Causality, the Coulomb field, and Newton's law of gravitation

F. Rohrlich[a)]
Department of Physics, Syracuse University, Syracuse, New York 13244-1130

(Received 23 July 2001; accepted 9 November 2001)

Causality in classical field theories must be inserted by hand by choosing the retarded solution. It is shown how apparent contradictions in the Coulomb gauge can be resolved and that a causal Coulomb field exists despite the appearance to the contrary. Similarly, it is shown how Newtonian gravitation leads from action-at-a-distance to a causal field when a first-order correction for space–time curvature is applied.
[DOI: 10.1119/1.1435345]

I. INTRODUCTION

In both electromagnetic theory and nonrelativistic gravitation theory, *action-at-a-distance* forces exist: Coulomb's law and Newton's law of gravitation. The electrodynamic fields have advanced solutions and retarded ones, but only the latter are *causal*. The existence of action-at-a-distance and causal forces implies an apparent self-contradiction in the formalism.

Classical physics does not contain an arrow of time, the only exception being the law of entropy for macroscopic systems. However, I have recently shown that this claim is not correct when the physical principle of causality is taken into account: the equations of motion of extended classical particles then *do* contain an arrow of time.[1]

The principle of causality states that no effect can occur outside the future light cone of its cause. This principle must be introduced as a condition. In electrodynamics, this condition is imposed by choosing the retarded solutions of Maxwell's differential equations. However, there seems to be a problem when the Coulomb gauge is used: it leads to an action-at-a-distance force. A paper by Brill and Goodman[2] discusses this problem, but they show only the consistency between the Coulomb gauge and causality and do not explicate the cause of the apparent contradiction. The text by Jackson[3] deals only with the special case of a flashing dipole.

For gravitation, causality is absent in Newtonian theory in which the force is strictly an action-at-a-distance force. After all, this theory involves a *static* interaction just like the time-independent Coulomb field. However, if one goes beyond Newtonian theory even *only to first order* to take into account space–time curvature—the linear approximation to Einstein gravitation theory—one can obtain a causal solution in the same way as in electrodynamics.

In Sec. II, I shall discuss causality for the electric field as a preliminary to the problem raised by the introduction of the Coulomb gauge. The latter is the subject of Sec. III. The gravitational case is dealt with in Sec. IV.

II. THE ELECTRIC FIELD

For the sake of completeness, I begin with a few well-known equations. It is sufficient to discuss only the vacuum form of Maxwell's equations. To minimize unimportant factors, I shall use Gaussian units with $c=1$. The inhomogeneous equations are

$$\boldsymbol{\nabla}\times\mathbf{B}-\dot{\mathbf{E}}=4\pi\mathbf{j}, \quad \boldsymbol{\nabla}\cdot\mathbf{E}=4\pi\rho. \tag{2.1}$$

They imply the differential equation of charge conservation,

$$\boldsymbol{\nabla}\cdot\mathbf{j}+\dot{\rho}=0. \tag{2.2}$$

The homogeneous Maxwell equations are

$$\boldsymbol{\nabla}\times\mathbf{E}+\dot{\mathbf{B}}=0, \quad \boldsymbol{\nabla}\cdot\mathbf{B}=0. \tag{2.3}$$

If we take the curl of the first part of Eq. (2.3) and substitute the time derivative of Eq. (2.1), we obtain the differential equation for the electric field,

$$\Box\mathbf{E}=4\pi(\partial\mathbf{j}/\partial t+\boldsymbol{\nabla}\rho), \tag{2.4}$$

where $\Box=\boldsymbol{\nabla}^2-\partial^2/\partial t^2$.

