Physics Notes

Note 10

September 1998

The physical origin of gauge invariance in electrodynamics and some of its consequences

Frank Gronwald and Jürgen Nitsch

Otto-von-Guericke-Universität Magdeburg
Institut für Allgemeine Elektrotechnik und Leistungselektronik
Postfach 4120
39016 Magdeburg
Germany

Abstract

The physical origin of the gauge invariance in Maxwell's theory is rooted in quantum physics and thus not accessible to classical electrodynamics. We provide an original and fundamental description of gauge invariance in electrodynamics to bridge this gap. Then we identify the dynamical components of the electromagnetic field, give an account of the concept of gauge fixing, and discuss solutions of Maxwell's equations in terms of gauge potentials.

1 Introduction

The backbone of classical electrodynamics are Maxwell's equations (see [1, 2], e.g.)

$$\nabla \cdot \boldsymbol{B} = 0\,, \qquad \nabla \times \boldsymbol{E} + \frac{\partial \boldsymbol{B}}{\partial t} = 0\,, \tag{1}$$

$$\nabla \cdot \boldsymbol{D} = \rho\,, \qquad \nabla \times \boldsymbol{H} - \frac{\partial \boldsymbol{D}}{\partial t} = \boldsymbol{J}\,. \tag{2}$$

They contain as sources the charge density ρ and current density $\boldsymbol{J}$ while the electromagnetic field itself is described by the field strengths $\boldsymbol{E}$ and $\boldsymbol{B}$. The electric excitation ("dielectric displacement") $\boldsymbol{D}$ and magnetic excitation ("magnetic field") $\boldsymbol{H}$ must be linked to the electromagnetic field strengths in order to make Maxwell's equations a determined set of partial differential equations. This is accomplished by the introduction of constitutive relations which are often assumed to be of the local and linear form

$$\boldsymbol{D} = \varepsilon\varepsilon_0 \boldsymbol{E} \qquad \text{and} \qquad \boldsymbol{B} = \mu\mu_0 \boldsymbol{H}\,. \tag{3}$$

With the constitutive relations at hand it is straightforward to decouple the Maxwell equations and to obtain wave equations for the field strengths. The result is ($c^2 := 1/\epsilon\epsilon_0\mu\mu_0$)

$$\begin{aligned} \triangle \boldsymbol{E} - \frac{1}{c^2}\frac{\partial^2 \boldsymbol{E}}{\partial t^2} &= \frac{1}{\varepsilon\varepsilon_0}\nabla \rho + \mu\mu_0 \frac{\partial \boldsymbol{J}}{\partial t}\,, && (4) \\ \triangle \boldsymbol{B} - \frac{1}{c^2}\frac{\partial^2 \boldsymbol{B}}{\partial t^2} &= -\mu\mu_0 \nabla \times \boldsymbol{J}\,. && (5) \end{aligned}$$

For the excitations analogous equations of motion are obtained with the aid of (3).

If we take the paragraph above for granted it seems as a mere mathematical convenience to introduce electromagnetic potentials φ and $\boldsymbol{A}$ in order to express the field strengths as

$$\boldsymbol{E} = -\nabla\varphi - \frac{\partial \boldsymbol{A}}{\partial t}\,, \qquad \boldsymbol{B} = \nabla \times \boldsymbol{A}\,. \tag{6}$$

It follows from this substitution that the homogeneous Maxwell equations (1) are automatically fulfilled. Moreover, the inhomogeneous Maxwell equations (2), together with (6), can be used to derive equations of motion for the electromagnetic potentials φ, $\boldsymbol{A}$. This yields

$$\begin{aligned} \triangle\varphi + \frac{\partial(\nabla \cdot \boldsymbol{A})}{\partial t} &= -\frac{\rho}{\varepsilon\varepsilon_0}\,, && (7) \\ \triangle \boldsymbol{A} - \frac{1}{c^2}\frac{\partial^2 \boldsymbol{A}}{\partial t^2} - \nabla\Big(\nabla \cdot \boldsymbol{A} + \frac{1}{c^2}\frac{\partial \varphi}{\partial t}\Big) &= -\mu\mu_0 \boldsymbol{J}\,. && (8) \end{aligned}$$

While the field strengths $\boldsymbol{E}$ and $\boldsymbol{B}$ are uniquely determined by a given set of potentials the converse is not true: Transforming the potentials according to

$$\delta_\epsilon \varphi = \varphi' - \varphi = -\frac{\partial \epsilon}{\partial t}\,, \qquad \delta_\epsilon \boldsymbol{A} = \boldsymbol{A}' - \boldsymbol{A} = \nabla \epsilon\,, \tag{9}$$

with an arbitrary function $\epsilon = \epsilon(\boldsymbol{r}, t)$, leaves the field strengths and the Maxwell equations (1), (2) invariant, as is easily verified. This invariance is what is usually called the gauge invariance of electrodynamics. The transformations (9) are the corresponding gauge transformations.

The following questions arise naturally:

- What is the physical origin of the electromagnetic gauge invariance with respect to the gauge transformations (9)?
- What field components of the set $\{\varphi, \boldsymbol{A}\}$ are truly physical, i.e., dynamically independent of the freedom to perform gauge transformations?
- What restrictions (= gauge fixing conditions) can be posed on the potentials in order to eliminate the gauge freedom (and to possibly make practical calculations easier)?

In most textbooks on electrodynamics these points are mentioned but not always clearly spelled out and explained. This leaves a certain amount of insecurity about how to use and deal with gauge invariance in formal and practical calculations. Therefore we will answer the questions above in the subsequent sections.

Recently, a stimulating article appeared which provided helpful background material in this regard [3, 4]. The present paper continues this direction and works out the physical and mathematical properties of the electromagnetic potentials from the very basic. In particular we will stress the non-dynamical character of both the scalar potential φ and the irrotational ("longitudinal") component of the vector potential $\boldsymbol{A}$. As a by-product we will also recognize that there is no problem with causality if the Coulomb gauge is applied to solve Maxwell's equations.

