IOP PUBLISHING EUROPEAN JOURNAL OF PHYSICS

Eur. J. Phys. **30** (2009) 1401–1416 doi:10.1088/0143-0807/30/6/018

Maxwell equations and the redundant gauge degree of freedom

Chun Wa Wong

Department of Physics and Astronomy, University of California, Los Angeles,
CA 90095-1547, USA

E-mail: cwong@physics.ucla.edu

Received 14 July 2009, in final form 12 August 2009
Published 12 October 2009
Online at stacks.iop.org/EJP/30/1401

Abstract

On transformation to the Fourier space $(\mathbf{k}, \omega)$, the partial differential Maxwell equations simplify to algebraic equations, and the Helmholtz theorem of vector calculus reduces to vector algebraic projections. Maxwell equations and their solutions can then be separated readily into longitudinal and transverse components relative to the direction of the wave vector $\mathbf{k}$. The concepts of wave motion, causality, scalar and vector potentials and their gauge transformations in vacuum and in materials can also be discussed from an elementary perspective. In particular, the excessive freedom of choice associated with the gauge dependence of the scalar and the longitudinal vector potentials stands out with clarity in Fourier spaces. Since these potentials are introduced to represent the instantaneous longitudinal electric field, the actual cancellation in the latter of causal contributions arising from these potentials separately in most velocity gauges becomes an important issue. This cancellation is explicitly demonstrated both in the Fourier space, and for pedagogical reasons again in spacetime. The physical origin of the gauge degree of freedom in the masslessness of the photon, the quantum of electromagnetic wave, is elucidated with the help of special relativity and quantum mechanics.

1. Introduction

Most textbooks on electromagnetism discuss the basic properties of the Maxwell equations and their solutions in ordinary spacetime $(\mathbf{r}, t)$ [1, 2]. Electromagnetic properties in vacuum and in materials are often nonlocal in spacetime, and are therefore awkward to describe. It is well known that many of these properties become more transparent in the Fourier space $(\mathbf{k}, \omega)$ reached by Fourier transforms. We present in this paper a pedagogical discussion of Maxwell equations in this Fourier space where their relative simplicity makes them more understandable at an introductory level. Electromagnetism is a vast subject. The main objective of this paper is to elucidate the mathematical and physical origins of the mysterious gauge degree of freedom

0143-0807/09/061401+16$30.00 © 2009 IOP Publishing Ltd Printed in the UK

in the choice of the electromagnetic scalar potential Φ and the longitudinal component $\mathbf{A}_\parallel$ of the vector potential.

It had taken a century of research, from the Coulomb force law of 1785 through the 1835 vector potential of Gauss to the Lorenz gauge of 1867, to establish the nonuniqueness of Φ and $\mathbf{A}_\parallel$ [1, 3]. Gauge transformations are discussed in almost all textbooks on electromagnetism, but the subject still appears mysterious to many students. In this paper, three issues are addressed to make gauge freedom more understandable. First, the longitudinal electric field $\mathbf{E}_\parallel$, known to be gauge independent, is shown to have fractional contributions from Φ and $\mathbf{A}_\parallel$ that differ in different gauges. Second, these two contributions to $\mathbf{E}_\parallel$ are found to be individually retarded in general, involving signals propagated at an *unphysical* gauge velocity that varies from 0 to ∞ depending on the gauge choice. However, these two retarded partial contributions are found to add to a total $\mathbf{E}_\parallel$ that is instantaneous, acting at a distance. Our conclusion thus differs from the conclusion presented or implied in many textbooks (e.g., p 242 of [1] and p 421 of [2]) and papers [4, 5] that $\mathbf{E}_\parallel$ propagates with the speed of light. Third, the physical origin of gauge freedom is traced to the masslessness of the photon, the quantum of electromagnetic waves, with the help of special relativity and quantum mechanics.

The paper is organized as follows. We begin, in section 2, with the spacetime Maxwell equations in materials in SI units. The Fourier electromagnetic fields are then introduced, and separated into longitudinal/transverse (L/T) components relative to the wave vector $\mathbf{k}$. The two Maxwell equations involving divergences are found to define longitudinal field components that are instantaneous in vacuum, acting at a distance. The two curl equations also involve the transverse fields.

Section 3 describes electromagnetic wave motion in linear isotropic materials. Electromagnetic waves involve transverse fields only. The electromagnetic wave equation contains a term dependent on both the time differential operator and a light speed $v \leqslant c$ in the material (c being the light speed in vacuum). Both features are needed for the transverse fields to describe events happening *causally* at the light speed of the material.

In section 4, the transverse part of a vector potential is used to describe the magnetic induction field. The longitudinal part of the vector potential is used together with a scalar potential to describe the instantaneous longitudinal electric field when either one of them alone can do the job. It is traditional to use a combination of both potentials. The permissible use of different combinations gives rise to an extra 'gauge' degree of freedom. Most of these gauge choices lead to potentials that are causal because they satisfy wave equations where signals are propagated with an arbitrary gauge velocity. We show explicitly that the resulting arbitrary gauge causal contributions from the two potentials cancel exactly to give a longitudinal electric field that is instantaneous, acting at a distance, as dictated by the Gauss law of electrostatics.

In section 5, our results are compared with the spacetime results of Brill and Goodman [4] and of Yang [5]. Although the technical analyses are similar, we find that the longitudinal electric field is instantaneous, not retarded, even though it is the sum of two retarded terms. The meaning of action at a distance is reviewed from the modern perspective of quantum mechanics and quantum field theory.

In teaching electromagnetism, it is important to present the cancellation of the arbitrary gauge causality in the physical instantaneous longitudinal electric field in an intuitive and appealing way. We achieve this objective in section 6 by finding a causality-cancelling combination in spacetime. Its gauge-dependent and causal parts associated with the scalar and longitudinal vector potentials are then found in $\mathbf{k}$ space. A final transformation back to spacetime shows why the simple expressions in $\mathbf{k}$ space look so complicated in spacetime. We also use the folding theorem of Fourier transforms as a qualitative test for instantaneity in time and localization in space of physical phenomena.

In section 7, the relations or transformations between different gauge choices are described. Even after a gauge is chosen, the scalar and longitudinal vector potentials can still vary by amounts proportional to the infinitely many gauge functions that are solutions of a certain homogeneous wave equation.

Finally, in section 8, we show that the gauge degree of freedom appears because light speed c in vacuum has the same value in all acceptable inertial frames called Lorentz frames. There is then always a longitudinal direction along which the magnetic induction vanishes, and the electric field component is instantaneous, acting at a distance. In the quantum description of light as a particle called the photon, this constancy of light speed comes from the masslessness of the photon. A massive photon, if it existed, would have a well-defined and dynamically meaningful longitudinal vector potential. As the photon loses its mass, its longitudinal vector potential decouples from the two transverse components, and becomes the redundant appendage called the gauge degree of freedom.

2. Longitudinal and transverse Maxwell equations in materials

The Maxwell equations for electromagnetic fields in ordinary materials in spacetime $(\mathbf{r}, t)$ in SI units, in the notation of Jackson ([1], p 248), are

$$\nabla \cdot \mathbf{B} = 0, \tag{1}$$

$$\nabla \times \mathbf{E} = -\partial_t \mathbf{B}, \tag{2}$$

$$\nabla \cdot \mathbf{D} = \nabla \cdot (\epsilon_0 \mathbf{E} + \mathbf{P}) = \rho, \tag{3}$$

$$\nabla \times \mathbf{H} = \nabla \times \left(\frac{1}{\mu_0} \mathbf{B} - \mathbf{M} \right) = \mathbf{J} + \partial_t \mathbf{D}. \tag{4}$$

Here $\partial_t \equiv \partial/\partial t$, while $\rho = \rho(\mathbf{r}, t)$, $\mathbf{J}$ are the external charge and current densities respectively. The vector fields $\mathbf{E} = \mathbf{E}(\mathbf{r}, t), \mathbf{D}, \mathbf{P}, \mathbf{B}, \mathbf{H}, \mathbf{M}$ are assumed to have well-defined second spacetime derivatives so that wave equations can be constructed. The polarization $\mathbf{P}$ and the magnetization $\mathbf{M}$ describe the responses of the material medium to the presence of external charges and currents. Note that the first two equations hold only in the absence of magnetic charges. When magnetic charges are present, an additional term ρ_m and $-\mathbf{J}_m$ should appear on the right-hand side of (1) and (2), respectively.

