relativistic QM, for spinless particles. Information is extracted from the

wave function, by acting on it with linear operators. Following Schrödinger's lead, let's replace the momentum *variable* $\mathbf{p}$ by the momentum *operator* $\frac{\hbar}{i}\vec{\nabla}$ and the parameter E by the energy operator $i\hbar\partial_t$, where $\hbar$ is Planck's constant, and $\vec{\nabla}$ is the usual 3-dimensional gradient operator. Applying both sides to the wave function ψ we have found the *Schrödinger equation*

$$i\hbar\partial_t\psi(t,\mathbf{x}) = -\frac{\hbar}{2m}\vec{\nabla}^2\psi(t,\mathbf{x}) + V(t,\mathbf{x})\psi(t,\mathbf{x}). \tag{22}$$

As a simple example, consider the plane wave

$$\psi(t,\mathbf{x}) = e^{i(\mathbf{k}\cdot\mathbf{x}-\omega t)} \tag{23}$$

which has momentum $\mathbf{p} = \hbar\mathbf{k}$ and energy $E = \hbar\omega$. It is not difficult to see that it solves the *free* Schrödinger equation, that is $V \equiv 0$, and $E = \frac{\mathbf{p}^2}{2m}$.

The probability of finding a particle in a given volume Ω is given by the formula

$$P_\psi(\Omega) = \int_\Omega |\psi(t,\mathbf{x})|^2 d^3x. \tag{24}$$

$|\psi|^2$ has the natural definition then of a *probability density*. Wave functions are always normalized so that the integral of the probability density over all space equals unity. However, this does not determine ψ uniquely. For example, $\psi' = e^{i\alpha}\psi$, for some real number α, (look familiar??) has the same probability density as ψ. Thus, the states ψ and ψ' are equivalent.

4.2 Phase Invariance

Because of this equivalence, we say that wave functions in QM have a *global phase symmetry*. Recalling the plane wave model, we notice that multiplying it by $e^{i\alpha}$ is equivalent to adding a phase α to the argument. However, we can go one step further.

If we believe the gauge principle (and why not?), then we can make this phase invariance a local property of the matter wave by sending $\alpha \to \alpha(t,\mathbf{x})$. After all, we still have $|\psi'|^2 = (e^{-i\alpha(t,\mathbf{X})}\psi^*)(e^{i\alpha(t,\mathbf{X})}\psi) = |\psi|^2$. But this leads to complications. Consider a function ψ which solves the free Schrödinger equation, and perform the local phase transformation

$$\psi(t,\mathbf{x}) \to \psi'(t,\mathbf{x}) = e^{i\alpha(t,\mathbf{X})}\psi(t,\mathbf{x}). \tag{25}$$

The Schrödinger equation should remain unchanged, since both states are equivalent. However, the transformed equation becomes (suppressing $(t, \mathbf{x})$ dependance)

$$i\hbar\partial_t(e^{i\alpha}\psi) = -\frac{\hbar^2}{2m}\vec{\nabla}^2(e^{i\alpha}\psi). \tag{26}$$

Clearly, when the derivatives act on the exponentials, new terms will be introduced. If we consider any single spacetime derivative, then we have

$$\begin{aligned} \partial_\mu(e^{i\alpha}\psi) &= i\partial_\mu\alpha e^{i\alpha}\psi + e^{i\alpha}\partial_\mu\psi \\ &\neq e^{i\alpha}\partial_\mu\psi. \end{aligned} \tag{27}$$