2 Origin of gauge invariance: The electromagnetic field as a gauge field

In the following we have to borrow some knowledge which belongs to the realm of quantum physics. Let us assume for the moment we already know how to formulate a theory of free electrons and positrons [5]: In quantum mechanics these particles are described by a so-called spinor Ψ, a mathematical quantity which consists of four complex components, $\Psi = (\Psi_1, \Psi_2, \Psi_3, \Psi_4)$. Any component Ψ_i is a field $\Psi_i(\boldsymbol{r}, t)$ which describes a distinct particle. More precisely, the four components comprise the description of an electron with spin up, an electron with spin down, a positron with spin up, and a positron with spin down. (Spin can be viewed as internal angular momentum of a particle. Given an arbitrary axis one can only measure two different values of electron or positron spin with respect to this axis. These values are called "spin up" and "spin down".)

It is possible to separate the complex Ψ into a real part $|\Psi|$ (which still consists of four components) and a phase part $\exp(j\theta)$ according to

$$\Psi = |\Psi| \exp(j\theta) = (|\Psi_1|, |\Psi_2|, |\Psi_3|, |\Psi_4|) \exp(j\theta) . \tag{10}$$

That we can actually separate the *same* phase from all four components is justified by the physical fact that the phase is not directly observable. Only phase *differences* are physically observable, for example in an interference experiment with electrons or positrons. This is a nontrivial result from quantum mechanics. Now the point is that *the unobservability of the phase θ is responsible for the gauge invariance of electrodynamics.*

We have to explain this, of course: Suppose we would like to assign at a certain position $\boldsymbol{r}$ to a certain time t a specific value to the phase θ, that is, some number of the interval $[0, 2\pi[$. Even if we cannot observe the complete spinor Ψ directly it does exist, at least in some mathematical space. This is sketched in Fig. 1, where one (arbitrary) component Ψ_i of the spinor Ψ is represented by a wiggly line. In order to give θ an absolute meaning we need some reference frame which tells us where $\theta = 0$, for example, is located. The choice of such a reference frame is not unique, as is shown in Fig. 1.

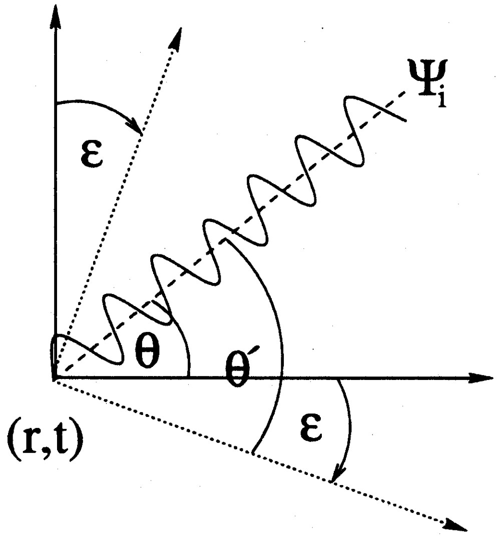

Figure 1: The value of the phase θ of Ψ_i does depend on the choice of a reference frame. A gauge transformation corresponds to the transformation of one reference frame to an equivalent one. It is mathematically expressed as a rotation in the plane about an angle ϵ.

There is a *gauge freedom* in choosing a reference frame. In fact, any two admissible reference frames are related by a rotation in the plane. The choice of one of the possible reference frames is said to be the choice of a specific *gauge*. In this language, the value of θ is *gauge dependent*. A gauge transformation of θ is mathematically expressed as

$$\delta_\epsilon \theta = \theta' - \theta = \epsilon , \tag{11}$$

since the change of θ, when shifting between two reference frames that differ by an angle ϵ, is just given by ϵ. We note that the difference of two phases θ_1, θ_2 (at *one* point ($\boldsymbol{r}$, t)) is gauge independent (=*gauge invariant*), i.e., independent of the reference frame chosen:

$$\delta_\epsilon(\theta_1 - \theta_2) = \delta_\epsilon\theta_1 - \delta_\epsilon\theta_2 = \epsilon - \epsilon = 0\,. \tag{12}$$

In the previous paragraph we concentrated on the phase at one point in spacetime. Now suppose we wish to describe how the phase θ changes between two different points of spacetime. To be specific we focus on two points $(\boldsymbol{r}, t)$ and $(\boldsymbol{r} + \boldsymbol{\Delta r}, t)$, with $\boldsymbol{\Delta r}$ infinitesimally small. In order to describe the change of θ we have to distinguish between the true physical change $D\theta/D\boldsymbol{r}$, which is observable by a suitable interference experiment, and the gradient $\boldsymbol{\nabla}\theta$. Remember that an observable change should be gauge independent. However, the definition of $\boldsymbol{\nabla}\theta$, which is given by the limiting process $(\theta(\boldsymbol{r} + \boldsymbol{\Delta r}) - \theta(\boldsymbol{r}))/\boldsymbol{\Delta r}$ for $\boldsymbol{\Delta r} \to 0$, involves two different reference frames (the one at $\boldsymbol{r} + \boldsymbol{\Delta r}$ and the one at $\boldsymbol{r}$), both of which can be arbitrarily chosen. The gradient $\boldsymbol{\nabla}\theta$ is not gauge invariant since a gauge transformation δ_ϵ acts according to

$$\delta_\epsilon(\boldsymbol{\nabla}\theta) = \boldsymbol{\nabla}(\delta_\epsilon\theta) = \boldsymbol{\nabla}\epsilon\,, \tag{13}$$

and this expression does not vanish in general. The reason for this complication is the fact that $\boldsymbol{\nabla}\theta$ includes a possible unphysical change of θ that is due to the arbitrary orientation of the reference frames at $(\boldsymbol{r}, t)$ and $(\boldsymbol{r} + \boldsymbol{\Delta r}, t)$ to each other. What we need to know, however, is the true physical change $D\theta/D\boldsymbol{r}$ which is obtained if *parallel* (="unchanged") reference frames are used.

