On the scale at which general relativity breaks down

D. H. Delphenich †

2801 Santa Rosa Drive, Kettering, OH 45440

Key words: Planck scale, quantum limit of general relativity, pre-metric electromagnetism, effective models for quantum electrodynamics, emergent gravity

The basis for the definition of the Planck scale as the proper scale at which general relativity ceases to be valid is criticized on the grounds that if general relativity is defined by the Lorentzian geometry of the spacetime manifold and the Lorentzian structure is obtained as a degenerate case of a more general quartic dispersion law for the propagation of electromagnetic waves then the proper scale at which general relativity breaks down would be the scale at which that degeneracy occurs, such as from vacuum birefringence due to vacuum polarization. This might imply that general relativity probably breaks down when one enters the cloud of vacuum polarization that surrounds any elementary charge, according to charge renormalization.

PACS 03.50.De, 04.20.Cv, 04.60.-m, 11.10.Lm, 12.20.-m

1 Introduction

Customarily, it has been assumed that general relativity, which attributes the presence of gravitation in spacetime to the curvature of a Lorentzian metric g that describes the fundamental geometric object, is valid down to the so-called "Planck scale." In fact, when Max Planck defined this scale of units in 1899 [1], it appeared as essentially an appendix to a paper on the thermodynamics of radiation. There, it came about as a matter of convenience when one wished to find a system of units in which the speed of light c, Newton's gravitational constant G, and two thermodynamic constants that were used in the paper would have value unity.

Although the use of "natural units," in which c and the Planck constant $\hbar$ have the value unity was widely used by theoreticians of the early Twentieth Century, the scale of length, time, mass, and temperature that Planck had defined remained largely unnoticed, probably due to the unphysical nature of the numbers. It was not until 1955 that John Wheeler [2] first suggested the possibility of using the Planck scale as the scale of quantum gravity, although he did not actually mention Planck in that discussion. The context of its introduction in that paper was that of trying to establish a scale at which one expects to find "geons," which is what Wheeler was calling the fundamental solutions to the Einstein-Maxwell field equations for gravitation and electromagnetism when they are coupled by including the Faraday stress-energy-moment tensor of the electromagnetic field in the source term to the Einstein equation for the Lorentzian metric that accounts for gravitation. (See also the discussion in [3], as well.)

Nowadays, the Planck scale of length is often defined by assuming that the Einstein-Hilbert action functional for g, which then gives the Einstein field equations for gravitation, takes on a value that is comparable to $\hbar$; this was also one of the heuristic

† E-mail: david_delphenich@yahoo.com.

possibilities that Wheeler examined in [2]. One thus defines a length scale $l_{\rm Pl} = \sqrt{\hbar G / c^3} = 1.62\times10^{-35}$ m, in which G is Newton's gravitational constant.

Sometimes the Planck scale is defined by means of dimensional analysis on the grounds that it is the only characteristic length that one can obtain from the relevant fundamental constants, but the weakness in all such arguments is the assumption that merely because physics does not know of any other fundamental constants they cannot exist, either; after all, it was not so long ago that the constants G, c, and h had not been known. In particular, when Planck defined the scale, essentially all that was known of quantum physics in that era was his work on blackbody radiation. It was some years later that experimental physics even established the rest of mass m_e of an electron, and when one adds that fundamental constant to $\hbar$ and c, one immediately derives a characteristic length $\lambda_C = \hbar/(m_e c) = 3.862 \times 10^{-13}$ m that one calls the *Compton wavelength* of the electron, which is much more accessible to experiments, and certainly much more characteristic of the quantum phenomena that have been established that way.

One can obtain other associated scales of physical dimensions from $l_{\rm Pl}$ and the basic relations of physics, such as:

Planck time:	$t_{\rm Pl} = l_{\rm Pl} / c$	$= 5.39 \times 10^{-44}$ s,
Planck frequency:	$f_{\rm Pl} = 1/ t_{\rm Pl}$	$= 1.86 \times 10^{43}$ Hz,
Planck energy:	$E_{\rm Pl} = h f_{\rm Pl}$	$= 1.22 \times 10^{19}$ GeV (1.95 GJ),
Planck mass:	$m_{\rm Pl} = E_{\rm Pl} / c^2$	$= 21.8$ ng.
Planck temperature:	$T_{\rm Pl} = E_{\rm Pl} / k$	$= 1.42 \times 10^{32}$ $^{\rm o}$K.

In order to give these numbers some scientific significance, apart from their intriguing mathematical aspects, one must identify some class of natural phenomena in which they would appear naturally. Here, one sees that the dynamically-oriented constants $E_{\rm Pl}$ and $T_{\rm Pl}$ have a certain mind-boggling unphysicality, unless, perhaps they apply to the very earliest history of the Big Bang singularity. In particular, the Planck energy scale is 16 orders of magnitude beyond the scale (~ TeV) of existing particle accelerators and 6 orders of magnitude beyond the scale of ultra-high cosmic rays [1]. One might well ask whether the very mathematical formalism of quantum field theory – i.e., perturbative power series expansions in coupling constants or $\hbar$ – is actually sufficiently fundamental in its description of elementary processes to survive an extrapolation through so many orders of magnitude.

There are two conceptually inconsistent extremes that relate to the experimental application of this hypothetical Planck scale. One sees that its subatomic dimensions as a length suggests a breakdown of general relativity at very small distances, although unless one is considering small neighborhoods of elementary matter, the only pathologies that physics reasonably expects to find in small volumes of interstellar space would relate to *low-energy* phenomena, such as the zero-point energy of the quantum electrodynamic vacuum. However, when one treats the Planck length as a minimum wavelength for electromagnetic waves, the effect is to introduce an exceedingly *high-energy* scale, which

[1] Although the Planck energy scale is far beyond the imaginable scale of elementary particle collisions, it does, however, correspond to the explosive energy of only about one-half a metric ton of TNT.

could only possibly relate to the physics of the very early ($< t_{Pl}$) Big Bang, unless one imagines that even a region of interstellar space resolves at some finer level of detail to a turbid caldera of high-energy phenomena, such as the formation and annihilation of wormholes in spacetime foam, that is somehow confined so completely below the Planck scale that its macroscopic effect is so placid as to be mundane.

The low-energy interpretation of the Planck scale of distance was once suspected to relate to the coupling of the cosmological constant Λ in general relativity to the zero-point energy of the zero-point quantum electromagnetic field. However, this immediately led to a serious contradiction in orders of magnitude [4-9]. On the one hand, Λ, which Einstein introduced [10] into general relativity as a means of stabilizing the field equations for gravitation has an experimentally established upper bound of $(3.4 \pm 0.4) \times 10^{-10}$ J/m^3 . [1]

On the other hand, since this is such a small number, it was reasonable to assume that Λ was proportional to the zero-point energy density ρ_{ZPE} of the electromagnetic field, as predicted by quantum electrodynamics [11]. Its existence as a natural phenomenon is, by now, well-established experimentally through the Casimir effect [12, 13], which has been verified for plate separations down to 100 nm (10^{-11} m).

However, as has been known for some time [3], if one tries to relate Λ to ρ_{ZPE} by introducing an ultraviolet wave number cutoff k_{UV} (or similarly, a minimum wavelength) to regularize quantum electrodynamics and assumes that this cutoff is on the order of the $1/l_{\mathrm{Pl}}$ then the resulting value of Λ becomes proportional to $E_{\mathrm{Pl}}^4 / c^3 = 3.1 \times 10^{111}$ J/m^3, which is then off by a factor of about 10^{120}. It seems only reasonable to view this as a definitive contradiction to the use of the Planck energy scale to explain the zero-point energy of electromagnetism. By comparison, the established lower bound for plate separation in the Casimir effect corresponds to a maximum cutoff wavelength and a maximum vacuum energy density of about $(1 \text{ eV})^4$; 1 eV is then roughly the rest energy of two electrons.