These fields and their first and second derivatives in spacetime are assumed to have the Fourier representations

$$\mathbf{E}(\mathbf{r}, t) = \int_{-\infty}^{\infty} \frac{\mathrm{d}\omega}{2\pi} \int \frac{\mathrm{d}^3 k}{(2\pi)^3} \, \mathrm{e}^{\mathrm{i}\mathbf{k}\cdot\mathbf{r} - \mathrm{i}\omega t} \tilde{\mathbf{E}}(\mathbf{k}, \omega), \tag{5}$$

$$\nabla \cdot \mathbf{E}(\mathbf{r}, t) = \int_{-\infty}^{\infty} \frac{\mathrm{d}\omega}{2\pi} \int \frac{\mathrm{d}^3 k}{(2\pi)^3} \, \mathrm{e}^{\mathrm{i}\mathbf{k}\cdot\mathbf{r} - \mathrm{i}\omega t} \mathrm{i}\mathbf{k} \cdot \tilde{\mathbf{E}}(\mathbf{k}, \omega), \tag{6}$$

etc. In (6), the spacetime differential operator ∇ acts directly on the Fourier basis function

$$\psi_{\mathbf{k},\omega}(\mathbf{r}, t) = \mathrm{e}^{\mathrm{i}\mathbf{k}\cdot\mathbf{r} - \mathrm{i}\omega t}. \tag{7}$$

In the theory of Fourier transforms, $\tilde{\mathbf{E}}(\mathbf{k}, \omega)$ may be taken to be the Fourier transform $\mathcal{F}$ of $\mathbf{E}(\mathbf{r}, t)$, namely

$$\tilde{\mathbf{E}}(\mathbf{k}, \omega) = \mathcal{F}\{\mathbf{E}\} = \int_{-\infty}^{\infty} \mathrm{d}t \int \mathrm{d}^3 r \, \mathrm{e}^{-\mathrm{i}\mathbf{k}\cdot\mathbf{r} + \mathrm{i}\omega t} \mathbf{E}(\mathbf{r}, t). \tag{8}$$

The representation (5) is then its inverse Fourier transform. The downside of using Fourier spaces is that it is necessary to perform this Fourier inversion if one wants the result in spacetime. So at an introductory level, Fourier spaces are most useful for the qualitative understanding of an electromagnetic concept for which the complete Fourier inversion back to spacetime is not essential.

In the Fourier space $(\mathbf{k}, \omega)$, the Maxwell equations simplify to the algebraic equations

$$\mathrm{i}\mathbf{k}\cdot\tilde{\mathbf{B}} = 0, \tag{9}$$

$$\mathrm{i}\mathbf{k}\times\tilde{\mathbf{E}} = \mathrm{i}\omega\tilde{\mathbf{B}}, \tag{10}$$

$$\mathrm{i}\mathbf{k}\cdot\tilde{\mathbf{D}} = \mathrm{i}\mathbf{k}\cdot(\epsilon_0\tilde{\mathbf{E}} + \tilde{\mathbf{P}}) = \tilde{\rho}, \tag{11}$$

$$\mathrm{i}\mathbf{k}\times\tilde{\mathbf{H}} = \mathrm{i}\mathbf{k}\times\left(\frac{1}{\mu_0}\tilde{\mathbf{B}} - \tilde{\mathbf{M}}\right) = \tilde{\mathbf{J}} - \mathrm{i}\omega\tilde{\mathbf{D}}. \tag{12}$$

These equations can be decomposed, term by term, into longitudinal (L or $\|$) and transverse (T or $\perp$) components that are respectively parallel and perpendicular to $\mathbf{k}$. In the rectangular coordinate system defined by the unit vectors $\mathbf{e}_1$, $\mathbf{e}_2$ and $\mathbf{e}_3 = \mathbf{e}_\mathbf{k} = \mathbf{k}/k$, an L/T decomposed vector takes the simple form

$$\tilde{\mathbf{E}} = \tilde{\mathbf{E}}_\| + \tilde{\mathbf{E}}_\perp, \tag{13}$$

where

$$\tilde{\mathbf{E}}_\| = \tilde{E}_k\mathbf{e}_\mathbf{k}, \qquad \tilde{\mathbf{E}}_\perp = \tilde{E}_1\mathbf{e}_1 + \tilde{E}_2\mathbf{e}_2. \tag{14}$$

Now in vector algebra, a vector can be transformed into other vectors by left or right multiplications with second-rank tensors. The nature of these transformations becomes particularly transparent when the second-rank tensors are written in the unit-tensor expansion

$$\mathbf{T} = \sum_{i,j=1}^{3} T_{ij}\mathbf{e}_i\mathbf{e}_j. \tag{15}$$

The unit tensor $\mathbf{e}_i\mathbf{e}_j$ that appears is a dyadic, i.e. a non-commuting product of two vectors [6]. As a two-sided object, a dyadic admits standard vector algebraic operations with vectors on both left- and right-hand sides, as well as sequential matrix products with other second-rank tensors. In this paper, we shall need only the two scalar products between the unit tensor/dyadic $\mathbf{e}_i\mathbf{e}_j$ and the unit vector $\mathbf{e}_m$:

$$\mathbf{e}_i\mathbf{e}_j\cdot\mathbf{e}_m = \mathbf{e}_i\delta_{jm}, \qquad \mathbf{e}_m\cdot\mathbf{e}_i\mathbf{e}_j = \delta_{mi}\mathbf{e}_j, \tag{16}$$

where δ_{jm} is a Kronecker delta. As a result,

$$\mathbf{T}\cdot\tilde{\mathbf{E}} = \sum_{i,j}\mathbf{e}_i T_{ij}\tilde{E}_j, \qquad \tilde{\mathbf{E}}\cdot\mathbf{T} = \sum_{i,j}\tilde{E}_i T_{ij}\mathbf{e}_j. \tag{17}$$

The L/T projection operators needed for the L/T separation of vector components are the tensors [7]

$$\tilde{\mathbf{I}}_\| = \mathbf{e}_\mathbf{k}\mathbf{e}_\mathbf{k}, \qquad \tilde{\mathbf{I}}_\perp = \tilde{\mathbf{I}} - \tilde{\mathbf{I}}_\| = \mathbf{e}_1\mathbf{e}_1 + \mathbf{e}_2\mathbf{e}_2. \tag{18}$$

The L/T separation of a vector field $\tilde{\mathbf{E}}$ can then be realized by using scalar and vector products:

$$\begin{aligned} \tilde{\mathbf{I}}_\|\cdot\tilde{\mathbf{E}} &= \mathbf{e}_\mathbf{k}\mathbf{e}_\mathbf{k}\cdot\tilde{\mathbf{E}} = \tilde{E}_\|\mathbf{e}_\mathbf{k} = \tilde{\mathbf{E}}_\| = \tilde{\mathbf{E}}\cdot\tilde{\mathbf{I}}_\|, \\ \tilde{\mathbf{I}}_\perp\cdot\tilde{\mathbf{E}} &= -\mathbf{e}_\mathbf{k}\times(\mathbf{e}_\mathbf{k}\times\tilde{\mathbf{E}}) = \tilde{\mathbf{E}}_\perp = \tilde{\mathbf{E}}\cdot\tilde{\mathbf{I}}_\perp. \end{aligned} \tag{19}$$

Written as a single equation,

$$\tilde{\mathbf{E}} = \tilde{\mathbf{E}}_\| + \tilde{\mathbf{E}}_\perp = \mathbf{e}_\mathbf{k}(\mathbf{e}_\mathbf{k}\cdot\tilde{\mathbf{E}}) - \mathbf{e}_\mathbf{k}\times(\mathbf{e}_\mathbf{k}\times\tilde{\mathbf{E}}), \tag{20}$$

this simple algebraic L/T separation is just the Helmholtz theorem of vector calculus, now reduced to simple vector algebraic projections in the Fourier space $(\mathbf{k}, \omega)$. We shall show in the appendix that an inverse Fourier transform of (20) gives the usual integral statement of the Helmholtz theorem in space. The scalar and vector products

$$\mathbf{k} \cdot \tilde{\mathbf{E}} = k\tilde{E}_\|, \qquad \mathbf{k} \times \tilde{\mathbf{E}} = \mathbf{k} \times \tilde{\mathbf{E}}_\perp \tag{21}$$

in the L/T separated form appear repeatedly in the Maxwell equations.