A priori, the notion of parallel reference frames is *not* defined in the space of the phase θ. Therefore we write as an ansatz

$$\frac{D\theta}{D\boldsymbol{r}} = \boldsymbol{\nabla}\theta - \boldsymbol{A} \tag{14}$$

with a so far unspecified vector function $\boldsymbol{A}$. The role of $\boldsymbol{A}$ is to define parallel frames at two points $(\boldsymbol{r}, t)$ and $(\boldsymbol{r} + \Delta\boldsymbol{r}, t)$. This is explained in Fig. 2.

We can derive the behavior of $\boldsymbol{A}$ under gauge transformations from the gauge independence of the observable $D\theta/D\boldsymbol{r}$. We use (13) and the definition (14) to obtain the conclusion

$$\delta_\epsilon\Big(\frac{D\theta}{D\boldsymbol{r}}\Big) = 0 \qquad \Longrightarrow \qquad \delta_\epsilon\boldsymbol{A} = \boldsymbol{\nabla}\epsilon \tag{15}$$

In an analogous way we can consider the physical change $D\theta/Dt$ of θ between two points $(\boldsymbol{r}, t)$ and $(\boldsymbol{r}, t + \Delta t)$. This leads to the introduction of a scalar function which we call φ. The field φ compensates a possible unphysical change of $\partial\theta/\partial t$,

$$\frac{D\theta}{Dt} = \frac{\partial\theta}{\partial t} + \varphi\,. \tag{16}$$

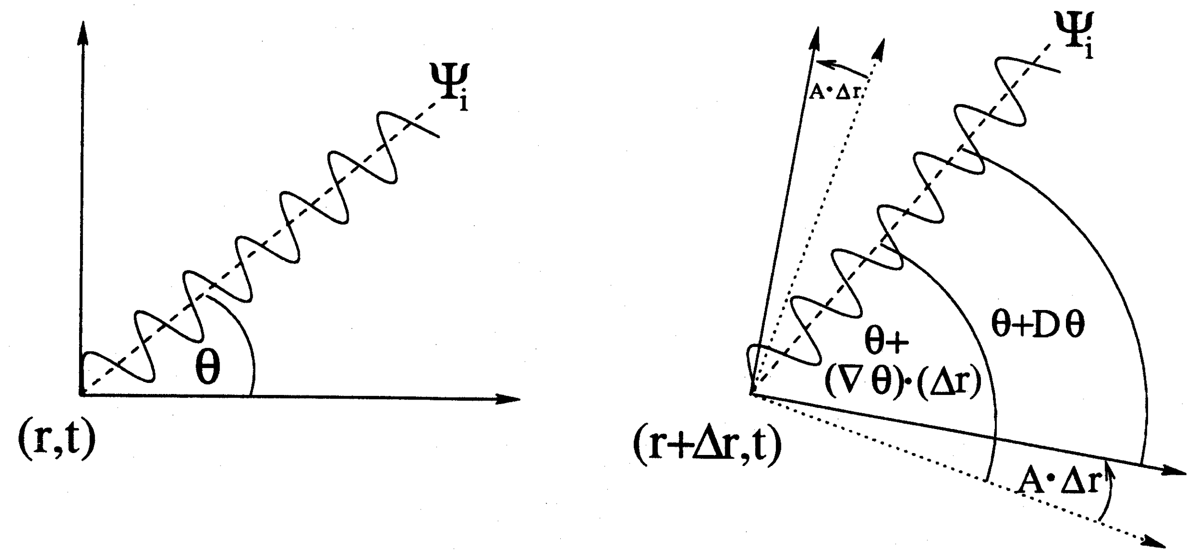

Figure 2: Determination of parallel reference frames: Suppose we start from some component Ψ_i at $(\boldsymbol{r}, t)$ and want to know how its phase changed while passing to $(\boldsymbol{r}+\Delta\boldsymbol{r}, t)$. We choose some reference frame at $(\boldsymbol{r}+\Delta\boldsymbol{r}, t)$ (here displayed as dotted frame). This gauge can be used to calculate the gradient $\boldsymbol{\nabla}\theta$. Now we obtain the reference frame at $(\boldsymbol{r}+\Delta\boldsymbol{r}, t)$ which is *defined* to be parallel to the one at $(\boldsymbol{r}, t)$ by a rotation of the dotted frame by an amount of $\boldsymbol{A}\cdot\Delta\boldsymbol{r}$. This is how $\boldsymbol{A}$ determines parallel reference frames and thus the true physical change $D\theta$. Note that the actual value of $\boldsymbol{A}$ depends on the choice of the dotted frame, i.e., $\boldsymbol{A}$ is gauge dependent.

For "experts" we remark that in (14) and (16) the change of the sign in front of the potentials is rooted in the Minkowskian structure of spacetime. It follows

$$\delta_\epsilon\left(\frac{D\theta}{Dt}\right) = 0 \qquad \Longrightarrow \qquad \delta_\epsilon\varphi = -\frac{\partial\epsilon}{\partial t}. \tag{17}$$

We recognize from the right hand side of (15) and (17) that the fields $\boldsymbol{A}$ and φ transform under gauge transformations exactly in the same way as the electromagnetic potentials do, compare equation (9).

So far, φ and $\boldsymbol{A}$ are introduced as a *necessary* means to describe the change of the phase of an electron or positron. Their actual values are not imposed by us but by nature. There are well-established physical principles that tell us how to make φ and $\boldsymbol{A}$ to true physical fields that are determined from physics. One (standard) way (see [5] or [1] Chap. 12, e.g.) is to use φ and $\boldsymbol{A}$ to construct kinetic energy terms together with source terms. During this construction it is decisive to know the transformation behavior (15) and (17) in order to derive gauge invariant expressions. Then an action principle is applied to obtain equations of motion. The result are the Maxwell equations (1), (2) and it is in this way recognized that φ and $\boldsymbol{A}$ *are* the electromagnetic potentials, introduced in a deductive and modern way. Therefore the (gauge) structure of electrodynamics is, in the end, a consequence of the freedom to choose certain arbitrary reference frames in quantum physics.