The second line is what we want. We have encountered a similar problem before, regarding co-ordinate invariance for GR. Then, we had to modify the partial derivative, making it covariant. Instead of changing co-ordinate frames, this time we are changing the phase. Let's attempt a similar manipulation. So, we seek a covariant derivative D_μ, such that under the transformation (25)

$$D_\mu\psi \to e^{i\alpha}D_\mu\psi, \tag{28}$$

so D_μ must transform as,

$$D_\mu \to D'_\mu = e^{i\alpha}D_\mu e^{-i\alpha}. \tag{29}$$

Consider, then, the object

$$D_\mu(x) = \partial_\mu + iA_\mu(x) \tag{30}$$

where $A_\mu(x)$ is some covector field. Then under the local phase rotation, (30) tell us that this covariant derivative transforms as

$$D_\mu \to D'_\mu = \partial_\mu + iA'_\mu \tag{31}$$

(since the partial derivative operator is position independent). So it is this covector field A_μ that must vary under phase changes. To find an explicit form for this variation, we simply require that (28) and (31) coincide. That is to say,

$$\begin{aligned} e^{i\alpha}(\partial_\mu + iA_\mu)\psi &= (\partial_\mu + iA'_\mu)e^{i\alpha}\psi \\ &= i\partial_\mu\alpha e^{i\alpha}\psi + e^{i\alpha}\partial_\mu\psi + iA'_\mu e^{i\alpha}\psi. \end{aligned} \tag{32}$$

Which implies

$$A'_\mu \;=\; A_\mu - \partial_\mu \alpha. \tag{33}$$

Thus, the modified Schrödinger equation

$$i\hbar(\partial_t + iA_0)\psi(t,\mathbf{x}) = -\frac{\hbar^2}{2m}(\vec{\nabla} + i\mathbf{A})^2\psi(t,\mathbf{x}) \tag{34}$$

is invariant under the simultaneous local transformations

$$\psi(t,\mathbf{x}) \to e^{i\alpha(t,\mathbf{x})}\psi(t,\mathbf{x}) \;\; and \;\; A_\mu(t,\mathbf{x}) \to A_\mu(t,\mathbf{x}) - \partial_\mu\alpha(t,\mathbf{x}). \tag{35}$$

4.3 Relation to Electromagnetism

The results of this local phase invariant are quite profound. We were forced to change our momentum operator into $\frac{\hbar}{i}(\vec{\nabla} + i\mathbf{A})$. Thus, changes in a particle's phase, which alters the 1-form A_μ, result in changes in its momentum. According to Newton's second law,

$$\mathbf{F} = m\ddot{\mathbf{x}} = \dot{\mathbf{p}} \tag{36}$$

($\dot{}$ signifies time derivatives) there must exist a force which performs these changes. That force is none other than EM. The equation

$$i\hbar\partial_t\psi(t,\mathbf{x}) = -\frac{\hbar^2}{2m}(\vec{\nabla} - \frac{iq}{\hbar}\mathbf{A})^2\psi(t,\mathbf{x}) + q\phi(t,\mathbf{x})\psi(t,\mathbf{x}) \tag{37}$$

where q is a particle's charge, is empirically known to govern the motion of a charged particle in an arbitrary EM field. This is equivalent to equation (35) if we multiply α and A_μ by $\frac{q}{\hbar}$. The gauge transformations now read

$$\psi(t,\mathbf{x}) \to e^{\frac{iq}{\hbar}\alpha(t,\mathbf{x})}\psi(t,\mathbf{x}) \;\; and \;\; A_\mu(t,\mathbf{x}) \to A_\mu(t,\mathbf{x}) + \frac{q}{\hbar}\partial_\mu\alpha(t,\mathbf{x}). \tag{38}$$

So, local phase invariance introduces EM interactions.

Interestingly, the above transformations are identical to Weyl's earlier gauge transformations after a) replacing the metric $g_{\mu\nu}$ by the wave function ψ, and b) setting the undetermined constant γ to $\frac{\hbar}{i}$. The first substitution tells us that EM is a phenomenon that accompanies matter fields, and not, as Weyl thought, the spacetime metric. Changing the constant γ from real

to imaginary, takes Weyl's conformal factor from the positive real axis to the unit circle in the complex plane. Weyl's non-physical path dependant vector lengths become the well proven path dependant matter-wave phases.