With L/T separation, the two Maxwell divergence equations become the scalar algebraic equations

$$\mathrm{i}k\tilde{B}_\| = 0, \tag{22}$$

$$\mathrm{i}k\tilde{D}_\| = \mathrm{i}k(\epsilon_0\tilde{E}_\| + \tilde{P}_\|) = \tilde{\rho} \tag{23}$$

for the $\|$ components. In free space, where $\tilde{P}_\|$ vanishes, the ω dependence of $\tilde{E}_\|$ is identical to that of $\tilde{\rho}$. Their time dependences in spacetime are therefore identical too. Hence $E_\|(\mathbf{r}, t)$ responds to the source charge density $\rho(\mathbf{r}, t)$ instantaneously, acting at a distance, as we shall show with more detail in section 6.

Each of the two Maxwell curl equations separates into individual T and L equations

$$\mathrm{i}\mathbf{k} \times \tilde{\mathbf{E}}_\perp = \mathrm{i}\omega\tilde{\mathbf{B}}_\perp, \tag{24}$$

$$\mathrm{i}\omega\tilde{\mathbf{B}}_\| = 0, \tag{25}$$

$$\mathrm{i}\mathbf{k} \times \tilde{\mathbf{H}}_\perp = \mathrm{i}\mathbf{k} \times \left(\frac{1}{\mu_0}\tilde{\mathbf{B}}_\perp - \tilde{\mathbf{M}}_\perp\right) = \tilde{\mathbf{J}}_\perp - \mathrm{i}\omega\tilde{\mathbf{D}}_\perp, \tag{26}$$

$$\tilde{\mathbf{J}}_\| - \mathrm{i}\omega\tilde{\mathbf{D}}_\| = 0. \tag{27}$$

The result $\tilde{B}_\| = 0$, stating the absence of magnetic charges, thus appears in two separate Maxwell equations (22) and (25). Equations (23) and (27) taken together give the continuity equation

$$-\mathrm{i}\omega\tilde{\rho} + \mathrm{i}k\tilde{J}_\| = 0. \tag{28}$$

In spacetime, $\mathbf{J} = \mathbf{v}\rho$, where $\mathbf{v}$ is the velocity of the charge. Then the continuity equation

$$\partial_t\rho + \nabla \cdot (\mathbf{v}\rho) = \partial_t\rho + \mathbf{v} \cdot \nabla\rho = \frac{\mathrm{d}}{\mathrm{d}t}\rho = 0 \tag{29}$$

describes charge conservation. Note in particular that charge conservation does not involve $\tilde{J}_\perp$.

3. Wave motion in linear isotropic materials

In linear materials, the contributions of charges and currents induced in the material are included through the dielectric tensor (or dyadic) $\tilde{\epsilon}$ and the permeability tensor $\tilde{\mu}$ [7]:

$$\tilde{\mathbf{D}} = \tilde{\epsilon} \cdot \tilde{\mathbf{E}}, \qquad \tilde{\mathbf{B}} = \tilde{\mu} \cdot \tilde{\mathbf{H}}. \tag{30}$$

We shall restrict ourselves to isotropic materials where the only preferred direction is $\mathbf{e_k}$. The chosen rectangular coordinate axes in $\mathbf{k}$ space are therefore also the principal axes of these physical properties. Thus,

$$\tilde{\epsilon} = \tilde{\epsilon}_\|\tilde{\mathbf{I}}_\| + \tilde{\epsilon}_\perp\tilde{\mathbf{I}}_\perp : \qquad \tilde{\mathbf{D}} = \tilde{\epsilon}_\|\tilde{\mathbf{E}}_\| + \tilde{\epsilon}_\perp\tilde{\mathbf{E}}_\perp, \tag{31}$$

where the dielectric functions $\tilde{\epsilon}_{\|}, \tilde{\epsilon}_{\perp}$ are scalar functions of k, ω only. To simplify subsequent formulae, we shall use a cruder approximation for magnetic properties:

$$\tilde{\mu}_{\|} = \tilde{\mu}_{\perp} = \tilde{\mu}(k, \omega): \qquad \tilde{\mathbf{B}} = \tilde{\mu}\tilde{\mathbf{H}}. \tag{32}$$

For these linear isotropic materials, (23) simplifies to

$$\tilde{E}_{\|} = -\frac{\mathrm{i}}{k\tilde{\epsilon}_{\|}}\tilde{\rho}. \tag{33}$$

Thus, $\tilde{E}_{\|}$ is completely determined in the Fourier space $(\mathbf{k}, \omega)$, and can be found in spacetime by calculating its (inverse) Fourier transform.

The two curl equations for the transverse fields are more complicated:

$$\mathrm{i}\mathbf{k} \times \tilde{\mathbf{E}}_{\perp} = \mathrm{i}\omega\tilde{\mathbf{B}}_{\perp}, \qquad \mathrm{i}\mathbf{k} \times \tilde{\mathbf{B}}_{\perp} = \tilde{\mu}(\tilde{\mathbf{J}}_{\perp} - \mathrm{i}\omega\tilde{\epsilon}_{\perp}\tilde{\mathbf{E}}_{\perp}). \tag{34}$$

After operating on these equations with $\mathrm{i}\mathbf{k}\times$, we can solve for the $\perp$ fields. The results are

$$\begin{aligned} \left(k^2 - \frac{\omega^2}{v^2}\right)\tilde{\mathbf{E}}_{\perp} &= \mathrm{i}\tilde{\mu}\omega\tilde{\mathbf{J}}_{\perp}, \\ \left(k^2 - \frac{\omega^2}{v^2}\right)\tilde{\mathbf{B}}_{\perp} &= \mathrm{i}\tilde{\mu}\mathbf{k} \times \tilde{\mathbf{J}}_{\perp}, \end{aligned} \tag{35}$$

where

$$v(k, \omega) = \frac{1}{\sqrt{\tilde{\mu}\tilde{\epsilon}_{\perp}}} \tag{36}$$

has the dimension of a velocity. These are inhomogeneous wave equations in the Fourier space $(\mathbf{k}, \omega)$, and v is the wave velocity or speed in the material for the given k, ω values.

To see that these equations describe waves, we go back to the associated partial differential equations in spacetime with the help of (6) under the simplifying assumption that $\tilde{\mu}, \tilde{\epsilon}_{\perp}$ and v are independent of k, ω:

$$\begin{aligned} \left(\nabla^2 - \frac{1}{v^2}\partial_t^2\right)\mathbf{E}_{\perp}(\mathbf{r}, t) &= \mu\partial_t\mathbf{J}_{\perp}(\mathbf{r}, t), \\ \left(\nabla^2 - \frac{1}{v^2}\partial_t^2\right)\mathbf{B}_{\perp}(\mathbf{r}, t) &= -\mu\nabla \times \mathbf{J}_{\perp}(\mathbf{r}, t). \end{aligned} \tag{37}$$

These partial differential equations are called wave equations, and their solutions, here the transverse fields, may be called wave functions. We shall not describe the properties of these wave equations and their solutions in spacetime, as these properties can be found in most textbooks of electromagnetism. We would only point out that it is the presence on the left-hand side of each partial differential equation of the term containing both the time derivative ∂_t ($-\mathrm{i}\omega$ in the Fourier space) and the wave speed v that makes the wave functions causal quantities involving signals propagating with the speed v. Without this term, the differential equations would have been reduced to Poisson equations. Then $\mathbf{B}_{\perp}(\mathbf{r}, t)$ would be instantaneous with $\mathbf{J}_{\perp}(\mathbf{r}, t)$ if μ is independent of t (or ω in the Fourier space). $\mathbf{E}_{\perp}(\mathbf{r}, t)$ too would be instantaneous with $\partial_t\mathbf{J}_{\perp}(\mathbf{r}, t)$, but not instantaneous with $\mathbf{J}_{\perp}(\mathbf{r}, t)$ itself. This is because $\partial_t\mathbf{J}_{\perp}(\mathbf{r}, t)$ at one time involves more than one time value of $\mathbf{J}_{\perp}(\mathbf{r}, t)$. A more detailed explanation will be given in section 6.