As for the possibility that the vacuum of deep space might be hiding a more violent undertow, by way of spacetime foam, this, too, seems unlikely, if the results of recent experiments are to be considered [14]. One theory was that interstellar particles might scatter off of the foam in a manner that would be analogous to the scattering of electrons and X-rays by atomic crystal lattices, and the effect might be measurable for sufficiently high-energy particles, such as ultra-high energy cosmic rays, which are really massive particles, such as protons. It was expected that this scattering of cosmic particles by the wormholes of spacetime foam would produce vacuum Čerenkov radiation, which might then be measurable on Earth. However, the experiments done by researchers in the field of ultra-high energy cosmic rays at the Pierre Auger Observatory (see [14], for example) seemed to indicate no such effect. Another theory is that gravitational waves might be affected by spacetime foam, as well, although their experimental detection in any form is still being pursued.

One begins to suspect that the Planck scale has nothing scientifically relevant to say about the low-energy phenomena of gravitation and electromagnetism. This probably originates in the basic flaws in the Dirac sea conception of the vacuum state that makes the UV cutoff necessary.

---

[1] This number can also be expressed as $(2.6 \pm 0.6 \text{ MeV})^4$, or about two protons per cubic meter.

As for the high-energy interpretation of the Planck scale, the usual way of justifying unphysically high energy scales in quantum physics is to invoke spontaneous symmetry breaking, which seems to be well-established in the context of the electroweak interaction, which was a learned borrowing from condensed matter. One then says that the physical phenomena that one is describing ceased to be valid once the energy scale of spontaneous symmetry breaking was reached on the downside. For instance, since the gauge fields of the strong interaction – viz., gluons – do not seem to possess a non-zero mass that might result from spontaneous symmetry breaking in the same way as for the gauge fields of the electroweak interactions (viz., the Higgs mechanism), it was proposed that the proper energy scale of spontaneous symmetry breaking to consider might be the Grand Unified Scale ($\sim 10^{16}$ GeV), at which the electroweak interaction was "unified" with the strong interaction by way of a larger gauge group. By analogy, if there is a scale at which the gravitational interaction is unified with the other three then this might reasonably be the Planck scale. The larger symmetry group was then broken after the early universe cooled past T_{Pl}, at which point, gravity "emerged" as a distinct interaction from the electro-weak-strong interaction.

The problem with such an argument is that in condensed matter physics spontaneous symmetry breaking also seems to be intimately related to the concept of critical point phase transitions, which happen only in some small neighborhood of the critical value at which they take place. Hence, the question is whether one might reasonably expect to find "signals" of that phase transition at many orders of magnitude before one reaches it, or would be that like looking for experimental evidence of ice melting at temperatures near $0\ ^{\text{o}}\text{K}$ under standard atmospheric pressure.

The alternative viewpoint on the breakdown of general relativity that will be suggested here is that since the Lorentzian metric of spacetime, which accounts for the presence of *gravity*, is actually due to the *electromagnetic* structure of the spacetime manifold – in particular, the dispersion law for the propagation of electromagnetic waves – it might be more experimentally promising to focus one's attention on the point at which that dispersion law (which is, in general, a homogeneous quartic polynomial in the frequency-wavenumber 1-form k) degenerates into the square of a homogeneous quadratic polynomial of Lorentzian type. Since this transition is associated with the disappearance of the birefringence that is associated with the quartic polynomial, which is assumed to be associated with vacuum polarization (among other things), this means that one is not addressing natural phenomena that ceased to occur 13.7 billion years ago, but ones that might manifest themselves in more reasonable scales of experimental phenomena, such as elementary particle collisions at attainable energies; for instance, Compton scattering or Møller scattering.

One finds many of the aspects of this sequence of model reductions being discussed in the name of quantum gravity, although in the context of the Planck scale. For instance, it subsumes the notions of modified Maxwell models [14], vacuum birefringence [15], emergent gravity and rainbow spacetimes [16], and Lorentz-symmetry breaking [17], although the class of natural phenomena that we are addressing pertain to the small neighborhoods of elementary charge distributions, which is probably more on the order of the Compton wavelength than the Planck length, rather than small neighborhoods of the Big Bang singularity. Hence, one expects that the phenomenological consequences of

the theory presented below are likely to be more experimentally tractable than those of Planck scale physics.

The structure of the present article is to first discuss the rudiments of pre-metric electromagnetism, which appears to be the most natural setting for describing the "emergence" of the Lorentzian structure of spacetime from something more fundamental. This approach is then applied to the case of the Heisenberg-Euler effective model for the interaction of elementary charges with external electromagnetic fields when one includes the one-loop corrections due to vacuum polarization. One sees that vacuum birefringence is a likely consequence of the resulting effective electromagnetic constitutive law and dispersion relation. In Section 4, we then review the manner by which charge renormalization requires that "bare" charges be "dressed" by a cloud of vacuum polarization and discuss the possibility that this cloud is also associated with vacuum birefringence. Gravity, in the sense of a Lorentzian metric, then emerges when one goes beyond the effective radius of that cloud. Inside the cloud, presumably, a more complicated situation prevails that might possibly relate to the problem of accounting for the stability of many elementary charge distributions, such as electrons and positrons.

2 Pre-metric electromagnetism

Before we discuss the pre-metric form of Maxwell's equations, let us first recall the metric form as it is presented in the relativistically-invariant language of differential forms on an orientable four-dimensional Lorentzian manifold M. We will then show how one generalizes to the pre-metric form and then how one obtains the Lorentzian metric from the dispersion law associated with those equations.

2.1 Metric form of Maxwell's equations

Ordinarily, one encounters Maxwell's equations of electromagnetism in vacuo in the Lorentz-invariant form:

$$dF = 0, \qquad d{*}F = {*}J, \qquad d{*}J = 0. \tag{2.1}$$

in which F is the electromagnetic field strength 2-form, d represents the exterior derivative operator, J represents the electric source current 1-form, and * represents the Hodge duality operator that is associated with the Lorentzian metric g that the spacetime manifold M is equipped with.

One must include the last equation in (2.1) since the second equation is over-determined and does not have to admit solutions for the arbitrary source current J, so one must specify a compatibility – or really, *integrability* – condition on the acceptable choices of J. One sees that this condition on J amounts to the conservation of charge under the motion of the charge distribution that defines it.

If one considers the component expressions for F and J in terms of the natural coframe field dx^μ defined on a local coordinate chart (U, x^μ), namely:

$$F = \tfrac{1}{2} F_{\mu\nu}\, dx^{\mu} \wedge dx^{\nu}, \qquad J = J_{\mu}\, dx^{\mu}, \tag{2.2}$$

then the Maxwell equations can also be written in component form as:

$$F_{\lambda\mu,\nu} + F_{\mu\nu,\lambda} + F_{\nu\lambda,\mu} = 0, \qquad (\sqrt{-g}\, F^{\mu\nu})_{,\nu} = J^{\mu}, \qquad (\sqrt{-g}\, J^{\mu})_{,\mu} = 0. \tag{2.3}$$

in which the comma refers to the partial derivative operator with respect to the coordinate in question and the factor of $\sqrt{-g}$ enters the equations as an artifact of the divergence operator. One can also use the covariant derivative defined by the Levi-Civita connection that comes from g, but, due to its vanishing torsion, when one anti-symmetrizes it the effect on differential forms is the same as the exterior derivative.