3 Dynamical and non-dynamical components of the electromagnetic field

The gauge invariance of electrodynamics, reflected by (15) and (17), leads to the conclusion that not all components of the set $\{\varphi, \boldsymbol{A}\}$ are dynamical. Arbitrary gauge transformations can be used to change the value of some field components. Let us identify these non-dynamical degrees of freedom.

To begin with we consider the scalar potential φ. Under a gauge transformation it transforms according to $\delta_\epsilon \varphi = -\partial\epsilon/\partial t$. The gauge parameter $\epsilon = \epsilon(\boldsymbol{r}, t)$ is a priori arbitrary and so is its time dependence $\partial\epsilon/\partial t$. Thus we conclude for the moment that φ is not dynamical since its value can pointwise be arbitrarily changed.[1]

Next we focus on the vector potential $\boldsymbol{A}$. It changes under a gauge transformation according to $\delta_\epsilon \boldsymbol{A} = \boldsymbol{\nabla}\epsilon$, i.e., it changes by a pure gradient, the rotation of which vanishes, $\boldsymbol{\nabla} \times \boldsymbol{\nabla}\epsilon = \boldsymbol{0}$. This motivates to use Helmholtz's decomposition theorem to decompose the vector potential $\boldsymbol{A}$ into a rotational part $\boldsymbol{A}^{\text{rot}}$ and an irrotational part $\boldsymbol{A}^{\text{irr}}$, [2]

$$\boldsymbol{A} = \boldsymbol{A}^{\text{rot}} + \boldsymbol{A}^{\text{irr}}, \quad \text{with} \qquad \boldsymbol{\nabla} \cdot \boldsymbol{A}^{\text{rot}} = 0, \quad \boldsymbol{\nabla} \times \boldsymbol{A}^{\text{irr}} = \boldsymbol{0}. \tag{18}$$

Such a decomposition is always possible if the vector field $\boldsymbol{A}$ is defined over the whole space and if $\boldsymbol{A}$ and its derivatives fall off sufficiently fast at infinity. Moreover, the decomposition is unique up to a constant vector.[3] It follows

$$\delta_\epsilon \boldsymbol{A} = \delta_\epsilon(\boldsymbol{A}^{\text{rot}} + \boldsymbol{A}^{\text{irr}}) = \boldsymbol{\nabla}\epsilon \simeq (\boldsymbol{\nabla}\epsilon)^{\text{irr}}, \tag{19}$$

where the wavy equal sign "$\simeq$" indicates "equal up to a vector which is constant in space". From the last equation we conclude

$$\boxed{\delta_\epsilon \boldsymbol{A}^{\text{rot}} \simeq \boldsymbol{0}, \qquad \delta_\epsilon \boldsymbol{A}^{\text{irr}} \simeq \boldsymbol{\nabla}\epsilon.} \tag{20}$$

We recognize that, up to a constant and in particular non-dynamical vector, the rotational component of the vector potential is gauge invariant while the irrotational component is gauge dependent.

[1] The attribute "arbitrary" is not quite correct since, so far, we have neglected possible boundary conditions, in particular the boundary condition (7). We will comment on this below and remark for the moment that boundary conditions certainly pose restrictions on a system but do not change the (non-)dynamical character of a field variable.

[2] In the physics literature it is common to denote the rotational part as *transverse* and the irrotational part as *longitudinal*.

[3] For a proof and a discussion of the Helmholtz theorem see §20 of [8]. The theorem is, in general, not valid if the domain of the vector field contains boundaries. In this case only specific boundary conditions can give meaning to the decomposition into rotational and irrotational parts. We would like to mention in this connection the construction of a specific antenna which, in a certain gauge, is characterized by a vector potential with both vanishing curl and divergence [9]. However, the vector potential is not trivial but, basically, of the form of an electric dipole field. This is possible due to the nontrivial topology of the domain of the vector potential in this specific example.

Therefore we can summarize by saying that *the dynamical electromagnetic degrees of freedom are given by the rotational part of the vector potential while the irrotational part together with the scalar potential are non-dynamical.* This means in particular that *the time derivatives of $\boldsymbol{A}^{\mathrm{irr}}$ and φ are of no physical relevance.*

We must stress that the arguments above are more of a plausible than rigorous nature. A strict derivation that leads to the results presented requires quite an amount of theoretical formalism. In this connection we mention the "Hamiltonian formalism" which is very well suited to investigate gauge systems like the electromagnetic field. For details the interested reader is referred to the excellent references [6, 7].

4 Gauge fixing in electrodynamics

The gauge freedom is often used to facilitate the calculation of an electromagnetic problem. Cleverly chosen gauge conditions can simplify equations which are otherwise difficult to solve.

Gauge conditions must satisfy the following point:

- They must be *admissible.* This means that any gauge potentials φ, $\boldsymbol{A}$ that satisfy the equations of motion but *not* necessarily satisfy the gauge fixing conditions can be transformed by a gauge transformation to gauge potentials φ', $\boldsymbol{A}'$ which *do* satisfy the gauge conditions.

Moreover, "good" gauge conditions assure that

- the gauge is fixed completely, that is, no nontrivial gauge transformations can be found that leave the gauge conditions invariant.