In many textbooks, the complete $\mathbf{E}, \mathbf{B}$ fields are expressed in a causal, actually retarded, form, when it is only their transverse parts $\mathbf{E}_{\perp}$ and $\mathbf{B}_{\perp}$ that are retarded, according to (37). Since (22) and (25) require $\mathbf{B}_{\|} = 0$, it is true that the complete $\mathbf{B} = \mathbf{B}_{\perp}$ is retarded. However, $\mathbf{E}_{\|}$ does not vanish in general. We shall show in the next section how the correct noncausal nature of $\tilde{E}_{\|}$ can be recovered from certain intermediate quantities that are both causal and dependent on an arbitrarily chosen 'gauge' velocity.

4. Scalar and vector potentials and their gauge degree of freedom

The Helmholtz theorem (20) shows that

$$\tilde{\mathbf{B}} = \tilde{\mathbf{B}}_\perp = \mathrm{i}\mathbf{k} \times \tilde{\mathbf{A}}_\perp, \quad \text{where} \quad \tilde{\mathbf{A}}_\perp = \frac{\mathrm{i}}{k^2}\mathbf{k} \times \tilde{\mathbf{B}}_\perp. \tag{38}$$

The wave equation (35) for $\tilde{\mathbf{B}}_\perp$ can then be written more simply as a wave equation for $\tilde{\mathbf{A}}_\perp$:

$$\left(k^2 - \frac{\omega^2}{v^2}\right)\tilde{\mathbf{A}}_\perp = \tilde{\mu}\tilde{\mathbf{J}}_\perp. \tag{39}$$

This wave equation depends on the light speed v in the medium, and is therefore causal, or more specifically retarded. We shall compare this wave equation for $\tilde{\mathbf{A}}_\perp$ to that for $\tilde{\mathbf{A}}_\parallel$, which is still to be found.

To find $\tilde{\mathbf{A}}_\parallel$, we begin by putting (38) into (24) to get one solution

$$\tilde{\mathbf{E}}_\perp = \mathrm{i}\omega\tilde{\mathbf{A}}_\perp \tag{40}$$

out of many mathematically permissible solutions. It would then be natural to introduce a $\parallel$ component $\tilde{\mathbf{A}}_\parallel$ by the equation $\tilde{\mathbf{E}}_\parallel = \mathrm{i}\omega\tilde{\mathbf{A}}_\parallel$, so that $\tilde{\mathbf{E}} = \mathrm{i}\omega\tilde{\mathbf{A}}$. However, the alternative expression $\tilde{\mathbf{E}}_\parallel = -\mathrm{i}\mathbf{k}\tilde{\Phi}$ defined by a mathematically equivalent scalar potential $\tilde{\Phi}$ can also be used.

Historically, different linear combinations of these equivalent forms were used that ultimately led to the inclusive expression [3]

$$\tilde{\mathbf{E}}_\parallel = -\mathrm{i}\mathbf{k}\tilde{\Phi} + \mathrm{i}\omega\tilde{\mathbf{A}}_\parallel, \tag{41}$$

and therefore

$$\tilde{\mathbf{E}} = -\mathrm{i}\mathbf{k}\tilde{\Phi} + \mathrm{i}\omega\tilde{\mathbf{A}}. \tag{42}$$

It is clear, however, that (41) contains too much freedom of choice. The choice $\tilde{\mathbf{A}}_\parallel = 0$, called the Coulomb gauge, has the advantage that the associated scalar potential called the Coulomb potential is, like $\tilde{\mathbf{E}}_\parallel$ itself, noncausal and instantaneous, acting at a distance.

A more general choice called the velocity gauge [5, 8] is defined by the *gauge condition*

$$\tilde{A}_\parallel^{(vg)} = \alpha\frac{\omega}{k}\tilde{\Phi}^{(vg)}, \tag{43}$$

where the gauge parameter

$$\alpha = \frac{1}{v_g^2} \tag{44}$$

can be written in terms of a gauge velocity v_g. Putting (43) into (41) yields

$$\tilde{E}_\parallel = -\frac{\mathrm{i}}{k}\left(k^2 - \frac{\omega^2}{v_g^2}\right)\tilde{\Phi}^{(vg)}. \tag{45}$$

Finally, the Gauss law (33) can be used to eliminate $\tilde{E}_\parallel = -\mathrm{i}\tilde{\rho}/k\tilde{\epsilon}_\parallel$ in favour of $\tilde{\rho}$ to give $\tilde{\Phi}^{(vg)}$ and $\tilde{A}_\parallel^{(vg)}$ as solutions of inhomogeneous wave equations

$$\tilde{\Phi}^{(vg)} = \frac{\tilde{\rho}}{\tilde{\epsilon}_\parallel\left(k^2 - \omega^2/v_g^2\right)}, \qquad \tilde{A}_\parallel^{(vg)} = \frac{\omega}{kv_g^2}\frac{\tilde{\rho}}{\tilde{\epsilon}_\parallel\left(k^2 - \omega^2/v_g^2\right)}. \tag{46}$$

We see that in general both $\tilde{\Phi}^{(vg)}$ and $\tilde{A}_\parallel^{(vg)}$ are causal, with the signal propagated between source and field positions at gauge velocity. Since the gauge velocity is an arbitrary parameter, this gauge causality is not a physical or gauge-independent property.

It is even more instructive to use (46) in (41) to give an expression,

$$\tilde{E}_{\|} = \frac{k^2}{k^2 - \omega^2/v_g^2}\tilde{E}_{\|} - \frac{\omega^2/v_g^2}{k^2 - \omega^2/v_g^2}\tilde{E}_{\|}, \tag{47}$$

that shows explicitly that the fractional shares residing in the $\tilde{\Phi}^{(vg)}$ and $\tilde{\mathbf{A}}_{\|}^{(vg)}$ terms vary with the arbitrary gauge velocity. There are thus infinitely many ways to cut the $\tilde{E}_{\|}$ cake, but it is always the same cake after all. The relation between $\tilde{\Phi}^{(vg)}$, $\tilde{\mathbf{A}}_{\|}^{(vg)}$ and $\tilde{\mathbf{E}}_{\|}$ is thus not one-to-one, but many-to-one. This is the redundant gauge degree of freedom.

Finally, the Lorenz gauge is realized by the choice $v_g = v = 1/\sqrt{\tilde{\mu}\tilde{\epsilon}_{\perp}}$, with the gauge velocity set to the physical wave velocity v of the transverse fields. Using this physical wave velocity as the gauge velocity does not make the resulting gauge causality any more physical than that in other choices of v_g for the instantaneous $\tilde{E}_{\|}$ field. The nonphysical nature of gauge causality, including that in the Lorenz gauge, is not sufficiently emphasized in many textbooks on electromagnetism.

Another noteworthy feature of the solution $\tilde{A}_{\|}$ given in (46) is that it satisfies the wave equation

$$\left(k^2 - \frac{\omega^2}{v_g^2}\right)\tilde{A}_{\|}^{(vg)} = \frac{1}{v_g^2\tilde{\epsilon}_{\|}}\tilde{J}_{\|}, \tag{48}$$

where $\omega\tilde{\rho}/k$ has been replaced by $\tilde{J}_{\|}$ with the help of the continuity equation (28). This wave equation is different from the wave equation (39) satisfied by $\tilde{A}_{\perp}$ unless two conditions are satisfied:

$$v_g = v \qquad \text{and} \qquad \tilde{\epsilon}_{\|} = \tilde{\epsilon}_{\perp}. \tag{49}$$

The Lorenz gauge in vacuum is so popular because these conditions are satisfied ($v_g = v = c$, and $\tilde{\epsilon}_{\|} = \tilde{\epsilon}_{\perp} = \epsilon_0$). Then all three components of $\tilde{A}$ satisfy the same wave equation. In linear isotropic materials such as plasmas, one finds that $\tilde{\epsilon}_{\|} \neq \tilde{\epsilon}_{\perp}$ in general, however. So the simplicity of the vacuum Lorenz gauge cannot be maintained for many linear materials.

5. Comparison with Brill and Goodman, and with Yang

The technique used in this paper depends on L/T separation, and is therefore the same technique as that used by Brill and Goodman [4] and by Yang [5]. The only difference is that we work in the Fourier space where all expressions are algebraic, and therefore much more transparent. The resulting simplicity is particularly striking in our discussion of the gauge degree of freedom where it is obvious that $\mathbf{A}_{\perp}$ is not involved at all. We are also able to extend the treatment to linear isotropic materials, a subject of interest in materials physics. For electromagnetism in free space already discussed in [4, 5], the important gauge independence of the electromagnetic fields **E, D, B** and **H** is correctly obtained by them, by us and by everybody else.