It had been observed, as early as 1922, by Kottler [18] that the only role that the Lorentzian metric plays in Maxwell's equations is in the definition of the * operator. In particular, *: $\Lambda^2 M \to \Lambda^2 M$ is the composition $\#_g \cdot i_g{}^{\wedge}\, i_g$ of two linear isomorphisms. The first one is $i_g{}^{\wedge}\, i_g : \Lambda^2 M \to \Lambda_2 M$, which takes 2-forms to bivector fields by essentially "raising the indices:"

$$i_g{}^{\wedge}\, i_g(F) = \tfrac{1}{2} F^{\mu\nu}\, \partial_{\mu} \wedge \partial_{\nu}, \tag{2.4}$$

with:

$$F^{\mu\nu} = g^{\mu\kappa} g^{\nu\lambda} F_{\kappa\lambda} = \tfrac{1}{2} (g^{\mu\kappa} g^{\nu\lambda} - g^{\mu\lambda} g^{\nu\kappa})\, F_{\kappa\lambda}\,. \tag{2.5}$$

Hence, one can define a fourth-rank contravariant tensor field χ_g that represents $i_g{}^{\wedge}\, i_g$ and has the local components:

$$\chi_g^{\kappa\lambda\mu\nu} = \tfrac{1}{2} (g^{\mu\kappa} g^{\nu\lambda} - g^{\mu\lambda} g^{\nu\kappa}). \tag{2.6}$$

Because it takes 2-forms to bivector fields, it will be anti-symmetric in the first and last pairs of indices, and will also be symmetric in the permutation of these pairs:

$$\chi_g^{\kappa\lambda\mu\nu} = -\, \chi_g^{\lambda\kappa\mu\nu} = -\, \chi_g^{\kappa\lambda\nu\mu} = \chi_g^{\mu\nu\kappa\lambda}\,. \tag{2.7}$$

The second linear isomorphism $\#_g : \Lambda_k M \to \Lambda^{n-k} M$ takes k-fields back to n–k-forms by way of the Poincaré duality that comes from the (pseudo-)*Riemannian* volume element $V_g \in \Lambda^4 M$ that one assumes exists on M; of course, this, in turn, assumes that M is orientable. The local form of the 4-form V_g is:

$$V_g = \sqrt{-g}\, V, \tag{2.8}$$

in which:

$$V = dx^0 \wedge dx^1 \wedge dx^2 \wedge dx^3 = \frac{1}{4!} \varepsilon_{\kappa\lambda\mu\nu} dx^{\kappa} \wedge dx^{\lambda} \wedge dx^{\mu} \wedge dx^{\nu}\,. \tag{2.9}$$

The four-form V also defines a volume element locally, but although the volume of a compact region of spacetime that it defines will remain unchanged under Lorentz transformations, it will generally change under general linear transformations by the

determinant of the matrix; in that sense, V is not "generally covariant." The multiplication by $\sqrt{-g}$ then makes V_g generally covariant.

The isomorphism $\#_g$ then takes the vector field $\mathbf{v} = v^\mu \partial_\mu$ and the bivector field $\mathbf{B} = 1/2\, B^{\mu\nu}\, \partial_\mu \wedge \partial_\nu$ to the 3-form $\#_g\mathbf{v}$ and 2-form $\#_g\mathbf{B}$, whose components are then:

$$v_{\lambda\mu\nu} = \sqrt{-g}\ \varepsilon_{\kappa\lambda\mu\nu}\, v^\kappa, \qquad B_{\mu\nu} = \tfrac{1}{2}\sqrt{-g}\ \varepsilon_{\kappa\lambda\mu\nu}\, B^{\kappa\lambda}, \tag{2.10}$$

Hence, one can think of the * isomorphism as defined by a doubly-covariant, doubly contravariant tensor field whose components are:

$$[*]^{\kappa\lambda}_{\mu\nu} = \tfrac{1}{4}\sqrt{-g}\ \varepsilon_{\mu\nu\alpha\beta}\,(g^{\alpha\kappa} g^{\beta\lambda} - g^{\alpha\lambda} g^{\beta\kappa}). \tag{2.11}$$

2.2 Pre-metric form of Maxwell's equations

The innovation that was introduced by Kottler, and later commented upon by Cartan [19], elaborated upon by Van Dantzig [20], and discussed by numerous other mathematicians and theoretical physicists (cf., Hehl and Obukhov [21] or some of the papers of the author – e.g., [22] – for more thorough lists of references) was that if one enlarges the scope of Maxwell's equations from the ones that pertain to the classical electromagnetic vacuum as a medium to the ones that pertain to more general electromagnetic media then the role of the isomorphism $i_g \wedge i_g$ is essentially the same as the role of an electromagnetic constitutive law, at least in the linear case.

Hence, if $\Lambda^2_x M$ is the six-dimensional vector space of 2-forms at the point $x \in M$ and $\Lambda_{2,x} M$ is the six-dimensional vector space of bivectors at the same point then a very general class of electromagnetic constitutive laws that are addressed by physics, more generally, is described by defining a diffeomorphism χ_x: $\Lambda^2_x M \to \Lambda_{2,x} M$, at each point; that is, an invertible differentiable map between the two vector spaces whose inverse is also differentiable. One can express such a constitutive law in local components in the form:

$$H^{\mu\nu} = \tfrac{1}{2}\chi^{\mu\nu\kappa\lambda}(x^\alpha, F^{\alpha\beta}) F_{\kappa\lambda}\,. \tag{2.12}$$

The bivector field:

$$\mathbf{H} = \tfrac{1}{2} H^{\mu\nu}(x^\alpha, F^{\alpha\beta})\ \partial_\mu \wedge \partial_\nu \tag{2.13}$$

does not actually transform tensorially under changes of local frame field unless the constitutive law is linear, which then makes the components of χ functions of only x, but not F. One calls **H** the *electromagnetic excitation* bivector field that is associated with the field strength 2-form F since the usual situation in the electrodynamics of continuous media (cf., e.g., [23, 24]) involves the formation of electric and magnetic dipoles in a medium as a result of the imposition of an electromagnetic field.

In addition to linearity, other considerations that one applies to electromagnetic constitutive laws are homogeneity, which relates to whether the $\chi^{\mu\nu\kappa\lambda}(x^\alpha, F^{\alpha\beta})$ are

functions of position x^α and isotropy, which relates to whether they are invariant under spatial rotations. Of course, both properties are usually well-defined only with respect to a choice of local frame field, such as the natural one that we have chosen, since otherwise one would be assuming that the manifold is an orbit of the action of the translation group or Euclidian rotation group, respectively. Moreover, the concept of spatial rotations implies not only the introduction of a Euclidian metric on a "spatial" sub-bundle of the tangent bundle $T(M)$, but also a choice of spatial sub-bundle.

In the absence of a metric, one must use a volume element V with the local form (2.9) instead of (2.8). Thus, the relevant Poincaré isomorphism $\# : \Lambda_k M \to \Lambda^{n-k} M$ must use V accordingly, and equations (2.10) become:

$$v_{\lambda\mu\nu} = \varepsilon_{\kappa\lambda\mu\nu} v^\kappa, \qquad B_{\mu\nu} = \tfrac{1}{2} \varepsilon_{\kappa\lambda\mu\nu} B^{\kappa\lambda}. \tag{2.14}$$

The pre-metric form of Maxwell's equations is obtained by first replacing the isomorphism $i_g \wedge i_g$ with χ and then the * operator with the composition $\kappa = \# \cdot \chi$, which has the local component representation:

$$\kappa^{\kappa\lambda}_{\mu\nu} = \tfrac{1}{2} \varepsilon_{\mu\nu\alpha\beta} \chi^{\alpha\beta\kappa\lambda} . \tag{2.15}$$

Once again, unless the constitutive law defined by χ is linear these functions do not define the local components of a fourth-rank tensor field. Note that, in effect, the previous factor of $\sqrt{-g}$, which is, of course, absurd now, has been "absorbed" into the constitutive isomorphism χ.

The pre-metric Maxwell equations are then of the form:

$$dF = 0, \qquad \delta \mathbf{H} = \mathbf{J}, \qquad \mathbf{H} = \chi(F), \qquad \delta \mathbf{J} = 0, \tag{2.16}$$

or, in local form:

$$F_{\lambda\mu,\nu} + F_{\mu\nu,\lambda} + F_{\nu\lambda,\mu} = 0, \qquad H^{\mu\nu}{}_{,\nu} = J^\mu, \qquad H^{\mu\nu} = \tfrac{1}{2} \chi^{\mu\nu\kappa\lambda} F_{\kappa\lambda} , \qquad J^\mu{}_{,\mu} = 0. \tag{2.17}$$

In (2.16), we have introduced the *divergence operator* δ: $\Lambda_k M \to \Lambda_{k-1} M$ that is defined by composing the exterior derivative operator d: $\Lambda^k M \to \Lambda^{k-1} M$ with the Poincaré isomorphism as follows:

$$\delta = \#^{-1} d \#. \tag{2.18}$$

Note that this operator is more intrinsic to the introduction of a volume element than the codifferential operator that one introduces on k-forms, which requires the introduction of an auxiliary metric for its definition, since the vanishing of divergence for a vector field is more intrinsically related to the fact that its flow preserves the volume element than any metric concepts.