We illustrate these points by means of the *Coulomb gauge* $\boldsymbol{\nabla} \cdot \boldsymbol{A}' = 0$. This gauge immediately implies $\boldsymbol{\nabla} \cdot \boldsymbol{A}'^{\mathrm{irr}} = 0$ since the divergence $\boldsymbol{\nabla} \cdot \boldsymbol{A}'^{\mathrm{rot}}$ identically vanishes by definition. Together with the defining condition $\boldsymbol{\nabla} \times \boldsymbol{A}'^{\mathrm{irr}} = \boldsymbol{0}$ we obtain the result

$$\boldsymbol{A}'^{\mathrm{irr}} \simeq \boldsymbol{0} \qquad \text{(in Coulomb gauge)} \tag{21}$$

From the previous section we know that this only fixes one out of two non-dynamical degrees of freedom because also the scalar potential φ is a non-dynamical variable. Hence it seems plausible to take as a further restriction the *temporal gauge* $\varphi' = 0$.

Let us check if both the temporal gauge and Coulomb gauge are generally admissible. We consider two arbitrary gauge potentials φ and $\boldsymbol{A}$. First we try to find a gauge transformation which transforms φ into $\varphi' = 0$. We have

$$\delta_\epsilon \varphi = \varphi' - \varphi = -\frac{\partial \epsilon}{\partial t} \quad \Longrightarrow \quad \varphi = \frac{\partial \epsilon}{\partial t}\,. \tag{22}$$

Integration yields

$$\epsilon(\boldsymbol{r}, t) = \int^t \varphi(\boldsymbol{r}, t')\, dt' + \tilde{\epsilon}(\boldsymbol{r})\,, \tag{23}$$

with a new function $\tilde{\epsilon} = \tilde{\epsilon}(\boldsymbol{r})$ which does not depend on time. Thus we find that all gauge parameters $\epsilon(\boldsymbol{r}, t)$ of the form (23) lead to the temporal gauge $\varphi' = 0$.

Now we also want to specify $\tilde{\epsilon}(\boldsymbol{r})$ such that $\boldsymbol{A}$ is transformed into $\boldsymbol{A}'$. We find from $\boldsymbol{\nabla} \cdot \boldsymbol{A}' = 0$ the conclusion

$$\delta_\epsilon \boldsymbol{A} = \boldsymbol{A}' - \boldsymbol{A} = \boldsymbol{\nabla}\epsilon \quad \Longrightarrow \quad \boldsymbol{\nabla} \cdot \boldsymbol{A} = -\boldsymbol{\nabla} \cdot \boldsymbol{\nabla}\epsilon\,. \tag{24}$$

We plug (23) into the last equation and obtain

$$\boldsymbol{\nabla} \cdot \boldsymbol{A}(\boldsymbol{r}, t) + \boldsymbol{\nabla} \cdot \boldsymbol{\nabla} \int^t \varphi(\boldsymbol{r}, t')\, dt' = -\boldsymbol{\nabla} \cdot \boldsymbol{\nabla}\tilde{\epsilon}(\boldsymbol{r})\,. \tag{25}$$

At first this does not look very promising since the right hand side with the freely choosable function $\tilde{\epsilon}(\boldsymbol{r})$ does only depend on $\boldsymbol{r}$ while the left hand side depends on both $\boldsymbol{r}$ and t. For arbitrary t it seems that we cannot find a function $\tilde{\epsilon}(\boldsymbol{r})$ to always satisfy (25). But let us investigate the time dependence of (25) a little closer. Differentiation with respect to time yields

$$\begin{aligned}
\frac{\partial}{\partial t}\Big(\boldsymbol{\nabla} \cdot \boldsymbol{A}(\boldsymbol{r}, t) + \boldsymbol{\nabla} \cdot \boldsymbol{\nabla} \int^t \varphi(\boldsymbol{r}, t')\, dt'\Big) &= \boldsymbol{\nabla} \cdot \Big(\frac{\partial \boldsymbol{A}}{\partial t} + \boldsymbol{\nabla}\varphi\Big) \\
&= -\boldsymbol{\nabla} \cdot \boldsymbol{E} \\
&= -\frac{1}{\varepsilon\varepsilon_0}\boldsymbol{\nabla} \cdot \boldsymbol{D}\,.
\end{aligned} \tag{26}$$

We remind us of the Maxwell equation $\boldsymbol{\nabla} \cdot \boldsymbol{D} = \rho$ and discover that the left hand side of (25) is constant in time if and only if the charge density vanishes, $\rho = 0$. In this case, reasonable boundary conditions presupposed, $\tilde{\epsilon}(\boldsymbol{r})$ is uniquely determined from (25). This, in turn, fixes $\epsilon(\boldsymbol{r}, t)$ via (23). Therefore we have the result that both the Coulomb gauge and temporal gauge are admissible and do fix the gauge completely if and only if $\rho = 0$.

To understand the situation for $\rho \neq 0$ we have to digress for a moment and return to the Maxwell equations (1), (2). One has to note that the equations $\boldsymbol{\nabla} \cdot \boldsymbol{B} = 0$ and $\boldsymbol{\nabla} \cdot \boldsymbol{D} = \rho$ are no dynamical equations but only *boundary conditions*: By means of the remaining Maxwell equations they are fulfilled for all times if they are fulfilled at one time. Therefore it is sufficient to calculate the solutions of $\boldsymbol{\nabla} \cdot \boldsymbol{B} = 0$ and $\boldsymbol{\nabla} \cdot \boldsymbol{D} = \rho$ at an initial time t_0 and then solve with these solutions as boundary conditions the equations $\boldsymbol{\nabla} \times \boldsymbol{E} + \partial \boldsymbol{B}/\partial t = \boldsymbol{0}$ and $\boldsymbol{\nabla} \times \boldsymbol{H} - \partial \boldsymbol{D}/\partial t = \boldsymbol{J}$.