The causality condition can now be imposed by choosing the retarded solution of Eq. (2.4). The retarded Green function is given by

$$\Box D_{\dot{R}}(x)=-\delta(x), \quad D_R=\frac{1}{4\pi r}\,\delta(t-r). \tag{2.5}$$

Equation (2.5) implies that for any function $f(t,\mathbf{x})$,

$$4\pi\int D_R(x-x')f(t',\mathbf{x}')d^4x'$$

$$=\int d^3x' f(t-|\mathbf{x}-\mathbf{x}'|,\mathbf{x}')/|\mathbf{x}-\mathbf{x}'|.$$

If we use this Green function, the retarded electric field (the word ''field'' is used here for what is often called ''field strength'') is

$$\mathbf{E}(x)=-4\pi\int D_R(x-x')[\partial\mathbf{j}(x')/\partial t'+\boldsymbol{\nabla}'\rho(x')]d^4x'. \tag{2.6}$$

$\mathbf{E}(x)$ is the field at $x=(t,\mathbf{x})$ that is generated by the current and charge densities at the retarded point $x'=(t',\mathbf{x}')$, where $t'=t-|\mathbf{x}\text{-}\mathbf{x}'|$. For the special case, $\partial\mathbf{j}/\partial t=0$, one obtains, using integration by parts,

$$\mathbf{E}(t,\mathbf{x})=-\int d^3\mathbf{x}'\rho(t-|\mathbf{x}-\mathbf{x}'|,\mathbf{x}')\boldsymbol{\nabla}(1/|\mathbf{x}-\mathbf{x}'|). \tag{2.7}$$

$\mathbf{E}(t,\mathbf{x})$ is the *retarded* Coulomb field; Eq. (2.7) shows the causal nature of $\mathbf{E}$ explicitly.

III. POTENTIALS AND GAUGES

Potentials are introduced to simplify the field equations. In particular, they make the homogeneous Maxwell equations into identities so that only the inhomogeneous equations need to be solved. They have no physical significance in classical physics, and only the fields are observable. Gauges

are introduced to put conditions on the potentials, because otherwise there is too much freedom for their choice. But no matter which gauge is used, the fields are defined by the potentials:

$$\mathbf{E}=-\boldsymbol{\nabla}\varphi-\dot{\mathbf{A}},\quad \mathbf{B}=\boldsymbol{\nabla}\times\mathbf{A}. \tag{3.1}$$

The best-known gauges are the Lorenz[4,5] gauge and the Coulomb gauge. The Lorenz gauge is also called the covariant gauge, because the potentials transform as components of a four-vector, $\mathbf{A}^{\mu}=(\varphi,\mathbf{A})$. This gauge is characterized by

$$\boldsymbol{\nabla}\cdot\mathbf{A}+\dot{\varphi}=0. \tag{3.2}$$

The inhomogeneous Maxwell equations in the Lorenz gauge become

$$\Box\mathbf{A}=-4\pi\mathbf{j},\quad \Box\varphi=-4\pi\rho. \tag{3.3}$$

Together with the gauge condition (3.2), they imply the differential law of charge conservation (2.2). In this gauge, there is no problem ensuring causality by choosing the retarded solutions of Eq. (3.3),

$$\begin{aligned}\mathbf{A}(x)&=4\pi\int D_R(x-x')\mathbf{j}(x')d^4x',\\ \varphi(x)&=4\pi\int D_R(x-x')\rho(x').\end{aligned} \tag{3.4}$$

These solutions for the retarded potentials yield the retarded fields by substitution into Eq. (3.1).

The Coulomb gauge (also called the radiation gauge or the transverse gauge) is not a covariant gauge. The potentials in this gauge do not form a four-vector. This gauge is characterized by

$$\mathbf{A}_{\parallel}=0. \tag{3.5}$$

This condition suggests that we should separate all three-vectors $\mathbf{V}$ into longitudinal (irrotational) and transverse (or solenoidal) components, $\mathbf{V}=\mathbf{V}_{\parallel}+\mathbf{V}_{\perp}$; this separation is unique. These components are defined by

$$\boldsymbol{\nabla}\cdot\mathbf{V}_{\perp}=0,\quad \boldsymbol{\nabla}\times\mathbf{V}_{\parallel}=0. \tag{3.6}$$