2.3 Emergence of the Lorentzian structure from the dispersion law

So far, the formulation of Maxwell's equations in the absence of a Lorentzian metric seems to take on the character of a merely mathematical refinement of the equations. However, it actually points to a subtle profundity in the theory of electromagnetism: Although a Lorentzian metric g – or really, its curvature – is usually regarded as a manifestation of the presence of *gravity* in the spacetime manifold, nonetheless, it is also a *consequence of the electromagnetic structure* of that manifold, in the sense of the electromagnetic constitutive law that one assumes to be in effect. That is, one still calls the hypersurfaces in the tangent spaces of M that are defined by the isotropic vectors *light cones*, and not gravity cones, since they have as much to say about the way that the spacetime manifold supports the propagation of electromagnetic waves as they do about the motion of things in the presence of gravitating bodies.

This fact also suggests that there might be something formally incorrect about the posing of the traditional Einstein-Maxwell unification problem that occupied so much of Einstein's later research, which strives to attain a field theory that is based upon some more general spacetime tensor field that subsumes both the fields g and F in some way, and can be shown to imply both the Einstein field equations for gravitation and Maxwell's equations for electromagnetism as special cases. Indeed, we see that, in another sense, the theories of electromagnetism and gravitation are already intimately linked by the manner in which the fact that spacetime supports the propagation of electromagnetic waves implies the appearance of light cones, which then accounts for the appearance of gravity.

This link is defined by the dispersion law that follows from the pre-metric Maxwell equations when one assumes that electromagnetic wave solutions take on some specific form. Indeed, it is essential to understand that the very concept of an electromagnetic wave is synonymous with "wave-like solution of the chosen field equations," which then implies that one must define what constitutes a wave-like solution more precisely. It is only in the case of linear, homogeneous constitutive laws, for which the field equations are systems of linear partial differential equations with constant coefficients, in which one can use the concept of plane-wave as a fundamental building block of all other wave solutions, by way of the Fourier transform. For nonlinear field equations, the dispersion law will generally depend upon the form of the wave-like solution chosen.

First, let us introduce an electromagnetic potential 1-form $A = A_\mu \, dx^\mu$ such that the first Maxwell equation is replaced with:

$$F = dA \qquad (F_{\mu\nu} = A_{\mu,\nu} - A_{\nu,\mu}). \tag{2.19}$$

Of course, such a potential 1-form does not have to exist, unless all closed 2-forms on M are exact, which says that its de Rham cohomology vanishes in dimension two. The non-vanishing of that cohomology would say that M has "two-dimensional" holes in it, which can lead to the appearance of magnetic monopoles or wormholes, and, by the Poincaré lemma, imply that the general closed 2-form F will only admit a local potential 1-form. Furthermore, A is never unique, but only up to the addition of an arbitrary closed 1-form; locally, this gauge invariance is defined by the addition of an arbitrary exact 1-form $d\lambda$.

By means of A, we can consolidate the four Maxwell equations for F, given $\mathbf{J}$, into two equations for A, given $\mathbf{J}$:

$$\Box_\kappa A = \mathbf{J}, \qquad \delta\mathbf{J} = 0, \tag{2.20}$$

in which we have introduced the *field operator* $\Box_\kappa : \Lambda^1 M \to \Lambda_1 M$, which is defined as the composition:

$$\Box_\kappa = \delta \cdot \chi \cdot d = \#^{-1} d \cdot \kappa \cdot d. \tag{2.21}$$

In local components, the first equation in (2.20) becomes:

$$\left[\bar{\chi}^{\mu\nu\kappa\lambda} \frac{\partial^2}{\partial x^\lambda \partial x^\nu} + \chi^{\mu\nu\kappa\lambda}{}_{,\nu} \frac{\partial}{\partial x^\lambda} \right] A_\kappa = J^\mu, \tag{2.22}$$

in which we have defined:

$$\bar{\chi}^{\mu\nu\kappa\lambda} = \chi^{\mu\nu\kappa\lambda} + \frac{\partial \chi^{\mu\nu\kappa\lambda}}{\partial F_{\alpha\beta}} F_{\alpha\beta} = \chi^{\mu\nu\kappa\lambda} + \frac{\partial \chi^{\mu\nu\kappa\lambda}}{\partial A_{\alpha,\beta}} A_{\alpha,\beta}\,. \tag{2.23}$$

This latter expression reduces to $\chi^{\mu\nu\kappa\lambda}$ for linear constitutive laws, while the second term the bracketed expression in (2.22) vanishes whenever the chosen law is homogeneous. This sort of form for the electromagnetic constitutive law is essentially the same as the form used in the so-called "modified Maxwell" theory of Lorentz symmetry breaking [14].

We can then say that the operator $\Box_\kappa$ has the matrix form:

$$\Box^{\mu\nu} = \bar{\chi}^{\mu\kappa\nu\lambda} \frac{\partial^2}{\partial x^\lambda \partial x^\kappa} + \chi^{\mu\kappa\nu\lambda}{}_{,\kappa} \frac{\partial}{\partial x^\lambda}\,. \tag{2.24}$$

The new form (2.22) of the pre-metric Maxwell equations now represents a set of four second-order partial differential equations for the four functions A_μ. Although the integrability condition on $\mathbf{J}$ reduces the number of independent components of $\mathbf{J}$ to three, this is compensated for by a choice of gauge for A, which reduces the number of independent components of A to three, as well. Note that in the absence of a metric, one cannot define the usual Lorentz gauge on A.

The manner by which one passes from the set of field equations (2.22) to the corresponding dispersion law is straightforward in the case of linear, homogeneous constitutive laws, because one is dealing with a linear, second-order differential operator with constant coefficients in the form of:

$$\Box^{\mu\nu} = \chi^{\mu\kappa\nu\lambda} \frac{\partial^2}{\partial x^\lambda \partial x^\kappa} \qquad (\chi^{\mu\nu\kappa\lambda} \text{ constants}),$$

and the methods of Fourier analysis can be applied using the concept of a plane-wave solution:

$$A_\mu(x^\nu) = a_\mu \exp[-ik_\nu x^\nu] \qquad (a_\mu\,,\, k_\nu \text{ constants})$$

as its basic building block.

By direct differentiation in this case, one finds:

$$A_{\mu,\,\nu} = -\,ik_\nu A_\mu\,, \tag{2.25}$$

which makes:

$$\Box^{\mu\nu} = \sigma[\,\Box_\kappa\,;\,k] = -\,\chi^{\mu\kappa\nu\lambda}\,k_\lambda\,k_\kappa\,. \tag{2.26}$$

This new operator $\sigma[\,\Box_\kappa\,;\,k] : \Lambda^1 M \rightarrow \Lambda_1 M$ is linear and algebraic – indeed, it is a matrix of homogeneous quadratic polynomials in the frequency-wave number 1-form k – and one calls it the *symbol* [1] of the field operator $\Box_\kappa$.

When one asks the question of whether the spacetime vacuum (i.e., the points outside the support of **J**) supports electromagnetic waves in this context the issue becomes whether there are characteristic values of k for which the system of linear algebraic equations:

$$\sigma[\,\Box_\kappa\,;\,k]^{\mu\nu}A_\nu = 0 \tag{2.27}$$

has non-trivial solutions for A_ν.

This problem immediately reverts to the question of whether the homogeneous eighth-degree polynomial equation in k_μ:

$$\mathrm{P}[k] = \det(\sigma[\,\Box_\kappa\,;\,k]) = 0 \tag{2.28}$$

has non-trivial solutions for k_μ. This polynomial P[k] is then called the *characteristic polynomial* of the second-order differential operator $\Box_\kappa$, and its zero locus is called the *characteristic hypersurface*. In the physics of waves, it defines the *dispersion law* for the waves in question: viz., an algebraic relation, whether implicit or explicit, between the frequency ω and the wave number k_i. In fact, although we have derived the dispersion law from the field equations, nevertheless, in experimental physics, one can define it directly by measurement and then infer the form of the field equations. In that sense, the dispersion law for a class of wave solutions is more physically fundamental than the field equations.