There are significant technical differences in the results for the Coulomb gauge. In [4] (p 833), an incorrect Coulomb-gauge condition $\nabla \cdot \mathbf{A}_{\perp} = 0$ is used. A nonzero term $\mathbf{A}^{(C,\Phi)}$ appears in [5] ((3.30) on p 745) when it should be zero because of the Coulomb-gauge condition $\nabla \cdot \mathbf{A} = 0$. These problems do not seem to affect the rest of the analyses in [4, 5].

We differ from these authors in the conclusion drawn from the similar technical analyses, however. Our conclusion in its simplest form is that the electrostatic potential of a stationary charge in vacuum is just the instantaneous Coulomb potential. So the electric field **E** generated from it is also instantaneous, acting at a distance. The current **J** of this stationary charge is zero. So the charge's vector potential **A** and its magnetic induction **B** both vanish. Here then is a simple example where an instantaneous interaction is unavoidable. Such a literal interpretation of the Maxwell equations in vacuum is accepted by many people, including Dirac [9]. Let us elaborate on our interpretation.

With the advent of quantum mechanics, the explanation of the instantaneous Coulomb interaction between two stationary charges in vacuum becomes more detailed. In this more modern picture, the Coulomb interaction arises from the exchange between the interacting charges of a virtual photon, the quantum of electromagnetic waves. This virtual photon is actually a combination of a time-like virtual photon and a longitudinal virtual photon [10], conforming to the requirement of Maxwellian electromagnetism, because these photons are respectively the quanta of the scalar and longitudinal vector potentials.

A virtual photon in quantum mechanics, like other virtual particles, does not conserve energy. It can exist momentarily only because of the Heisenberg uncertainty principle applied to the energy–time pair of quantized dynamical variables. Thus, virtual particles do not satisfy the energy–momentum relation of real particles 'on the energy shell'. In other words, virtual particles are 'off the energy shell'. This virtuality is why the potential or field can exist in all space around the source, instantaneously. Only real particles 'on the energy shell' that move with a speed $v \leqslant c$ can carry observable messages or signals. Virtual particles cannot. In particular, they cannot be detected as energy-violating objects, otherwise energy nonconservation would be observed. In other words, virtual particles are given the liberty of violating both energy conservation and causality in exchange for being forever unobservable directly.

In other respects, real and virtual particles are treated alike. For example, the same mathematical technique for describing particle emission and absorption is used in quantum mechanics for both real and virtual particles, and for both neutral or charged particles. For charged particles, Dirac has explicitly written [9]: 'Whenever an electron is emitted, the Coulomb field around it is simultaneously emitted, forming a kind of *dressing* for the electron. Similarly, when an electron is absorbed, the Coulomb field around it is simultaneously absorbed'. These processes are caused, for example, by the Hamiltonian (the quantized energy operator) which 'applies just to one instant of time', the instant of emission or absorption.

In quantum field theory, the physical vacuum ceases to be merely empty space. It is instead a sea of vacuum fluctuations, pervasive in all space and lasting for all times. An individual vacuum fluctuation exists momentarily, appearing and disappearing as if by magic. Taken together, vacuum fluctuations form a relatively smooth fabric on which the instantaneous Coulomb interaction of macroscopic electromagnetism can be realized.

The present-day picture of the physical vacuum is even more sophisticated. In the presence of charges, the vacuum acts like ordinary materials in that its permittivity ϵ_0 and permeability μ_0 can also change for sufficiently large changes in the spacetime scale of the physical phenomena involved ([11], pp 339–40). That is, they are also functions of $(\mathbf{k}, \omega)$, or equivalently of the energy scale M of the phenomena. Experiments have shown that the effective electromagnetic interaction strength e^2/ϵ_0, where e is the electronic charge, has increased by 7% from macroscopic physics where $M \approx 0$ to phenomena at the mass scale of $M = 100$ GeV. It is even expected to increase by another factor of more than 2 by the time the so-called grand unification scale of 10^{16} GeV is reached. This is the energy regime where the weak and strong interactions too can be expected to have the same effective coupling strength [12].

It thus appears that the Maxwell equations in vacuum do describe charges interacting instantaneously via the $\mathbf{E}_\parallel$ field, a physical process that is not subject to the relativistic limitation of all measurable speeds to c or less. However, the Maxwellian theory does not explain why this classical picture of action at a distance should hold.

In contrast, [4, 5] almost 40 years apart both conclude that the entire $\mathbf{E}(\mathbf{r}, t)$ in vacuum, including $\mathbf{E}_\parallel(\mathbf{r}, t)$, is retarded. Their conclusion is based on the assumption that since each of the four contributions (from the three vector potential components and one scalar potential) to the Lorenz gauge $\mathbf{E}(\mathbf{r}, t)$ is individually retarded, their total $\mathbf{E}(\mathbf{r}, t)$ must be retarded as well. This assumption is incorrect because, as we have shown in section 4, the two retarded contributions from $\tilde{\Phi}^{(vg)}$ and $\tilde{A}_\parallel^{(vg)}$ actually add up to an instantaneous $\tilde{\mathbf{E}}_\parallel$ for any gauge. The gauge independence of this instantaneous field component is an important feature of Maxwellian electromagnetism.

6. The instantaneous and gauge-independent $\mathbf{E}_\parallel$

The unexpected metamorphosis of two gauge causal terms in vacuum into an instantaneous interaction in $\tilde{\mathbf{E}}_\parallel$ is sufficiently important in teaching electromagnetism to merit an explicit demonstration of how it materializes in spacetime itself. We begin by relating the Helmholtz and Poisson equations in space:

$$\left(\nabla^2 + \frac{\omega^2}{v_g^2}\right) \frac{\mathrm{e}^{\pm \mathrm{i}\omega R/v_g}}{4\pi R} = -\delta^3(\mathbf{R}) = \nabla^2 \frac{1}{4\pi R}, \tag{50}$$

where $\mathbf{R} = \mathbf{r} - \mathbf{r}'$. The Helmholtz solutions are the retarded (upper + sign) and advanced (lower − sign) Green functions of Jackson [1], (6.40) on p 244. The Poisson Green function is just the special case when $\omega/v_g = 0$. It is neither retarded nor advanced, but instantaneous, as in electrostatics.

In the Fourier space $\mathbf{k}$, (50) becomes the trivial algebraic identity (writing the Poisson side first)

$$k^2 \frac{1}{k^2} = 1 = \left(k^2 - \frac{\omega^2}{v_g^2}\right) \frac{1}{k^2 - \omega^2/v_g^2}. \tag{51}$$

Multiplying the entire equation by $\tilde{\rho}/\epsilon_0 k^2$ and eliminating the charge density $\tilde{\rho}$ on favour of the longitudinal current density $\tilde{J}_\parallel$ from the continuity equation (28) in the second term on the Helmholtz side of the equation, we get

$$\frac{1}{\epsilon_0 k^2}\tilde{\rho} = \frac{1}{\epsilon_0 k^2}\left(k^2\tilde{\rho} - \frac{\omega k}{v_g^2}\tilde{J}_\parallel\right)\frac{1}{k^2 - \omega^2/v_g^2}. \tag{52}$$

A final multiplication by $-\mathrm{i}k$ gives

$$\begin{aligned} \tilde{E}_\parallel(\mathbf{k}, \omega) = \tilde{E}_\parallel^{(C)}(\mathbf{k}, \omega) &= -\frac{\mathrm{i}k}{\epsilon_0 k^2}\tilde{\rho} \\ &= -\mathrm{i}k\tilde{\Phi}^{(vg)} + \mathrm{i}\omega\tilde{A}_\parallel^{(vg)}, \end{aligned} \tag{53}$$

where

$$\tilde{\Phi}^{(vg)} = \frac{\tilde{\rho}}{\epsilon_0\left(k^2 - \omega^2/v_g^2\right)}, \qquad \tilde{A}_\parallel^{(vg)} = \frac{\tilde{J}_\parallel/v_g^2}{\epsilon_0\left(k^2 - \omega^2/v_g^2\right)}. \tag{54}$$