Historically, these correlations were first pointed out by Schödinger and London in the 1920's, though only in a rather tentative manner. Then, in 1929, Weyl published the paper *Electron and Gravitation* which introduced many now fundamental concepts, including a derivation of EM from the gauge principle. Before its release, Weyl published a short summary to which Pauli, upset by the mathematician's intrusion into physics, replied,

> I admire your courage; since the conclusion is inevitable that you wish to be judged, not for your success in pure mathematics, but for your true but unhappy love for physics.

However, after reading the whole article, Pauli wrote back saying,

> Here I must admit your ability in Physics. Your earlier theory with $g'_{ik} = \lambda g_{ik}$ was pure mathematics and unphysical. Einstein was justified in criticizing and scolding. Now the hour of your revenge has arrived.

5 Gravity as a Gauge Theory

The procedure given above, to construct locally gauge invariant systems, may be generalized to more complex symmetries. Doing so results in the introduction of more complicated fields, and hence, new forces. We have briefly seen how GR is a gauge (co-ordinate) independent theory. Let us develop this more formally.

Note: To properly understand the gauge structure of GR requires much more mathematical machinery than this paper shall develop. Instead, this section shall sketch the proper approach to formulate a gauge invariant theory of gravitation.

5.1 Vierbein Formalism

The Minkowski spacetime metric, as a matrix, is $\eta_{mn} = diag(-1, 1, 1, 1)$. Latin indices will be used to denote co-ordinates in this Minkowski basis, while Greek indices are reserved for arbitrary reference frames. According

to the Equivalence Principle, we may always choose our reference frame, at each point in spacetime, as a Minkowskian one, so that there appears to be no gravitational force! By re-expressing the Minkowski basis $\{x^m\}$ in terms of a general one $\{x^\mu\}$, we see the effect of gravity. By formulating our physics in the $\{x^m\}$ basis, they are independent of any relabelling $\{x^\mu\}$. This is the gauge invariance of gravity. (Proceeding in this manner is particularly useful for constructing *spinors* in curvilinear co-ordinates.)

The effects of gravity are contained in the changes of $\{x^m\}$ from point to point. Expressing this basis in terms of the general one, we have

$$dx^m = h^m_\mu(x)dx^\mu \quad where \quad h^m_\mu(x) \equiv \partial_\mu x^m.$$

The transformation matrix $h^m_\mu(x)$ is called the *vierbein*. Likewise, we define the inverse vierbein by

$$dx^\mu = h^\mu_m(x)dx^m \quad where \quad h^\mu_m(x) \equiv \partial_m x^\mu.$$

We see, then, that an arbitrary spacetime metric can always be written

$$g_{\mu\nu}(x) = \eta_{mn}h^m_\mu(x)h^n_\nu(x). \tag{39}$$

Lorentz transformations Λ^n_m (rotations in spacetime) are equivalent to a change of basis. By requiring invariance under local Lorentz transformations, we must introduce the covariant derivative for the $\{x^m\}$ basis, which depends on some new field ω_μ,

$$D_m = h^\mu_m D_\mu = h^\mu_m(\partial_\mu + i\omega_\mu), \tag{40}$$

with x dependance suppressed. To understand this operator in curved spacetime, consider the expression $h^m_\mu D_\nu v_m$. Though the calculations are beyond the scope of this paper, the end result is that

$$h^m_\mu D_\nu v_m = \partial_\nu v_\mu + \Gamma^\lambda_{\mu\nu}v_\lambda \quad with \quad \Gamma^\lambda_{\mu\nu} \equiv h^m_\mu D_\nu h^\lambda_m. \tag{41}$$

From which we may define the familiar

$$\nabla_\nu v_\mu \equiv \partial_\nu v_\mu + \Gamma^\lambda_{\mu\nu}v_\lambda.$$

If there is no torsion ($\Gamma^\lambda_{\mu\nu} = \Gamma^\lambda_{\nu\mu}$), then $\Gamma^\lambda_{\mu\nu}$ is the Christoffel connection. One important consequence of this approach is to free the covariant derivative

from the notion of parallel displacement. Instead the more fundamental gauge principle is used.