[1] More precisely, one should call it the *principal symbol* of the operator, although it is usually only in the case of linear differential operators that symbols relate to any but the highest-order derivatives.

Actually, at this point of the analysis, equation (2.28) is trivial, as it it true for all k. This is due to the fact that the linear map $\sigma[\Box_\kappa; k]$ is simply the composition $i_k \cdot \chi \cdot e_k$ of three linear maps: The map $e_k : \Lambda^1 M \to \Lambda^2 M$ takes A to $k \wedge A$ and the map $i_k : \Lambda_2 M \to \Lambda_1 M$ takes $\mathbf{B}$ to $i_k\mathbf{B} = B^{\mu\nu} k_\nu \partial_\mu$. Neither of them are invertible, though, unless one restricts the domain in each case, since for each choice of k the kernel of e_k consists of the one-dimensional linear subspaces $[k]$ of $\Lambda^1 M$ that are spanned by all scalar multiples of k and the kernel of i_k consists of the two-dimensional linear subspaces of $\Lambda_2 M$ that are spanned by bivectors whose components $B^{\mu\nu}$ satisfy the four linear equations in six unknowns $B^{\mu\nu} k_\nu = 0$. Hence, one must first restrict the domain of e_k to a three-dimensional linear subspace that is transverse to $[k]$, which gets mapped isomorphically by χ into a three-dimensional subspace of $\Lambda_2 M$. One then intersects this with a four-dimensional subspace of $\Lambda_2 M$ that is transverse to the kernel of i_k , with a possible further reduction in dimension from three to two or one. Ultimately the dimension of the subspace of $\Lambda^1 M$ that one must reduce to in order to make $\sigma[\Box_\kappa; k]$ invertible onto its image depends upon the nature of the constitutive law χ.

We shall continue to use the notation P[k] to describe the polynomial that results from the reduction of $\Lambda^1 M$ to the subspace on which $\sigma[\Box_\kappa; k]$ is one-to-one, as we shall have no further use for the unreduced expression.

In the case of conventional linear optics [23, 25-27], the constitutive law is purely dielectric in character (i.e., the medium is magnetically isotropic and homogeneous), and linear subspace of $\Lambda^1 M$ is two-dimensional, being the intersection of the three-planes defined by:

$$k \wedge A = 0, \qquad A(\mathbf{k}) = 0, \qquad (\mathbf{k} = \eta^{\mu\nu} k_\nu \partial_\mu), \tag{2.29}$$

the latter condition is derived from the choice of Lorentz gauge $dA = \partial_\mu A^\mu = 0$ when A represents a plane-wave.

The restriction of the matrix $\sigma[\Box_\kappa; k]^{\mu\nu}$ to a three-dimensional subspace becomes a 3×3 matrix, which makes the characteristic polynomial a homogeneous sextic polynomial in k and the restriction to a two-dimensional subspace gives a 2×2 matrix and a homogeneous quartic polynomial in k.

The latter case is what one encounters in conventional linear optics in the course of Fresnel analysis. In general, the Fresnel (wave) hypersurface is a complicated self-intersecting surface in three-dimensional projective space, whose homogeneous coordinates are defined by $k_\mu = (\omega, k_i)$ and whose inhomogeneous coordinates $n_i = k_i / \omega$ have the units of indices of refraction (i.e., one over velocity). The points of self-intersection give rise to the phenomenon of *conical refraction.*

When one fixes the components k_i of the spatial wave covector the characteristic polynomial becomes a (generally inhomogeneous) polynomial in ω^2. Hence, it will generally have as many roots – up to multiplicity – as the degree of the polynomial, although they will also be sign pairs since one only encounters powers of ω^2; the two signs then describe the possible directions of the wave normal covector $n_i = k_i / \omega$ for the plane in question. In Fresnel analysis, one encounters either two roots (up to sign) of multiplicity one or one root (up to sign) of multiplicity two. The latter condition is

referred to as *birefringence* – or *double refraction* – and manifests itself in optical media as producing two different refracted images from the same object according to the state of polarization of the electromagnetic waves it emits.

The next issue to address becomes the factorizability of the polynomial into the product of homogeneous quadratic polynomials. In the homogeneous quartic case of linear optics, if this factorization is possible then P[k] will be of the form:

$$\mathrm{P}[k] = g(k,k)\bar{g}(k,k), \tag{2.30}$$

in which $g(k, k)$ and $\bar{g}(k,k)$ define homogeneous quadratic polynomials of Lorentzian type; i.e., in some frame, their components are $\eta^{\mu\nu}$ = diag[+1, –1, –1, –1], although not necessarily simultaneously. This sort of situation is often referred to as birefringence, although a better choice of term, following [28], is *bimetricity*, since birefringence can still be present in the case where P[k] does not factorize.

The zero locus of P[k] in projective space $\mathbf{R}\mathrm{P}^3$, when it takes the form (2.30), consists of the union of two intersecting surfaces that are generally ellipsoidal. In $\mathbf{R}^4$, they will be intersecting cones through the origin with ellipsoidal cross-sections.

Finally, the factorization of P[k] can degenerate to:

$$\mathrm{P}[k] = g(k, k)^2. \tag{2.31}$$

In such a case, the two cones coalesce into a single one and the characteristic equation that follows from the constitutive law becomes the usual definition of a light cone:

$$g(k, k) = 0. \tag{2.32}$$

Ultimately, one must understand that the form of the dispersion law defined by P[k] for plane waves depends solely upon the nature of the constitutive law defined by χ.

All of the aforementioned gives one an intuition as to how to proceed in the more involved case of nonlinear or inhomogeneous constitutive laws. As mentioned above, in the case of nonlinear constitutive laws, the dispersion law will generally depend upon the form of the chosen wave-like solution, since Fourier analysis no longer applies. One then finds that the form of the chosen wave-like solution can also be crucial in obtaining dispersion laws that are mathematically tractable in terms of physical interpretation.

Various approaches to defining wave-like solutions of nonlinear field equations have been defined. For instance, one can simply continue to use plane-wave solutions, with the understanding that they can no longer be linearly superposed to obtain general solutions.

The Hadamard approach [29] to treating wave motion is particularly suited to the demands of nonlinear waves. One regards a wave as a singular hypersurface, across which the field ψ in question suffers a jump discontinuity $[\psi]$ of a particular type that is defined by a sufficiently differentiable function on the hypersurface. The characteristic equation then follows from an analysis of the Cauchy problem using the singular hypersurface as the initial Cauchy hypersurface; this approach has been preferred by Hehl and Obukhov [21] for pre-metric electromagnetism.

One can also generalize the notion of plane-wave in a manner that is widely used in geometrical optics by replacing the constant *amplitude* vector field a_μ with a more general vector field $a_\mu(x)$ that defines shape of the wave envelope and replacing the linear function $k_\mu\, x^\mu$ with a more general *phase function* $\theta(x)$, whose level hypersurfaces are then referred to as *isophases.* Depending upon the type of coordinate system chosen, the isophases might take the form of moving hyperplanes, concentric spheres or ellipsoids, etc.; in the case of concentric spheres the components $a_\mu(x)$ generally take on a $1/r$ dependency.

The frequency-wave number 1-form k is then obtained from the differential of the phase function:

$$k = d\theta. \tag{2.33}$$

Hence, k is the linear part of θ; since the constant part is arbitrary, it can be set to zero.

One then defines a broad variety of wave-like 1-forms on M by the definition:

$$A_\mu(x) = e^{-i\theta(x)}\, a_\mu(x). \tag{2.34}$$

Upon differentiation, one obtains:

$$A_{\mu,\,\nu} = -\, ik_\nu A_\mu + e^{-i\theta(x)}\, a_{\mu,\,\nu}\,. \tag{2.35}$$

One then invokes the *geometrical optics (or eikonal) approximation*, which consists of assuming that the absolute values of the components $a_{\mu,\,\nu}$ are small when compared to those of k_ν:

$$A_{\mu,\,\nu} \approx -\, ik_\nu A_\mu\,. \tag{2.36}$$

One notes that the only difference between this and the corresponding result (2.25) for plane-waves is in the nature of k_ν, which is not required to be constant, at this point, although one must always have:

$$dk = 0 \qquad (k_{\mu,\,\nu} = k_{\nu,\,\mu}); \tag{2.37}$$

hence, k is irrotational.