We are now ready to invert the Fourier transform in $\mathbf{k}$, as defined by

$$f(\mathbf{r}) = \mathcal{F}^{-1}\{\tilde{f}(\mathbf{k})\} = \int \frac{\mathrm{d}^3k}{(2\pi)^3}\,\mathrm{e}^{\mathrm{i}\mathbf{k}\cdot\mathbf{r}}\tilde{f}(\mathbf{k}). \tag{55}$$

We shall need the convolution or folding theorem [13] and the inverse Fourier transform from (50)

$$\mathcal{F}^{-1}\{\tilde{f}(\mathbf{k})\tilde{g}(\mathbf{k})\} = [f * g](\mathbf{r}) = \int \mathrm{d}^3r'\, f(\mathbf{r}-\mathbf{r}')g(\mathbf{r}'), \tag{56}$$

$$\mathcal{F}^{-1}\left\{\frac{1}{k^2-\omega^2/v_g^2}\right\} = \frac{\mathrm{e}^{\pm\mathrm{i}\omega r/v_g}}{4\pi r}. \tag{57}$$

A straightforward calculation then yields the desired connection between the results in the instantaneous Coulomb gauge and the general velocity gauge in the space $(\mathbf{r}, \omega/t)$, where ω/t means either ω or t:

$$\mathbf{E}_{\parallel}^{(C)}(\mathbf{r}, \omega/t) = -\nabla\Phi^{(vg)}(\mathbf{r}, \omega/t) + \mathrm{i}\omega\mathbf{A}_{\parallel}^{(vg)}(\mathbf{r}, \omega/t), \tag{58}$$

where the $\mathrm{i}\omega$ factor in the second term on the velocity-gauge side stands for the time differential operator $-\partial_t$ if the space is $(\mathbf{r}, t)$. (58) is the desired instantaneous and gauge-independent longitudinal electric field.

The functions in the $(\mathbf{r}, \omega)$ space that appear in (58) can be obtained by straightforward (inverse) Fourier transformations:

$$\begin{aligned}
\tilde{\mathbf{E}}_{\parallel}^{(C)}(\mathbf{r}, \omega) &= -\nabla \int \mathrm{d}^3r' \frac{\tilde{\rho}(\mathbf{r}', \omega)}{4\pi\epsilon_0|\mathbf{r}-\mathbf{r}'|};\\
\tilde{\Phi}_{\parallel}^{(vg)}(\mathbf{r}, \omega) &= \int \mathrm{d}^3r' \frac{\mathrm{e}^{\pm\mathrm{i}\omega|\mathbf{r}-\mathbf{r}'|/v_g}\tilde{\rho}(\mathbf{r}', \omega)}{4\pi\epsilon_0|\mathbf{r}-\mathbf{r}'|},\\
\tilde{\mathbf{A}}_{\parallel}^{(vg)}(\mathbf{r}, \omega) &= \int \mathrm{d}^3r' \frac{\mathrm{e}^{\pm\mathrm{i}\omega|\mathbf{r}-\mathbf{r}'|/v_g}\tilde{\mathbf{J}}_{\parallel}(\mathbf{r}', \omega)/v_g^2}{4\pi\epsilon_0|\mathbf{r}-\mathbf{r}'|},
\end{aligned} \tag{59}$$

where $\tilde{\mathbf{J}}_{\parallel}$ is from the Helmoltz theorem (A.6) in space

$$\begin{aligned}
\tilde{\mathbf{J}}_{\parallel}(\mathbf{r}', \omega) &= -\nabla' \int \mathrm{d}^3r'' \frac{\nabla''\cdot\tilde{\mathbf{J}}_{\parallel}(\mathbf{r}'', \omega)}{4\pi|\mathbf{r}'-\mathbf{r}''|}\\
&= -\mathrm{i}\omega\nabla' \int \mathrm{d}^3r'' \frac{\tilde{\rho}(\mathbf{r}'', \omega)}{4\pi|\mathbf{r}'-\mathbf{r}''|}.
\end{aligned} \tag{60}$$

The corresponding functions in $(\mathbf{r}, t)$ are easily obtained:

$$\begin{aligned}
\mathbf{E}_{\parallel}^{(C)}(\mathbf{r}, t) &= -\nabla \int \mathrm{d}^3r' \frac{\rho(\mathbf{r}', t)}{4\pi\epsilon_0|\mathbf{r}-\mathbf{r}'|};\\
\Phi_{\parallel}^{(vg)}(\mathbf{r}, t) &= \int \mathrm{d}^3r' \frac{\rho(\mathbf{r}', t \mp |\mathbf{r}-\mathbf{r}'|/v_g)}{4\pi\epsilon_0|\mathbf{r}-\mathbf{r}'|},\\
\mathbf{A}_{\parallel}^{(vg)}(\mathbf{r}, t) &= \int \mathrm{d}^3r' \frac{\mathbf{J}_{\parallel}(\mathbf{r}', t \mp |\mathbf{r}-\mathbf{r}'|/v_g)/v_g^2}{4\pi\epsilon_0|\mathbf{r}-\mathbf{r}'|},\\
\mathbf{J}_{\parallel}(\mathbf{r}', t_r) &= \partial_t\nabla' \int \mathrm{d}^3r'' \frac{\rho(\mathbf{r}'', t_r)}{4\pi|\mathbf{r}'-\mathbf{r}''|}.
\end{aligned} \tag{61}$$

They agree with those found in [4, 5].

From the perspective of teaching electromagnetism at an introductory level, the solutions of the Helmholtz equation given in (50) may be considered too advanced a result. The folding theorem (56) is quite elementary, however, once Fourier transforms are used. The folding theorem alone enables the concepts of locality in space and instantaneity in time to be discussed at an introductory qualitative level, without the need to actually work out the precise inverse transforms. This is possible because the folding theorem shows that two functions with the same $\mathbf{k}$ dependence must be localized relative to each other in space. This is so because both functions are the same function in space. Similarly, two functions with the same ω dependence must be instantaneous in time. Conversely, two functions with different $\mathbf{k}$ (or ω) dependences must be nonlocal (or non-instantaneous) in space (or time).

Let us then use this qualitative instantaneity/localization test on certain ω/$\mathbf{k}$ dependent expressions we have constructed in the preceding sections. Consider first electromagnetism in free space, where ϵ_0 and μ_0 are independent of $\mathbf{k}$ or ω in the Maxwellian theory respectively. Then $\tilde{E}_{\|}(\mathbf{k}, \omega) = -\mathrm{i}\tilde{\rho}(\mathbf{k}, \omega)/k\epsilon_0$ of (33) and $\tilde{\rho}(\mathbf{k}, \omega)$ have the same ω dependence but different $\mathbf{k}$ dependences. Hence, $\mathbf{E}_{\|}(\mathbf{r}, t)$ is instantaneous with, but not localized at, the charge density $\rho(\mathbf{r}, t)$. That is, $\mathbf{E}_{\|}(\mathbf{r}, t)$ is action at a distance. By the same test, both the transverse fields $\tilde{\mathbf{E}}_{\perp}(\mathbf{k}, \omega)$ and $\tilde{B}_{\perp}(\mathbf{k}, \omega)$ are non-instantaneous (i.e. causal) and nonlocal in spacetime relative to their sources.

In linear materials, where $\tilde{\epsilon}$ and $\tilde{\mu}$ depend on $\mathbf{k}$ and ω respectively, the test shows that in $\tilde{E}_{\|} = -\mathrm{i}\tilde{\rho}/k\tilde{\epsilon}_{\|}$, the instantaneity between $E_{\|}$ and ρ is lost. The external charge density ρ gives rise instantaneously to the field $D_{\|}$. The polarized response of the material is retarded relative to the external ρ, but as the polarization charges appear, they generate their electric fields simultaneously and instantaneously, in exactly the way Dirac has described.

7. Gauge transformations

Even after a gauge is chosen, the resulting potentials are still not uniquely determined, as we shall now demonstrate.