Again, this section is by no means a rigorous derivation of GR from the gauge principle, rather, it has been presented to highlight the analogies between EM and GR as gauge theories. The inclined reader should consult [3] for a complete account. Also suggested is Utiyama's paper, *Invariant Theoretical Interpretation of Interaction*, translated in [2], where a general method for constructing gauge invariant interactions is developed, with EM, Yang-Mills and GR as worked examples. Though published after Yang and Mills historic paper, it was in fact, written one year prior.

5.2 Comparison of Gauge Groups

Gauge theory has a very natural formulation in terms of groups. Recall that a group, G is a set of elements x satisfying

i) $\exists\ e \epsilon G$ such that $ex = x$, $\forall\ x \epsilon G$
ii) $\forall\ x \epsilon G$, $\exists\ y \epsilon G$ such that $yx = e$
iii) If $x, y \epsilon G$, then $xy \epsilon G$

Some known aspects of the gauge group structures of EM and gravitation will be presented to highlight the differences between the two as gauge theories.

Recall the gauge principle states that "systems invariant under a global group of transformations, should remain invariant when that group is considered locally". For EM, that group is, of course, the phase transformation $e^{i\alpha(t,\mathbf{X})}$. Each element may be represented as a point on the unit circle in the complex plane, formally this group is known as $U(1)$. It has a commutative structure and a compact topology. Also, it depends on a single parameter $\alpha(t, \mathbf{x})$. On the other hand, GR is invariant under change of co-ordinates; that is rotations and translations in spacetime. These transformations form a group, as well, called the *Poincaré group* (PG). There are 10 independent parameters: six for the Lorentz transformations and four for spacetime translations. Although translations are commutative, the (generalized) rotations are not, thus PG has a non-commutative structure. Furthermore, PG is non-compact since translations are unbounded.

Some other notable distinctions between the forces are the following. In EM, the nature of interactions is determined by the sign of the charge q, while for gravity all matter is attractive. Apart from the charge, all EM

objects arise from the 1-form (also called the connection) A_μ, while in GR it is not the connection $\Gamma^\lambda_{\mu\nu}$ but the 2-form $g_{\mu\nu}$ which plays the primary role.

The most remarkable fact, though, is the similarity between the two forces. The gauge principle is truly fundamental to the nature of each force.

6 Conclusions

Perhaps the most important result is that we have found the necessary condition to make symmetries local. It appeared in GR, and again for EM: the covariant derivative. The gauge principle, presented near the outset of the paper, should include the provision:

> *that partial derivatives be replaced by covariant ones, which depend on some new vector field.*

These fields correspond to the four known force fields. By generalizing the procedure of section 4.2, we may develop gauge invariant theories for the nuclear forces as well.

Clearly, the gauge principle's role in modern physics is key. From it, we can determine the nature of the force fields, as well as their interactions with matter. It serves as the base upon which unification theories are built. Whether a unified theory for all interactions will ever be developed is hard to say. That the gauge principal will play a central role in it is hard to deny.

7 References

1. C. Cohen-Tannooudji, et.al, *Quantum Mechanics.* John Wiley and Sons, Toronto, 1977.
2. L. O'Raifeartaigh, *The Dawning of Gauge Theory.* Princeton University Press, New Jersey, 1997.
3. P. Rammond, *Field Theory: A Modern Primer.* Addison-Wesley Publishing, Don Mills, 1989.
4. E. Scholz, *Hermann Weyl's* Raum-Zeit-Materie *and a General Introduction to His Work.* Birkhäuser Verlag, Boston, 2001.
5. H. Weyl, *Space Time Matter.* Dover Publications, New York, 1950.