Inserting the form (2.36) into the field equations (2.20), with $\mathbf{J} = 0$, gives the following form for the symbol of $\Box_\kappa$:

$$\sigma[\,\Box_\kappa;\, k_\alpha,\, k_{\alpha,\beta},\, A_\alpha]^{\mu\nu} =$$

$$-i\left[\left(\chi^{\mu\lambda\kappa\nu} - i\frac{\partial\chi^{\mu\lambda\beta\nu}}{\partial A_{\alpha,\kappa}}A_\alpha k_\beta\right)k_{\kappa,\lambda} + \chi^{\mu\lambda\kappa\nu}{}_{,\lambda}k_\kappa - i\left(\chi^{\mu\lambda\kappa\nu} - i\frac{\partial\chi^{\mu\nu\kappa\lambda}}{\partial A_{\alpha,\beta}}A_\alpha k_\beta\right)k_\kappa k_\lambda\right]. \tag{2.38}$$

This simplifies somewhat if one restricts the form of θ to make k_μ constants, but this really brings one back to plane waves, in effect. In such an event:

$$\sigma[\square_\kappa; k_\alpha, A_\alpha]^{\mu\nu} = -i\left[\chi^{\mu\lambda\kappa\nu}{}_{,\lambda}k_\kappa - i\left(\chi^{\mu\lambda\kappa\nu} - i\frac{\partial\chi^{\mu\nu\kappa\lambda}}{\partial A_{\alpha,\beta}}A_\alpha k_\beta\right)k_\kappa k_\lambda\right], \tag{2.39}$$

which now gives the symbol of the field operator that is associated with plane waves when the medium is inhomogeneous and nonlinear.

For a homogeneous, but nonlinear medium it takes the form:

$$\sigma[\square_\kappa; k_\alpha, A_\alpha]^{\mu\nu} = -\left(\chi^{\mu\lambda\kappa\nu} - i\frac{\partial\chi^{\mu\nu\kappa\lambda}}{\partial A_{\alpha,\beta}}A_\alpha k_\beta\right)k_\kappa k_\lambda, \tag{2.40}$$

which differs from the linear case (2.26) by the dependency upon A_α and the fact that the nonlinear contribution to $\chi^{\mu\lambda\kappa\nu}$ raises the degree of the characteristic polynomial:

$$\mathrm{P}[k, A] = \det(\sigma[\square_\kappa; k_\alpha, A_\alpha]) \tag{2.41}$$

to twelve, before reductions. After reductions, it is an inhomogeneous sextic polynomial in k.

One can still treat the analysis of the characteristic polynomial as if it were defined by:

$$\mathrm{P}[k, A] = -\bar{\chi}^{\mu\lambda\kappa\nu}(k, A)k_\kappa k_\lambda, \tag{2.42}$$

with:

$$\bar{\chi}^{\mu\lambda\kappa\nu}(k, A) = \chi^{\mu\lambda\kappa\nu} - i\frac{\partial\chi^{\mu\nu\kappa\lambda}}{\partial A_{\alpha,\beta}}A_\alpha k_\beta. \tag{2.43}$$

Note, in particular, the fact that the components $\bar{\chi}^{\mu\lambda\kappa\nu}(k, A)$ are now complex-valued functions.

The reduction of the domain of $\sigma[\square_\kappa; k_\alpha, A_\alpha]$ to a subspace on which it is injective is, of course, somewhat complicated by the fact that the map it defines is not actually linear, any more.

3 Effective models in quantum electrodynamics

One of the most perplexing obstacles [1] to the reconciliation of the general relativistic theory of gravitation with the quantum field theory of the other three fundamental interactions is the radically differing mathematical formalisms that one must adopt for each.

[1] For a discussion of some of the other issues, see the comments of Padmanabhan [30].

General relativistic gravitation is a true *field theory*, in the sense that it poses a system of partial differential equations for the fundamental field – viz., the Lorentzian metric g of spacetime – and then examines the phenomenological consequences of the field equations by posing boundary-value and initial-value (i.e., Cauchy) problems in various limiting cases that pertain to reasonable physical assumptions. This is similar to what one does in potential theory, classical electromagnetism, heat flow, and continuum mechanics.

Quantum field theory, due to the speculative, "over-the-horizon" nature of the subatomic structures and phenomena relative to the macroscopic world of the experimenter, has long since decided not to start by speculating on what the fundamental system of partial differential equations would be, much less examining the phenomenological nature of the boundary-value and initial-value problems. One reason is simply that such a model would have to be based upon some deeper knowledge of the internal structure of elementary matter, and the only knowledge that physics has comes from high-energy particle scattering experiments. Hence, the problem is analogous to that of constructing models of the Earth's interior by means of seismic-wave scattering data that one measures on the surface. Another reason is that one simply expects that the system of partial differential equations would be coupled and nonlinear in such way as to be analytically intractable.

Therefore, quantum field theory skips over the issue of the fundamental system of field equations and immediately passes from the Cauchy problem of propagating an incoming scattering state at a finite initial time t_0 to a later, but still finite, final time t_1 to the scattering approximation of letting t_0 go to $-\infty$ while t_1 goes to $+\infty$. This comes from approximating the interaction of the particles in the initial state to a small time interval in which all of the nonlinearity gets localized, while the asymptotic scattering states are then related by a linear operator that can be represented by a scattering Green function. Furthermore, one usually passes to "momentum space" by means of the Fourier transform, imposes various causality constraints, and obtains a momentum-space propagator for the Fourier transform of the scattering operator.

Since this momentum-space propagator for the incoming scattering state is usually presumed to be too complicated to construct directly, one then resorts to perturbation expansions of the momentum-space propagator, which are usually either vertex expansions in powers of the coupling constant or loop expansions in powers of $\hbar$. The way that one obtains the various terms can be defined by either canonical field operator methods – i.e., second quantization – or functional integral methods. Either starting point seems to produce the same perturbation expansions, along with their associated Feynman diagrams, so to experimentalists the diagrams are more fundamental than the theories that spawned them. One must note that the asymptotic expansions do not necessarily converge if one goes to high enough powers of the expansion parameter, and what one is ultimately defining in momentum space is not a nonlinear differential operator, but a linear pseudo-differential operator. Hence, one can see that a mere inversion of the Fourier transform will not produce a useful system of nonlinear field equations, at least, not partial differential equations.

The ways that quantum field theory then attempts to close the loop of the scientific method and bring the theory back into the realm of experiments and established theories in classical limits are to derive scattering data – such as differential and total cross

sections, particle lifetimes, and branching ratios – from the resulting propagator, and to derive *effective* field theories [31] from the loop expansions.

Although the introduction of loops into the perturbation series introduces unphysical infinities that must be removed through regularization and renormalization, which we shall discuss in the next section, for now we simply say that what an effective field theory gives one is some inkling of how to extend the classical model for the interaction by introducing quantum corrections. For instance, one can correct the classical Coulomb potential for a spherically-symmetric electric charge with quantum terms (cf., Berestetski, Lifschitz, and Pitaevski [32]). The effective field theory is generally thought of as something that becomes valid in some low-energy limit, while the energy scale is often defined by spontaneous symmetry breaking.

One of the early effective models for quantum electromagnetism, and one of the most enduring ones, was obtained by Heisenberg and Euler [33] in 1936 and recast by Schwinger [34] in 1951. It started with the quantum-electrodynamical formulation of the interaction of an electric charge with an external electromagnetic field when one includes the "one-loop" quantum corrections to the propagator for the interaction that come from the polarization of the electromagnetic vacuum state in the realm of large field strengths for the combined electromagnetic field of the charge and external field.