Consider the general gauge change or transformation

$$\begin{aligned}
&\tilde{\mathbf{A}}_{\|} \to \tilde{\mathbf{A}}'_{\|} = \tilde{\mathbf{A}}_{\|} + \mathrm{i}\mathbf{k}\tilde{\chi}, \\
&\tilde{\Phi} \to \tilde{\Phi}' = \tilde{\Phi} + \mathrm{i}\omega\tilde{\chi}, \\
&\alpha \to \alpha' = \alpha + \beta.
\end{aligned} \tag{62}$$

The change $\Delta\tilde{\mathbf{A}}_{\|} = \mathrm{i}\mathbf{k}\tilde{\chi}$ comes from the spatial vector field $\nabla\chi(\mathbf{r}, t)$ generated from a scalar field $\chi(\mathbf{r}, t)$ whose dimensional unit of measurement is chosen to be consistent with the dimension of $\mathbf{A}(\mathbf{r}, t)$. The change $\Delta\tilde{\Phi} = \mathrm{i}\omega\tilde{\chi}$ then has the same dimension as $\tilde{\Phi}$. Furthermore, the resulting electric field component

$$\tilde{E}'_{\|} = -\mathrm{i}k\tilde{\Phi}' + \mathrm{i}\omega\tilde{A}'_{\|} = \tilde{E}_{\|} \tag{63}$$

remains unchanged, i.e. gauge independent, for any gauge function $\tilde{\chi}$.

We should now mention the indispensable role $A_{\|}$ plays indirectly in the Lorentz covariance of the Maxwell equations discovered by Einstein in 1905 [14]. This Lorentz covariance refers to the invariance in form of the Maxwell equations under a Lorentz transformation to another special inertial frame called a Lorentz frame (where the time variable also changes in order to keep light speed at its universal value c). One can find in most textbooks on electromagnetism how this Lorentz covariance is assured if the 4-potential $\tilde{A}_{\alpha} = (\tilde{\mathbf{A}}, \mathrm{i}\tilde{\Phi}/c)$ is a 4-vector in the Minkowski four-dimensional spacetime. It then transforms under Lorentz transformations like any other 4-vector such as $k_{\alpha} = (\mathbf{k}, \mathrm{i}\omega/c)$. $\tilde{A}_{\|}$ is a needed component of this 4-potential. Furthermore, the new potential defined by (62), namely $\tilde{A}'_{\alpha} = \tilde{A}_{\alpha} + \mathrm{i}\tilde{\chi}k_{\alpha}$, is also a 4-vector, transforming correctly under Lorentz transformations. Electrodynamics, including quantum electrodynamics, becomes mathematically simpler and physically more transparent when expressed in terms of these 4-potentials.

The yet undetermined gauge function $\tilde{\chi}$ must also satisfy the gauge condition

$$\tilde{A}'_{\|} = (\alpha + \beta)\frac{\omega}{k}\tilde{\Phi}'. \tag{64}$$

Used with the old gauge condition (43), this equation can be simplified into the inhomogeneous wave equation

$$[k^2 - (\alpha + \beta)\omega^2]\tilde{\chi} = -\mathrm{i}\omega\beta\tilde{\Phi}. \tag{65}$$

For $\beta = 0$, the gauge condition or choice (43) is unchanged. The driving term on the right-hand side of (65) then vanishes. The resulting homogeneous wave equation is satisfied either when $\tilde{\chi} = 0$ (an uninteresting possibility or a trivial solution), or when $\tilde{\chi} = \tilde{\chi}_0 \neq 0$, but $k^2 = \alpha\omega^2$. Since k^2 comes from the space differential operator ∇^2, while ω^2 comes from the time differential operator ∂_t^2, the second possibility is satisfied only by the solutions $\chi_0(\mathbf{r}, t)$ of the homogeneous wave equation in spacetime. The functions $\tilde{\chi}_0(\mathbf{k}, \omega)$ that now appear in (65) are their Fourier transforms. These gauge functions define a class of gauge changes called restricted gauge transformations.

The gauge change (62) even allows for a change of the gauge condition if $\beta \neq 0$. Then an additional gauge function that is the solution of the inhomogeneous wave equation (65), namely

$$\tilde{\chi} = -\mathrm{i}\frac{\omega\beta\tilde{\Phi}}{k^2 - (\alpha + \beta)\omega^2}, \tag{66}$$

is needed in addition to those for restricted gauge transformations that do not change the gauge condition. These more general gauge transformations are called unrestricted gauge transformations.

Other gauges have been used with gauge conditions specified in spacetime. For such gauges, our treatment in the Fourier space $(\mathbf{k}, \omega)$ does not appear to offer any advantage. These special gauges have been discussed in the recent review by Jackson [8].

8. The physical origin of the gauge degree of freedom

Why should there be a gauge degree of freedom? We follow the explanation given by Zee ([11], pp 168, 171, 456), and elaborate on it.

The physical origin of the gauge degree of freedom lies at one of the most striking features of the Maxwell equations, a feature abstracted by Einstein into one of the two pillars of his special theory of relativity. As he wrote in his epochal 1905 paper on 'The electrodynamics of moving bodies' [14], 'light is always propagated in empty space with a definite velocity c which is independent of the state of motion of the emitting body'. In other words, the direction of light propagation in vacuum might change in a different Lorentz frame of reference. Light speed in vacuum will never decrease, and certainly will not decrease to zero. Hence, in every Lorentz frame where a light beam appears, there is a unique direction $\hat{\mathbf{e}}_{\mathbf{k}} = \hat{\mathbf{e}}_{\parallel}$ along which it propagates. Each of the electromagnetic fields $\mathbf{E}_{\perp}, \mathbf{B}_{\perp}$ responsible for its propagation is always made of two independent components in the perpendicular directions $\mathbf{e}_1$ and $\mathbf{e}_2$ in that Lorentz frame. As we say in optics, light has only two directions of polarization. We shall explain below how these two polarization states materialize in quantum mechanics as the two transverse states of the intrinsic spin of the photon, the quantum of electromagnetic waves.

When relativity is applied to the kinematics of a point particle of mass m, velocity $\mathbf{v}$, momentum $\mathbf{p}$ and energy E in a Lorentz frame, Einstein obtained the famous energy–momentum relation

$$E^2 - \mathbf{p}^2 = m^2, \tag{67}$$

for the relativistic momentum

$$\mathbf{p} = \gamma m \mathbf{v}, \qquad \text{where} \quad \gamma = \frac{1}{\sqrt{1 - v^2/c^2}}. \tag{68}$$

This relativistic kinematics restricts the velocity v of all massive particles of finite momentum or energy to finite values of γ, and to velocities less than c:

$$v = c\sqrt{1-\gamma^{-2}} < c. \tag{69}$$

If light is considered a particle, its invariant velocity c will require that $\gamma = \infty$. When Compton scattering was discovered [15], light was found to have indeed the attributes of a particle of finite energy E and momentum $\mathbf{p}$. These attributes are consistent with (68) only if the mass of a photon is exactly zero.

A photon in a beam of light described by the Fourier basis function $\psi_{\mathbf{k},\omega}(\mathbf{r}, t)$ of (7) has a wave vector $\mathbf{k}$ and frequency ω. The quantum wave–particle duality of the photon is then expressed by the quantum relations

$$\mathbf{p} = \hbar\mathbf{k}, \quad \text{and} \quad E = \hbar\omega, \tag{70}$$

where $\hbar$ is the reduced Plank constant. The photon, like any other particle, also has an orbital angular momentum vector $\boldsymbol{\ell} = \mathbf{r} \times \mathbf{p}$. This vector must therefore lie on the plane perpendicular to $\mathbf{e}_{\mathbf{k}} = \mathbf{e}_3$, with only two independent components ℓ_1 and ℓ_2, just like $\mathbf{A}_\perp$, in any Lorentz frame. We thus see that the absence of a physically meaningful and therefore gauge-independent $A_\parallel$ component has the same physical origin as the absence for the massless photon of the component $\ell_3 = \ell_\parallel$ of its orbital angular momentum vector, namely the fact that the massless photon always moves with light speed c in vacuum in any Lorentz frame.

The inaccessibility of the longitudinal components of these vectors $\mathbf{A}$ and $\boldsymbol{\ell}$ may appear a little unsatisfactory, but modern physics, i.e. special relativity and quantum mechanics, now come to the rescue. Consider first a massive particle that can only move with speed $v < c$. It is then possible to go to a Lorentz frame where the particle is at rest. In this rest frame, the particle's orbital angular momentum vanishes because its momentum $\mathbf{p}$ vanishes. Now it has been found in quantum mechanics that all massive particles at rest have an intrinsic spin or intrinsic angular momentum vector $\mathbf{s}$ whose length parameter s can only take one of the permissible values of $s = 0, 1/2, 1, 3/2, \ldots$ (in units of $\hbar$). In addition, its projection s_z along any coordinate axis $\mathbf{e}_z$ in space can only have values that differ from one another by an integer (in units of $\hbar$). Since s_z is a projection of $\mathbf{s}$, it also satisfies the constraint $|s_z| \leqslant s$. Both constraints are satisfied by the $2s + 1$ possible values of s_z in the range $-s \leqslant s_z \leqslant s$. For example, a particle with $s = 1$ can only have the three z-projections of $s_z = 1, 0$ or -1. Furthermore, in the rest frame of a massive particle, there is no preferred direction in space. All three components of its intrinsic spin vector $\mathbf{s}$ are dynamically equivalent.