Now, if we work in terms of potentials, the homogeneous Maxwell equations, and in particular the constraint $\boldsymbol{\nabla} \cdot \boldsymbol{B} = 0$, are trivially fulfilled. We are then left with the only constraint

$$\begin{aligned}
\boldsymbol{\nabla} \cdot \boldsymbol{D} &= \epsilon\epsilon_0 \boldsymbol{\nabla} \cdot \boldsymbol{E} \\
&= \epsilon\epsilon_0 \boldsymbol{\nabla} \cdot \Big(-\boldsymbol{\nabla}\varphi - \frac{\partial \boldsymbol{A}}{\partial t}\Big) = \rho\,.
\end{aligned} \tag{27}$$

In the Coulomb gauge $\nabla \cdot \boldsymbol{A} = 0$ the constraint (27) reduces to

$$\nabla \cdot \nabla \varphi = \Delta \varphi = -\frac{\rho}{\epsilon \epsilon_0}, \tag{28}$$

the solution of which is well-known from electrostatics,

$$\varphi(\boldsymbol{r}, t) = \frac{1}{4\pi \epsilon \epsilon_0} \int \frac{\rho(\boldsymbol{r}', t)}{|\boldsymbol{r} - \boldsymbol{r}'|} d^3 r'. \tag{29}$$

It follows that the gauge choice

$$\nabla \cdot \boldsymbol{A} = 0, \qquad \varphi(\boldsymbol{r}, t) = \frac{1}{4\pi \epsilon \epsilon_0} \int \frac{\rho(\boldsymbol{r}', t)}{|\boldsymbol{r} - \boldsymbol{r}'|} d^3 r', \tag{30}$$

is compatible with the boundary condition $\nabla \cdot \boldsymbol{D} = \rho$. It generalizes the choice of Coulomb gauge plus temporal gauge of the vacuum case $\rho = 0$: We can repeat the calculation (22)–(26) with the condition $\varphi' = 0$ replaced by $\varphi' = \frac{1}{4\pi\epsilon\epsilon_0} \int \frac{\rho}{|\boldsymbol{r}-\boldsymbol{r}'|} d^3 r'$. Then we obtain in place of (23), (25), and (26) the following three equations: The equation for the original gauge parameter $\epsilon(\boldsymbol{r}, t)$ is given by

$$\epsilon(\boldsymbol{r}, t) = \int^t \Big(\varphi(\boldsymbol{r}, t') - \varphi'(\boldsymbol{r}, t') \Big) dt' + \tilde{\epsilon}(\boldsymbol{r}), \tag{31}$$

while the determing equation for the secondary parameter $\tilde{\epsilon}(\boldsymbol{r})$ turns out to be

$$\nabla \cdot \boldsymbol{A}(\boldsymbol{r}, t) + \nabla \cdot \nabla \int^t \varphi(\boldsymbol{r}, t')\, dt' = -\nabla \cdot \nabla \tilde{\epsilon}(\boldsymbol{r}) + \frac{1}{\epsilon \epsilon_0} \int^t \rho(\boldsymbol{r}, t')\, dt'. \tag{32}$$

The time dependence of (32) is trivial since

$$\begin{aligned} &\frac{\partial}{\partial t}\Big(\nabla \cdot \boldsymbol{A}(\boldsymbol{r}, t) + \nabla \cdot \nabla \int^t \varphi(\boldsymbol{r}, t')\, dt' - \frac{1}{\epsilon \epsilon_0} \int^t \rho(\boldsymbol{r}, t')\, dt' \Big) \\ &\qquad = -\frac{1}{\epsilon \epsilon_0} \Big(\nabla \cdot \boldsymbol{D} - \rho(\boldsymbol{r}, t) \Big) = 0 \quad . \end{aligned} \tag{33}$$

This shows that the gauge (30) is admissible and does fix the gauge completely.

The Coulomb gauge $\nabla \cdot \boldsymbol{A} = 0$, the temporal gauge $\varphi = 0$, and its generalization (29) are natural and transparent since they directly restrict the undynamical degrees of freedom. Among the unlimited number of other admissible gauges we yet want to mention the popular *Lorentz gauge* which is given by

$$\nabla \cdot \boldsymbol{A} + \frac{1}{c^2} \frac{\partial \varphi}{\partial t} = 0. \tag{34}$$

It is admissible but does not fix the gauge completely since it is still possible to perform gauge transformations as long as the gauge parameter ϵ obeys

$$\Delta \epsilon - \frac{1}{c^2} \frac{\partial^2 \epsilon}{\partial t^2} = 0. \tag{35}$$

From a formal point of view, the Lorentz gauge has the disadvantage of involving the unphysical quantity $\partial\varphi/\partial t$. The reason why the Lorentz gauge is nevertheless widespread is its usefulness in solving the Maxwell equations. This will be considered next.

5 Gauge dependent equations of motion - Coulomb gauge and Lorentz gauge

In order to solve the Maxwell equations one usually does not rely on the gauge invariant wave equations (4) and (5) but rather takes advantage of the gauge dependent equations of motion (7) and (8). We now discuss the solutions of (7) and (8) for both the Coulomb gauge and the Lorentz gauge.

We begin with the *Coulomb gauge* $\boldsymbol{\nabla} \cdot \boldsymbol{A} = 0$. In this case the equations of motion reduce to

$$\Delta\varphi = -\frac{\rho}{\varepsilon\varepsilon_0}, \tag{36}$$

$$\Delta \boldsymbol{A} - \frac{1}{c^2}\frac{\partial^2 \boldsymbol{A}}{\partial t^2} + \boldsymbol{\nabla}\frac{1}{c^2}\frac{\partial\varphi}{\partial t} = -\mu\mu_0 \boldsymbol{J}. \tag{37}$$

We already encountered the solution of (36),

$$\varphi(\boldsymbol{r},t) = \frac{1}{4\pi\epsilon\epsilon_0}\int\frac{\rho(\boldsymbol{r}',t)}{|\boldsymbol{r}-\boldsymbol{r}'|}d^3r', \tag{38}$$

and now understand that this is not the solution of a dynamical equation but the solution of a boundary condition which fixes a gauge. In this connection it is worth noting that in (38) *there is no violation of causality involved.* We recall that the time dependence of φ is of *no* dynamical relevance. The "instantaneous" solution (38) refers to an instantaneous adjustment of the physically irrelevant reference frames discussed in Sec. 2. The reader in doubt might try to derive from (38) an "instantaneous" electric field via $\boldsymbol{E} = -\boldsymbol{\nabla}\varphi - \partial\boldsymbol{A}/\partial t$. That such an undertaking will not work can immediately be seen from (4) where the gauge invariant wave equation for $\boldsymbol{E}$ with phase velocity c is displayed.