After renormalization of the infinity that the loop introduces, the Heisenberg-Euler model then integrates out the higher-energy field modes (i.e., two or more loops) as being passive to the interaction and obtains the effective field Lagrangian $L = L_0 + L_1$, in which:

$$L_0 = \tfrac{1}{4} F_{\mu\nu} F^{\mu\nu} = \tfrac{1}{2} \kappa(F, F) \equiv \tfrac{1}{2} I_1 \tag{3.1}$$

is the classical Maxwellian Lagrangian and:

$$L_1 = \frac{\alpha}{2\pi} \int_0^\infty d\eta \frac{e^{-\eta}}{\eta^3} \left\{ (E_c^2 - \tfrac{1}{3}\eta^2 I_1) - i\eta^2 I_2 \frac{\cos\left(\frac{\eta}{E_c} \sqrt{I_1 - iI_2} + c.c. \right)}{\cos\left(\frac{\eta}{E_c} \sqrt{I_1 - iI_2} - c.c. \right)} \right\} \tag{3.2}$$

is the Heisenberg-Euler one-loop quantum correction.

In this latter expression $\alpha = 1/137$ is the fine-structure constant and:

$$I_2 = \tfrac{1}{2} F_{\mu\nu} {}^*F^{\mu\nu} = V(F, F). \tag{3.3}$$

In the metric theory of electromagnetism, the scalar functions I_1 and I_2 are the only gauge-invariant and Lorentz-invariant expressions that one can construct from F.

The expression for L_1 also involves a critical value $E_c = m_e^2 c^3 / e\hbar$ for the magnitude of the electric field strength at which the electromagnetic vacuum polarizes by creating a particle-anti-particle pair; numerically, it amounts to 1.3×10^{18} V/m; the corresponding critical value B_c of magnetic field strength amounts to 4.4×10^{13} Gauss.

Generally, the expression (3.2) is used in a simplified form by expanding it in a Taylor series in α and truncating the series after the first term, since that would be

consistent with the one-loop nature of the quantum formulation. This gives, in the limit of field strengths that are less than the critical field strength:

$$L_1 \approx \frac{\alpha}{360\pi E_c^2}\left(I_1^2 + \frac{7}{4} I_2^2 \right). \tag{3.4}$$

From this expression, one can obtain a one-loop quantum correction to the electromagnetic constitutive law for the classical vacuum. Note that the multiplicative constant in front of the parentheses has a numerical magnitude of about 10^{-38} m^2/V^2, which gives one an idea of the extent to which the quantum correction affects electromagnetic phenomena when the field strengths are much lower than the critical value.

The classical constitutive law is linear, isotropic, and homogeneous, and defined by two constants: ε_0, which is the electric permittivity of the vacuum and μ_0, which is its magnetic susceptibility. The resulting dispersion law is simply:

$$\omega(k_i) = c(\delta^{ij} k_i k_j)^{1/2} \qquad [c^2 = 1/(\varepsilon_0 \mu_0)] \tag{3.5}$$

which can be expressed as:

$$\eta^{\mu\nu} k_\mu k_\nu = 0 \qquad (\eta^{\mu\nu} = \mathrm{diag}[+1, -c^2, -c^2, -c^2]). \tag{3.6}$$

Thus, we have the conventional Lorentzian structure of Minkowski space being defined by the dispersion law for plane-waves in the classical vacuum.

With the one-loop quantum correction, the classical constitutive law becomes:

$$D^i = -\varepsilon(F)\, \delta^{ij} E_j + \gamma(F)\, B^i, \tag{3.7}$$
$$H_i = +\gamma(F)\, E_i + (1/\mu(F))\, \delta_{ij} B^j, \tag{3.8}$$

in which the new electric permittivity and magnetic susceptibilities are:

$$\varepsilon(F) = \left(1 + \frac{\alpha}{360\pi}\frac{I_1}{E_c^2}\right)\varepsilon_0, \qquad \mu(F) = \left(1 + \frac{\alpha}{360\pi}\frac{I_1}{E_c^2}\right)^{-1}\mu_0, \tag{3.9}$$

while the electromagnetic coupling coefficient is:

$$\gamma(F) = \frac{7\alpha}{360\pi}\frac{I_2}{E_c^2}. \tag{3.10}$$

This constitutive law is no longer linear or isotropic, in general; it is however, what some [35] call *bi-isotropic*. It does remain homogeneous, in the sense that the only dependency of the parameters upon spacetime position is by way of the electromagnetic field.

The characteristic polynomial for waves of the form (2.34) that follows from this in the geometrical optics approximation (see, [23, 25-27], for example) is homogeneous and quartic in k_μ and factorizes into a bi-metric form:

$$\mathrm{P}[k] = (\eta^{\kappa\lambda} + \varepsilon_1 T^{\kappa\lambda}(F))\,(\eta^{\mu\nu} + \varepsilon_2 T^{\mu\nu}(F))\,k_\kappa k_\lambda k_\mu k_\nu\,, \tag{3.11}$$

in which the ε_1 and ε_2 are small factors whose precise form is derived in [28] and:

$$T^{\mu\nu}(F) = L_0\,\delta^{\mu\nu} - H^{\mu\kappa}F_{\kappa\nu} \tag{3.12}$$

are the components of the Faraday stress-energy-momentum tensor for F, expressed in pre-metric form.

Hence, the characteristic hypersurfaces in the cotangent spaces to spacetime that are defined by the vanishing P[k] consist of pairs of conical homogeneous quadrics that are defined by the vanishing of $(\eta^{\kappa\lambda} + \varepsilon_1 T^{\kappa\lambda}(F))k_\kappa\, k_\lambda$ and $(\eta^{\mu\nu} + \varepsilon_2 T^{\mu\nu}(F))k_\mu\, k_\nu$ independently. These both represent quantum perturbations to the Lorentzian light cones that coalesce for sufficiently small field strengths compared to the critical values.

One sees that in regions of spacetime in which the electromagnetic field strengths are close to the critical values one would expect the Lorentzian metric to resolve into a product of Lorentzian metrics, while the characteristic Lorentzian quadrics – i.e., the light cones – resolve into pairs of quadrics. This phenomenon is generally referred to as *vacuum birefringence* [15], and would imply that an electromagnetic wave that propagates through such a region is diffracted through one of two possible angles, depending upon its state of polarization.

4 Charge renormalization

As mentioned above, the momentum space Green functions that one obtains for typical particle scattering processes will generally be unphysically infinite. This can imply other unphysical infinities in such particle parameters as mass and charge.

In the case of electromagnetism, some of these infinities existed in the classical theory, such as the infinite self-energy of point-like charge distributions, while others come about as a result of quantum considerations. In particular, the "Dirac sea" conception of the quantum electromagnetic vacuum state has always suffered from such unphysical infinities, due to the fact that since the negative energy spectrum goes to negative infinity, unless all of the energy states below some finite negative energy level are occupied the vacuum will be unstable; in effect, a positron at rest will tend to accelerate to infinity in the absence of external forces. Hence, Dirac proposed that the electromagnetic vacuum state has all of the negative energy levels – i.e., holes – occupied by electrons. Of course, this would suggest that the vacuum state had infinite charge and mass from the infinitude of electrons that one would need to accomplish this feat.

The process by which quantum electrodynamics attempted to remedy this glaring inconsistency in the theory was two-fold: First, the definition of the momentum-space propagator was made finite by a process of "regularization," either by introducing a

momentum (i.e., wave-number) cutoff or by passing to spatial dimensions that were not integers. Then, the field Lagrangians is "renormalized" by replacing some of its infinite parameters, such as mass, charge, and the field itself, with rescaled parameters that would render the resulting expressions independent of the value of the cutoff.

Generally, the scaling factor in this renormalization is infinite, just as Heisenberg had first proposed a "subtraction of infinities" in order to correct the infinite mass and charge of the Dirac sea. Despite the absence of mathematical rigor, the usual reason given for the practical utility of regularization and renormalization is that, in a sense, it is the resulting renormalized expression that has the physical significance, not how one obtained it. In effect, renormalization works like an "error-correcting algorithm," whose input is the wrong answer and whose output is the right one.

In particular, the charge of an electron needs to be renormalized. The input to the renormalization algorithm is the point-like "bare" charge distribution, while the output is the "dressed" charge. This takes the form of a cloud of polarized vacuum surrounding the point at which the bare charge is located. Indeed, if the renormalized charge is the one with all of the physical meaning – i.e., the one that affects the experiments – then one wonders why one needs the bare charge, at all. If renormalization works like an error-correcting algorithm then it was the wrong answer, anyhow.