The spin vector $\mathbf{s}$ is also called an angular momentum vector, because it is found to have all the properties of the particle's orbital angular momentum vector $\boldsymbol{\ell}$ when the particle is moving. Only then can the two angular momenta of a moving particle be added together to form a total angular momentum $\mathbf{j} = \boldsymbol{\ell} + \mathbf{s}$, as Pauli [16], and Uhlenbeck and Goudsmit [17] found in their explanation of the anomalous Zeeman effect. (The anomalous Zeeman effect has its origin in the unexpected doubling of the number of atomic energy levels in certain atoms in a magnetic field. It is caused by the intrinsic spin of the electron.) So for massive particles in motion, all three components of their orbital and total angular momenta are dynamically equivalent as well.

Massive particles of intrinsic spin $s = 1$ are called vector bosons, vector because its $\mathbf{s}$ vector has three quantized s_z values, matching the number of components of a vector in space. Any particle whose intrinsic spin s is an integer is called a boson, named after Bose, the famous Indian contemporary of Einstein. The photon too has been found to be a vector boson. If the photon were massive, it too would have a rest frame where there is no preferred direction in space. Its three intrinsic spin projections s_z would be just the three polarization states. Then all three components of its vector potential **A** would be accessible and dynamically equivalent, and the gauge redundancy would disappear. Quantum mechanics would thus work its magic in the massive photon's rest frame, turning the ugly toad of gauge redundancy into a handsome prince of dynamical wholesomeness.

So, as a massive photon becomes massless, it acquires the universal light speed c in all Lorentz frames. It loses its rest frame. Condemned to perpetual motion, both its orbital and intrinsic angular momentum vectors become forever confined to the transverse plane perpendicular to its direction of motion. Its $\tilde{A}_\parallel$ too decouples from the two dynamical transverse components. It is made accessible to dynamics only by sharing the function with the scalar potential. It becomes our redundant gauge degree of freedom.

Appendix. Helmholtz theorem

The usual and customary integral form of the Helmholtz decomposition of a vector field in space into a gradient part and a curl part can be obtained by evaluating the inverse Fourier transform of (20). The ω variable is actually not involved, and will be suppressed here. Hence, we consider the inverse Fourier transform $\mathcal{F}^{-1}$, as defined by (5) and reproduced here without its ω or t dependence as

$$\mathbf{E}(\mathbf{r}) = \mathcal{F}^{-1}\{\tilde{\mathbf{E}}(\mathbf{k})\} = \int \frac{\mathrm{d}^3k}{(2\pi)^3}\, \mathrm{e}^{\mathrm{i}\mathbf{k}\cdot\mathbf{r}}\tilde{\mathbf{E}}(\mathbf{k}). \tag{A.1}$$

Such inverse Fourier transforms can handle the unit vector $\mathbf{e}_\mathbf{k}$ readily only in the original vector/scalar form $\mathbf{k}/k$. So we begin by writing the vector field $\tilde{\mathbf{E}}(\mathbf{k})$ in (20) in the alternative dyadic form

$$\tilde{\mathbf{E}}(\mathbf{k}) = \tilde{\mathbf{E}}_\parallel + \tilde{\mathbf{E}}_\perp = \mathbf{k}\frac{1}{k^2}(\mathbf{k}\cdot\tilde{\mathbf{E}}) - \mathbf{k}\times\frac{1}{k^2}\;(\mathbf{k}\times\tilde{\mathbf{E}}). \tag{A.2}$$

The inverse Fourier transform $\mathbf{E}(\mathbf{r})$ of (A.2) can be found readily by using inverse transforms such as

$$\mathcal{F}^{-1}\{\mathbf{k}\} = -\mathrm{i}\nabla, \tag{A.3}$$

$$\mathcal{F}^{-1}\{\mathrm{i}\mathbf{k}\cdot\tilde{\mathbf{E}}\} = \nabla\cdot\mathbf{E}(\mathbf{r}), \tag{A.4}$$

$$\mathcal{F}^{-1}\left\{\frac{1}{k^2}\right\} = \frac{1}{4\pi|\mathbf{r}|}. \tag{A.5}$$

Each term on the right-hand side in (A.2) contains three $\mathbf{k}$ factors. The middle $1/k^2$ factor can be treated in two equivalent ways giving the same nonlocal kernel. The nonlocal kernel comes either from the convolution or folding theorem (56) applied to a product of functions of k in the Fourier space $\mathbf{k}$, or from the nonlocal operator $1/k^2$ in space that requires an integral over $\mathbf{r}'$, as done in quantum mechanics. In both cases, one finds the integral Helmholtz theorem in space, with

$$\begin{aligned}\mathbf{E}_\parallel(\mathbf{r}) &= -\nabla\int \mathrm{d}^3r'\frac{1}{4\pi|\mathbf{r}-\mathbf{r}'|}\nabla'\cdot\mathbf{E}_\parallel(\mathbf{r}'),\\ \mathbf{E}_\perp(\mathbf{r}) &= \nabla\times\int \mathrm{d}^3r'\frac{1}{4\pi|\mathbf{r}-\mathbf{r}'|}\nabla'\times\mathbf{E}_\perp(\mathbf{r}').\end{aligned} \tag{A.6}$$

References

[1] Jackson J D 1999 *Classical Electrodynamics* 3rd edn (New York: Wiley)

[2] Griffiths D J 1999 *Introduction to Electrodynamics* 3rd edn (Upper Saddle River, NJ: Prentice-Hall)

[3] Jackson J D and Okun L B 2001 Historical roots of gauge invariance *Rev. Mod. Phys.* **73** 663–80

[4] Brill O L and Goodman B 1967 Causality in the Coulomb gauge *Am. J. Phys.* **35** 832–7

[5] Yang K H 2005 The physics of gauge transformations *Am. J. Phys.* **73** 742–51

[6] Schwinger J, DeRaad Jr L I, Kimball A M and Tsai W 1998 *Classical Electrodynamics* (Reading, MA: Perseus Books) p 29

[7] Ichimaru S 1973 *Basic Principles of Plasma Physics* (Reading, MA: Benjamin) pp 37–9

[8] Jackson J D 2002 From Lorenz to Coulomb and other explicit gauge transformations *Am. J. Phys.* **70** 917–28 (arXiv: physics/0204034)

[9] Dirac P A M 1978 *Directions in Physics* (New York: Wiley) pp 31–3

[10] Sakurai J J 1967 *Advanced Quantum Mechanics* (Reading, MA: Addison-Wesley) pp 250–6

[11] Zee A 2003 *Quantum Field Theory in a Nutshell* (Princeton: Princeton University Press)

[12] Wilczek F 1997 The future of particle physics as a natural science *Critical Problems in Physics* ed V L Fitch, A R Marlow and M A E Dementi (Princeton: Princeton University Press) pp 281–306

[13] Wong C W 1991 *Introduction to Mathematical Physics* (New York: Oxford University Press) p 159

[14] Einstein A 1905 Zur Elektrodynamik bewegter Körper *Ann. Phys.* **17** 891–921
Lorentz H A, Einstein A, Minkowski H and Weyl H 1905 *The Principle of Relativity* (New York: Dover) pp 37–65 (Engl. transl.)

[15] Compton A H 1923 A quantum theory of the scattering of X-rays by light elements *Phys. Rev.* **21** 483–502

[16] Pauli W 1925 Über den Zusammenhang des Abschlusses der Elecktronengruppen im Atom mit der Komplexstrucktur der Spektren *Zeitschr. Phys.* **31** 765–83 (Pauli's study of the anomalous Zeeman effect culminated in the formulation of the Pauli exclusion principle in this paper.)

[17] Uhlenbeck G E and Goudsmit S 1925 Ersetzung der Hypothese vom unmechanischen Zwang durch eine Forderung bezüglich des inneren Verhaltens jedes einzelnen Elecktrons *Naturwissenschaften* **13** 953–4