Now we use (38) and the continuity equation

$$\boldsymbol{\nabla}\cdot\boldsymbol{J} + \frac{\partial\rho}{\partial t} = 0 \tag{39}$$

to write (37) in the form

$$\begin{aligned}\Delta \boldsymbol{A} - \frac{1}{c^2}\frac{\partial^2 \boldsymbol{A}}{\partial t^2} &= -\mu\mu_0\Big(\boldsymbol{J} + \frac{1}{4\pi}\boldsymbol{\nabla}\int\frac{\boldsymbol{\nabla}'\cdot\boldsymbol{J}(\boldsymbol{r}',t)}{|\boldsymbol{r}-\boldsymbol{r}'|}d^3r'\Big)\\ &=: -\mu\mu_0(\boldsymbol{J}-\boldsymbol{J}^{\mathrm{irr}})\\ &=: -\mu\mu_0\boldsymbol{J}^{\mathrm{rot}}.\end{aligned} \tag{40}$$

This wave equation has the retarded solution

$$\boldsymbol{A}^{\mathrm{rot}}(\boldsymbol{r},t)) \simeq \boldsymbol{A}(\boldsymbol{r},t) = \frac{\mu\mu_0}{4\pi}\int\frac{\boldsymbol{J}^{\mathrm{rot}}(\boldsymbol{r}',t-\frac{|\boldsymbol{r}-\boldsymbol{r}'|}{c})}{|\boldsymbol{r}-\boldsymbol{r}'|}d^3r'. \tag{41}$$

The currents

$$\boldsymbol{J}^{\text{rot}} := \boldsymbol{J} + \frac{1}{4\pi}\boldsymbol{\nabla}\int \frac{\boldsymbol{\nabla}'\cdot\boldsymbol{J}(\boldsymbol{r}',t)}{|\boldsymbol{r}-\boldsymbol{r}'|}d^3r'\,, \tag{42}$$

$$\boldsymbol{J}^{\text{irr}} := -\frac{1}{4\pi}\boldsymbol{\nabla}\int \frac{\boldsymbol{\nabla}'\cdot\boldsymbol{J}(\boldsymbol{r}',t)}{|\boldsymbol{r}-\boldsymbol{r}'|}d^3r' \tag{43}$$

are, respectively, the rotational and irrotational component of the current $\boldsymbol{J}$,

$$\boldsymbol{J} = \boldsymbol{J}^{\text{rot}} + \boldsymbol{J}^{\text{irr}}\,, \qquad \text{with} \quad \boldsymbol{\nabla}\cdot\boldsymbol{J}^{\text{rot}} = 0\,,\ \ \boldsymbol{\nabla}\times\boldsymbol{J}^{\text{irr}} = \boldsymbol{0}\,, \tag{44}$$

as is easily verified. It is worth noting that, as was also pointed out in [3, 4], the components $\boldsymbol{J}^{\text{rot}}$, $\boldsymbol{J}^{\text{irr}}$ are in general extended over the whole space, even if $\boldsymbol{J}$ is a localized source. Such an a priori counterintuitive behavior can be formally explained by the necessary description of a localized source current in terms of distributions or non-differentiable functions: If a localized current $\boldsymbol{J}$ is plugged into (42), (43) the divergence operator $\boldsymbol{\nabla}'\cdot$ produces singular terms which lead to the extension of $\boldsymbol{J}^{\text{rot}}$, $\boldsymbol{J}^{\text{irr}}$ into the whole space.

Equations (38) and (41) establish the general solution of Maxwell's equations in the Coulomb gauge. We note that the wave equation (40) and its solution (41) are purely rotational: Taking the divergence of (40) or (41), respectively, yields identically zero on both sides. That there is no nontrivial wave equation for the irrotational component $\boldsymbol{A}^{\text{irr}}$ is in accordance with the fact that in the Coulomb gauge $\boldsymbol{A}^{\text{irr}}$ is dynamically restricted to zero, $\boldsymbol{A}^{\text{irr}} \simeq \boldsymbol{0}$.

One might ask what happened to the irrotational component of the current $\boldsymbol{J}^{\text{irr}}$ since it does not enter the solution (41). The answer is that $\boldsymbol{J}^{\text{irr}}$ determines the change in time of the boundary condition (36) via the continuity equation (39),

$$\boldsymbol{\nabla}\cdot\boldsymbol{J} = \boldsymbol{\nabla}\cdot(\boldsymbol{J}^{\text{rot}}+\boldsymbol{J}^{\text{irr}}) = \boldsymbol{\nabla}\cdot\boldsymbol{J}^{\text{irr}} = -\frac{\partial\rho}{\partial t}\,. \tag{45}$$

Now we turn to the solution of (7) and (8) if the *Lorentz gauge* is applied. Substitution of $\boldsymbol{\nabla}\cdot\boldsymbol{A} + \partial\varphi/c^2\partial t = 0$ yields

$$\Delta\varphi - \frac{1}{c^2}\frac{\partial^2\varphi}{\partial t^2} = -\frac{\rho}{\varepsilon\varepsilon_0}\,, \tag{46}$$

$$\Delta\boldsymbol{A} - \frac{1}{c^2}\frac{\partial^2\boldsymbol{A}}{\partial t^2} = -\mu\mu_0\boldsymbol{J}\,. \tag{47}$$