There is a subtlety in the concept of vacuum polarization that must be addressed at this point: It is one thing to say that in the presence of a sufficiently strong external field strength a sufficiently high energy photon will split into an electron-positron pair (or perhaps muon-anti-muon, etc.) and another thing to say that in the neighborhood of static elementary charges when the electric field strength reaches a critical value the space around it polarizes. In order to reconcile the two notions, one should probably think of the field of the elementary charge as interacting with the zero-point field around it as if it were a photon, which then polarizes at the critical value of field strength.

This suggests that in the eyes of quantum physics there is probably more significance to extended electron models than there is to point-like ones. Indeed, it was already established by experiment in quantum mechanics that electrons were wavelike entities with a characteristic wavelength given by the Compton wavelength. Insofar as wave functions generally tend to be of finite spatial support, this suggests that one might regard the electron as being defined by the region of space in which vacuum polarization comes about, which one then imagines will probably have a characteristic dimension on the order of the Compton wavelength.

If this cloud of vacuum polarization that defines any elementary charge distribution is associated with vacuum birefringence then one sees that the length scale at which general relativity – i.e., Lorentzian geometry – breaks down is most likely the Compton scale, which is much more experimentally tractable than the Planck length seems to be.

5. Discussion

The basic gist of the foregoing presentation was that many of the phenomena that are commonly expected in the name of quantum gravity are more tangibly represented in the realm of small neighborhoods of elementary charges than in the neighborhood of the Big Bang singularity. In particular, one finds applications for the notions of Lorentz

symmetry breakdown due to vacuum birefringence and the emergence of the gravitational field from a more fundamental level of theory and phenomena, namely, the electromagnetic constitutive law that pertains to the interiors of such elementary charge distributions. Indeed, it seems misguided to search for the emergence of gravitation from a single event that happened 13.7 billion years ago when gravity manages to "emerge" from stellar – and indeed, planetary – nucleons and electrons on a regular basis to this very day.

Therefore, the methods of pre-metric electromagnetism seem to simultaneously embody promising paths for not only the unification of the theories of classical electromagnetism and gravitation, but also the reconciliation of classical electromagnetism with its quantum counterpart. Moreover, since the physical phenomena that relate to the theory are more experimentally tractable than those at the Planck scale, as well as more amenable to analogue models from other experimentally established topics in physics, one should expect more rapid progress in the near future when it comes to refining the theory with the results of experiments.

A possible class of experiments that might serve to probe the structure of the cloud of polarization surrounding elementary charges would be that of polarized Compton scattering, in which one measures the change in polarization of photons as they scatter from electrons to see if the vacuum birefringence affects the outcome. Now, the Compton wavelength for an electron corresponds to a photon wavelength in the high gamma range, and, in fact, only twice the critical wavelength for polarization, one suspects that one would have to look for such an effect in the scattering of photons by either electrons of high-momenta or slow-moving elementary charges of larger mass.

So far, polarized Compton scattering seems to be more of an issue in astrophysics, although in that milieu one is not free to determine the initial states of photons, but must live with the ones that are emitted by stellar bodies and events.

References

[1] M. Planck, Sitz. d. Preuss. Akad. Wiss. Berlin **5**, 440-480 (1899).

[2] J. A. Wheeler, Phys. Rev. **97** (2), 511-536 (1955); reprinted in Geometrodynamics, edited by J. A. Wheeler (Academic Press, New York, 1962).

[3] R. J. Adler, Preprint arXiv:1001.1205 (2010).

[4] S. Weinberg, Rev. Mod. Phys. **61**, 1-23 (1989).

[4] T. Padmanbhan, Phys. Rep. **380**, 235-320 (2003).

[6] G. Mahajan, S. Sakar, and T. Padmanabhan, Phys. Lett. B **641**, 6-10 (2006).

[7] G. E. Volovik, Int. J. Mod. Phys D**15**, 1987-2010 (2006).

[8] B. Broda and M. Szanecki, in Proceedings of the Grassmannian Conference in Fundamental Cosmology (Grasscosmofun'09), 14-19 September 2009, Szczecin, Poland. Preprint arXiv:0910.5145.

[9] J. Rafelski, L. Labun, Y. Hadad, and P. Chen, lecture given at the Tenth Workshop on Non-Perturbative Quantum Chromodynamics held at l'Institut Astrophysique de Paris June 8-12, 2009. Preprint arXiv:0909.2989.

[10] A. Einstein, Sitz. d. Preuss. Akad. Wiss. Berlin **1**, 142-152 (1917); English translation in The Principle of Relativity (Dover, Mineola, NY, 1952).

[11] P. Milonni, The Quantum Vacuum: An introduction to quantum electrodynamics (Academic Press, Boston, 1994).
[12] H. G. B. Casimir, Kon. Ned. Akad. Weten. Proc. **51**, 793 (1948).
[13] V. M. Mostpanenko and N. N. Trunov, The Casimir Effect and its Applications (Clarendon Press, Oxford, 1997).
[14] F. R. Klinkhamer and M. Risse, Phys. Rev. D **77**:016002 (2007).
[15] W. Dittrich and H. Gies, Talk given at Workshop on Frontier Tests of Quantum Electrodynamics and Physics of the Vacuum, Sandansky, Bulgaria, June 9-15, 1998. Preprint arXiv:hep-ph/9806417.
[16] C. Barceló, S. Liberati, and M. Visser, Living Rev. Rel. **8**, 12 (2005).
[17] V. A. Kostelecky and M. Mewes, Phys. Rev. D **66**:056005 (2002).
[18] F. Kottler, Sitz. Akad. Wien IIa, **131**, 119-146 (1922).
[19] E. Cartan, On manifolds with an affine connection and the theory of relativity (English translation by A. Ashtekar of a series of French articles from 1923 to 1926) (Bibliopolis, Napoli, 1986).
[20] D. van Dantzig, Proc. Camb. Phil. Soc. **30**, 421-427 (1934).
[21] F. W. Hehl and Y. N. Obukhov, Foundations of Classical Electrodynamics (Birkhäuser, Boston, 2003).
[22] D. H. Delphenich, Ann. d. Phys. (Leipzig) **14**, 347 (2005).
[23] L. D. Landau, E. M. Lifschitz, and L. P. Pitaevskii, Electrodynamics of Continuous Media, 2nd edition (Pergamon, Oxford, 1984).
[24] E. J. Post, Formal Structure of Electromagnetics (Dover, Mineola, NY, 1997).
[25] M. Born and E. Wolf, Principles of Optics (Pergamon, Oxford, 1980).
[26] R. K. Luneburg, Mathematical Theory of Optics (The University of California Press, Berkeley, 1964).
[27] M. Kline and I. W. Kay, Electromagnetic Theory and Geometrical Optics (Wiley-Interscience, New York, 1965).
[28] M. Visser, C. Barcelo, and S. Liberati, 2002 Festschrift in honour of Professor Mario Novello (Preprint arXiv.org gr-qc/0204017, 2002).
[29] J. Hadamard, Leçons sur la propagation des ondes et les equations de l'hydrodynamique (Chelsea, New York, 1949).
[30] T. Padmanabhan, in Festschrift volume in honour of Professor J. V. Narlikar, eds. Naresh Dadhich and Ajit Kembhavi (Kluwer, Amsterdam, 1999).
[31] G. V. Dunne, From Fields to Strings: circumnavigating theoretical physics, edited by M. Shifman, A. Vainshtein, and J. Wheater (World Scientific, Singapore, 2004).
[32] V. B. Berestetskii, E. M. Lifschitz, and L. P. Pitaevskii, Quantum Electrodynamics, 2nd edition (Elsevier, Amsterdam, 1984).
[33] W. Heisenberg and H. Euler, Zeit. f. Phys. **98**, 714-732 (1936).
[34] J. Schwinger, Phys. Rev. **82**, 664-679 (1951); reprinted in: Selected Papers on Quantum Electrodynamics, edited J. Schwinger (New York: Dover, 1958).
[35] I. V. Lindell, Differential Forms in Electromagnetics (IEEE Press, New Jersey, 2004).