The Coulomb gauge condition (3.5) thus implies $\mathbf{A}=\mathbf{A}_{\perp}$. From the Maxwell equations (2.1) and the definition of the fields in terms of the potentials, Eq. (3.1), one deduces not only the differential equations for the potentials,

$$\Box\mathbf{A}_{\perp}=-4\pi\mathbf{j}_{\perp},\quad \boldsymbol{\nabla}^2\varphi=-4\pi\rho, \tag{3.7}$$

but also the relation

$$4\pi\mathbf{j}_{\parallel}=\boldsymbol{\nabla}\dot{\varphi}. \tag{3.8}$$

For later discussion, it is important to realize that $\mathbf{j}\|$ is not a component of $\mathbf{j}$, but a highly nonlocal function of it,

$$\mathbf{j}_{\parallel}(\mathbf{x},\mathbf{t})=-\frac{1}{4\pi}\boldsymbol{\nabla}\int\frac{d^3\mathbf{x}'}{|\mathbf{x}-\mathbf{x}'|}\boldsymbol{\nabla}'\cdot\mathbf{j}(\mathbf{x}',t). \tag{3.9}$$

The differential charge conservation law also breaks up into two parts,

$$\boldsymbol{\nabla}\cdot\mathbf{j}_{\perp}=0,\quad \boldsymbol{\nabla}\cdot\mathbf{j}_{\parallel}+\dot{\rho}=0. \tag{3.10}$$

We are now ready to examine the question of causality in the Coulomb gauge. There is obviously no problem for the vector potential because from Eq. (3.7), we have

$$\mathbf{A}_{\perp}(x)=4\pi\int D_R(x-x')\mathbf{j}_{\perp}(x')d^4x', \tag{3.11}$$

which yields the retarded field $\mathbf{E}_{\perp}(x)=-\partial\mathbf{A}_{\perp}(x)/\partial t$. But for the scalar potential, Eq. (3.7) gives the instantaneous and therefore *not* the retarded potential

$$\varphi(\mathbf{x},t)=\int\frac{d^3\mathbf{x}'}{|\mathbf{x}-\mathbf{x}'|}\rho(\mathbf{x}',t), \tag{3.12}$$

so that

$$\mathbf{E}_{\parallel}(x)=-\boldsymbol{\nabla}\varphi(\mathbf{x},t)=-\boldsymbol{\nabla}\int\frac{d^3x'}{|\mathbf{x}-\mathbf{x}'|}\rho(\mathbf{x}',t) \tag{3.13}$$

is also not retarded; it is an *instantaneous* (action-at-a-distance) field. The surprise here is that a relativistic theory with retarded initial conditions would lead to such a field. Is there a contradiction in the theory so that it is internally inconsistent?

The answer lies in the nature of $\mathbf{E}_{\parallel}$. Just like $\mathbf{j}_{\parallel}$, $\mathbf{E}_{\parallel}$ *is also a highly nonlocal function of* $\mathbf{E}$ with the identical defining relation as $\mathbf{j}_{\parallel}$ to $\mathbf{j}$ in Eq. (3.9):

$$\mathbf{E}_{\parallel}(\mathbf{x},t)=-\frac{1}{4\pi}\boldsymbol{\nabla}\int\frac{d^3\mathbf{x}'}{|\mathbf{x}-\mathbf{x}'|}\boldsymbol{\nabla}'\cdot\mathbf{E}(\mathbf{x}',t). \tag{3.14}$$

Using the second Maxwell equation (2.1), this relation implies Eq. (3.13) trivially. A comparison with Eq. (2.7) now shows that this *nonlocal relation of the longitudinal field to the total field is exactly responsible for turning the retarded field* $\mathbf{E}$ *due to* ρ *in Eq. (2.7) into an instantaneous (action-at-a-distance) field* $\mathbf{E}_{\parallel}$ *due to* ρ *in Eq. (3.13).*