This looks appealing since we obtain wave equations for both φ and $\boldsymbol{A}$. However, the symmetric appearance of the fields φ, $\boldsymbol{A}^{\text{irr}}$, and $\boldsymbol{A}^{\text{rot}}$ is a bit misleading since one has to keep in mind that *only* $\boldsymbol{A}^{\text{rot}}$ is a true dynamical variable. Also one should be aware that the gauge is not fixed completely, yet. The advantage of (46) and (47) is that they are straightforwardly solved. One obtains the retarded potentials

$$\varphi(\boldsymbol{r},t) = \frac{1}{4\pi\varepsilon\varepsilon_0}\int \frac{\rho(\boldsymbol{r}',t-\frac{|\boldsymbol{r}-\boldsymbol{r}'|}{c})}{|\boldsymbol{r}-\boldsymbol{r}'|}d^3r'\,, \tag{48}$$

$$A(r,t) = \frac{\mu\mu_0}{4\pi}\int \frac{J(r',t-\frac{|r-r'|}{c})}{|r-r'|}d^3r' , \quad (49)$$

as physically meaningful solutions of the Maxwell equations. The fact that the gauge is not fixed completely is not disturbing since, in classical electrodynamics, one is usually interested in the calculation of observables, i.e., gauge invariant quantities, from (48) and (49). The advantage of (49) is that, in contrast to (41), one can work with the (in general) localized current J in place of the extended component $J^{\rm rot}$, compare also [3, 4] for a discussion of this point.

The solution (49) can be split into a rotational and irrotational part,

$$A^{\rm rot}(r,t) = \frac{\mu\mu_0}{4\pi}\int \frac{J^{\rm rot}(r',t-\frac{|r-r'|}{c})}{|r-r'|}d^3r' , \quad (50)$$

$$A^{\rm irr}(r,t) = \frac{\mu\mu_0}{4\pi}\int \frac{J^{\rm irr}(r',t-\frac{|r-r'|}{c})}{|r-r'|}d^3r' . \quad (51)$$

The rotational part (50) is identical to the solution (41) which was obtained in the Coulomb gauge. This is not really surprising since we recognized in Sec. 3 that $A^{\rm rot}$ is, up to a constant vector, a gauge invariant quantity. In contrast to this, the Lorentz gauge

$$\nabla \cdot A = \nabla \cdot A^{\rm irr} = -\frac{1}{c^2}\frac{\partial\varphi}{\partial t} \quad (52)$$

is reflected in the solutions (48) and (51): If plugged into the Lorentz gauge condition (52) they simply yield the continuity equation.

6 Conclusion

Let us summarize the main results.

- The origin of the gauge invariance in electrodynamics is found in quantum physics. Gauge transformations exhibit the freedom to choose (unobservable) reference frames that are needed for the description of certain particles. The electromagnetic potentials φ and A are necessary means to determine the physical change of these particles.

- Only the rotational part $A^{\rm rot}$ of the vector potential is dynamically independent of the freedom to perform gauge transformations. The time development of the irrotational part $A^{\rm irr}$ and the scalar potential φ is of no physical relevance.

- The Coulomb gauge constraints the irrotational part $A^{\rm irr}$ dynamically to zero, $A^{\rm irr} \simeq 0$. In accordance with the boundary condition $\nabla \cdot D = \rho$ the Coulomb gauge is reasonably completed and fully fixes the gauge if φ is chosen as the electrostatic Coulomb potential, cf. (29). With this choice there is no instantaneous propagation of observable quantities involved. It is straightforward to solve in this gauge Maxwell's equations. The solution explicitly shows that only $\boldsymbol{A}^{\text{rot}}$ enters as a dynamical variable.

- The Lorentz gauge does not fix the gauge completely but leads directly to a solution of the Maxwell equations. In this gauge the non-dynamical character of $\boldsymbol{A}^{\text{irr}}$ and φ is not obvious. This disadvantage is not really important for practical calculations in classical electrodynamics.

As a rule one should keep in mind that there are no "right" or "wrong" admissible gauge choices. *Any* proper gauge will lead to the same values of the gauge invariant quantities $\boldsymbol{E}$, $\boldsymbol{B}$, $\boldsymbol{D}$, $\boldsymbol{H}$, ρ, and $\boldsymbol{J}$. But depending on an actual problem a certain gauge can be more appropriate than others.

Acknowledgement: The authors would like to thank Prof. G. Wollenberg for a very helpful discussion.

References

[1] **Jackson JD:** (1975) Classical Electrodynamics, 2nd ed. John Wiley & Sons, New York

[2] **Lehner G:** (1996) Elektromagnetische Feldtheorie für Ingenieure und Physiker, 3. Aufl. Springer, Berlin

[3] **Schwab AJ; Fuchs C; Kistenmacher P:** (1997) Zur Bedeutung der Quellenfeldkomponente des magnetischen Vektorpotentials A. Electr. Eng. 80: 81–85

[4] **Schwab AJ; Fuchs C; Kistenmacher P:** (1997) Semantics of the Irrotational Component of the Magnetic Vector Potential, A. IEEE Ant. and Prop. Mag. 39, No. 1: 46–51

[5] **Hatfield B:** (1992) Quantum Theory of Point Particles and Strings, 1st ed. Addison Wesley, Redwood City

[6] **Sundermeyer K:** (1982) Constrained Dynamics, Lecture Notes in Physics, Volume 169, Springer, Berlin

[7] **Henneaux M; Teitelboim C:** (1992) Quantization of Gauge Systems, 1st ed. Princeton University Press, Princeton

[8] **Sommerfeld A:** (1978) Mechanik der deformierbaren Medien, reprint of the 6th ed. Harri Deutsch, Thun

[9] **Baum C.E:** (1991) Vector and Scalar Potentials Away from Sources, and Gauge Invariance in Quantum Electrodynamics, Physics Note 3, October 1991