When the retarded field $\mathbf{E}$ in Eq. (2.6) is separated into $\perp$ and $\|$ components, the field $\mathbf{E}_{\perp}$ is a retarded field in agreement with the field derived from Eq. (3.10), but the field $\mathbf{E}_{\parallel}$ is *also retarded*,

$$\mathbf{E}_{\parallel}(x)=-4\pi\int D_R(x-x')[\partial\mathbf{j}_{\parallel}(x')/\partial t'+\boldsymbol{\nabla}'\rho(x')]d^4x', \tag{3.15}$$

in apparent contradiction to Eq. (3.13). But when one eliminates $\mathbf{j}_{\parallel}$ and ρ in favor of the scalar potential φ [using Eqs. (3.8) and (3.7)], one finds

$$4\pi[\partial\mathbf{j}_{\parallel}(x)/\partial t+\boldsymbol{\nabla}\rho(x)]=\boldsymbol{\nabla}\ddot{\varphi}-\boldsymbol{\nabla}\boldsymbol{\nabla}^2\varphi=-\Box\boldsymbol{\nabla}\varphi. \tag{3.16}$$

And when this result is inserted into Eq. (3.15), the $\Box$ operator can be transferred to the D_R function by differentiation by parts. (The surface terms can be shown to vanish.) One thus obtains

$$\begin{aligned}\mathbf{E}_{\parallel}(x)&=\int D_R(x-x')\Box'\boldsymbol{\nabla}'\varphi(x')d^4x'\\ &=-\int\delta(x-x')\boldsymbol{\nabla}'\varphi(x')d^4x'=-\boldsymbol{\nabla}\varphi(x).\end{aligned} \tag{3.17}$$

Equation (3.17) is of course the same as (3.13) because of relation (3.12).

But (3.17) seems to express a "miracle:" it claims that *the retarded field* in Eq. (3.15) *equals the instantaneous field* in Eq. (3.13). The static nonretarded Coulomb field is the result of a "conspiracy" in which the relations between the sources $\mathbf{j}_{\parallel}$ and ρ and the scalar potential are such that the retarded Green function is eliminated. Just as we have seen in Eq.

(3.13), an instantaneous field is obtained because the nonlocality of $\mathbf{j}_\parallel$ exactly compensates the retardation effect. This compensation leads from the retarded field in Eq. (2.6) to the instantaneous one in Eqs. (3.13) and (3.17).

Two side remarks are in order here. First, this procedure teaches us the following trick: any function $f(x)$ can be made into a retarded integral in the following way:

$$\begin{aligned} f(x) &= \int \delta(x-x')d^4x' f(x') \\ &= -\int \Box' D_R(x-x')d^4x' f(x') \\ &= -\int D_R(x-x')d^4x' \Box' f(x'), \end{aligned}$$

where the surface integrals are assumed to vanish. Second, the above results could also have been obtained for an advanced rather than a retarded interaction; D_R would then be replaced by D_A, and the argument carries through in exactly the same way.

We can now summarize the lessons we have learned about the peculiarities of the Coulomb gauge.

(a) The Coulomb gauge condition (3.5) leads to the Coulomb potential (3.12), which is *not retarded*. The corresponding field, $\mathbf{E}_\parallel$ in Eq. (3.13), is therefore also not retarded and leads to an instantaneous (action-at-a-distance) interaction.

(b) This result is a consequence of the *nonlocal* relation between the longitudinal field $\mathbf{E}_\parallel$ and total field $\mathbf{E}$ in Eq. (3.14).

(c) From Eqs. (3.15) and (3.16) we see that the *nonlocality* of $\partial \mathbf{j}_\parallel / \partial t$ combines with the $\boldsymbol{\nabla}\rho$ term so as to *cancel the retardation exactly*, yielding an instantaneous interaction via the longitudinal field $\mathbf{E}_\parallel$.

(d) Finally, if $\boldsymbol{\nabla}\cdot\mathbf{j}_\parallel$ vanishes, the charge density ρ becomes time independent according to the equation for conservation of charge, Eq. (3.10), so that causality becomes a moot question.

IV. THE GRAVITATIONAL CASE

Historians tell us that Newton was quite unhappy over the fact that his law of gravitation implies an action-at-a-distance interaction over very large distances such as that between the sun and the earth. But he was unable to resolve this problem.[6] With the aid of general relativity, one can show that even a first-order correction to his law produces a causal interaction. I shall only sketch how this result arises, and I refer the reader to the excellent text by Rindler[7] or to other texts on general relativity for further details.

Newton's law of gravitation,

$$F = GmM/r^2, \tag{4.1}$$

violates causality, because it does not take into account that it takes a certain time for the interaction to travel from the source M to the mass m on which it acts. In Einstein's theory of gravitation (general relativity), interactions propagate with the speed of light, and gravitation is the result of a curvature of space–time due to the source M rather than due to a force. Its field equations are

$$R_{\mu\nu} = \kappa(T_{\mu\nu} - \tfrac{1}{2} g_{\mu\nu} T). \tag{4.2}$$

Here, $R_{\mu\nu}$ is the Ricci tensor (a contraction of the curvature tensor), κ is the coupling constant, $\kappa = 8\pi G$, and the matter tensor is assumed to be that of dust, $T_{\mu\nu}(x) = \rho u_\mu(x) u_\nu(x)$.

Because curvature depends on the derivatives of the metric tensor, $g_{\mu\nu}$, Einsteinian gravitation can be described to first order as small deviations from the metric for flat space–time, $\eta_{\mu\nu}$, so that

$$g_{\mu\nu} = \eta_{\mu\nu} + h_{\mu\nu}, \tag{4.3}$$

where $h_{\mu\nu}$ is small compared to $\eta_{\mu\nu}$. In this approximation, the field equation (4.2) becomes[7]

$$\Box h_{\mu\nu} = -2\kappa(T_{\mu\nu} - \tfrac{1}{2}\eta_{\mu\nu} T). \tag{4.4}$$

As in the electromagnetic case, one can impose causality by choosing the retarded solution,

$$\begin{aligned} h_{\mu\nu}(x) = 2\kappa \int & D_R(x-x') \\ & \times [T_{\mu\nu}(x') - \tfrac{1}{2}\eta_{\mu\nu}(x') T(x')] d^4x'. \end{aligned} \tag{4.5}$$

To show that this equation reproduces the equation of Newtonian gravitation theory, consider its static limit. In this limit, all components of $T_{\mu\nu}$ vanish except $T_{00} = \rho$. Equation (4.4) then reduces to (again using $c = 1$),

$$\boldsymbol{\nabla}^2 h_{00} = -8\pi G\rho, \tag{4.6}$$

while all the other $h_{\mu\nu}$ vanish. Equation (4.6) is of course exactly the equation for the Newtonian gravitational potential φ. But because the Newtonian and the Einsteinian gravitation theories have completely different pictures of gravitation, the symbols in the equation must be reinterpreted.[8] Thus, the component h_{00} of the correction to the Minkowski metric tensor is to be interpreted as the Newtonian potential (except for a factor of 2), $h_{00} = 2\varphi$. Equation (4.6) then becomes

$$\boldsymbol{\nabla}^2 \varphi = -4\pi G\rho, \tag{4.7}$$

and the Newtonian gravitational force acting on a mass m is $\mathbf{F} = m\boldsymbol{\nabla}\varphi$.[9] (Note that the gravitational potential has different dimensions than the electric one.)

Of course, one could play the same trick (3.14) as in the electromagnetic case. But it seems without justification here. Instead, let us return to Eq. (4.5), the *solution* of the differential equation for $h_{\mu\nu}$. We see that the *solution is retarded*. If we take the Newtonian limit, we find the desired result,

$$h_{00}(x) = -16\pi G \int D_R(x-x')[\rho(x')/2] d^4x'.$$

The Newtonian gravitational potential is therefore

$$\varphi = h_{00}/2 = -4\pi G \int \rho(t - |\mathbf{x}-\mathbf{x}'|, \mathbf{x}') d^3x' / |\mathbf{x}-\mathbf{x}'|, \tag{4.8}$$

so that the Newtonian gravitational force $\mathbf{F} = m\boldsymbol{\nabla}\varphi$ is now *retarded*. By keeping the time derivatives in the differential equation for h_{00}, the requirement for causality can be met. Because the Newtonian theory is entirely static, retardation is not possible until the correction due to deviations from Minkowski space is considered.

[a]Electronic mail: rohrlich@syr.edu

[1]F. Rohrlich, ''Causality and the arrow of classical time,'' Stud. Hist. Philos. Mod. Phys. **31**, 1–13 (2000).

[2]O. L. Brill and B. Goodman, ''Causality in the Coulomb gauge,'' Am. J. Phys. **35**, 832–837 (1967).

[3]J. D. Jackson, *Classical Electrodynamics* (Wiley, New York, 1975), 2nd ed., 3rd ed. (1999).

[4]This name refers to Ludwig Lorenz (1829–1891), who developed a formulation similar to Maxwell's equations a few years after Maxwell (see footnote 5). He is to be distinguished from H. A. Lorentz, 1853–1928. It is unfortunate that in the literature the condition (3.2) is almost without exception attributed to Lorentz. See E. Whittaker, *History of the Theories of Aether and Electricity* (Thomas Nelson, London, 1951).

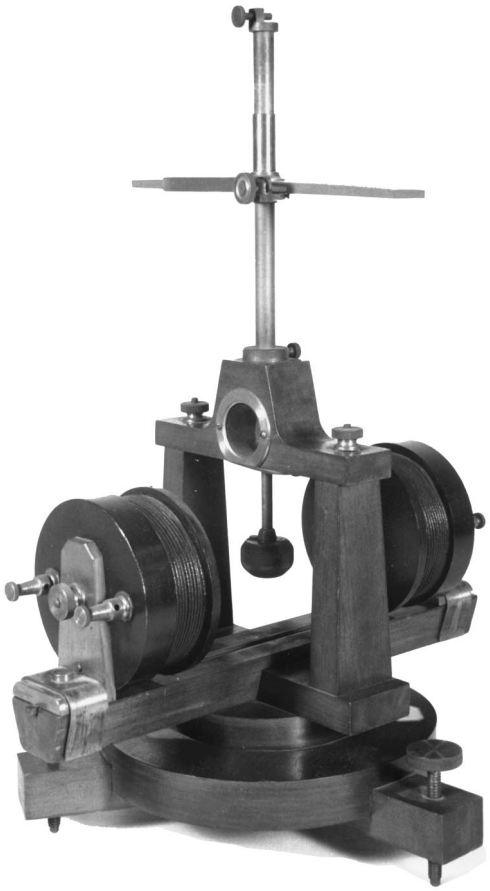
[5]L. Lorenz, ''The identity of the vibrations of light with the electric currents,'' Philos. Mag. **XXXIV**, 287 (1867). This paper is an English translation of his paper in Oversigt over det K. Danske Vidensk. Selsk. Forhandl. No. 1, 26 (1867); a German translation appeared in Ann. Phys. (Leipzig) **131**, 243–263 (1867). Note that H. A. Lorentz was only 14 years old at the time of this publication by Lorenz. Also note that a paper was just published that also points out the misattribution of this work of Lorenz to Lorentz: J. D. Jackson and L. B. Okun, ''Historical roots of gauge invariance,'' Rev. Mod. Phys. **73**, 663 (2001). I am indebted to David Jackson for this reference.

[6]This is the opinion arrived at by historians of science who studied Newton's personal papers including his vast correspondence with scholars of his time. See E. McMullin, *Newton on Matter and Activity* (University of Notre Dame Press, Notre Dame, 1978).

[7]W. Rindler, *Essential Relativity* (Springer, New York, 1977).

[8]F. Rohrlich, ''The logic of reduction: the case of gravitation,'' Found. Phys. **19**, 1151–1170 (1989).

[9]The difference in sign between this relation and the relation $\mathbf{F}=-q\boldsymbol{\nabla}\varphi$ in electrostatics comes from the fact that masses attract one another while like charges repel.



Wiedemann's Galvanometer. Gustav Wiedemann's version of the Helmholtz type of tangent galvanometer was developed in 1874. The coils on either side of the suspended compass needle could be moved back and forth to change the sensitivity. Damping was provided by the copper sphere surrounding the needle. Wiedemann employed Gauss and Weber's method of using a mirror attached to the lower part of the torsion fiber to act as an optical lever to increase the sensitivity of the instrument. This galvanometer is the most ambitious piece of apparatus built by Prof. Joseph Naylor of DePauw University ca. 1900. (Photograph and notes by Thomas B. Greenslade, Jr., Kenyon College)